

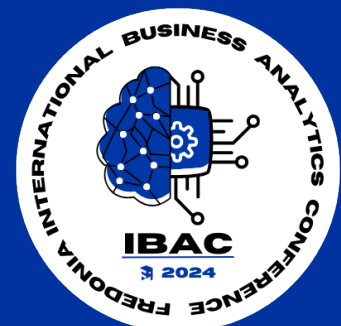


International Business Analytics Conference Proceedings

Fredonia, New York

Friday – Saturday, May 3 – 4, 2024

Conference Website: www.fredonia.edu/ibac
Conference Email: ibac@fredonia.edu



Fastenal

Fastenal, a global leader in industrial and construction supplies, excels in providing unparalleled solutions for businesses of all sizes. With a vast inventory of over 1 million products, including fasteners, safety equipment, tools, and more, Fastenal ensures customers have access to quality goods when and where they need them.



Leveraging a network of over 3,000 stores and 14 distribution centers worldwide, it prioritizes efficiency and reliability. Its commitment to innovation and quality is evident through advanced vending and inventory management systems, empowering clients to optimize operations. Fastenal's dedication to customer satisfaction and industry-leading service makes them the go-to partner for success in any project or endeavor.



For more information, visit <https://www.fastenal.com>



M&T Bank




Learning and growing is important.

At M&T Bank, we know how important it is to support education and lifelong learning. That's why we offer our time, energy and resources, and encourage others to do the same. [Learn more at mtb.com](https://mtb.com).

M&T Bank is proud to support SUNY Fredonia International Business Analytics Conference .

M&TBank

 Equal Housing Lender. ©2024 M&T Bank. Member FDIC.

Chautauqua County

Chautauqua County is supporting the International Business Analytics Conference through a tourism grant designed to “increase tourism, conventions, trade shows, special events, and other directly related or supporting activities.”

The County supports the "development of distinct themes or significant events and attractions that enhance the level of visitor experience," and ultimately the number of visitors to the County.

We thank Chautauqua County leaders for supporting this inaugural event.



Visit www.chqgov.com
for more information
on local attractions!

State University of New York at Fredonia

Founded in 1826, The State University of New York at Fredonia is one of the jewels of the SUNY system, known for its welcoming atmosphere, distinguished faculty, and beautiful campus. With affordable tuition and housing in a classic "college town," the State University of New York at Fredonia gives students the academic challenges of a selective university committed to developing their character and preparing them for a rewarding career.

Fredonia offers a complete college experience -- a "destination university" in a creative environment that is diverse, welcoming, and safe. It helps students become uniquely connected to classmates, professors, and the community. The vibrant campus features comfortable residence halls, delicious dining options, and an abundance of extracurricular options to keep your evenings and weekends as stimulating as your daily studies.

Fredonia has over 80 undergraduate and graduate programs, along with 50+ minors and 13 cooperative programs, designed to help students reach their full potential. Fredonia's 14:1 student faculty ratio provides the personal attention students expect in a small private school, delivered at a public college ranked by national publications like Money magazine, Princeton Review, Kiplinger's, and U.S. News & World Reporter for quality and affordability.



Visit www.fredonia.edu
for more information!

Conference at a Glance

The International Business Analytics Conference (IBAC) is pleased to announce its inaugural conference, hosted at SUNY Fredonia on Friday-Saturday, May 3-4, 2024. IBAC 2024 provides a unique platform that bridges the gap between academic and industry expertise in the field of business/data analytics through Academia-Industry Partnership. By bringing together industry practitioners and academicians, the conference facilitates meaningful discussions and foster collaboration in the ever-evolving world of business analytics. The IBAC 2024 conference welcomed submissions on a wide range of topics related to research and current industry use of business or data analytics, focusing on but not limited to the fields of Accounting, Business and Management, Computer Science/Management Information Systems, Economics, Finance and Insurance, Education, Healthcare, Marketing, Music and Entertainment, and Sport Management.

For example, submission specifically addressed research or current business practices in:

- Data-driven decision-making
- Predictive and prescriptive analytics
- Artificial Intelligence and Machine Learning in business
- Big data analytics and its applications
- Business Intelligence and data visualization
- Data mining and pattern recognition
- Text and sentiment analysis in business
- Supply chain analytics and logistics optimization
- Marketing and customer analytics
- Risk management and fraud detection
- Ethics and privacy in business analysis

Conference Committee Members

Organizing Committee

Conference Chair:

Kaustav Misra (Dean, School of Business)

Organizing Committee Chairs:

Program Chair (Academic):

Mark Nickerson

(School of Business - Accounting)

Justin Mindzak

(School of Business - Accounting)

Program Chair (Industry):

Charles Cornell (Center for Innovation and Economic Development)

Lisa Walters (School of Business - Management)

Ad-hoc Member: Syed Haider (Computer and Information Sciences)

Administrative Support:

Robyn Reger and Kristie Bobik
(School of Business)

Program Committee

Publicity Chair:

Cara Lanning (CIED)

Finance Chairs:

Linda Hall (School of Business - Accounting) and
Neepa Gaekwad Babulal (School of Business - Economics)

Event Management:

Lauri Gawronski (CIED), Mark Delcamp (Facilities Services), Jeff Walter and Katie Thies (FSA), Mark Suida (Campus Life), John McCune (Information Technology), Scott Saunders (Registrar), Kathy Forster (Residence Life) and Tim Murphy (College Foundation)

Conference Itinerary

Friday, May 3 rd		
8:00 a.m. - 12:00 p.m.	Registration & Information	MPR, Williams Center
8:00 a.m. - 9:00 a.m.	General Breakfast & Conference Inauguration Ceremony: President Kolison and Provost Starrett	MPR, WC
9:00 a.m. - 9:10 a.m.	Plenary Session: Greetings from President Stephen Kolison	MPR, WC
9:10 a.m. - 9:20 a.m.	Plenary Session: Greetings from Provost David A. Starrett	MPR, WC
9:30 a.m. - 10:50 a.m.	Concurrent Sessions	Williams Center
	Track 1: Machine Learning and Management	WC G103A
	Track 2: Applied Economics	WC G103B
	Track 3: Accounting and Auditing	WC G103C
	Track 4: Analytics in Business Education	WC S204A (Second Floor)
10:50 a.m. - 11:00 a.m.	Coffee Break	MPR, WC
11:00 a.m. - 12:30 p.m.	Round Table Discussion - Fastenal	MPR, WC
12:30 p.m. - 2:00 p.m.	Lunch	MPR, WC
1:00 p.m. - 2:00 p.m.	Keynote Speaker 1: Conversation with Karl Holz	MPR, WC
2:00 p.m. - 3:20 p.m.	Concurrent Sessions	Williams Center
	Track 5: Insights into Financial Markets	WC G103A
	Track 6: Student Research Projects	WC G103B
	Track 7: Advanced Business Analytics	WC G103C
	Track 8: Industry Insights	WC S204A (Second Floor)
3:20 p.m. - 3:30 p.m.	Coffee Break	MPR, WC
3:30 p.m.- 4:15 p.m.	Presentation by Cengage	MPR, WC
4:15 p.m. - 5:15 p.m.	Depart from MPR to Clarion Hotel and BREAK	
5:15 p.m. - 6:15 p.m.	Reception and Networking with Smooth Jazz	Clarion Hotel
6:15 p.m. - 7:15 p.m.	Dinner	Clarion Hotel
7:15 p.m. - 8:15 p.m.	Keynote Speaker 2: Linda Powell	Clarion Hotel
Saturday, May 4 th		
8:00 a.m. - 9:30 a.m.	Registration & Breakfast	MPR, WC
8:30 a.m. - 9:15 a.m.	Panel Discussion - Bridging the Gap	MPR, WC
9:30 a.m. - 10:50 a.m.	Concurrent Sessions	Williams Center
	Track 9: Applied Economics 2	WC G103A
	Track 10: Operational Strategies	WC G103B
	Track 11: Insights and Innovations in analytics	WC G103C
	Track 12: Management and Education	WC S204A (Second Floor)
10:50 a.m. - 11:00 a.m.	Coffee Break	MPR
11:00 a.m. - 12:00 p.m.	Fireside Chat with Dr. Ernest Fokoue	MPR, WC
12:00 p.m. - 1:30 p.m.	Lunch	MPR, WC
12:30 p.m. - 1:30 p.m.	Keynote Speaker 3: Chris Seferlis	MPR, WC
1:30 p.m. - 2:00 p.m.	Award Distribution and Concluding Remarks by Kaustav Misra	MPR, WC
2:30 p.m.	Depart from Campus!	

*** MPR - Multipurpose Room, WC - Williams Center ***

General Information

Conference Registration – Registration will be held in the Multipurpose Room (MPR) of the Williams Center, starting each day at 8:00am.

Badges – All conference registrants will receive a personalized badge when they check in at the registration desk. Please wear your badge at all times since will be checked at all sessions, meal functions and events. If you misplace your badge, please come to the registration desk for a replacement.

Connect with Us!

Annual Conference Website: <https://www.fredonia.edu/ibac>

Find the School of Business on Social Media!

- Facebook: <https://www.facebook.com/FredoniaSchoolofBusiness>
- Twitter: <https://twitter.com/FredBusAdmin>
- LinkedIn: <https://www.linkedin.com/in/fredonia-school-of-business-412106164/>

Special Assistance

Accessible Entrances Map

SUNY Fredonia is committed to making its entire campus accessible to all individuals, including those with disabilities. For further information regarding special needs, or if you have previously requested assistance for this conference, please visit the registration desk.

Copyright Notice

Authors maintain the copyright to their work presented at the International Business Analytics Conference (IBAC). Authors possess the right to present their work at the International Business Analytics Conference, even if it has been or will be presented elsewhere. Abstracts and papers may be published in the Conference Proceedings provided they have been presented at the International Business Analytics Conference, have not been previously published elsewhere, and the author does not decline publication.

To safeguard the interests of our Conference presenters, the International Business Analytics Conference adheres to the principles of the [Creative Commons Attribution 4.0 License](#).

Featured Speakers

Friday, May 3, 2024: Luncheon Keynote Speaker - Karl Holz

Karl Holz earned his bachelor's degree in Business Administration, is the former President of Disney Cruise Line and New Vacation Operations. A 22-year veteran of the Walt Disney Company, Karl has deep expertise in building organizational culture, customer experience, operations, strategic planning, international business and start-up ventures. In 2004, he was appointed President and COO of Disneyland Resort Paris, and shortly thereafter, became Chairman and CEO. Currently, Karl is a Senior Advisor at McKinsey & Company. He is also a board member for CRUISE SAUDI, Radisson Hotel Group and the Pro Football Hall of Fame Village Resort. Karl serves SUNY Fredonia as a current member of the Fredonia College Foundation Board of Directors, and was the recipient of the Fredonia Alumni Association's Outstanding Achievement Award in 2005.



Friday, May 3, 2024: Dinner Keynote Speaker - Linda Powell

Linda F. Powell is the Enterprise Head of Data Governance and Deputy Chief Data Officer at Bank of NY Mellon. She has over 20 years of experience in the finance industry including commercial banking, banking supervision, and supporting economic research. She spent most of her career with the Federal Reserve System. She was also the Chief Data Officer at the Treasury Department's Office of Financial Research and the Consumer Financial Protection Bureau. In 2018 she joined Citibank as the Global Head of Data Governance, Data Reporting, and End User Computing and joined Bank of NY Mellon in 2021. She has published several papers on the topics of data standards, metadata, and data strategy. She has a B.A. in Economics from Rutgers University and an M.S. in Quantitative Finance from George Washington University.



Saturday, May 4, 2024: Luncheon Keynote Speaker - Chris Seferlis

Chris Seferlis is a Technology Strategist at Microsoft where for the past 4+ years he has helped customers bring their technical strategies to reality with a heavy focus on how they are using their data. He is also a Visiting Professor at Boston University in the School of Computing and Data Sciences, teaching with a mix of practical and experiential discussions. A former CIO with over 20 years' experience in IT and a deep background in Data Warehousing and Business Intelligence, Chris brings a practical and theoretical approach to business technology challenges using a data driven mindset. Chris holds an MBA from the University of Massachusetts, is a Microsoft Certified Trainer, and recently completed the book A Practical Guide to Azure Cognitive Services and OpenAI - Leverage the power of AI to optimize operations, reduce costs, and deliver cutting-edge AI solutions.



Plenary Sessions: Panel Discussions, Fireside Chat and Roundtable

Roundtable by Fastenal

Friday, May 3rd, 2024 from 11:00am to 12:30pm

Title: The Application of Data Analytics in Decision Making

Panelists: Ryan Scott – Regional Vice President, Mike Pluchino – Regional Finance Manager, Josh Rice - Lean Solutions Manager, Vince DeRosa – District Manager, and – Cody Enger – District Sales Specialist, Andrew – Bagley Recruiter

Presentation by Cengage

Friday, May 3rd, 2024 from 3:30pm to 4:15pm

Title: Unlock the power of data with Cengage's Business Analytics presentation on MindTap

Discover how our cutting-edge platform enhances student engagement, provides real-world scenarios, and fosters critical thinking skills. Boost student success and prepare them for a data-driven world. Don't miss out on this transformative learning experience!

Panel Discussion

Saturday, May 4th, 2024 from 8:30am to 9:15am

Title: Bridging the Gap Between Data Analytics Pedagogy and Industry Demands

Panelists: Megan Johnson - Assistant Professor of Mathematics, Adam Cook -Associate Professor of Economics, Shahin Mehdipour Ataee – Assistant Professor of Computer Science, Emily Fay – M &T Senior Data Scientist, Rachel Hettenbaugh (M&T Business Analytics & Reporting), Anna Pierce (M&T Risk Analyst)

In the dynamic landscape of industry, the role of data analysis is pivotal. As educators, researchers, and practitioners, it's imperative to continually innovate our approaches to teaching and learning in this field. The primary objective of our panel discussion is to foster a collaborative dialogue among experts in data analytics education and industry practitioners. Through sharing experiences, methodologies, and case studies, we aim to elucidate effective strategies for enhancing data analytics pedagogy and ensuring graduates possess the requisite skills and competencies for the job market.

Fireside Chat by Dr. Ernest Fokoue

Saturday, May 4th, 2024 from 11:00am to 12:30pm

Title: Rediscovering the Quintessential Building Blocks of Effective Statistical Data Analysis

The gist of this, or rather the spirit, is to provide the audience with a mini tour of some of the ubiquitous methods and techniques along with practices that have remained at the core of analytics, despite the thunder/barrage of fancy extensions parading as better.

Gold Sponsors

Accountants + Mentors

BS&P

Adding value

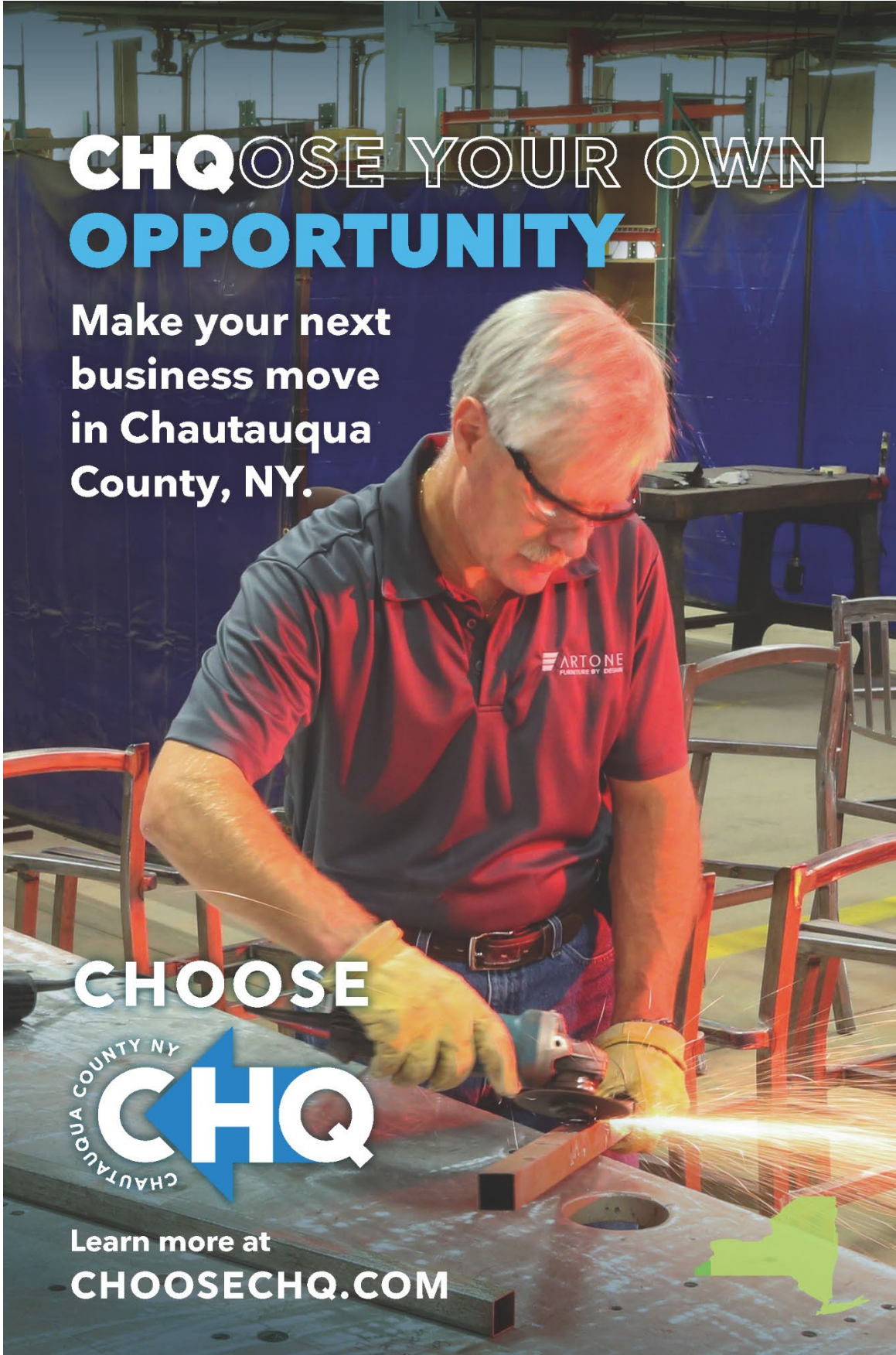
We are proud to sponsor the 2024 International Business Analytics Conference at SUNY Fredonia. At BS&P, we are continually bringing in fresh perspectives through our Leadership program, internships and shadowing opportunities—creating a culture of growth and learning. Adding value to businesses, employees and the community—it's what we do.

If you are interested in learning more about the opportunities at BS&P, scan the QR Code.



Tax | Consulting | Audit | Cross Border | SOC Engagements
Wealth Management | Human Resources

726 Exchange St, Suite 822, Buffalo, NY 14210 | P: (716) 854-5034 | BSPcpa.com




CHOOSE YOUR OWN OPPORTUNITY

Make your next business move in Chautauqua County, NY.

CHOOSE
CHAUTAUQUA COUNTY NY
CHQ

Learn more at
CHOOSECHQ.COM



Silver Sponsors

Cengage

Unlock the power of data with Cengage's Business Analytics presentation on MindTap!

Discover how our cutting-edge platform enhances student engagement, provides real-world scenarios, and fosters critical thinking skills. Boost student success and prepare them for a data-driven world. Don't miss out on this transformative learning experience!

For more information, visit <https://www.cengage.com>



Higher Ed Course
Materials for Easier
Learning

At Cengage, we're here for you.



Lawley

INSURANCE | EMPLOYEE BENEFITS

PROUD TO PARTNER WITH
FREDONIA STATE UNIVERSITY
INTERNATIONAL BUSINESS ANALYTICS CONFERENCE

lawleyinsurance.com

Bronze Sponsors

FIRST AFFILIATED

of the Chautauqua Region, LLC

Joseph A. Cerrie

9 East Main Street 716-672-2108
Fredonia, NY 14063 Cell: 716-467-4853
jcerrie@ma-adviser.com Fax: 716-484-4810

Investment advisory services provided by
NewEdge|Advisors, LLC doing business as
First Affiliated of the Chautauqua Region, LLC
as a registered investment adviser.

Securities offered through NewEdge Securities, Inc.
Member FINRA/SIPC. NewEdge Advisors, LLC and
New Edge Securities, Inc. are wholly owned subsidiaries
Of NewEdge Capital Group, LLC.



**CCE Offers creative IT-
Solutions for companies of all
sizes**

**Our team of experts will work
with you to understand your
specific needs and design
services that meet those needs**

**ccesoftware.com
jpiede@ccesoft.com
716 679 8229**



**LAW OFFICES OF
SHEILA STARKEY HAHN, PC**

**Serving the immigration needs
of companies and individuals since 1998
sheila@sheilahahn.com**



Conference Tracks

Concurrent Sessions (Friday - 9:30am) Tracks 1-4

Track 1: Machine Learning and Management

Room: WC G103A (9:30am -10:50am)

Moderator: Gabby Resch (*Ontario Tech University*)

1.1 Customer Churn Prediction with Machine Learning

Presented by: Bahareh Rahmani (*Saint Louis University*)

Co-authors:

A. Maazallahi (*Saint Louis University*)

Y. M. Meda (*Saint Louis University*)

B. M. S. Bokka (*Saint Louis University*)

H. Jamallamudi (*Saint Louis University*)

P. Norouzzadeh (*Saint Louis University*)

E. Snir (*Washington University in Saint Louis*)

This article explores the prediction of customer churn in the banking sector using various machine learning models. It comprehensively analyzes extensive customer data, employing techniques including linear regression, decision trees, KNN, Naive Bayes, and Random Forest. The study emphasizes visual data interpretation through scatter plots, histograms, and box plots. It concludes with an assessment of ensemble methods, highlighting the superiority of Random Forest in predicting churn, offering crucial insights for customer retention strategies in banking.

1.2 An Empirical Analysis of the Functionalities and Confidence Scoring Mechanisms in Leading Large Language Models

Presented by: Mohammad Nasim (*Northwood University*)

This research paper provides a comprehensive empirical analysis of the functionalities and confidence scoring mechanisms in leading large language models (LLMs) such as GPT, BERT, and others. The unprecedented capabilities of these models in understanding and generating human-like text have positioned them as central to advancing artificial intelligence and its applications across various domains. Despite their widespread use, there remains a significant gap in understanding the breadth of functionalities these models offer and how they quantify and communicate the confidence in their outputs. Through a methodical examination, this study aims to bridge this gap by assessing the range of tasks LLMs can perform—from text generation and summarization to question answering and sentiment analysis—and evaluating their mechanisms for confidence scoring, including softmax probabilities, Bayesian methods, ensemble techniques, and calibration approaches. The analysis is grounded on a robust research design involving benchmarking tasks, user studies, error analysis, and comparative studies across models. The findings not only shed light on the strengths and limitations of current LLM functionalities and confidence mechanisms but also offer insights into their practical implications for users and developers. By highlighting areas for improvement, this study

contributes to enhancing the reliability, transparency, and user trust in LLMs, paving the way for more robust and accountable AI systems.

1.3 Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism

Presented by: Soumik Banerjee (*Canisius University*)

Co-authors:

Gopal Nath (*Murray State University*)

Yawei Wang (*Montclair State University*)

Austin Coursey (*Murray State University*)

Krishna K. Saha (*Central Connecticut State University*)

Productivity losses caused by absenteeism at work cost U.S. employers billions of dollars each year. In addition, employers typically spend a considerable amount of time managing employees who perform poorly. By using predictive analytics and machine learning algorithms, organizations can make better decisions, thereby increasing organizational productivity, reducing costs, and improving efficiency. Thus, in this paper we propose hybrid optimization methods in order to find the most parsimonious model for absenteeism classification. We utilized data from a Brazilian courier company. In order to categorize absenteeism classes, we preprocessed the data, selected the attributes via multiple methods, balanced the dataset using the synthetic minority over-sampling method, and then employed four methods of machine learning classification: Support Vector Machine (SVM), Multinomial Logistic Regression (MLR), Artificial Neural Network (ANN), and Random Forest (RF). We selected the best model based on several validation scores, and compared its performance against the existing model. Furthermore, project managers may lack experience in machine learning, or may not have the time to spend developing machine learning algorithms. Thus, we propose a web-based interactive tool supported by cognitive analytics management (CAM) theory. The web-based decision tool enables managers to make more informed decisions, and can be used without any prior knowledge of machine learning. Understanding absenteeism patterns can assist managers in revising policies or creating new arrangements to reduce absences in the workplace, financial losses, and the probability of economic insolvency.

Track 2: Applied Economics

Room: WC G103B (9:30am -10:50am)

Moderator: Rachid Belhachemi (*Le Moyne College*)

2.1 Textual Analysis of Beige Books to Predict Regional Economic Changes

Presented by: Charlie Smith (*University of South Alabama*)

Beige Books, compiled by the Federal Reserve, contain anecdotal observations about current economic conditions from a diverse set of business leaders. This research uses natural language processing (NLP) to investigate the predictive potential of Beige Book sentiment on regional differences in the U.S. macroeconomy. The methodology involves the application of a random forest classifier to predict whether a Federal Reserve district experienced economic growth from

the previous month based on multiple sentiment features, including mean and median sentiment scores, interquartile range, and sentiment near key trigrams. Validation techniques are employed to assess model robustness, and the results suggest Beige Books are able to make regional economic predictions substantially greater than the level of a guess, with the AUC score on a receiver operating characteristic curve equaling .746. Enhancements are proposed for future research through advanced sentiment analysis methodologies using large language models like GPT and BERT and by comparing predictions made by alternative machine learning algorithms.

2.2 The Demand for Money for EMU: A Flexible Functional Form Approach

Presented by: Neepta Gaekwad Babulal (*State University of New York at Fredonia*)

Co-author: William Barnett (*University of Kansas*)

Monetary aggregates have a special role under the “two pillar strategy” of the ECB. Hence, the need for a theoretically consistent measure of monetary aggregates for economic union (like EU and EMU) is difficult to obliterate. This paper analyzes the monetary assets for EMU. Firstly, an aggregation is done for the monetary services for the economic union, EMU-11 countries, that is, Estonia, Finland, France, Germany, Ireland, Italy, Luxembourg, Malta, Netherlands, Slovakia, and Slovenia. The Divisia monetary aggregation approach (Barnett 1980, 2003, 2007) is adopted, which is consistent with the index number theory and microeconomic theory. For the first time in this paper a Multilateral Divisia monetary aggregate index is created. The multilateral Divisia monetary aggregate for EMU-11 is more informative and a good signal of the economic trends when compared to the corresponding simple sum aggregate. Secondly, the monetary assets for EMU-11 are analyzed in the framework of a representative consumer’s utility function, using the locally flexible functional form, the Minflex Laurent Indirect utility function (Barnett 1983). The analysis of the elasticities with respect to their user costs (income elasticity, price elasticity, elasticity of substitution) shows that: (i) Transaction balances (TB) and Deposits redeemable at notice (DRN) are income elastic (ii) The DRN has displayed a large variation in price elasticity (iii) the monetary assets are not perfect substitutes in EMU-11. The simple sum monetary aggregation assumes all the monetary assets as perfect substitutes, hence a simple sum monetary aggregate distorts the measurement of the monetary aggregate.

2.3 Optimal Forest Management for Interdependent Products: A Nested Dynamic Bioeconomic Model and Application to Bamboo

Presented by: Tong Wu (*Cornell University*)

Co-authors:

C.-Y. Cynthia Lin Lawell (*Cornell University*)

David R. Just (*Cornell University*)

Jiancheng Zhao (*Zhejiang Academy of Forestry*)

Zhangjun Fei (*Boyce Thompson Institute and Cornell University*)

Qiang Wei (*Nanjing Forestry University*)

Sustainable forest management is an important issue worldwide. Forests supply the world’s population with timber and non-timber forest products, including renewable products such as fruits, nuts, and maple syrup that can be harvested at more frequent intervals than the trees themselves. We develop a nested dynamic bioeconomic model and dynamic structural econometric model of the management of forests that generate interdependent products that

differ in their growth cycles, rates of growth, lengths of growing periods, and potential harvest frequency. We apply our model to detailed daily panel data we have collected and constructed on bamboo shoot and bamboo stem harvesting decisions made by bamboo farmers, in order to assess the optimality of their bamboo forest management strategies and to understand the beliefs and perceptions of bamboo farmers that underlie and rationalize their management strategies. We plan to use our dynamic model to simulate, analyze, and design policies and institutions to improve sustainable forest management. Our novel dynamic bioeconomic model has important implications for the sustainable management of forests worldwide, particularly when the forests produce products that grow on trees, are renewable, and can be harvested at more frequent intervals than the trees themselves.

Track 3: Accounting and Auditing

Room: WC G103C (9:30am -10:50am)

Moderator: Justin Mindzak (*State University of New York at Fredonia*)

3.1 Generational Resistance to Emerging Data Analytic Tools Among Accounting and Finance Professionals

Presented by: Cheryl Moore (*Mercyhurst University*)

Just half a decade ago researchers were positing that information on data analytics, AI, ChatGPT and other technological advancements were under-researched in the Accounting and Finance fields. Fast forward 5 years and a literature review on accounting and emerging technologies reveals a plethora of studies focused on acceptance, use, timing, resistance, and education to name a few.

This study seeks to identify and explain why some accountants are resistant to change and delay or forego acceptance of new data analytical methods that can enhance the user's data and increase the efficiency of the worker's duties. Utilizing the Status Quo Bias Theory independently and using this theory in conjunction with the Big Five personality dimensions, several studies investigate the reaction of Excel users and the participant's unwillingness to relinquish the comfort of this analytical tool and the resistance to adopt a new one. This study will continue to expand on this research by looking at generational differences of acceptance and adoption in the accounting and finance fields.

3.2 Leveraging Deep Learning for Accounting Fraud Detection

Presented by: Victoria Gonzalez (*The State University of New York at Buffalo*)

Co-authors:

Sai Teja Reddy (*The State University of New York at Buffalo*)

Chitral Patil (*The State University of New York at Buffalo*)

Haimonti Dutta (*The State University of New York at Buffalo*)

Fraudulent financial reporting encompasses deliberate misstatements or omissions in financial statements, aiming to mislead stakeholders and regulators, resulting in a breach of Generally Accepted Accounting Principles (GAAP). This research introduces a novel predictive model for detecting accounting fraud, leveraging publicly available financial data extracted from 10-K

filings. Specifically, a Multi-Layer Perceptron (MLP) is constructed with Random Under-Sampling (RUS) (called RUS MLP) which has comparable performance to traditional machine learning algorithms designed for fraud detection such as logit and probit models. Our findings reveal that advanced deep learning techniques can be used to fortify financial systems against fraudulent activities.

3.3 Is Artificial Intelligence (AI) a welcoming Tool to Conduct Risk Assessment and Analytical Procedures?

Presented by: Linval Frazer (*The State University of New York at Old Westbury*)

Co-author: Jeffrey D'Amico (*The State University of New York at Old Westbury*)

The integration of Information Technology, and the manipulation and use of big data in companies' operations pose a challenge to auditors, and their ability to reduce audit risk to an acceptable level. The use of Artificial Intelligence as a tool to conduct risk assessment and analytical procedures is evaluated in this paper. It presents information on how AI can be used to improve the auditing process. The paper concludes that the use of AI will enhance auditing procedures and address some of the limitations the profession has faced for many years. It is not expected that AI or any technology will obliterate audit risk completely. However, if used effectively as a tool to support the audit objective, AI can help reduce audit risk.

Track 4: Analytics in Business Education

Room: WC S204A (9:30am -10:50am)

Moderator: Itauma Itauma (*Northwood University*)

4.1 Enhancing Business Analytics Education: Strategies for Success

Presented by: Joseph Porter (*Nazareth University*)

In this session, I will share proven strategies for teaching Business Analytics, and I will:

1. Clarify the Distinction: I will clarify the disparity between generic analytics (i.e. identifying patterns in COVID data) and Business Analytics (i.e. identifying trends in sales, inventory, customer acquisition, etc.).
2. Address Data Quality: I will emphasize the significance of data integrity by differentiating between clean and dirty data (i.e. inconsistent or missing COVID data).
3. Analytics Framework: I will outline the evolution from Descriptive Analytics (i.e. past sale of a specific product) to Predictive Analytics (i.e. A/B Testing and Regression forecast), and finally to Prescriptive Analytics (i.e. actionable insights on revenue maximization and cost reduction under various constraints).
4. Visualization: Exploring an array of visualization tools such as line graphs, bar charts, and pie charts, I will demonstrate their efficacy in aiding decision-makers to identify trends and patterns.
5. Collaboration and Certification: I will highlight the importance of collaborative assignments and the attainment of certifications from prominent platforms like Google, Cloud, and Salesforce.

4.2 Empowered Decision-Making: Bridging the Gap Between Business Simulations and Generative AI

Presented by: Joseph Kuvshnikov (*Gannon University*)

Business simulations have transformed the world of business education. Initial explorations in bridging the gap between business simulations and large language model generative artificial intelligence have resulted in empowered decision making. Business simulations have long been embraced as a powerful tool in helping participants hone their strategic business decision-making skills (Kurtz, 2003; Schwarz, 2009). Sources of pedagogical power include: experiential learning (Humphreys, Bakir, & Babb, 2022), the opportunity to try new strategies in a low-risk environment (Mubaraz & Mezrar, n.d.), the opportunity to practice business management and decision-making in a collaborative team-based environment (Anderson & Lawton, 2009), and business simulations allow participants to develop strategic and operational decision-making skills (Levant, Coulmont, & Sandu, 2016). Combining business simulations and generative AI has led to significant new pedagogical synergies. Faculty/facilitators can use live generative artificial intelligence prompts to populate discussions with a myriad of considerations, strategies, and pros and cons of various business decisions (Moser, 1986; Ferreira, González-González, & Adamatti, 2021; Paschen, Wilson, & Ferreira, 2022). The speed at which content is delivered makes generative artificial intelligence ideal for either synchronous or asynchronous learning environments. The purpose of this paper is to report on explorations in the ways faculty can coach simulation participants in prompting AI and discerning how to apply the results to specific decision-making contexts. Generative AI rapidly equips faculty to embrace each teachable moment with simulation and real-world applications. This explorative initiative report contains several examples of prompts applied in the context of the Income|Outcome (ASI) business simulation.

4.3 So Many Decisions: How Higher Education is Utilizing Analytics

Presented by: Andrea Rodgers (*University of Miami*)

Analytics has come to forefront in the business environment over the last decade or so. It is used across all industries and segments. One of the primary ways it is used is for decision-making. The creation of dashboards and other visual tools has made it convenient for CEO's, directors, and others to track key indicators in one place. One industry that the use of analytics is the most widespread and used for a multitude of reasons is in higher education. In higher education, institutions use analytics in the classroom, educating students about what it is, how to use it as well as other aspects including data collection, statistical analysis and reporting and data management. They not only teach it, but also use it in practice outside of the classroom, in the business of education, for decision-making. This paper aims to discuss the ways and how data is used within institutions for decision-making. Being used across the entire institution, not only at the institutional level, but also by the individual colleges and schools. A comparison of types of institutions will also be discussed. Particularly, comparing and contrasting a large, public institution with a small, private institution.

4.4 The Use of Generative AI in Quantitative Coursework

Presented by: Lisa Walters (*State University of New York at Fredonia*)

Co-authors:

Mark Nickerson (*State University of New York at Fredonia*)

Linda Hall (*State University of New York at Fredonia*)

So-Jin Yu (*State University of New York at Fredonia*)

Generative Artificial Intelligence (GAI) has firmly established itself as a transformative force, poised to revolutionize various industries. As its influence continues to expand, educators are presented with an unprecedented opportunity to harness AI's potential in crafting comprehensive syllabi that incorporate students' engagement with AI. How can educators effectively utilize AI to develop syllabi that not only integrate AI but also encourage its constructive utilization among students? What essential principles should guide this process to ensure both motivation and active participation?

This session aims to showcase the creation of a syllabus tailored for an accounting course, illustrating how AI can be employed to generate initial drafts and subsequently refine them to yield robust educational materials. Additionally, the session will delve into the generation of detailed lesson plans for three distinct topics drawn from the syllabus. Following the syllabus model, the prompts for these lesson plans will be fine-tuned to enhance specificity and structure, resulting in more coherent and actionable plans.

Participation from attendees will be solicited to foster engagement and stimulate interest throughout the session. Moreover, the session will culminate in the dissemination of practical tips and best practices applicable to pedagogical endeavors, alongside a discussion of the inherent limitations of GAI in educational contexts.

Concurrent Sessions (Friday – 2:00pm) Tracks 5-8

Track 5: Insights into Financial Markets

Room: WC G103A (2:00 p.m. -3:20 p.m.)

Moderator: Linval Frazer (*The State University of New York at Old Westbury*)

5.1 Predicting Bankruptcy: Ask the Employees

Presented by: Kristina Lalova (*Michigan State University*)

Co-author: John Knopf (*University of Connecticut*)

The purpose of the paper is to show that employees have information on an upcoming bankruptcy years before actual bankruptcy filings. We test our model using employee information against other bankruptcy models from the literature using a dataset from 2008 to 2020. We build a new model to reflect employees' attitudes before bankruptcy filings and include key variables from the already established bankruptcy models in the literature in our model. We find that our model more accurately predicts bankruptcy two to three years before the actual bankruptcy filing, while the other models are more accurate in the year prior to the bankruptcy. For two and three years prior to a bankruptcy filing, the model outperforms the existing models in in- and out-of-sample tests. We create neural network models consisting of reviews and ratings separately and show that textual reviews provide additional predictive

power for bankruptcy filings on top of rating and financial information.

5.2 Interpretable Effect Analysis of Exogenous Market Indicators in Stock Market Prediction

Presented by: Max Yun (*Sacred Heart University*)

Geared with Machine Learning (ML) and Deep Learning (DL) techniques, recent stock market prediction studies achieve phenomenal prediction accuracies despite the inherent complexity and difficulty of time series stock price data. However, obtaining a reliable human-friendly interpretability of the prediction outcome is still challenging due to the black-box nature of ML and DL models. Human-friendly interpretability in stock price prediction models helps investors discover the underlying reasons of the prediction outcome and enables a prompt response to unpredictable market prices. Among the types of stock market studies, technical analysis can instantly reflect constantly changing stock prices and avoid drawbacks of fundamental analysis. However, technical analysis using internal price-derived technical indicators is criticized as pseudoscience because of its subjectivity and self-fulfilling prophecy. Besides, input feature space expansion by price-based technical indicators can cause data snooping and sample bias problems. A remedial solution is to use exogenous technical market indicators. Volatility Index (VIX), as the most popular exogenous technical market indicator, expands the information beyond classic price and volume data to a variety of financial information. There is much research on the effectiveness of VIX in stock market prediction, but previous studies fail to observe important characteristics of time series price data. This study focuses on temporality and collective behaviour characteristics of time series data to investigate the effect of VIX in stock market prediction. The proposed piecewise best feature subset selection using Savitzky–Golay smoothing enhances the human-friendly interpretability that reflects temporality and collective behaviour of VIX.

5.3 Manufactured Doubt and Stock Market Implications

Presented by: Sabrina Davis (*Pace University*)

Co-author: Leigh Anne Novak Donovan (*Pace University*)

Organizations deal with uncertainty (Thompson 1967). The way the tobacco industry executives handled extreme pressure and uncertainty has been partially reviewed in the public health domain. Fama et al (1969) introduced evidence of how stock returns respond to information. However, no one has examined how stock markets react to disinformation. This research is designed to add to the existing body of knowledge on manufacturing doubt. Manufactured doubt is a term that has appeared within the healthcare literature and is usually examined in hindsight in areas such as smoking, fossil fuel, and climate change. The research we propose examines the effects of manufactured doubt on cigarette sales and the stock market by examining reactions to positive and negative public announcements divided into 3 categories (1) scientific announcements, (2) regulatory changes and (3) data litigation. This sample contains 40 national announcements from June 6th, 1954 to April 29th, 2021 and 28 cigarette manufacturers. We compared each company's stock price 5 days prior to and 5 days after the announcement. The data includes the daily stock price for 121 days prior to the event study dates for each manufacture, for each announcement date. We calculated the raw percentage price changes, prior to the 11-day window to see if there was any abnormal movement brought on by the announcement. We used the standard event study methodology (Brown and Warner, 1985) with

the Abnormal Stock Returns formula $ARit = Rit - E (Rit)$. The results show that the negative announcements lowered the stock price and positive announcements increased the stock price was proven. The results were both statistically significant and economically significant and show implications for manufactured doubt strategies used by companies.

5.4 The Evolution of the Financial Technology Network

Presented by: Alisher Mansurov (*Nipissing University*)

Co-author: Marc Pilon (*HEC Montréal*)

The diffusion of financial technologies (FinTech) in financial services has drawn worldwide engagement. However, there is little known on the involvement of various types of organizations in the advancement of FinTech. In this study, we apply a network analysis on 23,000 FinTech news articles from 2008-2022 to illustrate the network of FinTech organizations and examine its characteristics. We also apply a textual analysis on the content of the news articles to identify trends in FinTech and distinguish between the collaboration elements of various organizations. We find that incumbent financial services providers are central in the FinTech network, while the prominence of technology giants, regulatory bodies, and FinTech start-ups has grown. Moreover, ties between similar organizations are stronger, although these connections have weakened over time. Our textual analysis results provide evidence for a shifting focus in FinTech over time and significant differences in associations between various groups of organizations.

Track 6: Student Research Projects

Room: WC G103B (2:00 p.m. -3:20 p.m.)

Moderator: Megan Johnson (*State University of New York at Fredonia*)

6.1 Franciscan and Money

Presented by: Fr. Jason Wooleyhan (*Saint Francis University*)

As our community of Franciscans age and decrease in number, we also have young men entering annually who are able to work and support the community. What does this mean for our sources of income as we change over time?

6.2 An Overview of Sport Analytics Methods

Presented by: Jacob Minniefield II (*State University of New York at Fredonia*)

Co-author: Reneta Barneva (*State University of New York at Fredonia*)

The competitive and performance landscape of sport is changing due to the power of data analytics. Research shows how data analytics in sports can help coaches and players improve performance along with other metrics that are part of their respective sport. Coaches, players, and other researchers can customize training plans using data analytics, which enhances athlete performance and reduces injury risks. Training regimens are tailored to each participant's requirements by analyzing data, metrics, and other significant markers while in play. This promotes how athletes are in a physical state while performing and how they can improve by using this data found in research.

Data analytics in sports has its roots in basic statistics and observational analysis. But as technologies have emerged and the number of sources to research data has increased—including wearable sensors, video tracking systems, and other performance monitoring wearables. The potential for data analytics in sports seems limitless as we move forward. The predictive power of analytics is about to reach previously unheard-of heights with the introduction of AI and other computer sourced research methods. With AI being such a used source across the world today, AI advancements within sport will become more common in the upcoming years. With my research conducting multiple ways that shows how great data analytics and other methods of data research can benefit coaches, athletes and other parts of sport, the combination of sports and data analytics is a mutually beneficial for both as it enhances competition and spectator engagement while raising the bar for athletics through data insights. The competitive advantage in this digitally advanced era goes to those who can use data to find success in sports and use the data to their advantage.

6.3 E-Commerce Supply Chain Risk Mitigation and Online Sales Performance

Presented by: Aayush Sharma Giri (Brock University)

Co-authors:

Shuai Yuan (Brock University)

Anteneh Ayanso (Brock University)

This paper explores the relationship between a retailer's commitment to risk mitigation through the adoption of web features and functionalities and its impact on online sales performance. While previous research has examined the impact of various retail service offerings on online sales performance, this study adopts a focused approach by investigating the impact of transactional, logistics, and post-sales service offerings. The resulting model proposes three major hypotheses, which are empirically tested using hierarchical multiple linear regression and further assessed for robustness using the binary logistic regression technique. A sample of 398 top retail companies operating in North America is used for the study. The findings reveal a positive relationship between a retailer's commitment to risk mitigation and its online sales performance. Specifically, the results indicate that retailers who proactively adopt risk mitigation strategies are more likely to be categorized as leaders rather than followers in the market. This study contributes to the existing literature on retail and risk mitigation by highlighting the importance of proactive risk mitigation through the adoption of web features and functionalities in the context of online retail services. It also provides guidance for practitioners and managers in assessing their market position and directing their risk mitigation strategies accordingly.

6.4 The impact of post-pandemic travel on Environmental, Social, and Governance frameworks

Presented by: Ankita Sinha (*University of Buffalo*)

Co-author: Namratha Pulluru (*University of Buffalo*)

The COVID-19 pandemic significantly disrupted the global travel and tourism industries, warranting a detailed analysis of its impacts on Environmental, Social, and Governance (ESG) criteria. This study analyzes how key factors affected by changes in travel patterns influence ESG

performance.

The study shows the complex interactions between the pandemic effects and the travel industry's ESG outcomes by employing structural equation modeling to evaluate the impact of the volume of commercial flights, travel costs, infrastructure expansion, employment and unemployment rates, and environmental impacts, among other related variables on ESGs. Initial findings indicate that a pronounced increase in commercial flight activity and a rise in commercial growth impact employment. These changes profoundly affect the 'Environmental' and 'Social' components of ESG, reflecting changes in carbon footprints and overall community well-being. Additionally, the study assesses the 'Governance' aspect, considering how industry responses and recovery strategies align with sustainable and responsible business practices. This paper enriches the discourse on the pandemic's repercussions on travel and tourism through the lens of ESG criteria, offering critical insights for stakeholders aiming to navigate the recovery phase while prioritizing sustainability and resilience. The analysis highlights the immediate ESG challenges and frames a discussion around long-term strategies for enhancing sustainability and governance after COVID-19, contributing to a more robust and ethical industry framework.

Track 7: Advanced Business Analytics

Room: WC G103C (2:00 p.m. -3:20 p.m.)

Moderator: **Joseph Porter** (*Nazareth University*)

7.1 3D Methods for Geospatial Business Analytics

Presented by: **Gabby Resch** (*Ontario Tech University*)

The ability to analyze and visualize geospatial data is a crucial skill for making sense of geopolitical phenomena (e.g. climate migration), global economic issues (e.g. supply chain disruptions), and public health crises (e.g. pandemics). Various contemporary geographic information systems provide innovative methods for analyzing, visualizing, and mapping data, from integrated geographically weighted regression tools to map-centered interactive dashboards to scrollytelling story maps. The recent adoption of 2.5D methods (e.g. "spike maps") in domains like data journalism, coupled with mixed reality/metaverse hype, has led to a flurry of interest in 3D tools for geospatial analytics. Industry leading platforms like Esri's ArcGIS now enable integration with game development software like Unity and Unreal Engine, making it possible to prepare immersive 3D dashboards and virtual reality interfaces for real-life business analytics applications. That said, little consensus has formed around appropriate use of 3D tools, ranking of visualization methods, development best practices, etc. It also remains to be seen what these trends will mean for business analytics/intelligence research and practice, where big data and AI-driven approaches have led to novel processing, aggregation, mining, and forecasting techniques, but far less attention has been given to interactive visualization methods. This paper will share insights from multiple iterations of a course on 3D methods for GIS that I developed and have taught in a Business and IT program. It will discuss which methods are intuitive, which ones confuse students, and which ones hold promise for business analytics.

7.2 Spatial Big Data and Shopping Mall Analytics: an exploration beyond the hype

Presented by: **Tony Hernandez** (*Toronto Metropolitan University*)

Co-author: **Joe Aversa** (*Toronto Metropolitan University*)

Mobile location data (MLD) in shopping mall analytics grew substantially during the pandemic. During this time of widespread business and consumer uncertainty, several major data vendors packaged MLD, a form of spatial big data, and marketed their data solutions to commercial landlords and their tenants with the promise of providing unparalleled insights into rapidly changing shopping behaviours. The ability to geofence commercial properties and extract data on visitors to reveal shopping patterns over time at a level of resolution not previously thought plausible created a rapid adoption of MLD-based analytics. This paper critiques MLD-based shopping mall analytics for a small set of major malls in the US for the pre-, during and post-pandemic period. Through a detailed examination of pathing level data, the analysis highlights the significant challenges associated with MLD and discusses the balance between the promise and the hype. The paper demonstrates the varied metrics and insights that can be developed using MLD data. However, significant issues related to data accuracy and coverage, bias and representativeness, inference and subjectivity, and the growing concerns over consumer privacy and the surveillance state are also explored. The paper identifies several key questions that remain to be addressed and form the basis of a future research agenda.

7.3 Exploring the Potential of Analytics to Better Inform Individuals Decisions: A Conceptual Approach

Presented by: Satya Chattopadhyay (*University of Scranton*)

Co-authors:

Richard O'Hara (*University of Scranton*)

Ramsha Siddiqui (*University of Scranton*)

The widespread embrace of analytics across disciplines, driven by the surge in available data, has become a cornerstone in decision-making. Operational analytics, marketing analytics, financial analytics, and business analytics are now integral components of academic programs in business. However, the predominant focus has been on managerial decision-making, benefits primarily measured in economic terms. This leaves a gap in practice and research of enhancing consumer decision-making using personal analytics. This study addresses this gap by exploring the use of Artificial Intelligence (AI) to empower individuals in making better decisions that consider economic as well as other criteria. This can be done by leveraging analytics to access previous personal experience data and relevant external data feeds over time.

Examining three trends—data proliferation, increased computing power, and advancements in AI and machine learning—this research explores their collective potential to aid individuals in making better, more informed decisions amid the growing complexity of daily choices. The study delves into the use of AI to analyze previous experiences, including individual preferences, learned preferences, and stream of received social media feeds, to develop a dynamic decision-making application targeted at individuals.

By addressing gaps in existing literature and available AI-driven tools, this research contributes valuable insights into the practical implementation of AI tools for individual decision-making. It proposes a framework that considers the multifaceted nature of human choices, emphasizing the adaptability of available tools to evolving needs and preferences across diverse contexts of decisions made on a day-to-day basis.

7.4 A Total Data-to-Model-to-Decision (TDMD) framework for Business Analytics

Presented by: Anthony Lee (*Eastern University*)

In our digitized and interconnected world, data has emerged as a critical currency in business. Achieving success in today's world necessitates transforming into a quant-savvy entity or professional, requiring a broad spectrum of quantitative skills across big data, science, statistics, analytics, business knowledge, and technology. The shift toward enhanced Speed-to-Understanding, propelled by AI and machine learning, is reshaping decision-making capabilities for quant-focused companies.

As the Business Analytics curriculum lead, I surveyed diverse sources of quantitative literature, textbooks, both degree and non-degree educational programs, and various social media outlets, finding a wealth of information. However, the insights often present fragmented and conflicting views such as the ongoing debate on terminologies regarding whether "Data Science" is a subset of "Data Analytics," or vice versa. This calls for a more comprehensive framework to navigate this complex quantitative landscape.

At this conference, I aim to present my distilled findings into a two-dimensional schematic: the Total Data-to-Model-to-Decision (TDMD) framework. Leveraging Systems Engineering and over 30 years of industry analytics-centric experience, this framework integrates a four-step decisioning process (Access, Analyze, Assess, Act) across three core disciplines (Data, Math-Modeling, Business) and data roles (Creators, Curators, Consumers). This TDMD model not only "follows the data" but also enables both small and large enterprises to calibrate their quantitative teams and equip them with a fitting skill set.

Track 8: Industry Insights

Room: WC S204A (2:00 p.m. -3:20 p.m.)

Moderator: **Mohammad Nasim** (*State of Michigan*)

8.1 Global Supply Chain Analysis of Electric Vehicles

Presented by: Robert Cutlip (*Fairmont State University*)

Co-authors:

Rebecca Giorcelli (*Fairmont State University*)

Hunter Baker (*Fairmont State University*)

Michael Gaskill (*Fairmont State University*)

Joshua Hite (*Fairmont State University*)

Austin Kay (*Fairmont State University*)

The global demand for electric vehicles (EVs) is rapidly increasing, and the supply chain plays a crucial role in meeting the increased demand. The EV battery supply chain includes the extraction of raw materials, the transportation of those raw materials to facilities, battery production, and eventual disposal or recycling of used batteries being phased out of the market. The process is further complicated by other variables such as the battery components being sourced from a variety of different countries, and final products are assembled in different countries from where the raw materials are extracted.

The extraction of raw materials for EV batteries, such as cobalt, is typically done in more developing countries such as the Democratic Republic of the Congo. Other minerals used in EV

batteries, such as nickel and lithium, can be sourced from Australia and Chile. These raw materials are then transferred to other countries for processing and refinement. The EV supply chain involves a wide range of stakeholders, some of which include mining companies, refining, and processing companies, and the vehicle manufacturers themselves. As the demand for EVs increases, there will be a significant amount of pressure being put on the supply chain, which makes it crucial for sustainability and resilience. In this research, the supply chain of battery and magnet mining and processing was analyzed using descriptive, predictive, and prescriptive analytical models. Our results indicate that diversification of the supply chain of rare earth oxides and metals will improve supply chain resilience in the near term.

8.2 Strategizing Success: Enhancing Engineering Consultancy with Data-Driven Proposal Analytics

Presented by: Dennis Voss (*Crozier: Consulting Engineers*)

In the competitive realm of engineering consulting, strategic planning and decision-making are crucial for business success. This presentation delves into the transformative impact of leveraging proposal analytics for enhanced performance. By utilizing a comprehensive set of metrics — including proposal numbers, win/capture rates, proposal value, and cross-selling — we conduct a multidimensional analysis to inform targeted strategies for each department. Our approach involves rigorous monthly, quarterly, and annual analyses that guide not only immediate adjustments but also long-term strategic forecasting. A particular focus is on predicting 12-month revenue based on won proposals. This data-driven methodology enables precise planning and significantly improved business outcomes.

A key innovation is the win prediction tool, which leverages five years of proposal data to forecast success rates. Project managers input potential proposal details, and the tool evaluates these against historical data to predict success probabilities. It highlights strengths and weaknesses, allowing managers to recalibrate strategies to increase win rates. This tool has become instrumental in refining our approach to proposal development and strategic bidding. Furthermore, we've developed a proposal/project online map that provides managers with historical and geographical data, enabling better-informed decision-making. This presentation will showcase the methodology, implementation, and tangible results of integrating sophisticated analytics into our business strategies, underlining the power of data in driving growth and operational excellence in the engineering consulting sector.

8.3 Words Matter: Evaluating the Communication of Data in Press Releases

Presented by: Travis Brodbeck (*Siena College*)

Co-author: Necip Doganaksoy (*Siena College*)

Founded in 1980, the Siena College Research Institute (SCRI) conducts regional, statewide and national surveys, of experts and the public, on economic, political, cultural, and social issues. These surveys conducted primarily via telephone interviewing provide ample findings for stakeholders to act upon and the media to discuss. Following the conclusion of a poll, SCRI creates a press release based on the analysis of the data that is distributed to journalists and media entities. Like the childhood game of telephone, the meaning of the findings of empirical survey data are translated into a narrative that is then adapted by members of the press for local and national stories. In a consolidated and often politicized media landscape, an important

question arises for this non-partisan organization: could SCRI's press release bias how the press covers the poll?

In an evaluation of press releases and the polling data tied to them, this paper looks at the biases that can creep into natural language in the form of a press release. The practical implications of this research provide a framework for research entities to evaluate their communication of findings to the public with respect to what the empirical analysis of data demonstrates. Using sentiment analysis and other text processing tools powered by artificial intelligence, this paper demonstrates how to visualize the sentiment of press releases and compare them to historical trends based on survey data. Additionally, this analysis compares the performance of human coding of sentiment versus the output generated by ChatGPT to provide insights into the reliability of generative AI tools for text processing.

8.4 Machine Learning Approaches for Predicting Wine Quality from Chemical Properties in Azure Machine Learning Studio

Presented by: Renu Mutha (*Northwood University*)

Co-author: Itauma Itauma (*Northwood University*)

This paper presents a study that uses Azure Machine Learning to predict wine quality based on its chemical properties. The dataset used in this study includes various chemical properties of wine along with quality evaluations, where the target variable is the wine quality rating. We evaluate several classification models, including VotingEnsemble, Random Forest, Decision Tree, and Logistic Regression. The objective of this study is to develop an accurate and efficient model for predicting wine quality based on chemical characteristics. The VotingEnsemble Classification method outperforms other methods with an accuracy of 87.5%. The findings of this study can be used by winemakers to improve their production process and enhance the quality of wine.

Concurrent Sessions (Saturday – 9:30am) Tracks 9-12

Track 9: Applied Economics 2

Room: WC G103A (9:30am -10:50am)

Moderator: Joseph Kuvshnikov (*Gannon University*)

9.1 Organic Farming Transitions: A Dynamic Bioeconomic Model

Presented by: Michael A. Meneses (*Cornell University*)

Co-authors:

Clare L. Casteel (*Cornell University*)

Miguel I. Gómez (*Cornell University*)

David R. Just (*Cornell University*)

Ravi Kanbur (*Cornell University*)

David R. Lee (*Cornell University*)

C.-Y. Cynthia Lin Lawell (*Cornell University*)

We develop a dynamic bioeconomic model of a farmer's decisions regarding the use of synthetic compounds (e.g., synthetic fertilizers and pesticides) and the transition from conventional to organic management. Our crop production model accounts for newly documented interrelationships among synthetic compound use, soil health, and crop yields. In particular, new insights from soil science show that the use of synthetic compounds can be harmful to beneficial soil microbes that improve agricultural yields by enhancing crop nutrient use, stress tolerance, and pest resistance. We characterize and solve for a "fully informed" farmer's optimal synthetic compound use strategy, and for whether and how a farmer should transition from conventional to organic farming. These solutions are compared to those from a "misinformed" model in which the farmer is not aware of the interactions between synthetic compound use, soil health, and crop yields, allowing us to assess how gaining knowledge of these interactions might be expected to change farmers' synthetic compound use strategies and, ultimately, their decisions around adopting organic management. We identify and discuss agricultural and economic conditions under which farmers can be expected to voluntarily reduce their reliance on synthetic compounds, and possibly even adopt organic management, upon learning of the benefits associated with stewardship of their soil's microbiome. We apply our model to farmer-level pesticide-use panel data to estimate parameters governing farmers' current understanding of the interrelations between soil microbes, pesticides, and crop yields, and to examine possible effects of extension programs targeting farmers' understanding of soil microbes.

9.2 Does your fandom really matter? The Effects of Heterogeneous COVID-19 NFL Stadium Attendance Restrictions

Presented by: Adam Cook (*State University of New York at Fredonia*)

Using a unique panel data collected and aggregated from Pro Football Reference (<https://www.pro-football-reference.com>), Sports Media Watch (<https://www.sportsmediawatch.com>), the 506Sports archives 2009-2022 (<https://archive.506sports.com/wiki/NFL>), NFL Weather (<https://www.nflweather.com>), NFL Penalties (<https://www.nflpenalties.com>) and Australia Sports Betting (<https://www.aussportsbetting.com>) websites, I test the impact of NFL fan attendance using heterogeneously-applied COVID-19 stadium attendance restrictions during the 2020 NFL season as an instrumental variable for fan attendance. I analyze how varying stadium attendance during this period affected various outcome variables including total, home and away team penalties, final score point differentials, TV viewership, Nielsen ratings, and total home and away team scores. Additionally, I estimate the effect of stadium attendance on point spread and over/under outcomes in NFL sports betting markets. I find that fan attendance significantly affected both home and away team penalties, away team scoring (while having no effect on home team scores), TV viewership and Nielsen ratings. The IV estimates indicate that for every additional 20000 fans in the stadium, away team score decreases by approximately 1 point. This result indicates the contribution of the fanbase to home team success.

9.3 Flexible Spending Account Allocations: Exploring Factors that Contribute to Allocation Decisions

Presented by: Melissa Waite (*State University of New York at Brockport*)

Co-author: Mustafa Canbolat (*State University of New York at Brockport*)

This study explores optimal Flexible Spending Account (FSA) allocations. An employer-provided benefit permitting employees to contribute pre-tax dollars to cover eligible medical expenses not paid by insurance, health care FSA funds are “use it or lose it,” requiring participating employees to estimate anticipated health expenses. Eligible expenses include co-pays, dental and vision care, medical devices, and over-the-counter items such as allergy medicine, pain relievers, sunscreen, DNA testing, acne creams, feminine hygiene, birth control, and first aid supplies, among others. With the passage of the 2020 CARES Act expanding the range of eligible items to include more over-the-counter medications and consumer goods, expending annual FSA funds is made easier, as consumers can deplete their FSA accounts at the end of the year with the purchase of eligible OTC items.

This study expands the field of estimating optimal contributions to FSAs by exploring two research questions: (1) Will subjects who are aware of reimbursable items and tax savings be more likely to contribute the optimal amount to an FSA? (2) Will subjects who experience a large, unexpected health expenditure change their ordering behavior (FSA contribution amount) in the future? These questions are tested using simulation experiments with a 2x2 factorial design to explore decision-maker behavior.

9.4 The heterogeneous effects of COVID-19 on young consumer purchasing patterns: An examination using causal random forest

Presented by: Youngran Choi (Embry-Riddle Aeronautical University)

The coronavirus (COVID-19) pandemic prompted governments worldwide to implement lockdown measures that resulted in an increase in online purchases. Young consumers, who are familiar with browsing online platforms, may respond differently than older consumers to the pandemic in terms of online purchases. The generalized random forest (GRF) was employed to identify the heterogeneous causal effects of COVID-19 on the online shopping habits of young consumers. The attributes the GRF analysis identified as sensitive to COVID-19 in inducing online purchases include: gender based on employment, race, television ownership, and specific types of grocery products (i.e., dry foods and frozen foods). Understanding the heterogeneity in the effects of COVID-19 that induce young consumers to shop online can provide practitioners with insights on how to prioritize and target high-impact groups with specific characteristics while improving management and policy strategies for future pandemics or other extended periods of disruption of in-store businesses.

Track 10: Operational Strategies

Room: WC G103B (9:30am -10:50am)

Moderator: Anthony Lee (*Eastern University*)

10.1 Analyzing the Predictive Patterns in Time Series of Building Electricity Consumption

Presented by: Hsien-Tseng Wang (*Lehman College, City University of New York*)

Co-author: Di Wu (*Lehman College, City University of New York*)

The predictability of time series is an important component that reflects the inherent

information contained in it and can be used to assist in evaluating the performance of forecasting models. Typically, the performance measure of a forecasting model outputs the probability of success but unfortunately does not provide an informative understanding for improvability in prediction accuracy. Recognizing this challenge, our research focuses on the intrinsic predictability, which is an intuitive metric that quantifies predictability levels in a time series. The intrinsic predictability of time series can be very informative in deciding whether the time series is predictable, or a chosen model fails to achieve accurate forecasting results. In our analysis, we utilize building electricity consumption datasets obtained from a higher education institute for evaluating the intrinsic predictability. Our goal is to investigate the feasibility of a unified metric that includes information entropy to result in the quantification of intrinsic predictability of time series data. By connecting our findings with wider fields of business analytics, we expect that the results of our study will not just help improve forecasting methodologies but also provide valuable insights for optimizing energy consumption strategies in broader business sectors.

10.2 Business Improvement Districts and the Homeless: Can Businesses Actually Help?

Presented by: Shahryar Gheibi (*Siena College*)

Co-author: Wonhyung Lee (*State University of New York at Albany*)

Contrary to the conventional perception that business improvement districts (BIDs) deploy a strong-arm approach toward chronically homeless populations, there is empirical evidence that several BIDs take a more holistic, long-term approaches that, not only are based on the framework of public space management and security, but also of service delivery and policy advocacy. While previous research revealed such a wide range of strategies that BIDs implement, there is limited understanding of what type of factors affect BIDs' attitude toward homelessness.

To examine the factors associated with BIDs' various approaches to homelessness, this study utilizes machine learning to analyze survey data conducted with BIDs regarding their experiences of homelessness and approaches to homeless populations at a district level. Our results suggest that while collaboration with law enforcement (a strong-arm strategy) is a reactive approach to solving issues generated by the homeless, the mere presence of homelessness serves as a significant driver for BIDs to consider homelessness a social phenomenon and adopt more holistic strategies such as establishing outreach programs.

10.3 Global Supply Chain Analysis of Pharmaceuticals

Presented by: Robert Cutlip (*Fairmont State University*)

Co-authors:

Jacob Morris (*Fairmont State University*)

Eugene Barbery (*Fairmont State University*)

Isaac, Stankus (*Fairmont State University*)

Rebecca Giorcelli (*Fairmont State University*)

The pharmaceutical industry will always be an evolving field, with new drugs and treatments always being researched and developed. In the context of the U.S. pharmaceutical supply chain, Analytics, Data Science, and Artificial Intelligence (AI) can be used to highlight the true extent of China's dominance in the U.S. pharmaceutical supply chain, despite the media obfuscating the reports. Utilizing AI, data science, and BI, pharmaceutical companies will have a better

understanding of United States supply chain to mitigate the risk of over-reliance on China for pharmaceutical imports. Descriptive analytical analysis indicated that the United States has experienced a 98.7% increase in imports from China since 2017. In 2021, the U.S. received 11,848,640 Kgs in antibiotics, and precursors from countries that have signed the Chinese Belt and Road Initiative (BRI). Total imports of antibiotics and precursors from these countries accounted for approximately 49% of the total U.S. imports. Predictive analytics of global supply indicated a heightened reliance on China and its BRI partners, particularly in antibiotics and its precursors.

Prescriptive analytics was also conducted based on data and trends from the descriptive and predictive analytical models. One recommendation is to establish Standards for Pharmaceuticals: The U.S. should coordinate their strengths with other leading countries to develop and enforce global standards for pharmaceutical products. This will prevent China from dominating the pharmaceutical industry and ensure the safety and efficacy of pharmaceuticals worldwide. Secondly, diversifying their Pharmaceutical Supply Chains to reduce the reliance on Chinese pharmaceutical supplies by encouraging the development of alternative partners for precursors, raw materials, and finished/manufactured products. Third, the United States should commit to stronger engagement in strategic partnerships with other countries to counter China's influence, particularly with the countries already involved within the BRI.

10.4 Global Supply Chain Analysis of Semiconductors

Presented by: Rebecca Giorcelli (*Fairmont State University*)

Co-authors:

Robert Cutlip (*Fairmont State University*)

Hayden Andersen (*Fairmont State University*)

Nicole Hall (*Fairmont State University*)

Brandon Arnal-Glasscock (*Fairmont State University*)

Global demand for semiconductors is rapidly increasing due to the use of electronic circuits in many consumer and defense-related products. Semiconductors are an integral part of an electronic circuit resident in many everyday devices such as cars, appliances, and phones, with the United States being the largest consumer of semiconductors next to China. Taiwan is the world leader for semiconductor manufacturing with 63% of the market share and their largest manufacturer, TSMC (Taiwan Semiconductor Manufacturing Company), produces 54% of the world's supply. However, advanced semiconductors are the most vital product and are used in laptops, cars, medical equipment, and phones. Taiwan absolutely dominates the market of advanced semiconductor manufacturing, being home to 90% of the manufacturing capacity. Compared to Taiwan and TSMC, other notable producing countries and manufacturers include South Korean company Samsung at 17% market share and Chinese company SMIC at 5%. As the demand for semiconductors increases, there will be a significant amount of pressure being put on the supply chain, which makes it crucial for sustainability and resilience. In this research, the supply chain of semiconductors was analyzed using descriptive, predictive, and prescriptive analytical models. Analysis indicated that Taiwan is the largest manufacturer of semiconductors, although they are not dominant in all steps of the supply chain. Many other countries are involved in the design, manufacturing, and processing, such as the USA with design or Japan with wafer production. Thus, continued diversification of the semiconductor supply chain will improve supply chain resilience in the near term.

Track 11: Insights and Innovations in Analytics

Room: WC G103C (9:30am -10:50am)

Moderator: Satya Chattopadhyay (*University of Scranton*)

11.1 Option Valuation with Conditional Heteroskedastic Hidden Truncation Models

Presented by: Rachid Belhachemi (*Le Moyne College*)

While asymmetric mixture models improve option pricing over generic pricing models, mispricing remains due to their inability to capture the effect of economic factors on price levels. This paper uses the hidden truncation normal \Leftarrow HTN \Rightarrow distribution introduced by Arnold et al. (1993) and the NGARCH model of Engle and Ng (*J Finance*, 48:1749–1778, 1993) to price options. Compared to the Black–Scholes model, the HTN -NGARCH option pricing model has extra parameters linked to economic dynamics and with economic interpretations. The model integrates some stylized facts underlying option prices such as a time-varying price of risk, non-normal innovations, asymmetry, and kurtosis. The model can be estimated by maximum likelihood. With an application to market data, we show that the HTN-NGARCH model accurately prices index options and captures adequately the smirk of implied volatility.

11.2 Missing Data Imputation – A Two-Phase Model

Presented by: Di Wu (*Lehman College, City University of New York*)

Co-author: Hsien-Tseng Wang (*Lehman College, City University of New York*)

In the domain of business analytics, data science, and information science, the challenge of addressing missing data is ever-present and often complicated to both scholars and practitioners. This paper introduces a novel two-phase model designed to navigate this challenge with precision and clarity. In phase 1, a supervised approach is deployed to comprehensively understand the unique characteristics of the specific dataset. This initial phase conducts an exhaustive evaluation of existing missing data imputation methods, including both traditional statistical techniques and advanced machine learning-based approaches. These methods are rigorously assessed to determine their applicability to the dataset's specific attributes. In phase 2, the selected methods from the previous phase are applied to the dataset for the actual missing data imputation. The contribution of this two-phase model lies in its practicality and its ability to resolve the common dilemma faced by practitioners when confronted with a myriad of imputation methods. By adhering to a structured and data-driven approach, this model minimizes the uncertainty and complexity that often surround the method selection process. The outcome is the confident and informed selection of the most effective imputation method fitted to the unique characteristics of the given dataset. In doing so, this approach contributes not only to enhanced clarity but also to increased effectiveness in the domain of missing data imputation, thus strengthening the foundations of sound data science and rigorous research practices.

11.3 Measuring Online Customer Satisfaction Based on Customer Reviews: Topic Modeling Method (AI)

Presented by: Gehan Dhameeth (*Wells College*)

Companies spend an extensive amount of money trying to retain their customers. However,

organizations have lost customers as a result of poor loyalty. This phenomenon is predominantly noticeable among online customers. The need for a tool to measure online customer satisfaction that goes beyond the capabilities of existing tools has been the root cause of the issue mentioned above. This quantitative study aimed to explore the online customer satisfaction measurement dimensions ground a model that can be applied across industries to measure and predict online customer satisfaction by administering an online survey via social media to 462 respondents using supervised and unsupervised machine learning techniques combined with the topic modeling algorithm (Latent Dirichlet Allocation - LDA). This study discovered a significant relationship between the independent variables (navigation, playfulness, information quality, trust, personalization, and responsiveness) and the dependent variable (online customer satisfaction) using multiple linear modeling (LSM). Furthermore, navigability and playfulness were the most significant factors influencing online customer satisfaction relative to other independent variables. This study also discovered that the phenomenon occurs in both age groups, young and old customers. The limitations include being a cross-sectional study with challenges establishing external validity, having a narrow focus by only considering the retail sector (B2C), and being limited to the United States market.

11.4 Replacement or enhancement: Future of academic research in the emergence of AI

Presented by: Mary Han (*Toronto Metropolitan University*)

Validity and reliability are important for research as its result need to be generalizable. For decades, management research has been using limited number of samples ranging from thirty participants for qualitative method to ten thousand using quantitate method. These research draw results from limited samples size and have been published. Some are leading work in their discipline. AI is not newly emerged but its impacting in our life is. Ranging from our daily life to our management research, it is becoming more obvious as scholars are drawing from AI in their respective fields. Will AI overturn decades of research findings if we test them all with AI? Though it's not likely that all research will be rerun, but scholars have compared traditional research methods to AI based methods and found alarming results. Comparative research found the benefits of AI in academic research is not limited in size of sample, that AI based research can "undertake", "read" and "analysis" more data entries; speed of results generation, that it is faster and easier to input data and faster in generating results; but also, accuracy of findings, that results are actually more accurate than traditional methods. These have important implications to the future of academic research. What shall we do? Where do we go from here? Shall academic research proceed to replace logistic regression with XGBoost? Or developing some methods to combine, re-bundle or reconfigure the research models so that AI can enhance our research and results, if so, how? We conclude with theoretical and practical implications and directions for future research.

Track 12: Management and Education

Room: WC S204A (9:30am -10:50am)

Moderator: **Andrea Rodgers** (*University of Miami*)

12.1 Merging Project Management with Data-Driven Student Performance Evaluation: A Transformative Approach

Presented by: **Eren Akdur** (*Ithaca College*)

In project management, the Schedule Performance Index (SPI) is a crucial metric that assesses the efficiency of work completion and time utilization throughout a project's lifecycle. This metric serves as a valuable tool for project managers, offering a quantitative measure of schedule adherence and aiding in timely decision-making to ensure project success.

In the educational context, leveraging SPI involves adapting this project management principle to gauge student progress and success. By aligning educational objectives with planned learning outcomes, educators can employ SPI as a measure to assess how effectively students are advancing through their courses. This adaptation requires a clear delineation of learning objectives, meticulous planning, and the incorporation of data analytics to track student performance over time. Utilizing SPI in education thus empowers educators to provide timely feedback, predict potential challenges, and optimize interventions, contributing to a more adaptive and proactive learning environment.

This presentation explores the innovative fusion of Data Analytics and Project Management for effective decision-making in student success and progress track. The strategic integration of Project Management's Schedule Performance Index (SPI) into data-driven methodologies for measuring student course success serves as the cornerstone of this transformative approach.

12.2 Leveraging Large Language Models for Predictive Analytics in Student Performance Classification

Presented by: **Itauma Itauma** (*Northwood University*)

In today's competitive educational landscape, accurately predicting student performance is crucial for enhancing academic outcomes and ensuring student success. Traditional machine learning (ML) techniques have been widely employed for student performance classification, but their effectiveness is often limited by the structured nature of their input data. This paper explores the application of large language models (LLMs), specifically GPT, powered by the OpenAI API, for predictive analytics in student performance classification.

The Higher Education Students Performance Evaluation dataset, encompassing personal questions, family questions, and education habits, is utilized as the foundation for the study. The performance of GPT-based LLMs is compared against traditional ML algorithms, such as logistic regression and decision trees, in predicting students' end-of-term performances. LLMs have been shown to capture and analyze the nuances of unstructured text data, leading to more accurate performance predictions.

LLMs can provide valuable insights into the factors influencing student performance, enabling educators to tailor interventions and support strategies to address individual student needs. Additionally, LLMs can be continuously updated with new data, ensuring their adaptability and relevance in the ever-changing educational landscape.

In conclusion, this paper explores the transformative potential of LLMs in predictive analytics for student performance classification. By leveraging the power of LLMs, educational

institutions can gain a deeper understanding of student factors, optimize resource allocation, and ultimately enhance student success rates.

12.3 Introducing Music Analytics to Undergraduate Students

Presented by: Reneta Barneva (*State University of New York at Fredonia*)

Co-authors:

Anthony Casuccio (*State University of New York at Fredonia*)

Stuart Shapiro (*State University of New York at Fredonia*)

In the last decade, the music industry was profoundly transformed by the new technologies and platforms and the accumulation of big data. Several new business models have emerged. Data became an asset and a key aid for song producers, artists, labels, listeners, and professionals across all fields of the music ecosystem. In particular, it is widely used for music recommendation, song analysis, promotional strategies, decision-making, and pricing.

Music analytics skills and expertise are in high demand and the specialists are very well paid. There are several music analytics tools in the market, such as Soundcharts, Chartmetric, and Mordor Intelligence, that provide information to stakeholders. The music analytics applications extend to other fields such as gaming, social media, and films.

The supply of music analytics education and training is limited, both nationally and internationally. There are very few academic programs that offer music analytics as a major or minor, as most of these programs are in specialized and highly priced universities. The majority of them are at master's level.

Therefore, there is a need for a more comprehensive and integrative program that covers the various dimensions and disciplines related to music analytics. In this presentation, we will propose an approach towards music analytics education in the settings of a small liberal arts college.

12.4 ato: A Phonetic Alphabet to Teach Reading and Spelling!

Presented by: Edward McClenathan (*State University of New York at Fredonia*)

Despite our best efforts, the U.S. ranks very low in Reading literacy. On that educators, political leaders, and the public agree. Something must be done.

Proposal: A teaching phonetic alphabet that works and is also spell-friendly. The alphabet is called ato which stands for Adjusted Traditional Orthography.

History: ITA, a Pitman alphabet was tried for a while in Fredonia schools. (It was an odd alphabet that worked well at first, but didn't feature spelling.) The ato alphabet, which did emphasize spelling, was part of an experiment with West Seneca pupils. A full report can be found in a University of Buffalo dissertation: A Comparison of Three Alphabetic Media Used in Beginning Reading Instruction. (The three alphabets were Traditional Orthography, Initial Teaching Alphabet, and my ato alphabet.) Dr. Richard Sheil (SUNY Fredonia Emeritus) an expert in linguistics, believed in ato so much that he made the alphabet available for printing.

What's needed: A team to develop, publish and promote the ato Reading Instruction Method. (The success of ato would bring national attention to SUNY Fredonia and to the persons involved.)

Best Paper Awards

Best Student Paper Award –

Textual Analysis of Beige Books to Predict Regional Economic Changes

: Charlie Smith (University of South Alabama)

Best Faculty Research Award –

Predicting Bankruptcy: Ask the Employees

: Kristina Lalova (Michigan State University) & John Knopf (University of Connecticut)

Fredonia School of Business

The Fredonia School of Business prepares future business leaders by providing the knowledge, skills, and real-world experience necessary to compete in a global business environment. We offer a wide range of business programs to prepare you for success in Accounting, Public Accountancy, Finance, Management, Marketing, Business Administration, Music Industry, Sport Management, and Economics. All programs in the School of Business will introduce you to the latest technology, make you aware of the global and cultural components of business, and provide you with experiential learning and practical internship opportunities. Our curriculum emphasizes critical thinking, communication skills, business ethics, and professionalism.

The School of Business sustains the tradition of excellence at Fredonia, which is consistently ranked among the finest public universities in the Northeast -- and among the most affordable options in higher education. Small class sizes lead to close interaction with our faculty and the opportunity to conduct research at the undergraduate level - a hallmark of a Fredonia education. In addition, all students have opportunities for internships, available locally, regionally and across the nation.

Our students have access to the Center for Innovation & Economic Development (CIED), which supports start-up companies committed to Western New York. The Center is a New York State certified business incubator operated by SUNY Fredonia. Formerly called the Fredonia Technology Incubator (FTI), the CIED provides entrepreneurs with access to work-ready space, business consulting, mentoring and training, professional services, assistance with business formation, networking opportunities, and connections to university resources and student interns. Students work with entrepreneurs on market research, business plan development, operations and record keeping, and unique problem-solving initiatives. The Center expands the entrepreneurial aspects of our curriculum, enhances internship and employment possibilities, and inspires students to start businesses of their own.



Customer Churn Prediction with Machine Learning

¹A. Maazallahi, ¹Y. M. Meda, ¹B. M. S. Bokka, H. Jamallamudi, ¹P. Norouzzadeh, ²E. Snir, ¹B. Rahmani

¹Saint Louis University, Computer Science, Saint Louis, MO

²Washington University in Saint Louis, Olin Business School, Saint Louis, MO

Corresponding Author:

Dr. Bahareh Rahmani, Bahareh.Rahmani@slu.edu

Abstract

This article explores the prediction of customer churn in the banking sector using various machine learning models. It comprehensively analyzes extensive customer data, employing techniques including linear regression, decision trees, KNN, Naive Bayes, and Random Forest. The study emphasizes visual data interpretation through scatter plots, histograms, and box plots. It concludes with an assessment of ensemble methods, highlighting the superiority of Random Forest in predicting churn, offering crucial insights for customer retention strategies in banking.

1. Introduction

In the contemporary banking landscape, customer churn represents a significant challenge, directly impacting financial stability and long-term customer relationships. This study delves into this issue, utilizing a robust dataset of banking customers' profiles. The dataset's breadth and depth provide a rich foundation for analysis, aiming to reveal underlying patterns and predictors of churn. This introduction sets the stage for exploring various machine learning methodologies and their applications in predicting customer churn, highlighting the critical role of data-driven strategies in enhancing customer retention and fostering business growth in the banking sector.

In recent years, the banking sector has increasingly turned to machine learning for insights into customer retention. This shift is driven by the need for more sophisticated predictive analytics capable of navigating complex customer behavior patterns. Machine learning models such as KNN, Naive Bayes, and Random Forest have been at the forefront of this transformation. Each model offers unique perspectives on customer data, with Random Forest gaining prominence for its accuracy in churn prediction. However, despite advancements, challenges remain, notably in model interpretability and data privacy. Addressing these challenges is essential for the next wave of predictive analytics in banking, an endeavor that this study contributes to.

In our study, we employed a robust methodology to predict customer churn in the banking sector. The analysis began with a diverse dataset from Kaggle, encompassing various aspects of banking services. We utilized an array of data visualization techniques, including scatter plots, histograms, density plots, and box plots, to uncover key patterns. The core of our methodology involved the application of multiple machine learning models: linear regression, decision trees, KNN, Naive Bayes, and Random Forest. The focus was on evaluating the predictive accuracy of these models.

Our findings reveal that among these machine learning models, Random Forest emerged as the most effective in predicting customer churn. This superiority is attributed to its ability to handle large datasets and its robustness against overfitting. The discussion highlights the implications of these findings for the

banking sector, emphasizing the potential for enhanced predictive capabilities in customer retention strategies.

2. Related Works

The study "Regression-Based Machine Learning Framework for Customer Churn Prediction in Telecommunication Industry" presents a comprehensive machine learning framework using regression analysis to predict customer churn in the telecom sector. The research evaluates nine different regression models, focusing on computational efficiency and predictive accuracy. It emphasizes the importance of accurate churn prediction in maintaining a competitive edge in the telecom industry, particularly due to the high costs associated with acquiring new customers compared to retaining existing ones [1].

In the field of customer churn prediction, various approaches and methodologies have been explored. For instance, a study conducted by Wai-Ho Au proposed a data mining algorithm called Data Mining by Evolutionary Learning (DMEL) for classification problems. This method involves an evolutionary approach to search through possible rule spaces, beginning with the generation of an initial set of rules and iteratively finding higher-order rules. It focuses on identifying viable rules and estimates the likelihood of predictions to rank customers according to their likelihood to churn. However, it struggles with handling high-dimensional data [2].

Bart Larivière proposed using random forests techniques for predicting customer churn. This method involves two types of random forests: one used for binary classification and the other for models with linear dependent variables. These techniques have shown better performance than ordinary linear regression and logistic regression models, although they do not consider the correlation between variables [2].

Another significant approach is the use of deep learning techniques, such as the BiLSTM-CNN architecture. This method leverages the strengths of bidirectional LSTM and convolutional neural networks to extract features from a customer's data and predict churn. The BiLSTM-CNN architecture is particularly effective for capturing long-term dependencies in the input sequence, which is crucial for predicting customer churn [3].

In the context of dataset handling, it's important to note that accuracy is often used as a primary metric in churn prediction studies. However, due to the typically unbalanced nature of datasets in this field, accuracy may not always be a suitable metric. Techniques like oversampling and dimensionality reduction through feature selection have been employed to address these challenges and improve prediction efficiency [4].

The article "Customer churning analysis using machine learning algorithms" by B. Prabadevi, R. Shalini, and B.R. Kavitha examines the application of various machine learning techniques in predicting client churn. It underscores the importance of client retention programs for businesses. The study evaluates different algorithms like stochastic gradient booster, Random Forest, KNN, and Logistic Regression, providing insights into their relative effectiveness without specifying exact performance metrics. This research offers a significant contribution to strategies for enhancing customer retention through early churn detection [5].

The study "Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review" discusses the significance of churn prediction in various service-based industries. It emphasizes how crucial customer retention is due to the higher costs associated with acquiring new customers. The review covers various machine learning and deep learning techniques used in churn prediction, including algorithms like Naive Bayes, SVM, KNN, logistic regression, decision trees, Random Forest, AdaBoost, XGBoost, and artificial neural networks. It also discusses performance metrics beyond accuracy, highlighting their importance in churn prediction models [6].

In parallel, Sigme delves into the effectiveness and nuances of using machine learning for customer churn analysis. These studies extend the application of machine learning from theoretical models to practical solutions, underscoring its role in enhancing predictive accuracy in diverse sectors, including banking.

Their collective contributions lay a foundational framework for our understanding of churn dynamics, significantly informing the methodologies and interpretations in our study [7].

The article "Customer churn prediction in telecom sector using machine learning techniques" explores the application of classification algorithms like Random Forest, KNN, and Decision Tree Classifier for predicting customer churn in the telecom industry. The study emphasizes the development of efficient churn prediction models to aid in customer retention strategies. It stands out for its high accuracy in predicting churn, providing valuable insights for the telecom sector to reduce customer turnover [8].

Tsai & Lu proposed a hybrid neural network composed of ANN and self-organizing maps (SMO), where data are clustered by SOM and then used for ANN training. This hybrid approach has shown optimal accuracy in test settings. Additionally, deep learning methods like neural networks have been increasingly used in churn prediction due to their ability to adaptively extract higher-level features and reduce the workload of manual feature engineering [9].

Each of these approaches offers unique advantages and challenges, contributing to the evolving landscape of customer churn prediction methodologies.

3. Data Description

The dataset¹, integral to our study, encompasses extensive details about banking customers' engagement with various financial products and services. It includes specifics on account types such as savings, checking, credit card, and loan accounts, reflecting the diversity of financial products utilized by customers. This part of the dataset offers insights into the range and depth of customers' financial involvement with the bank.

The dataset includes various attributes essential for churn prediction. 'RowNumber' serves as a unique identifier for each record. 'CustomerId' and 'Surname' provide personal identification details, while 'CreditScore' reflects the customer's creditworthiness. 'Geography' and 'Gender' offer demographic insights, and 'Age' is crucial for understanding customer lifecycle stages. 'Tenure' indicates the duration of the customer's relationship with the bank. 'Balance' and 'NumOfProducts' show the depth of financial involvement. 'HasCrCard' and 'IsActiveMember' offer insights into the customer's engagement with the bank's services. 'EstimatedSalary' provides an economic background, and 'Exited' indicates whether the customer has left the bank. These data has been collected in France, Spain and Germany.

The scatter plot matrix Figure 1 provides a visual examination of the relationships and distributions among the quantitative features of the churn modeling dataset. From this matrix, we observe that age has been excluded due to its linear relationship with credit score, which itself exhibits a Gaussian distribution. The remaining quantitative attributes—CreditScore, Tenure, Balance, NumOfProducts, and EstimatedSalary—do not display distinct linear relationships, suggesting more complex interactions. Clusters are not readily apparent, indicating that customer behaviors and attributes are diverse and possibly influenced by a range

¹ <https://www.kaggle.com/datasets/shubh0799/churn-modelling>

of factors. The qualitative attributes, while not displayed in this scatter plot matrix, add additional dimensions to the analysis.

The histograms for the dataset's features **Error! Reference source not found.** reveal various distributions: Credit Score forms a symmetric, unimodal distribution, suggesting a majority of customers have a credit score around the mean. Tenure shows a relatively uniform spread, indicating customers are equally likely to have any tenure length. Balance has a left skew, with most customers holding a moderate balance and fewer with very low balances.

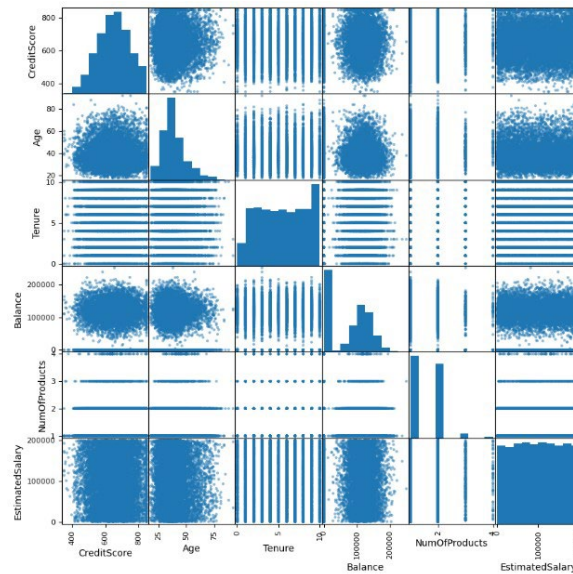


Figure 1 A scatter plot matrix visually examines relationships among quantitative features in the churn modeling dataset, revealing complex interactions and diversity in customer behaviors and attributes, with age excluded due to its linear relationship with credit score.

NumOfProducts exhibits a discrete distribution, with a concentration of customers holding one to two bank products. Lastly, Estimated Salary is uniformly distributed, indicating salaries are spread evenly across the range without bias towards any specific income level.

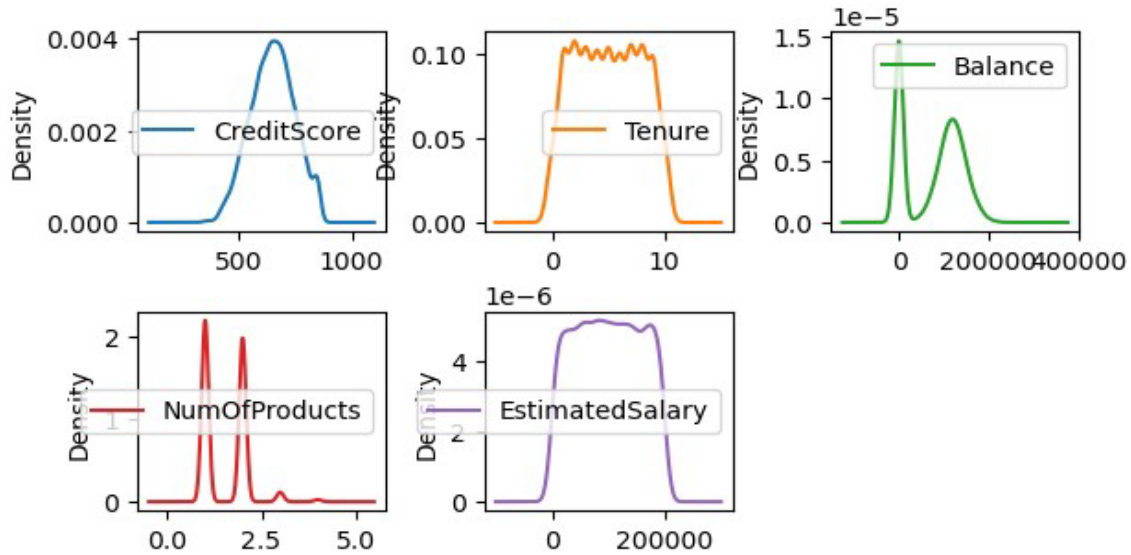


Figure 2 Density plots visually depict data: Credit Score is normal with most in the median, Tenure is uniform, Balance clusters around a central value, NumOfProducts shows fewer products for most, and Estimated Salary is evenly distributed across salary levels.

The Balance variable displays a concentration of values around a central balance amount with a tapering distribution towards higher values. The NumOfProducts plot reveals that customers tend to have a fewer number of products, with a stark drop as the number increases. Lastly, the Estimated Salary plot suggests an even distribution across various salary levels, indicating no significant concentration within any salary range.

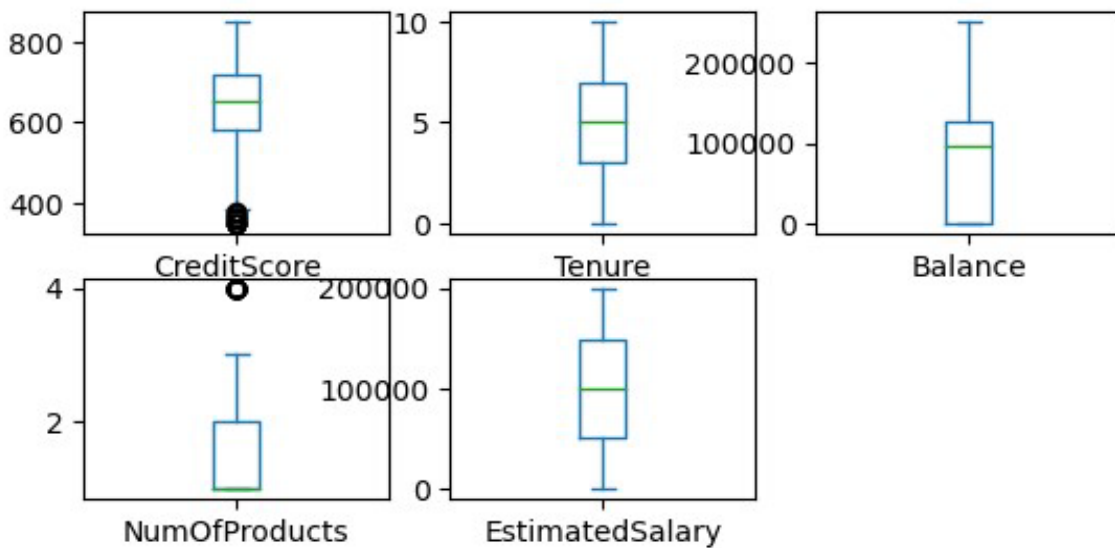


Figure 3 Box plots reveal key statistics: Credit Score has lower outliers, skewing towards higher scores; Tenure is evenly distributed; Balance concentrates within the interquartile range; NumOfProducts shows low product counts for most; Estimated Salary is uniformly distributed, reflecting diverse income levels among customers.

The box plots for the dataset features Figure 3 convey key statistical insights: Credit Score displays outliers on the lower end, indicating a few customers with scores significantly below the average. The large outliers with low credit scores skew the data. Tenure is evenly spread across its range, with quartiles distributed

equally, reflecting a uniform duration of customer engagement with the bank. Balance shows a concentration of values in the interquartile range without lower outliers, whereas NumOfProducts indicates most customers have a low product count. Estimated Salary is uniformly distributed, similar to Tenure, with a wide range indicating diverse income levels among customers. The mentioned variables are dependent values to predict.

4. Classification Methods

4.1 Linear Regression

The linear regression with only Age predicting customer churn is used as an initial analysis. The linear model, summarized by the equation $y=0.01095741x - 0.22278197$, fails to capture the complexity of churn behavior, suggesting the need for more sophisticated models.

Using a cutoff of 0.5 for classifying outcomes, the vast majority of customers are classified as staying with the bank. The confusion matrix reveals the model's performance. It has 7677 true negatives and 73 true positives, indicating successful predictions for non-churners and churners, respectively. However, there are 286 false positives and 1964 false negatives, highlighting a substantial number of incorrect predictions, particularly in identifying actual churners, which underscores the model's limitations in predicting churn accurately.

4.2 Decision Tree

The Decision Tree model Figure 4, known for its superior performance in classification tasks, has shown promising results. By utilizing an ensemble of decision trees, it mitigates the overfitting issues often seen in individual trees and enhances the overall prediction accuracy. The model's strength lies in its ability to handle the complexity and non-linearity in the data, making it a robust choice for predicting customer churn.

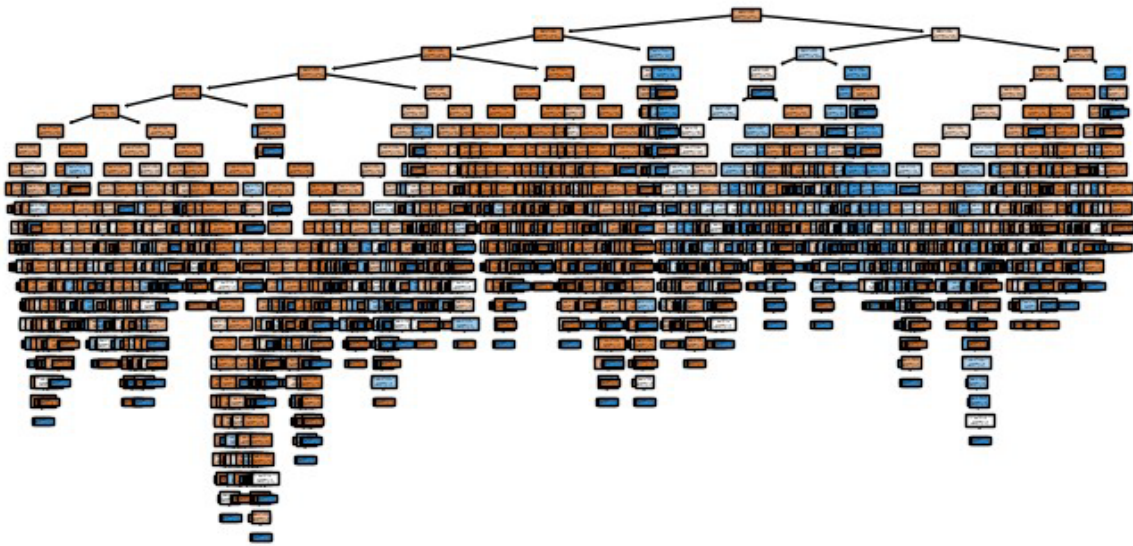


Figure 4 The Decision Tree model, known for superior classification performance, mitigates overfitting using decision tree ensembles, handling data complexity and non-linearity effectively for predicting churn. The Decision Tree achieved a commendable 77% accuracy, with a pure dataset (Gini value of 0.3263), serving as a reliable baseline model for churn prediction, though further accuracy improvement is possible.

The Decision Tree model exhibited a commendable 77% accuracy for the testing data, indicating a strong probability in churn prediction. The root node's Gini value of 0.3263 suggests that the dataset is relatively pure, enabling the Decision Tree to provide a reliable prediction. Although there's room for improvement in accuracy, the Decision Tree serves as a solid baseline model.

4.3 K-Nearest Neighbors

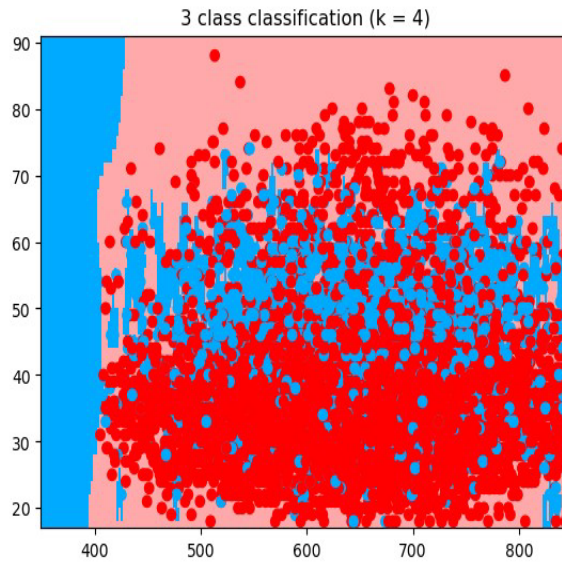


Figure 5 The KNN algorithm's optimal number of neighbors appears to be around four, striking a balance between avoiding overfitting and achieving good generalization, though it struggled with mixed clusters and proved unsuitable for this project.

The KNN algorithm's performance Figure 5 as evaluated by the elbow graph suggests that the optimal number of neighbors is around four. This balance helps to avoid overfitting that can occur with a very low k value. However, the model's performance does not significantly improve with increasing neighbors beyond this point, indicating a good level of generalization has been achieved.

In practice, the KNN model struggled with the dataset, showing mixed clusters which made accurate prediction challenging. Due to these limitations and the model's sensitivity to the proximity of data points from other clusters, it was deemed unsuitable for this project. 4.4 Naive Bayes

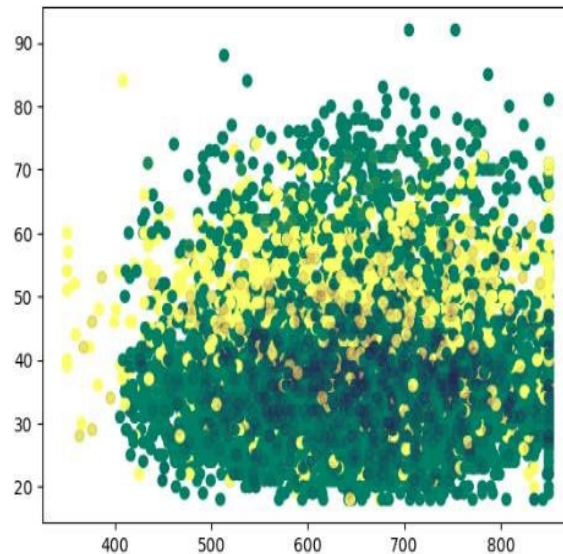


Figure 6 The Naïve Bayes model exhibits reasonable accuracy in the training data but with potential errors due to outliers, and while stability is seen in the test data's elbow graph from four clusters, refining for outliers is computationally excessive, making it unsuitable for this dataset's predictive goals.

The Naïve Bayes model's scatter plot Figure 6 reveals a reasonable prediction accuracy within the training data, indicated by the 'X' marks clustering around the correct outcomes. However, the presence of several outliers suggests potential predictive errors. The elbow graph for the test data demonstrates stability in accuracy starting from four clusters, but the computation required to refine the clustering for outliers is considered excessive. Therefore, it's concluded that the Naïve Bayes model is not the best fit for this dataset and its predictive goals.

4.5 Random Forest

The Random Forest algorithm employs a collection of decision trees to make predictions. Each tree votes, and the most common outcome is chosen. This ensemble method outperforms a single decision tree by reducing overfitting through result averaging. Implementing 100 trees in our analysis yielded an 80.35% accuracy. Compared to the 77% accuracy achieved with a single decision tree in a prior assessment, Random Forest demonstrates a notable improvement. Thus, for predicting the 'Exited' variable, Random Forest proves to be the more accurate model Figure 7.

5. Conclusion

The investigation into customer churn within the banking sector, underpinned by a robust dataset, has culminated in significant discoveries. Through meticulous analyses utilizing histograms, density plots, and box plots, this study illuminates the distributions and relationships of key quantitative variables. Scatter plot evaluations specifically dictated the exclusion of 'Age' due to its minimal predictive value.

Subsequently, the study ventured through various predictive models, where Random Forest and decision trees were identified as the most promising, with the former achieving an 80.35% accuracy rate, a noteworthy improvement over the 77% accuracy of the latter. The binary nature of the dataset—churn versus retention—was pivotal in guiding the choice of Random Forest, an ensemble method that demonstrated robustness against overfitting and sensitivity to individual decision trees.

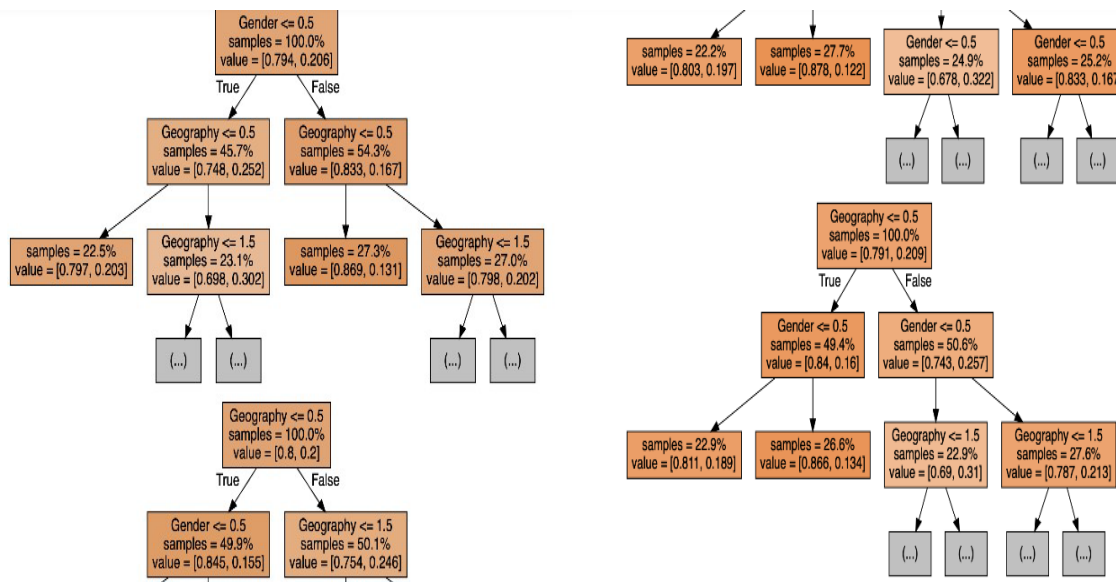


Figure 7 The chart highlights the superiority of the Random Forest algorithm, which uses an ensemble of decision trees to predict the 'Exited' variable. Employing 100 trees, it achieves an 80.35% accuracy, showcasing a significant improvement over a single decision tree's 77% accuracy in prior assessments.

This analysis has not only provided insights into feature distributions and relationships but has also underscored the importance of model selection in predicting customer churn. The findings advocate for the strategic use of ensemble methods in predictive analytics, reinforcing their potential for operational application within the financial sector to mitigate customer attrition.

References

- [1] Ele, S. I., Alo, U. R., Nweke, H. F., & Ofem, O. A. Regression-Based Machine Learning Framework for Customer Churn Prediction in Telecommunication Industry. *Journal of Advances in Information Technology*, 14(5), 1046, 2023.
- [2] Johnny, C., & Mathai, P. Customer Churn Prediction: A Survey. *International Journal of Advanced Research in Computer Science*, 8(5), 2017.
- [3] Khattak, A., Mehak, Z., Ahmad, H., Asghar, M. U., Asghar, M. Z., & Khan, A. Customer churn prediction using composite deep learning technique. *Scientific Reports*, 13(1), 17294, 2023.
- [4] Nhu, N. Y., Van Ly, T., & Truong Son, D. V. Churn prediction in telecommunication industry using kernel Support Vector Machines. *PLoS ONE*, 17, 2022.
- [5] Prabadevi, B., Shalini, R., & Kavitha, B. R. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4, 2023.
- [6] Raj, A., & Vetrithangam, D. Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review. *Proceedings of International Conference on Computational Intelligence and Sustainable Engineering Solution, CISES 2023*.
- [7] Simge E. Churn Prediction using Machine Learning, Soner Yıldırım. (2020, April 13). *Churn Prediction with Machine Learning. Towards Data Science*, 2020.

[8] Wagh, S. K., & Wagh, K. S. Customer Churn Prediction in Telecom Sector Using Machine Learning Techniques. SSRN Electronic Journal, 2022.

[9] Wu, H. A High-Performance Customer Churn Prediction System Based on Self-Attention. SSRN Electronic Journal, 2022

An Empirical Analysis of the Functionalities and Confidence Scoring Mechanisms in Leading Large Language Models

Mohammad Nasim
AF, Northwood University
Midland, Michigan, USA
nasimm48@northwood.edu

Shanshan Zhu
CSP, Mount Holyoke College
South Hadley, MA, USA
zhu36s@mtholyoke.edu

Itauma Itauma
AP-DC, Northwood University
Midland, Michigan, USA
itaumai@northwood.edu

Abstract— The proliferation of Large Language Models (LLMs) like OpenAI's GPT series, Google's LaMDA, and others has marked a significant advancement in artificial intelligence. These models are useful for a variety of applications, such as conversational bots and sophisticated problem-solving tools, because they can produce text that appears human because they have been trained on large datasets. The functionality and Confidence Scoring benchmarking techniques that support these models are empirically analyzed in this research. We examine the operational capabilities and architectural subtleties—such as token restrictions and tokenization techniques—that are essential to characterizing the functionality and reach of language learning models. Using a range of benchmarks, including HelloSwag, TruthfulQA, MMLU, DROP, MMMU and ARC, which are all intended to measure distinct facets of language comprehension and reasoning, our research methodologically evaluates diverse models. The findings reveal significant variability in model capabilities, highlighting strengths in areas like reasoning and weaknesses in handling biased data or maintaining factual accuracy. Additionally, we examine the settings and controls like temperature and top-p controls that influence the text output's coherence and randomness. In addition to helping to comprehend the state of LLMs now, this research will help shape future advances in AI by ensuring that models are more moral, effective, and compliant with legal requirements and user requirements. In order to fully utilize AI while upholding essential moral and security criteria, LLMs must continue to evolve and be benchmarked.

Index Terms— *Large Language Models (LLMs), Artificial Intelligence (AI), Tokenization, Model Benchmarking, Performance Evaluation, Text Generation, Natural Language Processing (NLP), Computational Linguistics, Ethical AI, AI Regulation, Model Variability, Dataset Evaluation, Unsupervised Learning, Multi-modal Data, Reasoning Tests*

I. INTRODUCTION

Large Language Models (LLMs) are a game-changing advancement in artificial intelligence (AI) that use massive datasets to produce text that closely resembles human language in terms of nuance and complexity. This advancement is exemplified by models such as OpenAI's GPT-3 and GPT-4, Google's LaMDA and PaLM, and Meta's LLaMA, among others, each distinguished by their distinct structures and capabilities. This study aims to conduct an empirical analysis of these models, with a particular emphasis on their capabilities and the confidence rating system, which serves as a benchmark for evaluating the models' effectiveness and usefulness in practical settings. The ability of LLMs to comprehend and produce text is usually the basis for their evaluation, as it is essential for a variety of applications, from advanced research assistants to automated customer care. These models' fundamental mechanism is tokenization, which divides the input text into digestible chunks like words, subwords, or characters. The model's ability to handle context and its maximum input length are both impacted by this procedure, which also has an impact on the model's output quality and range of possible tasks. Furthermore, as LLMs become larger and more sophisticated, benchmarking becomes crucial to comprehending their strengths and weaknesses. Benchmarks like MMLU, DROP, MMMU,

ARC, HelloSwag, and TruthfulQA offer information on a model's capacity for understanding, reasoning, and knowledge application. By highlighting which models function best in particular scenarios and tasks, these benchmarks assist developers and users in selecting the most appropriate model for their requirements. As the use of LLMs increases, ethical and biased debates become more prevalent, underscoring the necessity for models that not only function well but also do so in a way that is morally upright and devoid of bias. In this paper, we explore how various LLMs address these issues, offering an insightful look at the state of the field now and outlining potential directions for future development. In order to guarantee that AI development complies with more general societal norms and legislation, our analysis seeks to provide stakeholders with the knowledge they need to use these technologies efficiently and responsibly.

II. LITERATURE SURVEY

Large Language Models (LLMs) have been the subject of much documentation in recent literature, which shows how they have progressed from simple text processing tools to sophisticated systems that can carry out a wide range of intricate tasks. The discipline of natural language processing (NLP) has been

transformed by transformer-based architectures, which were first made possible by Raffel et al. (2020) in their early research on models such as GPT. Subsequent research and more recent generations, such as GPT-3 and GPT-4 by OpenAI et al. (2021), have continued to push the envelope, exhibiting models with better contextual awareness and token capacity. In order to examine the understanding and predictive skills of these models, research by Zellers et al. (2019) introduced benchmarks such as the Hellaswag datasets, establishing a norm for evaluating the practical utility of LLM outputs. These benchmarks evaluate the capacity of a model to produce language that is both cohesive and contextually suitable, thereby offering a measurable indication of its linguistic proficiency. LLM research also focuses on investigating model efficiency and scalability, as highlighted by Big Data Demystified et al. (2024). Their research investigates the correlation between the size of a model, the amount of training data, and its performance. The findings indicate that larger models, albeit more proficient, necessitate a significantly greater amount of data and processing resources. This prompts significant inquiries on the ecological and financial consequences of implementing LLMs on a large scale. In addition, the use of benchmarks such as MMLU and DROP, as examined by Vongthongsri et al. (2024), signifies a significant change in our comprehension of the intricate capabilities of LLMs. The benchmarks primarily evaluate a model's capacity to handle intricate reasoning tasks and specialized knowledge, which are crucial for applications that go beyond basic text creation and are relevant in specific, high-stakes domains like as medical diagnostics or legal analysis. Bewersdorff et al. (2024) have examined recent progress in unsupervised and multi-modal learning, indicating a potential future where LLMs might combine knowledge from several types of input (such as text, voice, and images) to provide more comprehensive AI solutions. This signifies a substantial advancement towards AI systems that can comprehend the world in a more comprehensive manner, similar to human sensory and cognitive processes. To summarize, the literature on LLMs offers a thorough perspective on the technological progress and difficulties associated with these innovations. The continuous discourse between technological capabilities and ethical considerations is crucial in comprehending the appropriate and responsible evolution of LLMs to effectively satisfy future requirements.

III. RESEARCH DESIGN AND METHODOLOGY

A. Research Objective

This paper utilizes a comparative empirical analysis to assess the functionality and benchmarking methods of many

notable Large Language Models (LLMs) including GPT-4, LaMDA, and LLaMA. The LLMs possess a fundamental structure depicted in Figure 1.

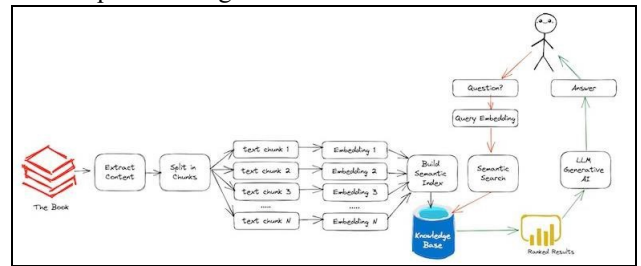


Fig 1 : LLM Architecture diagram

The methodology is designed to analyze the models by examining their architectural designs, operational capacities, and performance on several established benchmarks.

B. Methodological Framework

Our research utilizes a combination of quantitative studies and qualitative evaluations to provide a thorough grasp of LLM functionalities and benchmarking mechanisms. Below table shows the comparison of LLM's on data size and token limits (Table 1).

LLM	Training Data Size	Maximum Token Limit	Developer
GPT-3	570 billion tokens	4,096 tokens	OpenAI
GPT-4	Unknown (larger than GPT-3)	Up to 32,768 tokens	OpenAI
LaMDA 1.56	trillion code	Not specified	Google AI words +
PaLM	540 billion parameters	Up to 32,000 tokens	Google AI
Megatron-Turing NLG	530 billion parameters	Not specified	NVIDIA & Microsoft
Jurassic-1 Jumbo	178 billion parameters	Not specified	AI21 Labs
LLaMA 2 (70B)	70B parameters	4,096 tokens	Meta AI
BLOOM 1.7	trillion parameters	Not Hugging tokens	specified Face
WuDao 2.0	1.75 trillion parameters	Not specified	Beijing Academy of Artificial Intelligence

(BAAI)

Table 1 : Comparison of LLM's on data size and token limits

The methodology is structured as follows:

Model Selection and Data Collection: Our research began with the selection of a diverse set of LLMs, chosen based on their prevalence in academic literature and their varied uses in industry applications.

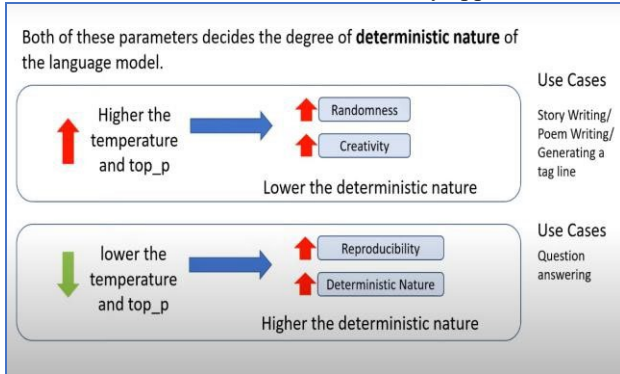


Fig 2 : key Parameters that can be adjusted to control the quality, diversity, and creativity of the generated text

For each model, detailed specifications were gathered, including information on training data size, tokenization methods, token limits, and model parameters like temperature & top_p. Few of these LLM's have Parameters that can be adjusted to control the quality, diversity, and creativity of the generated text as per Figure:2. Temperature controls text generation randomness. Top-P adjusts the set of candidate words for text. Number of Tokens controls the length of the generated text. Top-K influences text diversity by restricting token. When temperature=1 & top_p=1 then the response would be creative and diverse Figure:3. When temperature=0 & top_p=0 then the response would be more specific and concise Figure:4

```
prompt = "create a tag line for a coffee shop"

#high temperature and high top-p
for i in range(10):
    response = use_endpoint(prompt, temperature=1, top_p=1)
    print(response)

"Fuel up your day with every sip"
"Caffeine kick, cozy ambiance - sip your way to happiness at our coffee shop."
"Awaken your senses, one sip at a time."
"Where every cup is a new adventure in taste"
"Awaken your senses with every sip at our cozy coffee shop!"
"Fuel up for your day with our delicious brews"
"Savor every sip with our gourmet brews."
"Awaken your senses with every sip at our cozy cafe!"
"Awaken your senses one sip at a time"
"Fuel up on life with every sip"
```

Fig 3 : temperature=1 & top_p=1

```
#high temperature and low top-p
for i in range(10):
    response = use_endpoint(prompt, temperature=1, top_p=0)
    print(response)

"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"

#low temperature and low top-p
for i in range(10):
    response = use_endpoint(prompt, temperature=0, top_p=0)
    print(response)

"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
"Awaken your senses with every sip"
```

Fig 4 : temperature=0 & top_p=0

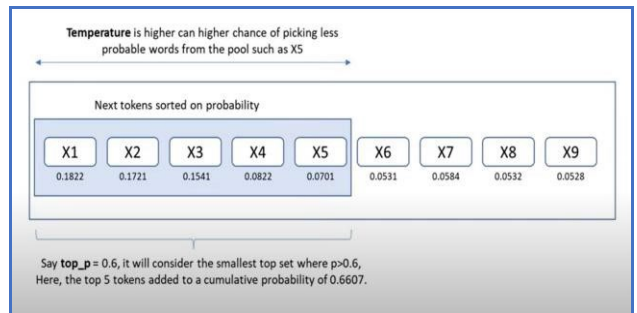


Fig 5 : New tokens sorted on probability.

New tokens are sorted on probability (Figure 5). But few of the LLM's are not exposing their API's /KPI's to understand their LLM's on confidence scoring for their responses

Benchmarking Approach: The core of our methodology lies in the systematic application of benchmarks that assess different capabilities of LLMs. We employed both established benchmarks such as HelloSwag for evaluating common-sense reasoning and newer assessments like DROP for complex reasoning tasks involving numerical data. Each benchmark was chosen to highlight specific aspects of LLM performance, including comprehension, reasoning, and the ability to generate coherent and contextually appropriate responses as mentioned in as shown in Figure 6.

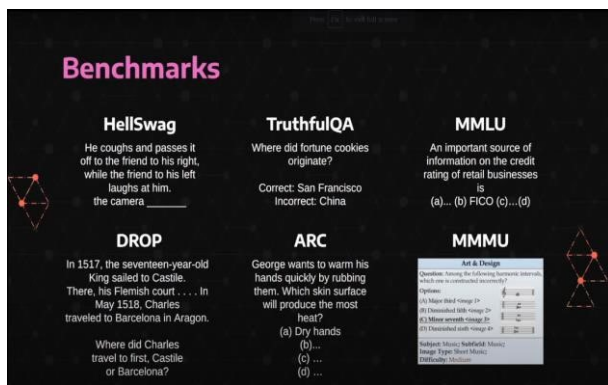


Fig 6 : LLM's Benchmarks.

Performance Metrics: The evaluation criteria were aligned with the objectives of each benchmark. Metrics such as accuracy, completeness of response, and time taken to generate responses were recorded. Additionally, qualitative assessments of the text's coherence and relevancy were conducted by a panel of NLP experts to provide deeper insights into each model's linguistic abilities.

Statistical Analysis: Data collected from these benchmarks were subjected to statistical analysis to identify significant performance trends and differences between models. Standard deviations, mean performance scores, and variance were calculated to ascertain consistency and reliability across different test conditions.

Ethical and Bias Assessment: Alongside performance evaluation, an ethical review was conducted to examine how each model handles issues of bias and fairness. Tests specifically designed to detect biases in gender, race, and socioeconomic status were applied, and models were scored based on their ability to generate unbiased content.

Methodological Challenges: Throughout the study, we addressed several methodological challenges, including varying definitions of tokens across models and the proprietary nature of some model architectures which limits the depth of analysis possible from outside the developing organization.

In summary, this research design and methodology combine quantitative performance assessments with qualitative expert analysis to provide a comprehensive understanding of LLM functionalities and their practical implications in real-world applications. This approach ensures a balanced evaluation that not only highlights technical capabilities but also considers the ethical dimensions of AI deployment.

IV. SYSTEM DEVELOPMENT AND RESULTS

A. Utilizing System Development

The system development for this study involved setting up a controlled environment to benchmark and evaluate the selected Large Language Models (LLMs). Each model was

tested using a consistent set of inputs across various benchmarks to ensure comparability of results. The outcomes presented here detail how each model performed across different tasks, providing insights into their strengths and limitations.

Development of Benchmarking Suite: We developed a benchmarking suite that incorporated a variety of tests, including HelloSwag, TruthfulQA, MMLU, DROP, and ARC. These benchmarks were selected to cover a broad spectrum of capabilities, from reasoning and comprehension to knowledge application in specialized domains. The suite was automated to run each test multiple times to minimize variability and ensure reliability in the results.

Model Configurations: Each LLM was configured according to its optimal operational settings as recommended by its developers. This included setting appropriate temperature and top-p values to balance creativity and coherence in the model outputs. For each model, we also configured the token limits to their maximum allowable inputs to test their performance at scale.

B. Result (Part A)

Benchmark's Performances of the top LLMs are shown in below Table 2.

Table 2 : Benchmark's Performance of the top LLM

Comprehension and Reasoning: In the HelloSwag benchmark, models like GPT-4 showed superior performance in predicting logical sequence continuations, reflecting their advanced understanding of context and causal relationships. However, models with smaller parameter sizes struggled with complex logical inferences.

Truthfulness and Bias: The TruthfulQA results highlighted that while most models performed well in general knowledge queries, they often replicated biases present in their training data. Notably, newer models like LaMDA demonstrated improved mechanisms to mitigate such biases, although no model was entirely free from this issue.

Specialized Knowledge: The MMLU tests, which assess domain-specific knowledge, showed that larger models generally had better coverage across diverse subjects. However, specialized models like Gemini Ultra were more efficient in certain domains, indicating a trade-off between breadth of knowledge and specialization.

Numerical Reasoning: In the DROP benchmark, the ability to handle numerical data varied significantly among the models. Gemini Ultra outperformed others in tasks requiring complex numerical manipulations, underscoring its sophisticated algorithmic design.

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Infect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 17 subjects (professional & academic) (Hendrycks et al., 2021a)	90.04% CoT@12*	79.13% CoT@8*	87.29% CoT@32	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
GSM8K Grade-school math (Cobbe et al., 2021)	94.4% Maj@32	86.5% Maj@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	83.2% 4-shot	32.6% 4-shot	52.9% 4-shot	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8% 4-shot	23.9% 4-shot	13.5% 4-shot
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	83.6% 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	74.4% 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New hold-out set with no leakage on web)	74.9% 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1 score) (Dua et al., 2019)	82.4 Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
Hellaswag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	95.3% 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	74.4 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

C. Result (Part B)

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. They collected over 800,000 humans pairwise comparisons to rank LLMs with the

Bradley-Terry model and display the model ratings in Elo-scale (Figure-7,8,9,10).

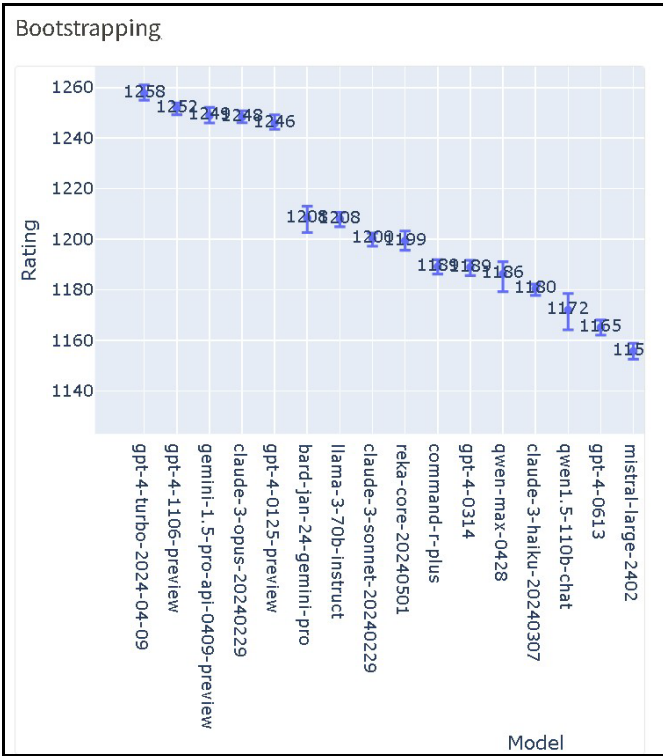


Fig 7 : Confidence Intervals on Model Strength (via Bootstrapping)

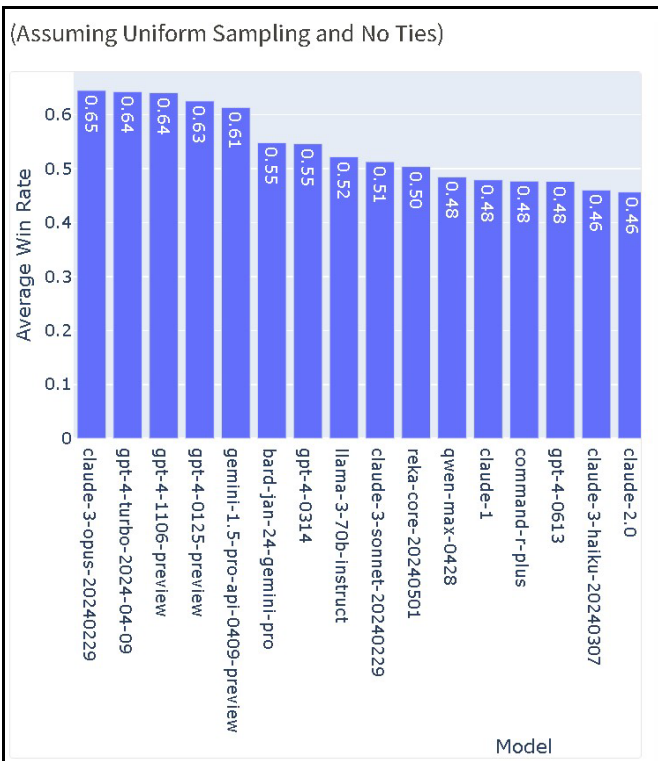


Fig 8 : Average Win Rate Against All Other Models (Assuming Uniform Sampling and No Ties)

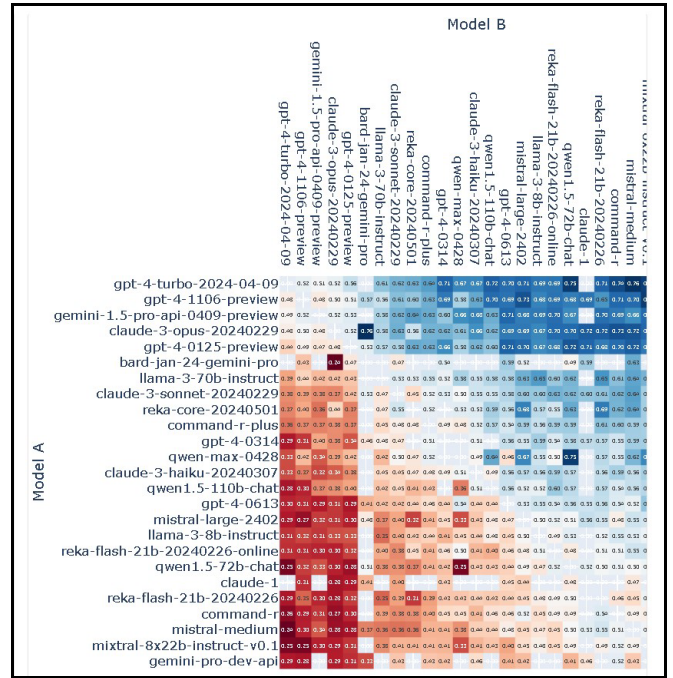


Fig 9 : Fraction of Model A Wins for All Notied A vs. B Battles

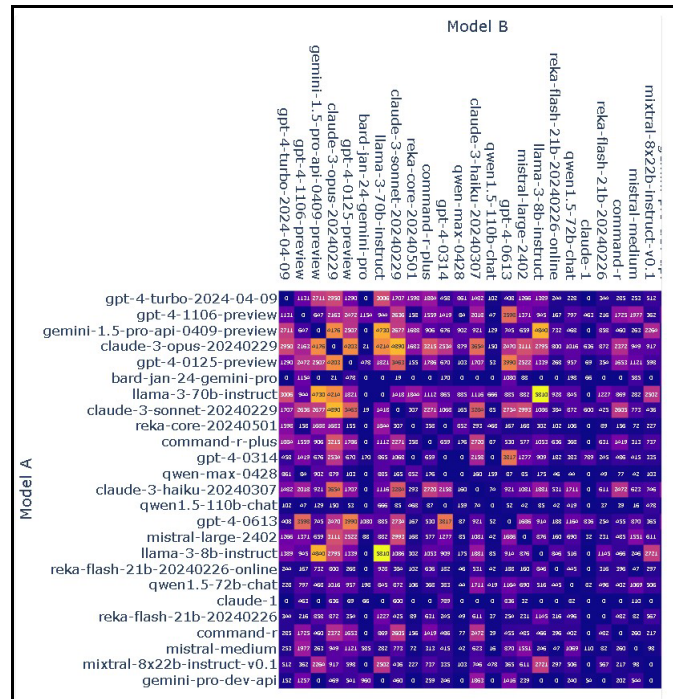


Fig 10 : Battle Count for Each Combination of Models (without Ties)

V. CONCLUSION

This empirical analysis of Large Language Models (LLMs) has underscored the significant strides made in the field of artificial intelligence, particularly in natural language processing. Our comparative study across several leading LLMs, including GPT-4, LaMDA, and LLaMA, has illuminated the nuanced capabilities and limitations of these models, providing a robust framework for their evaluation and application. Our findings reveal that while LLMs have reached unprecedented levels of linguistic sophistication and versatility, they continue to face challenges in areas such as bias mitigation and resource efficiency. The performance variability across different benchmarks—ranging from reasoning and comprehension to specialized knowledge tasks—highlights the importance of tailored model selection based on specific use cases.

The benchmarking processes employed in this study have not only demonstrated the current capabilities of these models but have also pointed towards areas requiring further research and development. Issues such as ethical AI development, bias in model outputs, and the environmental impact of training large-scale models are critical considerations that need to be addressed as we advance. It is imperative for the AI community to continue refining these models to ensure they are both powerful and responsible tools. Moreover, our analysis suggests a strong need for continuous improvement in the methodologies used for training and evaluating LLMs. Enhancing the transparency and fairness of these models, improving their efficiency, and expanding their adaptability to different linguistic and cultural contexts are essential steps towards their more effective integration into society.

In conclusion, the journey of LLM development is far from complete. While they present formidable capabilities, their responsible implementation is crucial for maximizing their benefits and minimizing potential harms. As this technology continues to evolve, it will require concerted efforts from developers, researchers, and policymakers alike to navigate the challenges and harness the opportunities presented by these advanced AI systems.

VI. REFERENCES

- [1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of FAccT.
- [2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
- [3] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300.
- [4] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.
- [5] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.
- [6] McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in Translation: Contextualized Word Vectors. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)
- [7] OpenAI. (2021). Introducing GPT-4. OpenAI Blog, <https://openai.com/index/gpt-4-research/>
- [8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research.
- [9] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.
- [10] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Proceedings of NIPS.
- [11] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv:1804.07461.
- [12] Big Data Demystified. (2024, Jan 17). Why you should build an LLM benchmark [English]. YouTube. <https://www.youtube.com/watch?v=DeIUJRd48fI&t=371s>
- [13] Ziegler, Z., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Christiano, P. (2019). Fine-Tuning Language Models from Human Preferences. arXiv preprint arXiv:1909.08593.
- [14] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).
- [15] Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2024, January 01). Taking the Next Step with Generative Artificial Intelligence: The Transformative Role of Multimodal Large Language Models in Science Education. arXiv [cs.AI]. <http://arxiv.org/abs/2401.00832v1>.
- [16] Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. The Cambridge Handbook of Artificial Intelligence.
- [17] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. Proceedings of ICLR.

- [18] Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Choi, Y. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing..
- [19] Vongthongsri, K. (2024, March 16). LLM Benchmarks: MMLU, HellaSwag, BBH, and Beyond [SNIPPET]. Confident AI. <https://confident.ai/blog/llm-benchmarks/>
- [20] Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2024, January 01). Taking the Next Step with Generative Artificial Intelligence: The Transformative Role of Multimodal Large Language Models in Science Education. arXiv [cs.AI]. Retrieved from <http://arxiv.org/abs/2401.00832v1>
- [21] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., & Stoica, I. (2024, March 7). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. [arXiv preprint arXiv:2403.04132v1] [cs.AI]. Retrieved from <http://arxiv.org/abs/2403.04132v1>

VII. AUTHORS PROFILE



First Author: Mohammad Nasim, currently a Senior IT Solution Architect at the State of Rhode Islands, brings over 25 years of specialized expertise in

Geographical Information Systems (GIS) and IT development. With a strong academic foundation, including a Doctorate in Computer Science Engineering and a Global Master of Science in Business Analytics, Nasim excels in leveraging GIS combined with advanced technologies like AI, ML, IoT, and big data for innovative solutions. His experience spans significant roles, including serving as a Solution Architect where he enhanced IT systems integration and data management for transportation infrastructure. Nasim's technical prowess is demonstrated through his extensive use of programming languages and software development frameworks, contributing to his ability to deliver comprehensive digital solutions in public service and enterprise settings. Notably, Nasim has been instrumental in projects that integrate aerial imaging, sensor data analysis, and smart city developments, showcasing his capability to drive digital transformation and operational excellence across various sectors. His career is also marked by contributions to international projects and professional growth through active involvement in professional associations like IEEE. As a seasoned professional, Nasim's interdisciplinary approach and commitment to technology innovation make him a pivotal member of our research

team, aiming to push the boundaries of digital architecture and geospatial intelligence.



Second Author: Shanshan (Shane) Zhu, a Bachelor of Arts candidate at Mount Holyoke College, majoring in

Computer Science and Experimental Psychology expertise. Shane's academic

pursuits cover essential domains such as Algorithms, Computer Vision, Data Structures, Machine Learning, and Software Graphic Design. Beyond classroom success, Shane's practical research as a laboratory research assistant at Mount Holyoke College and Brown

University. At Brown University and Google, she led the development of the conversation generation model

DiagramGPT, advancing the application of natural language processing (NLP) in artificial intelligence. At

Mount Holyoke College, her work focuses on immersive VR projects that enhance user experiences through

innovative technologies such as C, C#, Unity, and eyetracking technology. Her Neuro Science research also

explores the relationships between neuronal activation, inflammation, neurogenesis, and gut microbiota with

depressive cognitive biases. Additionally, Shane has interned as a Software Engineer, developing innovative

applications across healthcare, social media, networking, and facial recognition, notably enhancing intelligent

information processing. Her technical skills encompass full-stack programming, equipping her to address

challenges at the nexus of technology and psychology. As a co-researcher, Shane brings an interdisciplinary approach,

merging rigorous academic training with hands-on technological innovation, enhancing our project's potential

for groundbreaking insights in the study of technology's impact on psychological phenomena.



Third Author: Itauma Itauma, is a distinguished Assistant Professor and Division Chair at the DeVos Graduate School of Management, Northwood

University, with over 15 years of experience in educational technology and business

analytics. He holds a PhD in Instructional Design and Technology from Keiser University and a Master's in

Computer Science from Wayne State University. His expertise extends to leveraging data-driven strategies and

technology to enhance learning experiences across diverse student populations. In his academic roles, Dr. Itauma

has notably contributed to the field through positions like a Graduate Research Assistant at Wayne State University,

focusing on high-performance computing for image analysis, and as a Big Data Scientist Intern at Comrise,

utilizing Python and R for data modeling. Additionally, he served as a Technical Program Facilitator for Data

Analytics at Southern New Hampshire University, where he led curriculum development integrating technology and

analytics. Dr. Itauma's research interests include education analytics, HR analytics, and technology-enhanced learning. His work has been published in several peer-reviewed journals, emphasizing his commitment to advancing educational methodologies through innovative research. His significant contributions to instructional design and business analytics mark him as a leading figure in the integration of analytics into educational practices.

Optimal Forest Management Under Uncertainty: A Framework for Stochastic Dynamic Bioeconomic Modeling^{1,2}

Tong Wu,
David R. Just, C.-Y. Cynthia Lin Lawell, Jiancheng Zhao,
Zhangjun Fei, Ariel Ortiz-Bobea, Qiang Wei

Abstract

Sustainable forest management is an important issue worldwide. Forests supply the world's population with timber and non-timber forest products, including renewable products such as fruits, nuts, and maple syrup that can be harvested at more frequent intervals than the trees themselves. In this paper, we review, synthesize, and discuss the literature on forest economics and management; review our research that develops a nested dynamic bioeconomic model of the management of forests that generate interdependent products; and present a framework for stochastic dynamic bioeconomic modeling of optimal forest management under uncertainty. Our research and framework have important implications for the sustainable management of forests worldwide.

¹ Wu: Cornell University; tw494@cornell.edu. Just: Cornell University; drj3@cornell.edu. Lin Lawell: Cornell University; clinlawell@cornell.edu. Zhao: Zhejiang Academy of Forestry; jiancheng68@163.com. Fei: Boyce Thompson Institute and Cornell University; zf25@cornell.edu. Ortiz-Bobea: Cornell University; ao332@cornell.edu. Wei: Nanjing Forestry University; weiqiang@njfu.edu.cn.

² We thank Ryan Abman, H. Jo Albers, Jeff Arnold, Trevor Arnold, Joe Balagtas, Rachid Belhachemi, Susanna Berkouwer, Léa Bou Sleiman, Jacob Bradt, Daniel Brent, Mike Buffo, Marshall Burke, Jonah Busch, Virginia Callison, Luming Chen, Lauren Chenarides, Sahan Dissanayake, Ida Djenontin, Jackson Dorsey, Molly Doruska, Luc Esprabens, Paul Ferraro, Scott Francisco, Stephanie Fried, Teevrat Garg, Todd Gerarden, Dalia Ghanem, Tengda Gong, Gautam Gowrisankaran, Logan Hamilton, Ted

Helvoigt, Danae Hernández-Cortés, Jerzy Jaromczyk, Valerie Karplus, Jim Kiniry, Cathy Kling, Samuel Kortum, Elena Krasovskaia, Yusuke Kuwayama, Ashley Langer, Jim Lassoie, David R. Lee, David Lewis, Dingyi Li, Shanjun Li, Mengwei Lin, Clark Lundberg, Erin Mansur, Toni Marcheva, Shana McDermott, Brandon McFadden, Michael Meneses, Carlos Muñoz Brenes, Anjali Narang, Ishan Nath, Matias Navarro Sudy, Harry Nelson, José Nuño, Frederick Nyanzu, David Popp, Linda Powell, Carson Reeling, Brigitte Roth Tran, Ivan Rudik, Michelle Segovia, Jonathan Scott, Chris Severen, Jeff Shrader, Shuyang Si, Peter Smallidge, Aaron Smith, Charlie Smith, Christophe Spaenjers, Nick Tsivanidis, Calum Turvey, Wayne Walker, David Weinstein, Matthew Wibbenmeyer, Peter Woodbury, Weiguang Wu, Tianming Yen, Shuo Yu, and Hui Zhou for detailed and helpful comments. We also benefited from comments from seminar participants at Cornell University; and conference participants at the 2024 International Business Analytics Conference (IBAC); the Western Forest Economists (WFE) Annual Meeting; the Northeastern Agricultural and Resource Economics Association (NAREA) Annual Meeting; an Association of Environmental and Resource Economists (AERE) session at the Western Economic Association International (WEAI) Annual Conference; the Agricultural and Applied Economics Association (AAEA) Annual Meeting; the University of Michigan conference on Forests & Livelihoods: Assessment, Research, and Engagement (FLARE). We thank Qin Li, Lina Liu, Jianping Pan, and Boqing Yuan for helping us with the data collection; for providing us information about bamboo management; and for helping to host our visits to the Zhejiang Provincial Key Laboratory of Bamboo of Zhejiang Provincial Academy of Forestry, the Anji Forestry Technology Promotion Center, the Fumin Bamboo Shoot Specialized Cooperative, and the Tianlin Bamboo Shoot Specialized Cooperative, respectively. We thank Jianyang Lin, the Party Secretary for Xikou Forest Farm in Longyou County, for his exceptional support and organization for our interviews and short trips to the trade corporation in Longyou County. This research was supported by Cornell Center for Social Sciences Cloud Computing Solutions, and was conducted with support from the Cornell University Center for Advanced Computing. We received funding for our research from the USDA National Institute of Food and Agriculture (NIFA); a Cornell TREESPEAR Research Grant; and a Cornell University Graduate School Conference Grant. Our IRB protocol

(Protocol Number: IRB0148123) was granted exemption from IRB review according to Cornell IRB policy and under the Department of Health and Human Services Code of Federal Regulations 45CFR46.104(d). Just, Lin Lawell, and Ortiz-Bobeja are Faculty Fellows at the Cornell Atkinson Center for Sustainability. All errors are our own.

1. Introduction

The sustainable management of forests is a critical, timely, and important issue worldwide. Forests supply the world's population with timber as well as renewable non-timber forest products such as fruits, nuts, and maple syrup that can be harvested at more frequent intervals than the trees themselves. Unfortunately, the extent of the world's forests continues to decline as human populations continue to grow and the demand for food and land increases (FAO, 2005; Matthews, 2012; FAO, 2015).

Owing to intertwined feedback links between biological and economic systems, bioeconomic modeling is challenging, and there is a considerable need for studies that couple economic models of decision-making with biophysical models to provide policy-relevant implications (Kling et al., 2017). In this paper, we review, synthesize, and discuss the literature on forest economics and management; review our research in Wu et al. (2024) that develops a nested dynamic bioeconomic model of the management of forests that generate interdependent products; and present a framework for stochastic dynamic bioeconomic modeling of optimal forest management under uncertainty. Our research and framework have important implications for the sustainable management of forests worldwide.

2. Forest Economics and Management

Forest management is a dynamic problem because trees take time to grow. The seminal model of the optimal rotation time for a forest was developed by Faustmann (1849) and elaborated upon by Samuelson (1976). Since then, the Faustmann model has been extended in many ways (Newman, 1988), including to even-aged forest management (Jackson, 1980; Chang, 1983), uneven-aged forest management (Chang, 1981; Hall, 1983), externalities (Berck, 1981; Bowes, 1983; Calish et al., 1978; Hartman, 1976; Nguyen, 1979; Strang, 1983), taxation (Chang, 1982; Klemperer, 1979; Pearse, 1967; Rideout, 1982; Ollikainen, 1991), evolving stumpage price (Bare and Waggener, 1980; Gregersen, 1975; McConnell et al., 1983; Hardie et al., 1984; Newman et al., 1985), a onetime change of unchanged factors (Nautiyal and Williams, 1990), uncertainty (Chang, 1998), the intertemporal allocation of consumption (Deegen et al., 2011), rotation and thinning (Arimizu, 1958), optimal density (Amidon and Akin, 1968), net present discounted value of future payoff (Kilkki and Väisänen, 1969), forest production control (Hool, 1965), production control with Markov process (Hool, 1966), and thinning decisions (Amidon and Akin, 1968). The previous literature has also examined more complicated thinning decisions or combined thinning and rotation decisions (Amidon and Akin, 1968; Brodie et al., 1978; Brodie and Kao, 1979; Chen et al., 1980; Ritters et al., 1982).

In most of the forest economics literature, growth simulation models or yield models characterize the objective as the timber yield for tree species of interest in dynamic programming. Growth simulation models and yield models both describe the productivity of a tree standing as a function of multiple variables such as age, temperature, soil, rainfall, slope, and rooting depth

(Tyler, Macmillan, and Dutch, 1996). If the objective of the forest owners is instead profit maximization (Buongiorno, and Gillies, 2003; Kant and Alavalapati, 2014), then the market price of the timber also becomes a significant factor in the payoff function.

Sophisticated studies on forestry management utilizing dynamic optimization typically focus on developed countries (Ritters, 1982; Haight, 1985; Yousefpour and Hanewinkel, 2009). Pine and fir are two major types of tree species that researchers are interested in, due to their popularity in the western world and well developed productivity simulation models, and since these tree species are expensive to manage, requiring intensive thinning machinery and labor. Dynamic models have also been developed to study other topics relevant to forest resources, including apple tree pollination (Wilcox et al., 2024), organic farming (Meneses et al., 2024), agricultural groundwater management (Sears et al., 2019, 2024a, 2024b, 2024c), and agricultural disease control (Carroll et al., 2019, 2024a, 2024b; Sambucci et al., 2024; Yeh et al., 2024). Fewer studies have been carried out in developing countries and poor areas, where the need for sustainable forest management is particularly acute, and where different political structures, forestry contexts, objectives, and previous silvicultural practices demonstrate various research opportunities for forest management.

In Wu et al. (2024), we innovate on the previous literature by developing a novel nested dynamic bioeconomic model of the management of forests that generate interdependent products that differ in their growth cycles; and by analyzing bamboo forest management in particular. We apply our model to detailed daily panel data on bamboo shoot and bamboo stem harvests, in order to assess the optimality of bamboo farmers' forest management strategies and to understand the beliefs and perceptions that underlie and rationalize their management strategies. Both bamboo shoots and bamboo stems are valuable products. The harvesting of bamboo stems entails cutting down the bamboo plant, while the harvesting of bamboo shoots -- which grow annually from the bamboo plant -- does not. To solve for the optimal bamboo forest management strategy, our novel nested dynamic bioeconomic model nests an inner finite-horizon within-year daily dynamic programming problem that captures daily bamboo shoot growth within a season, inside an outer finite-horizon between-year annual dynamic programming problem that captures annual bamboo stem growth from year to year. Our nested dynamic bioeconomic model has important implications for the sustainable management of forests worldwide, particularly when the forests produce products that can be harvested at more frequent intervals than the trees themselves.

3. Modeling Framework

We present a framework for stochastic dynamic bioeconomic modeling of optimal forest management under uncertainty consisting of several key components.

A first key component of a bioeconomic model of forest management are biological production functions for the forest resources being managed, which may include the trees or plants themselves, as well as any products (such as fruits, nuts, etc.) that grow on the trees or plants. The choice of the functional form and/or parameter values for the biological production functions is best informed by relevant scientific information from biology, plant sciences, and forest science, and should be as specific as possible to the respective forest resource and species being managed. In

our research on bamboo forest management in Wu et al. (2024), for example, we model each of the products from a bamboo forest using a separate Chapman-Richards model (Richards, 1959), which is a flexible growth model for plants suggested by biological studies (Liu and Li, 2003) that has been used for bamboo (Yen, 2016), and we calibrate the parameters using data and information from previous studies of bamboo growth in the scientific, biological, and plant science literature.

A second key component are the choice variables a faced by the forest manager. Typically the choice variables likely include the harvesting decisions for the various forest resources being managed. In our research on bamboo forest management in Wu et al. (2024), for example, our action variables are the harvest decisions for each of the bamboo forest products we model.

A third key component is to characterize the sources of uncertainty. In our research on bamboo forest management in Wu et al. (2024), for example, we allow precipitation, prices, and the possibility of bamboo shoots death to all be stochastic. For both precipitation and prices, we use the empirical distribution of precipitation and prices in the data. For the possibility of bamboo shoots death, we calibrate the probability of death using data and information from previous studies of bamboo growth in the scientific, biological, and plant science literature.

A fourth key component are the state variables s . These state variables should include state variables affected by the choice variables, such as variables measuring the state and/or quantity of the forest resources being managed, and may also include variables related to the sources of uncertainty. In our research on bamboo forest management in Wu et al. (2024), our state variables included the number of each of the bamboo products we modeled, precipitation, and prices.

A fifth key component is to specify the per-period payoff (or per-period net benefits) $\pi(a, s)$ to the forest manager as a function of the action and state variables. In the case of forest businesses, the per-period payoff is typically the per-period profit. In our research on bamboo forest management in Wu et al. (2024), the per-period payoff was daily profit, as calculated by the total revenue of all the forest products being harvested that day, minus the total cost of harvesting all the forest products being harvested that day.

Combining the above five key components, one can construct the value function, which is the present discounted value (PDV) of the entire stream of per-period payoffs when the forest resource harvest decisions are chosen optimally, via the following Bellman equation (Bellman, 1954):

$$v(s) = \max_a \pi(a, s) + \beta E[v(s')|s, a]. \quad (1)$$

If the solution to the dynamic programming problem (1) yields an optimal strategy that differs from forest managers' actual decisions, then, to the extent that some of the differences between actual behavior and optimal strategy reflect possible sub-optimal behavior on the part of forest managers, the model may suggest ways to improve forest management and policy. Our research and framework have important implications for the sustainable management of forests worldwide.

References

- Amidon, E.L., and G.S. Akin. (1968). Dynamic programming to determine optimum levels of growing stock. *Forest Science*, 14(3), 287-291.
 Arimizu, T. (1958). Regulation of the cut by dynamic programming. *Journal of the Operations Research Society of Japan*, 1(4), 175-182.
 Bare, B.B., and T.R. Waggener. (1980). Forest land values and return on investment. *Forest Science*, 26(1), 91-96.

- Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6), 503-515.
- Berck, P. (1981). Optimal management of renewable resources with growing demand and stock externalities. *Journal of Environmental Economics and Management*, 8(2), 105-117.
- Bowes, M.D. (1983). Economic foundations of public forestland management. *Resources for the Future*.
- Brodie, J.D., and C. Kao. (1979). Optimizing thinning in Douglas-fir with three-descriptor dynamic programming to account for accelerated diameter growth. *Forest Science*, 25(4), 665-672.
- Brodie, J.D., D.M. Adams, and C. Kao. (1978). Analysis of economic impacts on thinning and rotation for Douglas-fir, using dynamic programming. *Forest Science*, 24(4), 513-522.
- Buongiorno, J., and J.K. Gilles. (2003). *Decision Methods for Forest Resource Management*. Academic Press, San Diego.
- Bystriakova, N., V. Kapos, I. Lysenko, and C.M.A Stapleton. (2003). Distribution and conservation status of forest bamboo biodiversity in the Asia-Pacific Region. *Biodiversity & Conservation*, 12(9), 1833-1841.
- Calish, S., R.D. Fight, and D.E. Teeguarden. (1978). How do nontimber values affect Douglas-fir rotations?. *Journal of Forestry*, 76(4), 217-221.
- Carroll, C.L., C.A. Carter, R.E. Goodhue, and C.-Y.C. Lin Lawell. (2019). Crop disease and agricultural productivity: Evidence from a dynamic structural model of verticillium wilt management. In W. Schlenker, editor, *Agricultural Productivity and Producer Behavior*, pages 217-249. University of Chicago Press, Chicago.
- Carroll, C.L., C.A. Carter, R.E. Goodhue, and C.-Y.C. Lin Lawell. (2024a). The economics of decision-making for crop disease control. Working paper, Cornell University. URL: http://clinlawell.dyson.cornell.edu/Vwilt_short_long_paper.pdf
- Carroll, C. L., C.A. Carter, R.E. Goodhue, and C.-Y.C. Lin Lawell. (2024b). Supply chain externalities and agricultural disease. Working paper, Cornell University. URL: http://clinlawell.dyson.cornell.edu/spinach_seed_paper.pdf
- Chang, S. (1982). An economic analysis of forest taxation's impact on optimal rotation age. *Land Economics*, 58(3), 310-323.
- Chang, S.J. (1981). Determination of the optimal growing stock and cutting cycle for an uneven-aged stand. *Forest Science*, 27(4), 739-744.
- Chang, S.J. (1983). Rotation age, management intensity, and the economic factors of timber production: Do changes in stumpage price, interest rate, regeneration cost, and forest taxation matter? *Forest Science*, 29(2), 267-277.
- Chen, C.M., D.W. Rose, and R.A. Leary. (1980). Derivation of optimal stand density over time—a discrete stage, continuous state dynamic programming solution. *Forest Science*, 26(2), 217-227.
- Deegen, P., M. Hostettler, and G.A. Navarro. (2011). The Faustmann model as a model for a forestry of prices. *European Journal of Forest Research*, 130(3), 353-368.
- Dong, D., Y. Tong, X. Yi., C. Zhang, and M. Gu. (2015). Harvesting cost and economics efficiency analysis of bamboo forests in various slope hill lands. *Journal of Bamboo Research*, 34(4), 23-27. [in Chinese]
- Faustmann, M. (1849). Calculation of the value which forest land and immature stands possess for forestry. *General Journal of Forestry and Hunting*. [in German]
- Food and Agriculture Organization [FAO]. (2005). Global Forest Resources Assessment 2005.
- Food and Agriculture Organization [FAO]. (2015). Global Forest Resources Assessment 2015.
- Gregersen, H.M. (1975). Effect of inflation on evaluation of forestry investments. *Journal of Forestry*, 73(9), 570-572.
- Haight, R.G. (1985). A comparison of dynamic and static economic models of uneven-aged stand management. *Forest Science* 31(4), 957-974.
- Hall, D.O. (1983). Financial maturity for even-aged and all-aged stands. *Forest Science*, 29(4), 833-836.
- Hardie, I.W., J.N. Daberkow, and K.E. McConnell. (1984). A timber harvesting model with variable rotation lengths. *Forest Science*, 30(2), 511-523.
- Hartman, R. (1976). The harvesting decision when a standing forest has value. *Economic Inquiry*, 14(1), 52-58.
- Jackson, D. H. (1980). *The Microeconomics of the Timber Industry*. Westview Press.
- Kant, S., and J.R.R. Alavalapati. (2014). *Handbook of Forest Resource Economics*. Abingdon: Routledge.
- Kilikki, P., and U. Väisänen. (1969). Determination of the optimum cutting policy for the forest stand by means of dynamic programming. *Acta Forestalia Fennica*, 102.
- Klemperer, W.D. (1979). Inflation and present value of timber income after taxes. *Journal of Forestry*, 77(2), 94-96.
- Kling, C.L., R.W. Arriitt, G. Calhoun, and D.A. Keiser. (2017). Integrated assessment models of the food, energy, and water nexus: a review and an outline of research needs. *Annual Review of Resource Economics*, 9(1), 143-163.
- Kusters, K., and B. Belcher. (2004). *Forest Products, Livelihoods and Conservation: Case studies of non-timber forest product systems. Volume 1 - Asia*. Center for International Forestry Research (CIFOR).
- Liu, Z.-G., and F.-R. Li. (2003). The generalized Chapman-Richards function and applications to tree and stand growth. *Journal of Forestry Research*, 14(1), 19-26.
- Matthews, R. (2012). The economic and employment benefits of forests. *Eco-Business*, 7 March 2012.
- McConnell, K.E., J.N. Daberkow, and I.W. Hardie. (1983). Planning timber production with evolving prices and costs. *Land Economics*, 59(3), 292-299.
- Meneses, M.A., C.L. Casteel, M.I. Gómez, D.R. Just, R. Kanbur, D.R. Lee, and C.-Y.C. Lin Lawell. (2024). Organic farming transitions: A dynamic bioeconomic model. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/organic_transitions_paper.pdf
- Nautiyal, J.C., and J.S. Williams. (1990). Response of optimal stand rotation and management intensity to one-time changes in stumpage price, management cost, and discount rate. *Forest Science*, 36(2), 212-223.
- Newman, D.H., C.B. Gilbert, and W.F. Hyde. (1985). The optimal forest rotation with evolving prices. *Land Economics*, 61(4), 347-353.
- Newman, D.H. (1988). *The optimal forest rotation: A discussion and annotated bibliography*. Gen. Tech. Rep. SE-48. Asheville, NC: US Department of Agriculture, Forest Service, Southeastern Forest Experiment Station.
- Nguyen, D. (1979). Environmental services and the optimum rotation problem in forest management. *Journal of Environmental Management*, 8(2), 127-136
- Ollikainen, M. (1991). The effect of nontimber taxes on the harvest timing—the case of private nonindustrial forest owners: A note. *Forest Science*, 37(1), 356-363.
- Pearse, P.H. (1967). The optimum forest rotation. *The Forestry Chronicle*, 43(2), 178-195.
- Richards, F. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, 10(2), 290-301.

- Rideout, D.B. (1982). Micro-Economic Properties of Four Local Forest Tax Methods. (doctoral dissertation). University of Washington, Seattle, United States.
- Riitters, K.H.; J.D. Brodie, and D.W. Hann. (1982). Dynamic programming for optimization of timber production and grazing in ponderosa pine. *Forest Science*, 28(3), 517-526.
- Sambucci, O., C.-Y.C. Lin Lawell, and T.J. Lybbert (2024). Pesticide spraying and disease forecasts: A dynamic structural econometric model of grape growers in California. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/PMI_dynamic_paper.pdf
- Samuelson, P.A. (1976). Economics of forestry in an evolving society. *Economic Inquiry*, 14(4), 466-492.
- Sears, L., D. Lim, and C.-Y.C. Lin Lawell. (2019). Spatial groundwater management: A dynamic game framework and application to California. *Water Economics and Policy*, 5(1):1850019.
- Sears, L., C.-Y.C. Lin Lawell, G. Torres, and M.T. Walter. (2024a). Adjudicated groundwater property rights: A structural model of the dynamic game among groundwater users in California. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/property_rights_struc_paper.pdf
- Sears, L., C.-Y.C. Lin Lawell, G. Torres, and M.T. Walter. (2024b). Moment-based Markov equilibrium estimation of high-dimension dynamic games: An application to groundwater management in California. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/mme_groundwater_paper.pdf
- Sears, L., C.-Y.C. Lin Lawell, and M.T. Walter. (2024c). Groundwater under open access: A structural model of the dynamic common pool extraction game. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/open_access_structural_paper.pdf
- Strang, W.J. (1983). On the optimal forest harvesting decision. *Economic Inquiry*, 21(4), 576-583.
- Tyler, A.L., D.C. MacMillan, and J. Dutch. (1996). Models to predict the General Yield Class of Douglas fir, Japanese larch and Scots pine on better quality land in Scotland. *Forestry: An International Journal of Forest Research*, 69(1), 13-24.
- Wilcox, S.W., D.R. Just, C.-Y.C. Lin Lawell, M.I. Gómez, and H. Grab. (2024). To (rent) bees or not to (rent) bees?: An examination of the farmer's question. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/rent_bees_paper.pdf
- Wu, T., D.R. Just, C.-Y.C. Lin Lawell, J. Zhao, Z. Fei, A. Ortiz-Bobea, and Q. Wei. (2024). Optimal forest management for interdependent products: A nested dynamic bioeconomic model and application to bamboo. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/Moso_bamboo_dynamic_paper.pdf
- Yeh, D.A., M.I. Gómez, and C.-Y.C. Lin Lawell. (2024). Sustainable pest management, beliefs, and behavior: A dynamic bioeconomic analysis. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/Maine_SWD_dynamic_bioeconomic_paper.pdf
- Yen, T.-M. (2016). Culm height development, biomass accumulation and carbon storage in an initial growth stage for a fast-growing Moso bamboo (*Phyllostachy pubescens*). *Botanical Studies*, 57(1), 10.
- Yousefpour, R., and M. Hanewinkel. (2009). Modelling of forest conversion planning with an adaptive simulation-optimization approach and simultaneous consideration of the values of timber, carbon and biodiversity. *Ecological Economics*, 68(6), 1711-1722.

Leveraging Deep Learning for Accounting Fraud Detection

Victoria Gonzalez¹, Sai Teja Reddy², Chitral Patil², Haimonti Dutta^{1,2}

¹Department of Management Science and Systems

²Institute for Artificial Intelligence and Data Science

The State University of New York at Buffalo, New York 14260

{vg38,saitejar,chitralp,haimonti}@buffalo.edu

February 2024

Abstract

Fraudulent financial reporting encompasses deliberate misstatements or omissions in financial statements, aiming to mislead stakeholders and regulators, resulting in a breach of Generally Accepted Accounting Principles (GAAP). This research introduces a novel predictive model for detecting accounting fraud, leveraging publicly available financial data extracted from 10-K filings. Specifically, a Multi-Layer Perceptron (MLP) is constructed with Random Under-Sampling (RUS) (called RUS MLP) which has comparable performance to traditional machine learning algorithms designed for fraud detection such as logit and probit models. Our findings reveal that advanced deep learning techniques can be used to fortify financial systems against fraudulent activities.

1 Introduction

According to the U.S. Securities and Exchange Commission (SEC, <https://www.sec.gov/>), businesses incur an annual loss of 5% of their revenue due to fraudulent activities [9]. Fraudulent financial reporting involves intentional misstatements or omissions in financial statements designed to deceive users, leading to non-compliance with Generally Accepted Accounting Principles (GAAP). This can occur through manipulating records, misrepresenting information, or misapplying accounting principles. It is very hard to detect fraud because regulatory bodies (such as auditing firms) often have limited resources. Furthermore, by the time a fraud is detected, significant damage may already have been done [4].

The objective of our study is to analyze and compare fraud-detection models to help investors, auditors, financial analysts, and regulators identify and detect misstatements. A variety of statistical methods have been employed for this task

(Bao et al. [4], Cecchini et al. [6] and Dechow et al. [8]), but previous research primarily focuses on identification of “red flag” firms – businesses exhibiting warning signs indicative of potential involvement in fraudulent activities. Our research, in contrast, examines the financial statements of *all* publicly traded U.S. companies that file 10K reports to the SEC. The Commission’s enforcement staff conducts investigations into possible violations of the federal securities laws and publishes yearly reports – Accounting and Auditing Enforcement Releases (AAERs) – which provide the list of accounting fraud cases used in this study.

To build state-of-the-art fraud detection models, we resort to deep learning – a branch of machine learning that helps find complex, nonlinear patterns from the data. Specifically, a Multi-Layer Perceptron (MLP) is constructed with Random Under-Sampling (RUS) – called RUS MLP – which outperforms popular statistical models such as Logistic and Probit Regression models. Automatic hyperparameter tuning framework, Optuna [1], is leveraged to optimize the MLP architecture. Our findings reveal that deep learning techniques are in fact able to detect fraud more consistently than statistical models popular in the community.

2 Related Work

In this section, we provide a brief review of existing models for fraud detection. Summers and Sweeney [10] designed a logit model that integrates financial metrics with indicators related to insider trading activities within companies. The model’s generalization performance remains uncertain without out-of-sample validation. Beneish [5] discerned 74 companies engaged in earnings manipulation and matched them against 2,332 non-manipulative counterparts from COMPUSTAT. Employing a probit model, he crafted a model capable of effectively differentiating between manipulative and non-manipulative entities. Cecchini et al. [6] used a support vector machines model with a financial kernel to detect fraudulent financial reporting with readily available financial data. Dechow et al. [8] developed a predictive model, using logistic regression and F-score, to assess the likelihood of earnings manipulation. Amiram et al. [3] introduced Benford’s Law to assess the accuracy of financial statements. This method measures how closely the distribution of financial numbers in a firm’s statements matches the expected theoretical distribution. The research demonstrated that errors in financial statements result in a greater deviation from this distribution. Alawadhi et al. [2] conducted research that incorporated a logistic model and three variants of a Support Vector Machine. Bao et al. [4] introduced an ensemble learning model that leveraged 28 raw financial items extracted directly from financial statements as fraud predictors. Their data was processed to study “serial fraud” – fraud that could span multiple years, differing from the objective of this study. Chakrabarty et al. [7] used the Adjusted Benford score (AB-score), which assesses the

likelihood of financial statement manipulation for a given firm in a year while accounting for peer comparisons and time-series trends.

3 Data

Our dataset contains all publicly traded U.S. companies in the period 1990 – 2019. Financial data for these companies was sourced from COMPUSTAT [11], and instances of financial misstatements were identified using the AAER dataset. This dataset includes 4,278 Accounting and Auditing Enforcement Releases (AAERs) issued by the SEC between 1982 and 2022. Within these releases, there are 1,816 firm misstatement events. It is relevant to note that fraud is recognized in the year it occurred, not the year of the AAER report. All firms with zero total assets in a given year were excluded. A total of 28 raw financial items and 14 financial ratios [4] were used as features. Table 1 shows the data distribution of fraud versus non-fraud instances. In the context of fraud detection, there is an inherent class imbalance, where there are fewer instances of fraudulent companies compared to non-fraudulent ones. This is rectified using *Random Under Sampling (RUS)*, wherein the same proportion of instances in the minority class is sampled from the majority class. All code and data used in this research are available from the following Github repository: <https://github.com/UB-Fraud-Detection/Data-Processing-and-Modelling>.

[//github.com/UB-Fraud-Detection/Data-Processing-and-Modelling](https://github.com/UB-Fraud-Detection/Data-Processing-and-Modelling).

Training Period	Instances	Misstatements
1990-2002	120,819	831
Testing Period	Instances	Misstatements
2003-2019	125,314	540

Table 1: Number of instances, and labels for training and testing.

The distribution of misstatements by year is presented in Figure 1.

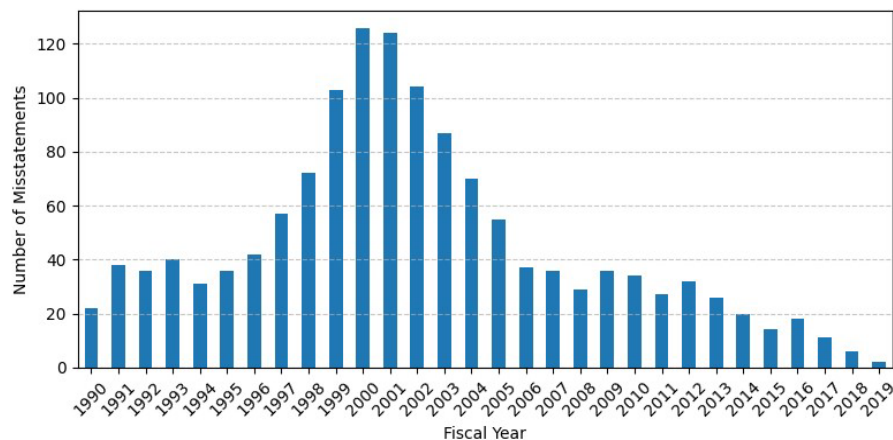


Figure 1: Number of accounting misstatements by year

4 Machine Learning Algorithms

Both linear (logit and probit) and non-linear models (Multi-Layer Perceptrons (MLP)s) are used to study the problem of fraud detection.

Logistic regression estimates the probability of an event through a Logit transformation, capturing the ratio of success to failure probabilities. Similarly, Probit regression, models binary outcomes by expressing the linear combination of predictors as the inverse standard normal distribution of the outcome's probability.

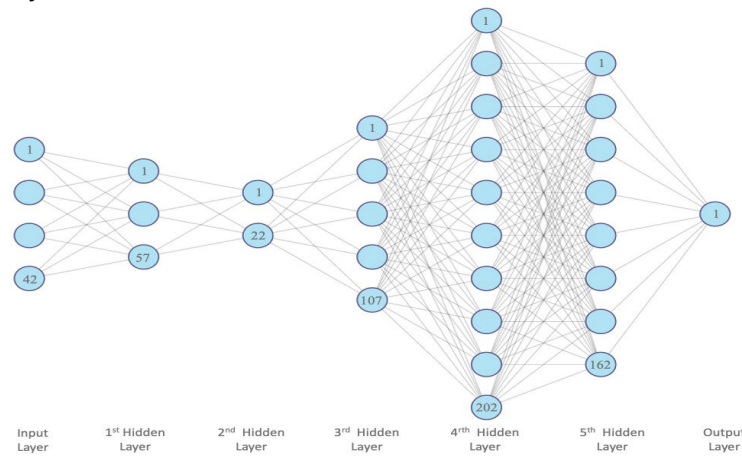


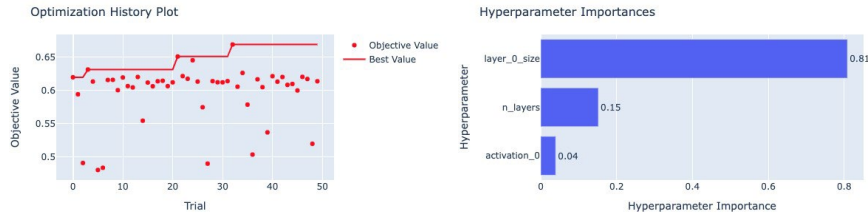
Figure 2: MLP Architecture

An MLP is a feed-forward artificial neural network, consisting of fully connected neurons with a nonlinear activation function (such as logistic, tanh or Rectified Linear Unit (ReLU)) organized in three distinct layers: the input layer, hidden layer(s), and output layer. The input layer accepts input signals for processing, while the output layer handles classification and regression tasks. The hidden layers, which determine the *depth* of the network, are situated between the input and output layers. Figure 2 illustrates the architecture of the MLP used in empirical analysis.

5 Empirical Results

The objective of this study was to explore whether non-linear models can detect fraudulent activity better than linear models (such as logit and probit). Since the dataset was inherently imbalanced, we allowed Random Under Sampling for both linear and non-linear models. Diverse network architectures were used to optimize the RUS MLP performance, leveraging Optuna [1] for hyperparameter tuning. Table 2 presents the parameter space explored by Optuna.

The best-performing model was a 5-layer MLP with specific neuron counts



(a) Results obtained after optimization (b) Importance of hyperparameters in the with Optuna models

Figure 3: Optimization of the Multilayer Perceptron with Optuna

Parameters	Values
No. of hidden layers	1 – 5
No. of neurons per layer	10 – 250
Activation functions	ReLU, Sigmoid
Loss Function	Binary Cross Entropy (BCE)
Optimizer	Stochastic Gradient Descent (SGD)
Learning Rate	0.01
Number of training epochs	150

Table 2: Parameter space of the MLP used in empirical research

and activation functions: 57 neurons in the first layer, 22 neurons in the second layer, 107 neurons in the third layer, 202 neurons in the fourth layer, and 162 neurons in the fifth layer. ReLU was applied in the first three layers, while sigmoid was used in the last two layers.

Features	RUS MLP	Logit	Probit
28 Raw Items + 14 Ratios	0.62 ± 0.01	0.57 ± 0.05	0.61 ± 0.0
14 Ratios	0.63 ± 0.02	0.57 ± 0.05	0.61 ± 0.0
28 Raw Items	0.55 ± 0.06	0.52 ± 0.07	0.57 ± 0.0

Table 3: Average AUC score and standard deviations for 10 training sessions.

Table 3 presents the average AUC scores and their corresponding standard deviations obtained from 10 randomized trials of training for the optimized MLP, Logit, and Probit models. Remarkably, the MLP model, especially when utilizing 14 financial ratios and the combination of ratios and raw items, demonstrates

significantly higher AUC compared to the Logit and Probit models. This observation highlights the effectiveness of the MLP approach, particularly when utilizing different sets of financial features. Interestingly, the Probit model outperforms the others only in the setting where all 28 raw financial items are used as features.

6 Conclusion

This study investigates the implementation of a RUS MLP for detecting accounting fraud in out-of-sample financial data. Employing Optuna, we optimized the MLP architecture, identifying the most effective hyperparameters for the model in this context. Our analysis reveals that deep neural network models have the potential to surpass the performance of traditional linear models.

Acknowledgements

Victoria Gonzalez gratefully acknowledges financial support for this research by the Fulbright U.S. Student Program, which is sponsored by the U.S. Department of State and the Fulbright Commission in Chile. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Fulbright Program, the Government of the United States, or the Fulbright Commission in Chile.

References

- [1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework, July 2019. arXiv:1907.10902 [cs, stat].
- [2] Abdullah Alawadhi, Jonathan M. Karpoff, Jennifer Lynch Koski, and Gerald S. Martin. The Prevalence and Costs of Financial Misrepresentation. *SSRN Electronic Journal*, 2020.
- [3] Dan Amiram, Zahn Bozanic, and Ethan Rouen. Financial statement errors: evidence from the distributional properties of financial statement numbers. *Review of Accounting Studies*, 20(4):1540–1593, December 2015.
- [4] Yang Bao, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, 58(1):199–235, 2020. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12292>.

- [5] Messod D. Beneish. The Detection of Earnings Manipulation. *Financial Analysts Journal*, 55(5):24–36, 1999. Publisher: CFA Institute.
- [6] Mark Cecchini, Haldun Aytug, Gary J. Koehler, and Praveen Pathak. Detecting Management Fraud in Public Companies. *Management Science*, 56(7):1146–1160, July 2010.
- [7] Bidisha Chakrabarty, Pamela C. Moulton, Leonid Pugachev, and Xu (Frank) Wang. Catch Me If You Can: In Search of Accuracy, Scope, and Ease of Fraud Prediction. *Working Paper*, November 2022.
- [8] Patricia M. Dechow, Weili Ge, Chad R. Larson, and Richard G. Sloan. Predicting Material Accounting Misstatements*: Predicting Material Accounting Misstatements. *Contemporary Accounting Research*, 28(1):17–82, March 2011.
- [9] Paul Munter. The Auditor’s Responsibility, October 2022.
- [10] Scott L. Summers and John T. Sweeney. Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis. *The Accounting Review*, 73(1):131–146, 1998. Publisher: American Accounting Association.
- [11] WRDS. Compustat - Fundamentals Annual, January 2023.

2024 International Business Analytics Conference

Empowered Decision-Making: Bridging the Gap Between Business Simulations and Generative AI

Joseph T. Kuvshnikov, CPA, CFE, MBA, PhD
Gannon University
Erie, PA, USA
Kuvshini002@gannon.edu
+1(814)871-7279

Extended Abstract

Business simulations have transformed the world of business education. Initial explorations in bridging the gap between business simulations and large language model generative artificial intelligence have resulted in empowered decision making. Business simulations have long been embraced as a powerful tool in helping participants hone their strategic business decision-making skills (Kurtz, 2003; Schwarz, 2009). Sources of pedagogical power include: experiential learning (Humphreys, Bakir, & Babb, 2022), the opportunity to try new strategies in a low-risk environment (Mubaraz & Mezrar, n.d.), the opportunity to practice business management and decision-making in a collaborative team-based environment (Anderson & Lawton, 2009), and business simulations allow participants to develop strategic and operational decision-making skills (Levant, Coulmont, & Sandu, 2016). Combining business simulations and generative AI has led to significant new pedagogical synergies. Faculty/facilitators can use live generative artificial intelligence prompts to populate discussions with a myriad of considerations, strategies, and pros and cons of various business decisions (Moser, 1986; Ferreira, González-González, & Adamatti, 2021; Paschen, Wilson, & Ferreira, 2022). The speed at which content is delivered makes generative artificial intelligence ideal for either synchronous or asynchronous learning environments. The purpose of this paper is to report on explorations in the ways faculty can coach simulation participants in prompting AI and discerning how to apply the results to specific decision-making contexts. Generative AI rapidly equips faculty to embrace each teachable moment with simulation and real-world applications. This explorative initiative report contains several examples of prompts applied in the context of the Income|Outcome (ASI) business simulation.

Introduction

Business simulations have long been embraced as a powerful tool in helping participants hone their ability to strategically use business analytics in decision-making (Kurtz, 2003; Schwarz, 2009). Some of the sources of pedagogical power include: experiential learning (Humphreys, Bakir, & Babb, 2022), the opportunity to try new analytical strategies in a low-risk environment (Mubaraz & Mezrar, n.d.), the opportunity to practice business analytics in a collaborative teambased environment (Anderson & Lawton, 2009), and business simulations allow

participants to develop strategic and operational decision-making skills (Levant, Coulmont, & Sandu, 2016).

Generative artificial intelligence (AI) is a relative newcomer to the world of business-decision making. According to ChatGPT 4, generative AI “can significantly enhance the teaching of business analytics through several key methods:” (ChatGPT 4, 2024)

1. **Data Generation and Simulation:** Generative AI can create large datasets or simulate realistic business scenarios that students can analyze. This helps them practice handling, interpreting, and drawing insights from real-world-like data, even if such data is not readily available.
2. **Personalized Learning:** By understanding individual student needs, generative AI can offer customized tutorials, exercises, and feedback. This allows students to focus on areas where they need improvement and progress more efficiently.
3. **Scenario Analysis:** Generative AI can create complex business scenarios involving different variables and outcomes. Students can then analyze these scenarios, develop strategic insights, and see how their decisions might play out in different contexts.
4. **Automated Assistance:** It can function as an intelligent tutor, providing instant feedback, answering questions, and guiding students through analytical problems. This allows for a more interactive and responsive learning experience.
5. **Real-time Visualization:** Generative AI can convert raw data into visually compelling charts, graphs, and dashboards. These visuals help students better understand data trends and the impact of their analytical decisions.
6. **Case Studies and Insights Generation:** By processing existing business cases and generating new insights, generative AI can enrich case study materials, helping students see broader applications of analytical principles.
7. **Collaboration and Team Dynamics:** AI can simulate team-based projects, analyze student interactions, and provide feedback on collaborative problem-solving in business analytics, fostering skills that are crucial for professional environments.

Integrating business simulations with generative AI produces educational synergies that elevate teaching quality (Babin, 2023; Taylor et al., 2021). Educators can utilize real-time prompts from generative AI to enrich discussions with diverse factors, strategies, and the advantages and disadvantages of different business choices (Moser, 1986; Ferreira, González-González, & Adamatti, 2021; Paschen, Wilson, & Ferreira, 2022). The rapid content delivery provided by generative AI makes it suitable for both synchronous and asynchronous learning formats.

These synergies empower students/participants to cultivate skills in business analytics, innovation, and discernment. Discernment is critical for business decision-making, as it involves

distinguishing between various options, evaluating their impacts, and making well-informed choices (OpenAI, 2024). According to ChatGPT 4 (2024), "discernment entails sorting through extensive data to separate reliable and relevant sources from irrelevant or biased ones, which is crucial for business decisions based on accurate and credible information." This definition highlights the strengths of generative AI. It sifts through the vast web of information to find relevant data, leaving it up to the user to identify what applies to their unique context and situation. Facilitators or faculty can assist participants in crafting effective questions and discerning which AI-generated results are relevant for specific scenarios. Context is crucial for determining which information is useful and which can be applied to the current set of decisions.

This paper aims to explore the educational synergies generated when participants of business simulations utilize generative AI to make essential decisions and complete tasks during the simulation process. The paper specifically focuses on how generative AI enriches the content provided to participants in a particular business simulation while also laying the groundwork for practicing context-based result discernment.

The simulation to be considered in this paper is Andromeda Simulation International's Income|Outcome. Income|Outcome replicates making strategic and operational decisions in a competitive team-based environment. Income|Outcome can be delivered in live face-to-face or remote distributed environments. Each team manages their own company and competes against other teams in an open market environment. This is the same business simulation used by The Coca-Cola Company, General Electric, Michelin, BASF, Hewlett Packard and other international business organizations to train their management. Further information about the simulation can be found on the world wide web at <http://www.income-outcome.com>. The Income|Outcome gameboard is provided in Figure 1.



Figure 1: The Income|Outcome Gameboard (Andromeda Simulations International)

The Income|Outcome Manual begins by explaining the “winner” of the simulation is the team that ends the simulation with the highest retained earnings. From the outset simulation participants can use generative artificial intelligence to populate their personal knowledge base

with ways real-world companies can increase retained earnings. An initial prompt using ChatGPT provides the following results:

ChatGPT 4 Prompt and Results (Personal conversation, 05/10/24):

“What are different strategies a company can use to increase retained earnings?”

A company can increase its retained earnings through a combination of strategic and financial measures:

1. **Revenue Growth:** Expanding market share, increasing sales, or developing new products and services to generate more revenue.
2. **Cost Reduction:** Streamlining operations, reducing production costs, and optimizing supply chains to lower overall expenses.
3. **Pricing Strategy:** Adjusting pricing models to better reflect market demand or offer value-added services to increase profit margins.
4. **Debt Management:** Reducing interest expenses by paying off high-interest debt or refinancing existing loans at lower rates.
5. **Efficient Inventory Management:** Minimizing inventory carrying costs by aligning inventory levels with demand forecasts.
6. **Investing in Technology:** Adopting new technologies to improve productivity and reduce manual labor, thereby saving costs.
7. **Market Diversification:** Expanding into new geographic markets or different customer segments to spread business risks.
8. **Customer Retention:** Improving customer loyalty through better service, reducing acquisition costs associated with gaining new customers.
9. **Tax Optimization:** Utilizing tax credits and deductions and restructuring finances to minimize tax liabilities.
10. **Share Buybacks:** Repurchasing shares to reduce the number of shares outstanding, increasing earnings per share and potentially raising the stock price.
11. **Mergers and Acquisitions:** Merging with or acquiring other businesses to expand capabilities and achieve economies of scale.

These strategies, individually or in combination, can help companies improve profitability and ultimately increase retained earnings.

It is at this point that simulation participants need to begin applying results to context. Results 1, 2, 3, 4, 5, 6, and 8 apply to Income|Outcome and can be used to guide future decision-making in

the simulation. Discussing the full set of results provides the simulation facilitator/faculty with excellent teaching moments. The facilitator can explain how the other options are viable retained earnings enhancement strategies in contexts outside of the simulation. **Strategic vs Operational Decisions**

Generative artificial intelligence aids in both strategic and operational decision-making by offering prompt results that highlight factors relevant to each type of decision. In this section this paper offers examples of prompts and their results at both decision-making levels.

Strategic Decisions

Strategy is the group of decisions that explain how an organization plans to fulfill its mission and vision. Strategic decisions affect the next three to five years of an organization’s life.

Income|Outcome explores strategic decision-making by asking participants to consider the pros and cons of five different scenarios. The scenarios can be seen in Table 1.

What happens if you:	Pros	Cons
<i>Maintain existing capacity?</i>		
<i>Expand capacity?</i>		
<i>Submit the minimum number of bids?</i>		
<i>Prepare a bid for every order?</i>		
<i>Target a relationship customer?</i>		

Table 1: Income|Outcome strategic decisions

Generative artificial intelligence can help the participant build a list of the pros and cons for each decision. A sample prompt with results is shown below.

ChatGPT 4 Prompt and Results (Personal conversation, 05/10/24):

“What are the pros and cons of a factory choosing to maintain existing capacity?”

<p>Choosing to maintain existing capacity for a factory can have various pros and cons:</p> <p>Pros:</p> <ol style="list-style-type: none"> Reduced Costs: Maintaining the current capacity avoids the immediate costs of expanding or reducing production facilities, including the expenses associated with new equipment or selling off excess machinery. Stability: Keeping production levels steady provides consistent, predictable output, making it easier to plan for inventory and supply chain management. Employee Morale: By avoiding major changes, employee morale may remain stable since jobs are less likely to be impacted by shifts in production capacity.
--

4. **Customer Relationships:** A steady output helps maintain reliable supply to existing customers, ensuring consistent order fulfillment.
5. **Utilizing Existing Resources:** The factory makes full use of existing resources without the potential waste involved in downsizing or the investment required for expansion.

Cons:

1. **Missed Opportunities:** Maintaining capacity might mean missing out on growth opportunities if market demand increases and the factory cannot meet higher production requirements.
2. **Obsolescence:** Current capacity could become less efficient over time if new technologies or production methods emerge, potentially putting the factory at a competitive disadvantage.
3. **Market Changes:** Market demand might decrease due to shifts in consumer preferences or industry trends, leaving the factory with underutilized capacity.
4. **Operational Inefficiency:** Without upgrading facilities, equipment may become outdated, leading to higher operational costs over time.
5. **Lack of Flexibility:** A static production capacity might limit the factory's ability to respond to fluctuations in market demand, leading to potential supply chain issues.

Evaluating these factors is crucial for a factory to align its production capacity with its long-term strategic goals.

It should be noted additional iterations of the same prompts sometimes returned different pros and cons or some different pros and cons. It appears that discernment on the part of the user is paramount. In addition, it is imperative the participant understands the decision-making context to understand the value of the results returned by ChatGPT 4. Not all results benefit decision making in the context of the simulation equally. Nevertheless, simulation facilitators/faculty can use the simulation platform to segue into discussions regarding more comprehensive strategic decisions. The simulation provides the context while generative artificial intelligence provides extended content.

Operational Decisions

Operational decisions cover the daily operations of an organization. In the Income|Outcome simulation the following are some of the tasks done each simulation month:

- Update Receivables ○ Update Liabilities
- Buy Inventory for Cash or Credit ○ Take Depreciation and Pay Expenses/Labor

- Consider Supply and Demand in the Simulation Environment
- Pay Advertising
- Participate in the Market

In Income|Outcome, accounts receivable and accounts payable are managed using monthly update checklists and market-based decisions within the simulation. Generative AI offers participants insights and recommendations on how these processes are handled in real-world scenarios. Management of accounts receivable is specifically addressed in this paper.

ChatGPT 4 Prompt and Results (Personal conversation, 05/10/24):
“What are some ways a company can manage accounts receivable?”

Managing accounts receivable effectively is crucial for a company's cash flow and financial health. Here are some strategies:

1. **Clear Credit Policies:** Establish and communicate clear credit terms with customers, including payment deadlines and penalties for late payments, to ensure consistency.
2. **Credit Checks:** Assess a customer's creditworthiness before extending credit to minimize the risk of non-payment.
3. **Invoicing Promptly:** Send invoices as soon as a product is delivered or a service is completed, helping customers remember and prioritize the payment.
4. **Early Payment Incentives:** Offer discounts for early payments to encourage customers to pay their invoices promptly.
5. **Regular Follow-Up:** Implement a system for following up on overdue accounts, starting with friendly reminders and escalating to formal collection actions if needed.
6. **Aging Reports:** Regularly review aging reports to monitor outstanding invoices and identify patterns of late payments, allowing you to address issues proactively.
7. **Payment Plans:** For customers facing financial difficulties, offer structured payment plans to recover overdue amounts while maintaining a positive relationship.
8. **Automation and Technology:** Utilize accounting software to automate invoicing and reminders, ensuring efficient and accurate tracking.
9. **Outsourcing Collections:** Consider outsourcing collections to third-party agencies for persistently delinquent accounts.
10. **Sales Team Collaboration:** Work with the sales team to ensure they understand the importance of credit policies and involve them in collection efforts.

11. **Legal Action:** As a last resort, pursue legal action or arbitration to recover debts, but weigh this against potential costs and customer relationships.

Implementing these practices can help companies improve their accounts receivable management and maintain healthy cash flow.

Once again considering context, results 3 and 9 apply specifically to the Income|Outcome environment. Faculty can use this foundation to explore real-world considerations. Participants now have a base upon which to build understanding.

Participants in the Income|Outcome simulation earn sales by winning bids at the market. Each team competes for orders and the team with the lowest bid price on each order wins that order. A list of sample orders is presented in Table 2.

Order	Quantity of units in order	Credit terms	Your total bid
#203	4 units	Cash	
#204	3 units	n/30 Days	
#205	3 units	n/90 Days	

Table 2: Sample Monthly Market Orders

The market process provides participants with another opportunity to draw on the power of generative artificial intelligence. The following is a sample decision that a team will make before approaching the market competition table. ChatGPT 4 Prompt and Results (Personal conversation, 05/10/24):

“What factors should a company consider when deciding between selling with shorter payment terms and lower prices or longer payment terms with higher prices?”

When deciding between shorter payment terms with lower prices and longer payment terms with higher prices, a company should evaluate several factors:

1. **Cash Flow Needs:** Shorter payment terms may be necessary if the company requires immediate cash flow to fund operations or manage debts. Longer terms might delay cash inflows, which could cause liquidity issues.
2. **Customer Relationships:** Longer payment terms might strengthen customer relationships by providing more flexibility, especially for clients with cash flow challenges. However, shorter terms could be better for customers who prefer immediate settlement.
3. **Market Competition:** If competitors offer longer payment terms, matching or exceeding those terms could help retain customers. Conversely, in competitive markets where price sensitivity is high, lower prices with shorter terms might attract more business.
4. **Credit Risk:** Longer payment terms increase the risk of default or delayed payments. A company should assess customers’ creditworthiness to minimize bad debt.

5. **Industry Norms:** Understanding industry standards for payment terms can guide decisions to ensure that the company's practices align with customer expectations and competitive norms.
6. **Administrative Costs:** Longer terms may incur higher administrative costs for tracking overdue payments, sending reminders, and managing collections.
7. **Financing Costs:** If the company needs to borrow due to delayed receivables, longer payment terms could increase financing costs.
8. **Discounts and Incentives:** Assess the feasibility of providing early payment discounts or incentives to encourage quicker payments under longer terms.
9. **Pricing Strategies:** Evaluate how higher prices for longer payment terms could impact overall sales volume and profitability. Consider whether customers would accept a higher price for greater flexibility.
10. **Economic Environment:** In economic downturns, longer terms may be necessary to sustain sales, even if it means higher credit risks. In stronger economies, shorter terms might be more feasible.
11. **Legal Implications:** Review any legal obligations related to payment terms, such as late fee structures, to ensure compliance with applicable regulations.

Carefully analyzing these factors will help a company balance its needs with market demands, customer relationships, and financial health.

Based on how the simulation is managed, all of these recommendations apply to both Income|Outcome and the real-world. Exploring the same decision by asking ChatGPT 4 with a differently worded prompt can bring additional insights. Consider the following prompt and results. ChatGPT 4 Prompt and Results (Personal conversation, 05/10/24):

“Is it smarter to sell with lower prices and shorter payment terms or higher prices with longer payment terms?”

The choice between lower prices with shorter payment terms or higher prices with longer payment terms depends on the specific business context and objectives. Both approaches have their advantages and trade-offs:

Lower Prices and Shorter Payment Terms:

- **Benefits:** Immediate cash flow, reduced credit risk, and minimized potential bad debt. Quick payment terms help stabilize cash flow and can offer a competitive advantage in pricesensitive markets.

- **Drawbacks:** Lower prices may reduce profit margins and potentially impact profitability. It might also attract price-sensitive customers who are less loyal and more likely to switch for better offers.

Higher Prices and Longer Payment Terms:

- **Benefits:** Increased revenue per sale through higher pricing, and longer terms can attract and retain customers who need more flexibility. It may also build customer loyalty by demonstrating a willingness to accommodate their financial needs.
- **Drawbacks:** Longer payment terms increase the risk of delayed payments or defaults, which can strain cash flow and require more rigorous credit management. It may also lead to higher administrative costs in managing outstanding receivables.

Which Approach is Smarter?

- **Cash Flow Needs:** Companies requiring immediate cash flow for operational purposes might benefit from shorter payment terms.
- **Credit Risk Management:** If customers have a history of delayed payments, shorter terms could minimize risks.
- **Market Norms and Competition:** Consider industry norms and what competitors offer. If longer terms are standard, aligning with those might be necessary.
- **Customer Relationships:** If fostering long-term customer relationships is important, longer terms could provide a customer-centric approach.
- **Pricing Strategy:** Assess the market's acceptance of higher prices. Longer terms are justified only if the customer perceives value in flexibility.

In conclusion, there isn't a one-size-fits-all answer. Companies must analyze their market, customer base, and financial situation to determine which approach aligns best with their strategic goals.

Ratio Analysis

A notable business analytics related strength of many business simulations is their requirement for participants to prepare and review financial ratios. At the end of each simulation month, Income|Outcome participants finalize their financial statements and calculate the "triangle of forces"—asset turnover, return on sales, and return on assets. These ratios are then benchmarked against other teams in the simulation to assess performance. Generative artificial intelligence also aids participants in exploring the factors influencing these ratios. For instance, they can query how to improve their company's asset turnover. ChatGPT 4 Prompt and Results (Personal

conversation, 05/10/24): “What are some strategies a company can use to increase asset turnover?”

To increase asset turnover, a company can adopt various strategies aimed at improving the efficiency of its assets and boosting revenue generation. Here are some approaches:

1. **Optimize Inventory Management:** Implement inventory optimization techniques, like just-in-time (JIT), to reduce holding costs and minimize excess inventory. This strategy ensures assets are not tied up unnecessarily.
2. **Enhance Sales and Marketing:** Boost sales through effective marketing campaigns, entering new markets, or developing new distribution channels. Higher sales can lead to better utilization of current assets.
3. **Asset Utilization:** Improve asset utilization by reducing downtime and ensuring production facilities operate at optimal capacity. Regular maintenance and preventive measures can minimize equipment breakdowns.
4. **Product Line Rationalization:** Review the product portfolio and discontinue underperforming or non-core products, allowing the company to focus on more profitable lines.
5. **Lease or Sell Idle Assets:** Unused or underutilized assets can be leased out or sold, freeing up capital and focusing resources on more productive assets.
6. **Improve Operational Efficiency:** Streamline operations and processes to improve productivity. This can include automation, workforce training, or workflow improvements.
7. **Reduce Accounts Receivable:** Shorten the payment terms to collect receivables more quickly. Improved cash flow allows reinvestment in other productive activities.
8. **Enhance Supplier Negotiations:** Negotiate favorable terms with suppliers to reduce lead times and improve inventory turnover.
9. **Leverage Technology:** Utilize technology to monitor and optimize asset performance, predict maintenance needs, and make informed strategic decisions.
10. **Outsource Non-Core Functions:** Outsource non-core operations to external partners to reduce fixed asset investments and focus on core business activities.

By carefully implementing these strategies, companies can achieve a more efficient asset turnover, leading to improved overall financial performance.

As with previous prompts and responses, facilitator faculty have the opportunity to help participants understand which options apply in their specific context. Ratio analysis is empowered.

Conclusion

Generative artificial intelligence is going to bring a paradigm shift to the world of business analytics education. Business simulations have done much to provide faculty with environments in which students can apply and practice classroom concepts and skill sets. Generative artificial intelligence now brings the world of business knowledge into the simulation environment. Faculty can demonstrate how to build prompts to retrieve powerful results. Faculty can then instruct students in how to discern which AI results to the context at hand. The coming paradigm shift will result in students being better equipped with the knowledge and discernment skills necessary for them to become next generation business leaders. **References**

- Anderson, P. H., & Lawton, L. (2009). Business simulations and cognitive learning: Developments, desires, and future directions. *Simulation & Gaming, 40*(2), 193-216.
- Babin, R. (2023). How can ai be used to enhance case simulations? In EDULEARN23 Proceedings (pp. 699-706). IATED.
- ChatGPT 4, personal communication, May 9, 2024.
- Ferreira, C. P., González-González, C. S., & Adamatti, D. F. (2021). Business simulation games analysis supported by human-computer interfaces: A systematic review. *Sensors, 21*(14), 4810.
- Humpherys, S. L., Bakir, N., & Babb, J. (2022). Experiential learning to foster tacit knowledge through a role play, business simulation. *Journal of Education for Business, 97*(2), 119125.
- Levant, Y., Coulmont, M., & Sandu, R. (2016). Business simulation as an active learning activity for developing soft skills. *Accounting Education, 25*(4), 368-395.
- Kurtz, J. (2003). Business wargaming: Simulations guide crucial strategy decisions. *Strategy & Leadership, 31*(6), 12-21.
- OpenAI. (2023). ChatGPT (September 25 version) [Large language model]. <https://chat.openai.com/chat>
- Moser, J. G. (1986). Integration of artificial intelligence and simulation in a comprehensive decision-support system. *Simulation, 47*(6), 223-229.
- Mubaraz, S., & Mezrar, B. (n.d.). Case of Business Simulations. *Sisällys/content*, 28.
- Paschen, J., Wilson, M., & Ferreira, J. J. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons, 63*(3), 403-414.
- Schwarz, J. O. (2009). Business wargaming: Developing foresight within strategic simulation. *Technology Analysis & Strategic Management, 21*(3), 291-305.

Taylor, S. J., Son, Y. J., Branke, J., Rose, O., & Sanchez, S. M. (2021, December). Using simulation and artificial intelligence to innovate: are we getting even smarter? In 2021 Winter Simulation Conference (WSC) (pp. 1-9). IEEE

Predicting Bankruptcy: Ask the Employees

Abstract

The purpose of the paper is to show that employees have information on an upcoming bankruptcy years before actual bankruptcy filings. We test our model using employee information against other bankruptcy models from the literature using a dataset from 2008 to 2020. We build a new model to reflect employees' attitudes before bankruptcy filings and include key variables from the already established bankruptcy models in the literature in our model. We find that our model more accurately predicts bankruptcy two to three years before the actual bankruptcy filing, while the other models are more accurate in the year prior to the bankruptcy. For two and three years prior to a bankruptcy filing, the model outperforms the existing models in in- and out-of-sample tests. We create neural network models consisting of reviews and ratings separately and show that textual reviews provide additional predictive power for bankruptcy filings on top of rating and financial information.

Keywords: bankruptcy filing and emergence prediction models, employees' attitudes, in and out-of-sample tests, neural network, Glassdoor

JEL Codes: G33, G41, C53

I. Introduction

The purpose of the paper is to test the predictive performance of established bankruptcy prediction models against a new model inclusive of employee information. Our basic hypothesis is that employee satisfaction shows up as a predictor of financial distress prior to financial statement data. Well before bankruptcy or even negative financial performance, managers and employees may be aware of significant problems. Although managers may be reluctant to disclose this information, workers may reveal problems through dissatisfaction with their jobs and the firm. However, we empirically test our model for predictability not causation. Whether employees are less satisfied because of an impending bankruptcy or whether employee satisfaction impacts the chances of bankruptcy is an interesting topic for further studies.

We document that employee satisfaction is a strong predictor of bankruptcy. Specifically, we find that the employee satisfaction model predicts bankruptcy more accurately than any of the existing financial information-based models in all years other than the year immediately prior to a bankruptcy filing. We additionally find that close to the bankruptcy filing date, models with inclusion of both financial statement and employee satisfaction data outperform models with inclusion of financial data only. Separately, we hypothesize that if a company is more likely to emerge from bankruptcy, employees are more likely to keep their jobs which would increase their level of satisfaction. We document that employee satisfaction predicts bankruptcy emergence and that companies with higher employee satisfaction are more likely to emerge from bankruptcy.

We show that employees have information on an upcoming bankruptcy years before actual bankruptcy filings. We test four key bankruptcy models from the literature using a dataset from 2008 to 2020 to show that each one contains unique information regarding the probability of bankruptcy filings. We also build a new model to reflect employees' attitudes and emotions before bankruptcy filings and include key variables from each of the four already established bankruptcy models in the literature in our model. We make three novel findings. First, employee satisfaction shows up as a predictor of bankruptcy prior to financial data-based models. Second, when we add employees' attitudes in each of the four bankruptcy models, we find improvement to their predictive performance, although it is a small improvement above the models' results. We conclude that in the year before the bankruptcy filing, financial statement and market information overwhelm any other information about the company. Third, employee satisfaction around the bankruptcy

filing predicts whether the company will emerge from bankruptcy. We perform several analyses including parameter estimation, bankruptcy classification rates, and out-of-sample analyses. We also compare the models based on model-fit criteria, such as the Receiver Operating Characteristics area (ROC area) and the Pseudo R-squared. Pseudo R-Squared and ROC are the highest for the Altman's Z-Score using our estimation of models one year prior to bankruptcy filings, but we more closely examine if this finding changes, as we move back in time from actual bankruptcy filings, and we make the following key conclusions.

While Altman's model exhibits the highest Pseudo R-Squared and ROC one year prior to bankruptcy filings, our model exhibits a more consistent ROC over the three years prior to bankruptcy filings and higher ROC two and three years prior to bankruptcy filings. We make the conclusion that employee satisfaction shows up as a predictor of bankruptcy prior to financial statement data, but that it doesn't outperform financial statement data one year prior to bankruptcy filings. In addition, when we add employee satisfaction to the other four models, their predictive power increases slightly close to bankruptcy filings.

We more closely compare the Altman's Z-Score and our model. In both Altman's and employees' satisfaction's models, as the sensitivity increases, the specificity decreases. The optimal threshold for both models is between 0.60 and 0.75, which is where we have the highest sensitivity and specificity. Additionally, we find that the employees' satisfaction's model classifies 76.89%, 76.52%, and 71.53% of the out-of-sample bankruptcies one, two, and three years before actual bankruptcy filings at the level of 0.5. We also find that the model has sensitivity of 86.54%, 81.65%, 89.91% one, two, and three years before actual bankruptcy filings and specificity of 62.15%, 68.69%, 43.46% one, two, and three years before actual bankruptcy filings indicating that the model has the lower percentage of Type I and Type II errors. We also report sensitivity analyses and provide the accuracy of the employees' satisfaction's model based on the various rating categories, at a given threshold level, in predicting out-of-sample defaults along with sensitivity and specificity. We are using the 0.5 threshold and report findings for one, two, and three years prior to bankruptcy filings. The results with the breakdown of ratings are consistent with the main findings and we find that the models exhibit a high correct classification rate and low occurrence of Type I and Type II errors.

In additional set of tests, we build neural networks using employee satisfaction ratings and employee satisfaction reviews separately to test the predictive performance of employee information alone without any financial statement information added onto the models. For the neural network with employee satisfaction reviews, we use 13,349 reviews in the year before the bankruptcy filing which include 6,906 actual bankrupt companies' reviews and 6,443 pseudo bankrupt companies' reviews. We clean the text from Stop Words and use the Bag of Words approach to tokenize our sentences and turn the text into numerical values. After doing so, we split our sample into training and testing data. We use a machine learning technique to classify the text and fit the data, fit the classifier on the resampled data, and make predictions based on the testing data. We find that the neural network with textual reviews provides an accuracy of 73.37% and a standard deviation of 1.16%. The ROC curve area comes out to be 0.75 which is higher than the ROC curve area for the neural network using employee satisfaction ratings of 0.64. The accuracy of the neural network with ratings is also lower and standard deviation is higher (61.66% and 2.37%, respectively). The results point to employee satisfaction textual reviews possessing superior information to employee satisfaction ratings one year before bankruptcy filings. In addition to this finding, we also show that employee satisfaction is predictive of bankruptcy emergences one year before bankruptcy filings, although the effect disappears as we move back two and three years before bankruptcy filings.

II. Literature Review and Hypotheses Development

Our paper adds to the literature on bankruptcy prediction models. Finance literature has determined several bankruptcy prediction models, such as [Altman's Z-Score \(1968\)](#), [Ohlson's Model \(1980\)](#), [Zmijewski's Model \(1984\)](#), and [Shumway's Model \(2001\)](#). Since the development of those models, researchers have made efforts to develop models with even greater predictive performance. Prior bankruptcy prediction models have employed financial ratios from financial statements before the bankruptcy filing, while more recent models have used financial market data, such as excess stock returns and stock return volatility, along with the application of the BlackScholes option-pricing model. The earliest studies on bankruptcy prediction have utilized univariate analyses which have focused on individual ratios and comparison of ratios of failed companies with those of successful firms. Those earliest studies have laid the groundwork for multivariate studies. Among the

univariate studies, those to be highlighted include [Merwin \(1942\)](#), [Chudson \(1945\)](#), and [Beaver \(1966\)](#). In his study of small manufacturers, [Merwin \(1942\)](#) found that three ratios are significant indicators of business failure – Net Working Capital to Total Assets, Net Worth to Total Debt, and the Current Ratio – even four or five years before failure. [Chudson \(1945\)](#) has tried to determine whether there is a normal pattern to prediction of bankruptcy and has reported that there is no such pattern, but there is a clustering of ratios within industry, size, and profitability groups. The most popular univariate study comes from [Beaver \(1966\)](#) in which he compares 30 ratios of 79 failed and 79 non-failed companies in 38 industries and tests their predictive abilities by classifying them under bankrupt and non-bankrupt firms. The author has given future researchers the idea to consider multiple ratios altogether as they might have higher predictive ability than single ratios.

The most popular multivariate study remains [Altman \(1968\)](#). In this paper, the author uses a multivariate discriminant analysis to develop a five-factor model to predict bankruptcy of manufacturing firms. The paper documents that the Z-Score predicts bankruptcy if the firm's score falls within a certain range. The model has high predictive accuracy one year before the merger, but the accuracy falls off two, three, four, and five years before business failure. Since Altman's study on bankruptcy prediction, models' number and complexity have increased. More recent papers have utilized logit and probit models, neural networks, and multivariate discriminant analysis, or some combination of those models. For example, [Mensah \(1983\)](#) has used both multivariate discriminant analysis and logit analysis to predict bankruptcy. In the 1980s, neural networks became the dominant method used. Neural networks analyze inputs to find patterns and develop a model capable of a decision-making process. During the training mode, in which the network learns the decision-making process, several sample cases are run, while during the testing mode the neural network model is validated using hold-out sample data.

Another strand of literature, which the paper contributes to, is related to the impact of employees' information content on various firm outcomes. Research utilizing Glassdoor has been growing due to its coverage of public firms and its presentation of the perception of a firm's work environment from employees' standpoint. For example, [Ji et al. \(2017\)](#) study the association between financial reporting and job satisfaction, company culture, and the opinions of senior leadership. The authors show that poorly implemented performance objectives lead to pressure and create a boiler room effect which has a negative impact on the corporate climate of the firm. This

is reflected in firms with lower levels of job satisfaction and lower levels of culture and values being more likely to be subjected to SEC fraud enforcement. [Green et al. \(2019\)](#) find support for the reviews being accurate assessments of and revealing value-relevant information about a firm's culture based on changes in employee morale ratings predicting future firm performance. The authors find that the return effect is concentrated among reviews from current employees and stronger among early firm reviews. They also determine that when they decompose employee morale ratings, the return effect is related to changes in categories such as *Career Opportunities*, but not *Work-Life Balance*. Additionally, in the first paper in my dissertation, I show that the similarity between acquirer and target employees' emotions in mergers and acquisitions is an important ingredient for the success of M&A deals, while in the second paper in my dissertation, I show that low similarity between private equity acquirers and their targets enhances post-merger integration and short- and long-term synergies. In this paper, we show that employees have information on upcoming bankruptcies by testing various already established models in the literature against a newly established one consisting of employees' attitudes and show that its predictive power is greater than the established models up to two and three years before bankruptcy filings, and that when a proxy for employee satisfaction is added to the established models, their predictive power increases.

Other more recent papers utilizing the Glassdoor dataset focus on themes varying from employee sentiment to gender diversity. [Marchetti \(2019\)](#) measures organizational culture using text reviews employees have posted on Glassdoor and finds that pre-deal acquirer-target culture compatibility is positively associated with the likelihood of a deal announcement and with superior stock returns. [Chen et al. \(2022\)](#) propose an employee sentiment index, complementing investor sentiment and manager sentiment indices, and find that high employee sentiment predicts low monthly (weekly) market returns significantly both in- and out-of-sample. [Campbell and Shang \(2021\)](#) show that information extracted from Glassdoor reviews can be used to develop measures with useful properties for measuring misconduct risk and those measures clearly discriminate between high and low misconduct firms. [Liu et al. \(2022\)](#) use Glassdoor data to show that firms offer higher quality maternity leave benefits in labor markets where female talent is relatively scarce. The main argument of the paper is that firms use non-wage benefits, specifically femalefriendly benefits, such as maternity leave, to increase gender diversity by efficiently

attracting women. [Welch and Yoon \(2021\)](#) find that high-ability managers allocate resources to ESG efforts in a way that enhances shareholder value. Their most important finding is that firms with highly rated managers and high ESG exhibit significantly higher future stock returns than firms with low ratings on both or firms with only a high ESG or a high employee opinion rating. In the spirit of those papers, I use Glassdoor data to show that employees can predict bankruptcy filings and emergencies.

Building on prior bankruptcy prediction and Glassdoor research, we build the following three hypotheses:

- i. Around bankruptcy announcements, financial information overwhelms employee satisfaction information, but employee satisfaction shows up as a predictor of bankruptcy prior to financial statement data.
- ii. Employee satisfaction around the bankruptcy filing predicts whether the company will emerge from bankruptcy.

III. Models and Data

In this paper, we utilize UCLA's LoPucki, Glassdoor, Compustat, and CRSP databases. We use UCLA's LoPucki data for bankruptcy filings' information, Glassdoor data for employees' satisfaction metrics, Compustat data for financial statement variables, and CRSP data for return variables. Our sample consists of 327 deals in the period between 2008 and 2020. In addition to including data on employees' satisfaction, we include profitability variables, liquidity variables, leverage variables, firm-size variables, other firm characteristics, such as standard deviation of the return, excess return, and logarithm of firm age, in accordance with prior bankruptcy prediction studies. We use a matched-pair methodology to compare the estimation procedures of the models we are testing. We construct a matched-pair sample based on industry, size (based on a company's total assets), and book-to-market. The match generated for each bankrupt firm is based on minimizing the absolute value of the ratio of the difference between the firm size and book-to-market of the bankrupt firm and that of the healthy firm to the firm size and book-to-market of the bankrupt firm. The models we include in our study are Altman (1968), Ohlson (1980), Zmijewski (1984), Shumway (2001), and we build a new model where we use some key variables from previous models and include employees' satisfaction proxy.

- Altman (1968): Multiple discriminant analysis with financial ratios

$$Z = \beta'X$$

where Z is the MDA score and X represents the variables listed. Cutoff point: $Z \geq 2.675$, classified as non-bankrupt. Under X , we include Net Working Capital/Total Assets, Retained Earnings/Total Assets, EBIT/Total Assets, Market Value of Equity/Book Value of Total Liabilities, and Sales/Total Assets.

- Ohlson (1980): Logit model with financial ratios

$$P = (1 + \exp\{-\beta'X\})^{-1}$$

where P is the probability of bankruptcy and X represents the variables listed. The logit function maps the value of $\beta'X$ to a probability bounded between 0 and 1. Under X , we include Ohlson's Size, Total Liabilities/Total Assets, Working Capital/Total Assets, Current Liabilities/Current Assets, OENEG (a dummy variable indicating if Total Liabilities exceed Total Assets), Operating Income/Total Liabilities, INTWO (a dummy variable indicating if Net Income was negative for the last 2 years, and 0 otherwise), and Change in Net Income.

- Zmijewski (1984): Probit model with financial ratios $P = \Phi(\beta'X)$

where P is the probability of bankruptcy and X represents the variables listed, and $\Phi(\bullet)$ represents the cumulative normal distribution function. The probit function maps the value of $\beta'X$ to a probability bounded between 0 and 1. Under X , we include Net Income/Total Liabilities, Total Liabilities/Total Assets, and Current Assets/Current Liabilities.

- Shumway (2001): Hazard model with both financial ratios and market data

$$P_{i,t} = (1 + \exp\{-y_{i,t}\})^{-1}$$

$$y_{i,t} = \alpha + \beta'X_{i,t-1} = \beta' \begin{bmatrix} X_{1,t-1} \dots X_{1,t-j} \\ \dots \dots \\ X_{n,t-1} \dots X_{n,t-j} \end{bmatrix}$$

where P is the probability of bankruptcy and X represents the variables listed. This is a multiperiod logit model, but instead of treating each firm-year as an independent observation, all prior values of the independent variables for a particular firm are included in the information set. n represents the number of independent variables, and j represents the number of time periods prior to time t for which data are available. Under X , we include Net Income/Total Liabilities, Total Liabilities/Total Assets, Relative Size, Cumulative Annual Return in prior year minus the valueweighted CRSP index in prior year, and Standard Deviation of Return in prior year.

○ Employees' Model: Hazard model with employees' data

$$P_{i,t} = (1 + \exp\{-y_{i,t}\})^{-1}$$

$$y_{i,t} = \alpha + \beta' X_{i,t-1} = \beta' \begin{bmatrix} X_{1,t-1} \dots X_{1,t-j} \\ \dots \dots \\ X_{n,t-1} \dots X_{n,t-j} \end{bmatrix}$$

where P is the probability of bankruptcy and X represents the variables listed. This is a multiperiod logit model. n represents the number of independent variables, and j represents the number of time periods prior to time t for which data are available. We keep only one firm-year observation for each bankrupt firm but all firm-year observations for surviving and pseudo firms. We go back three years before bankruptcy filings and track all variables included in our model up to three years of bankruptcy filing. Under X , we include Mean of Employees' Satisfaction, EBIT/Total Assets, Working Capital/Total Assets, Total Liabilities/Total Assets, Change in Net Income, Excess Return, Standard Deviation of Return, and Logarithm of the closing price of prior year.

IV. Results

[Insert Table 1 here]

Table 1 presents summary statistics for each variable used in the study over the period 2008-2020. The dataset contains 327 actual bankrupt firms, which we define as non-healthy, and 327 nonbankrupt firms, which we define as healthy. Panel A of Table 1 groups the variables into various categories going back three years before actual bankruptcy filings. The profitability variables include EBITTA (EBIT to Total Assets), Sales, NITA (Net Income to Total Assets), and CHIN (Change in Net Income), and they measure the ability of the firm to generate sufficient profits or

returns to remain a company that is going concern. The liquidity variables include WCTA (Working Capital to Total Assets), CLCA (Current Liabilities to Current Assets), FUTL (Funds from Operations to Total Liabilities), INTWO (a dummy variable signifying if the firm has a negative Net Income over the past two years), and they measure the ability of the firm to meet its short-term obligations. The leverage variables include RETA (Retained Earnings to Total Assets), MVETL (Market Value of Equity to Total Liabilities), TLTA (Total Liabilities to Total Assets), OENEG (a dummy variable indicating if Total Liabilities exceed Total Assets), and they measure the relative amount of debt and other obligations of the firm. The firm-size variables include OhlsonSize (calculated as the natural logarithm of Total Assets to the GNP Price-Level index) and RelativeSize (calculated as the natural logarithm of the Number of Outstanding Shares multiplied by the Year-End Share Price to Total Market Value), and they reflect the ability of larger firms to trade through difficult times and their ability to be less likely to go bankrupt. We also include other firm characteristics, such as the Excess Return, Standard Deviation of Excess Return, and we determine that bankrupt firms tend to have lower and more volatile returns and that they have substantially underperformed in their years prior to bankruptcy. In addition, those firms tend to be younger (as observed by their firm age). Panel B of Table 1 reports the mean and standard deviation of the sum of all rating categories up to three years before actual bankruptcy filings. As can be seen, the mean increases the closer we get in time to the bankruptcy filing and the standard deviation increases the closer we get in time to the bankruptcy filing. Panel C of Table 1 presents the breakdown of the sum of mean and standard deviation into different categories – *Career Opportunities, Compensation Benefits, Culture Values, Senior Leadership, Work-Life Balance, and Overall Rating* – up to three years before actual bankruptcy filings. One can observe that the rating category means are relatively low on a possible range between 1 and 5. Panel D of Table 1 presents the correlations between the Z-Score and the mean and standard deviation of sum of rating categories, and the individual employees' satisfaction rating categories. The correlation between the Z-Score and the mean and standard deviation of the sum and the various rating categories is low and negative one and two years before actual bankruptcy filings, while the correlation between the Z-Score and the mean and standard deviation of the sum and the various rating categories is low and positive three years before actual bankruptcy filings. This could be attributed to the fact that the predictive power of the Z-Score decreases the further we move back in time from actual bankruptcy filings. This

finding is documented by Altman (1968), but it is also what we find in our sample of bankruptcies between 2008 and 2020. Our main findings are discussed in the following few pages.

[Insert Figure 1 and Figure 2 here]

Figure 1 shows the yearly distribution of our bankruptcy sample from 2008-2020. As expected, 2009 and 2020 are the years with the most bankruptcies (63 and 51, respectively). Figure 2 shows a plot of the average marginal effects and the predictive margins with 95% confidence interval of the mean of sum of rating categories one, two, and three years before bankruptcy filings. The figures plot the predicted outcome at various levels of the mean one, two, and three years before. The probability of a bankruptcy filing decreases with the increase in firms' employee satisfaction on average. Those companies with lower employee satisfaction are more exposed to the risk of an actual bankruptcy filing.

[Insert Table 2 here]

Table 2 reports the parameter estimates from the various four models and our model discussed above. We use the same variables that have been used by the original authors in the authors' proposed setting except for Altman's model in which we use a multi-period logistic setting. We include one firm-year observation for each company filing for bankruptcy and all firm-year observations for matched control and surviving companies for all three years prior to bankruptcy filing. The discriminant analysis coefficients (both standardized and unstandardized) are presented in [Table A2](#) columns (1) and (2) in the Appendix. The estimated coefficients in [Table 2](#) from the models are bolded if there is any statistical significance. We set the dependent variable to be equal to 1 for bankrupt firms and to 0 otherwise. Therefore, if we observe a positive coefficient, it means that a higher value of that particular variable increases the likelihood of bankruptcy. We can see that the companies with lower profitability and liquidity, but higher leverage are more likely to file for bankruptcy, and smaller firms are more likely to file for bankruptcy. Our employee satisfaction model includes employee satisfaction proxy and profitability, liquidity, and leverage variables additionally. It also includes the lagged return and volatility of the return, and the firm-size variable as indicated by the company's stock price. We also compare the models based on model fit-criteria, such as the Receiver Operating Characteristics area (ROC area) and the Pseudo R-squared. Pseudo

R-Squared is the highest for the Altman's Z-Score using our estimation of models one year prior to bankruptcy filings, but we more closely examine if this finding changes as we move back in time from actual bankruptcy filings. The models' performance based on ROC is examined in Figure 3, while in Figure 4 we document the optimal cutoff points for the various models.

[Insert Figure 3 and Figure 4 here]

In Figure 3, one can observe that Altman's model has the highest ROC (0.9084), while the employees' satisfaction model is the second-best performing model one year before bankruptcy filings (0.8867). We further examine how the ROC changes from three to one years before bankruptcy filings for the Altman's and employees' satisfaction models in the following paragraphs and document that the ROC of the employees' satisfaction model is higher than the ROC of the Altman's model two and three years before bankruptcy. Figure 4 shows the optimal cutoff point for the various models and the relationship between sensitivity and specificity. Sensitivity is the correct classification of true default and specificity is referred to as the correct classification of true non-default. As can be seen, in both the Altman's and employee satisfaction's models, as the sensitivity increases, the specificity decreases. The optimal threshold for both models is between 0.60 and 0.75, which is where we have the highest sensitivity and specificity. We perform additional analyses of the sensitivity and specificity of the two models in the coming tables.

[Insert Table 3 here]

[Insert Figure 5 and Figure 6 here]

Table 3 presents parameter estimates for the four models with inclusion of the employee satisfaction proxy, similar to the approach in Table 2. We use the same variables that have been used by the original authors in the authors' proposed setting except for Altman's model in which we use a multi-period logistic setting. The discriminant analysis coefficients (both standardized and unstandardized) are presented in [Table A2](#) columns (3) and (4) in the Appendix. Even though the R-squared for all models increases with the addition of the employee satisfaction proxy, we cannot make the conclusion that the addition of employee satisfaction information improves the predictive performance of the above-mentioned bankruptcy prediction models. Employee satisfaction is negatively associated with the probability of bankruptcy filings in all models, although the coefficient is not statistically significant. Figure 5 shows a plot of the ROC curves for

the four models with inclusion of the employee satisfaction proxy. The ROC with the employee satisfaction proxy increases in comparison to the ROC of the models without the employee satisfaction proxy. The Altman's ROC increases from 0.9084 to 0.9182. The greatest improvement in the ROC, however, is in Zmijewski's model – the ROC increases from 0.5391 to 0.5959. Figure 6 plots the optimal cutoff point for the models in Table 3. The optimal cutoff point for Altman's model is now in the range between 0.50 and 0.75, but similar to the findings in Figure 4, specificity decreases as sensitivity increases.

[Insert Table 4 here]

[Insert Figure 7 here]

We move onto reporting results of bankruptcy classification rates and out-of-sample forecast accuracy. In Tables 4, 5, 6, and 7, we use all actual model econometric techniques and variables each author proposed in each respective paper. In Table 4, we report the rate of Types I and II errors for the five models tested in the paper. The table shows the incidence of Type I errors (classifying a bankrupt firm as healthy) and Type II errors (classifying a healthy firm as bankrupt), according to model scores. For example, under the Altman model if companies with model scores above the 85th percentile are classified as bankrupt, and those with model scores below the 85th percentile are classified as healthy, Types I and II error rates are 73.39% and 2.03%, respectively.

In comparison to Altman's model, under the employee satisfaction model if companies with model scores above the 85th percentile are classified as bankrupt, and those with model scores below the 85th percentile are classified as healthy, Types I and II error rates are 77.68% and 3.74%, respectively. Figure 7 plots the Total error rate based on different percentiles. Both the classification rate plot for Altman's model and the classification rate plot for the employee satisfaction model follow the same behavior which is also reflected in the results in Table 4.

[Insert Table 5 here]

[Insert Figure 8 here]

In Table 5, we report the rate of Types I and II errors for the four models with the inclusion of the employee satisfaction proxy. Under the Altman model if companies with model scores above the 85th percentile are classified as bankrupt, and those with model scores below the 85th percentile are classified as healthy, Types I and II error rates are 74.31% and 3.05%, respectively. Under the

Ohlson model if companies with model scores above the 85th percentile are classified as bankrupt, and those with model scores below the 85th percentile are classified as healthy, Types I and II error rates are 76.76% and 2.34%, respectively. Figure 8 shows the classification rates for the four models with inclusion of employee satisfaction. Altman's and Shumway's classification rate curves behave consistently with and without the employee satisfaction proxy.

[Insert Table 6 here]

The out-of-sample performance of each model is summarized in Table 6. We estimate the bankruptcy probability for each observation in the out-of-sample period. These observations in the out-of-sample period are ranked by the estimated bankruptcy probability, and we group observations into deciles on this basis. Then, the actual number of bankruptcies from among those firms are classified in each decile. We report both the numbers of actual and predicted bankruptcy filings classified into deciles by their estimated probability of bankruptcy. As one can see, the employee satisfaction model has an increasing number of bankruptcies from the bottom to the top decile. The top decile contains 49 of the actual bankruptcies, while the predicted bankruptcies have been approximately 53.8. Altman's model has 38 actual bankruptcies in the top decile with approximately 38.16 predicted bankruptcies. The predictions for this model most closely match the actual bankruptcies. We further conduct an out-of-sample performance analysis for the various models with the inclusion of the employee satisfaction proxy in those models.

[Insert Table 7 here]

Table 7 reports the out-of-sample performance for the various models with the inclusion of the employee satisfaction proxy. Similar to what we do in Table 6, in Table 7, we report the actual number of bankruptcies from among the firms classified in each decile. Similar to Table 6, we report the actual and predicted numbers for each model. The addition of the employee satisfaction proxy improves the out-of-sample performance in the top decile for Ohlson's, Zmijewski's, and Shumway's models. The actual and predicted bankruptcy filings are identical (51 actual and 52.99 predicted, 51 actual and 51.34 predicted, 52 actual and 52.63 predicted for Ohlson's, Zmijewski's, and Shumway's models, respectively).

[Insert Table 8 and Table 9 here]

Table 8 reports sensitivity analyses and provides the accuracy of Altman's and employee satisfaction's models, at a given threshold level, in predicting out-of-sample defaults along with sensitivity and specificity. We use a multi-period logit setting for both models in this table. We include one firm-year observation for each company filing for bankruptcy and all firm-year observations for matched control and surviving companies for all three years prior to bankruptcy filing. In Table A3, we present error rates but in a discriminant analysis setting for both models, and in this setting, we also find more balanced and lower stratified and unstratified total error rates of the ratings' model persisting three years before the bankruptcy filing. We define sensitivity as the true classification of default firms when the firms actually defaulted, and specificity as the true classification of non-defaulted firms as non-defaulted. The two types of uncertainty are Type I and Type II errors. The Type I error classifies a bankrupt company as non-bankrupt, while the Type II error classifies the non-bankrupt firm as bankrupt. We can see that Altman's model classifies 70.54%, 61.00%, 63.40% of the out-of-sample bankruptcies one, two, and three years before actual bankruptcy filings at the threshold level of 0.5. The threshold level of 0.5 indicates that the model can differentiate between defaulted and non-defaulted firms. Additionally, we can see that the employee satisfaction model classifies 76.89%, 76.52%, and 71.53% of the out-of-sample bankruptcies one, two, and three years before actual bankruptcy filings at the level of 0.5. We can also see that the model has sensitivity of 86.54%, 81.65%, 89.91% one, two, and three years before actual bankruptcy filings and specificity of 62.15%, 68.69%, 43.46% one, two, and three years before actual bankruptcy filings indicating that the model has the lower percentage of Type I and Type II errors.

In Table 9, we report the sensitivity analyses and provide the accuracy of the employee satisfaction's model based on the various rating categories, at a given threshold level, in predicting out-of-sample defaults along with sensitivity and specificity. We are using the 0.5 threshold and report findings for one, two, and three years prior to bankruptcy filings. The results with the breakdown of ratings are consistent with the main findings in Table 8 and we find that the models exhibit a high correct classification rate and low occurrence of Type I and Type II errors.

[Insert Figure 9 here]

Figure 9 plots the ROC curve from the tests in Table 8. The ROC score for Altman's model is 0.9084, 0.7808, 0.7075 one, two, and three years before bankruptcy filings, while the ROC score for the employee satisfaction model is 0.8867, 0.8280, 0.8212 one, two, and three years before bankruptcy filings. We can conclude that the ROC score for the employees' satisfaction model is more consistent, even though it still decreases the further we go back from the time of the actual bankruptcy filing.

[Insert Table 10 and Figure 10 here]

[Insert Table 11 and Figure 11 here]

To provide additional proof that employee satisfaction ratings and reviews provide bankruptcy prediction information above financial information, we build a neural network using both employee satisfaction reviews and employee satisfaction ratings. Table 10 and Figure 10 provide results of a neural network with employee satisfaction reviews. For this neural network, we use 6,906 actual failed company reviews and 6,443 matched pseudo company reviews present in our sample for one year before bankruptcy filing described in the data section to predict bankruptcy filing. We use the pros, cons, and feedback sections in Glassdoor to make our bankruptcy filing predictions. Table 11 and Figure 11 provide results of neural network with employee satisfaction ratings. For this neural network, we use 6,906 actual failed company reviews and 6,443 matched pseudo company reviews present in our sample for one year before bankruptcy filing described in the data section to predict bankruptcy filing. We use the career opportunities, compensation benefits, culture values, senior leadership, work-life balance, and overall rating ratings present in Glassdoor to make our bankruptcy filing predictions. In the neural network with employee satisfaction reviews, we split our sample into training and test datasets where our training dataset includes 10,679 reviews and our test dataset includes 2,670 reviews, while in the neural network with employee satisfaction ratings, we split our sample into training and test datasets where our training dataset includes 10,679 ratings and our test dataset includes 2,670 ratings. Approximately 60% of our data is training data, while 40% of our data is testing data.

We use the Synthetic Minority Oversampling Technique to oversample the minority class. This procedure was first introduced by Chawla et al. (2002) and has been used in machine learning

research thereafter. It can be used to create as many synthetic examples for the minority class as are required. It allows us to use random under-sampling to trim the number of examples in the majority class and to oversample the minority class to balance the class distribution after. We use the CatBoostClassifier as a machine learning model to fit the data, we fit the classifier on the resampled data, and we make bankruptcy predictions based on the test set. We calculate the confusion matrix for each neural network and compute the models' accuracy. We use the same approach for both neural networks. For the neural network with employee satisfaction reviews, we clean the textual reviews in Python from Stop Words and use the Bag of Words approach to apply tokenization to all sentences. Our results show that the neural network with textual reviews produces the higher accuracy of 73.37% in comparison to the neural network with ratings that produces accuracy of 61.66%. Additionally, the neural network model with textual reviews shows the better fit as evident by the ROC curve. The ROC for the neural network with reviews is 0.75, while the ROC for the neural network with ratings is 0.64. We conclude that textual reviews provide additional information necessary for bankruptcy prediction and allow for more accurate bankruptcy filing predictions.

[Insert Figure 12 and Figure 13 here]

[Insert Table 12 and Table 13 here]

We also perform additional survival analyses to examine predictions of companies' emergence from bankruptcies and determine that companies with high employee satisfaction are more likely to emerge from a bankruptcy based on mean of sum of various rating categories and individual rating categories. The mean of sum of rating categories has the highest predictive power one year before bankruptcy filings, while the mean of various individual categories has the highest predictive power two years before bankruptcy filings. We also compare these findings with tests with the Z-Score. Companies with low Z-Score are less likely to emerge from a bankruptcy based on the Z-Score one year before bankruptcy filings, while companies with high Z-Score are more likely to emerge from a bankruptcy based on the Z-Score two and three years before bankruptcy filings. Therefore, we can conclude that employee satisfaction is predictive of bankruptcy emergences.

V. Conclusion

We test four key bankruptcy models from the literature using a dataset from 2008 to 2020 to show that each one contains unique information regarding the probability of bankruptcy filings. We also build a new model to reflect employees' attitudes and emotions before bankruptcy filings and include key variables from each of the four already established bankruptcy models in the literature in our model. The model generated shows that employee satisfaction information shows as a predictor of bankruptcies before financial statement information. Right around bankruptcy filings, financial information overwhelms other information about companies including employee satisfaction information. We document that employee satisfaction reviews provide information relevant for bankruptcy predictions on top of information provided by employee satisfaction ratings. We also document that employee satisfaction predicts bankruptcy survival (emergence) using additional analyses to our main analyses. Our findings suggest that employees sense financial difficulties and problems in the companies they work for years before they show up on financial statements. The findings could be beneficial for policy makers on imposing more successful internal controls that allow for employees to have a more direct say in companies' disclosure of performance.

References

- [1]Altman, Edward. I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23, 589–609.
- [2]Beaver, William H., 1966. Financial ratios as predictors of failure. *Journal of Accounting Research*, 71–111.
- [3]Campbell, Dennis & Ruidi Shang, 2022. Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews. *Management Science*, 68(9): 7034-7053.
- [4]Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002. SMOTE: Synthetic minority over-sampling technique.
- [5]Chen, J., Guohao Tang, Jiaquan Yao, & Guofu Zhou, 2023. Employee sentiment and stock returns. *Journal of Economic Dynamics & Control*, 149.
- [6]Chudson, Walter A., 1945. A survey of corporate financial structure. Volume Title: The pattern of corporate financial structure: a cross-section view of manufacturing, mining, trade, and construction. *NBER*.
- [7]Green, T. Clifton, Ruoyan Huang, Quan Wen, & Dexin Zhou, 2019, Crowdsourced employer reviews and stock returns. *Journal of Financial Economics*, 134(1): 236-251.
- [8]Ji, Yuan, Oded Rozenbaum, & Kyle Welch, 2022. Corporate culture and financial reporting risk: Looking through the Glassdoor. Available at SSRN: [Corporate Culture and Financial Reporting Risk: Looking Through the Glassdoor by Yuan Ji, Oded Rozenbaum, Kyle Welch :: SSRN](#).
- [9]Liu, Tim, Christos Makridis, Paige Ouimet, & Elena Simintzi, 2023. The distribution of nonwage benefits: Maternity benefits and gender diversity. *The Review of Financial Studies*, 36(1): 194-234.
- [10]Marchetti, Arianna, 2019. Firms of a feather merge together: The coordination benefits of compatible cultures. Available at SSRN: [Firms of a Feather Merge Together: The Coordination Benefits of Compatible Cultures by Arianna Marchetti :: SSRN](#).
- [11]Mensah, Yaw M., 1983. The differential bankruptcy predictive ability of specific price level adjustments: some empirical evidence. *The Accounting Review*. Vol. LVIII, No. 2.
- [12]Merwin, 1942. Financing small corporations in five manufacturing industries, 1926-36. *NBER*.
- [13]Ohlson, James, 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18, 109–131.
- [14]Shumway, Tyler, 2001. Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business* 74, 101–124.
- [15]Zmijewski, Mark E., 1984. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research* 22 (Suppl.), 59–82.

[16]Welch, Kyle & Aaron Yoon, 2021. Do high-ability managers choose ESG projects that create shareholder value? Evidence from employee opinions. *Review of Accounting Studies*.

Table 1: Summary Statistics

Table 1 presents summary statistics for the full bankruptcy sample from 2008 through 2020. The final dataset contains 327 bankruptcies and 327 non-bankrupt firms. EBITTA = earnings before interest and taxes to total assets; Sales = sales to total assets; NITA = net income divided by total assets; CHIN = change in net income from year of to year before; WCTA = working capital to total assets; CLCA = current liabilities to current assets; FUTL = income from operations after depreciation divided by total liabilities; INTWO = 1 if net income was negative for the previous two years, 0 otherwise; RETA = retained earning to total assets; METL = market equity to total liabilities; TLTA = total liabilities to total assets; OENEG = 1 if total liabilities exceed total assets, 0 otherwise; TLMTA = total liabilities to market value of total assets; Ohlson's Size = $\log(\text{total assets}/\text{GNP price-level index})$, the index assumes a base value of 100 for 1968; Relative Size = $\log(\text{the number of outstanding shares multiplied by year-end share price then divided by total market value})$; Price = \log of closing price at end of previous fiscal year; SDReturn = standard deviation of excess return; LagExReturn = lagged excess return; FirmAge = $\log(\text{years for which firm has traded})$. Panel A presents the above-mentioned bankruptcy models' characteristics one, two, and three years before. Panel B presents mean and variability summary statistics for the 327 bankruptcies, while Panel C presents mean and variability of breakdown of various rating categories one, two, and three years before bankruptcy filings. Panel D presents correlations between Z-Score and overall ratings and rating categories.

Panel A. Bankruptcy Models' Characteristics One, Two, and Three Years Before

Variable	Observations	Mean	Standard Deviation
<i>Profitability Variables</i>			
EBITTA_1	654	-0.07530	0.325769
Sales_1	528	1936.804	8580.106
NITA_1	654	-0.14671	0.344535
CHIN_1	654	-0.25264	7.626269
<i>Liquidity Variables</i>			
WCTA_1	654	0.003879	0.398063
CLCA_1	654	1.129634	2.739787
FUTL_1	654	-0.10756	0.549372
INTWO_1	654	0.401109	0.490577
<i>Leverage Variables</i>			
RETA_1	654	-0.60907	1.894407
MVETL_1	654	0.791590	2.604564
TLTA_1	654	0.628457	0.554343
OENEG_1	654	0.164510	0.371081
<i>Firm-Size Variables</i>			
OhlsonSize_1	654	1.601003	1.857073
RelativeSize_1	654	-1.58703	1.561818

Variable	Observations	Mean	Standard Deviation
<i>Other Firm Characteristics</i>			
Price_1	654	0.736131	1.237024
ExcessReturn_1	654	-0.22312	0.644702
SDReturn_1	654	0.036120	0.035597
FirmAge_1	654	1.361375	1.276400
<i>Profitability Variables</i>			
EBITTA_2	654	-0.01884	0.155762
Sales_2	654	1398.292	8099.896
NITA_2	654	-0.06706	0.173948
CHIN_2	654	0.19339	3.418142
<i>Liquidity Variables</i>			
WCTA_2	654	0.110971	0.203358
CLCA_2	654	0.574281	0.778111
FUTL_2	654	-0.04879	0.886334
INTWO_2	654	0.402958	0.490946
<i>Leverage Variables</i>			
RETA_2	654	-0.44544	1.527194
MVETL_2	654	1.175064	3.958065
TLTA_2	654	0.631944	0.381070
OENEG_2	654	0.107209	0.309665
<i>Firm-Size Variables</i>			
OhlsonSize_2	654	1.921487	1.694350
RelativeSize_2	654	-1.365870	1.317351
<i>Other Firm Characteristics</i>			
Price_2	654	1.376031	1.150767
ExcessReturn_2	654	-0.13428	0.720694
SDReturn_2	654	0.032141	0.034412
FirmAge_2	654	1.282631	1.257304

Variable	Observations	Mean	Standard Deviation
<i>Profitability Variables</i>			
EBITTA_3	654	-0.01803	0.211587
Sales_3	654	774.2690	2155.300
NITA_3	654	-0.04780	0.254363
CHIN_3	654	-0.46575	11.12245
<i>Liquidity Variables</i>			
WCTA_3	654	0.096723	0.199287
CLCA_3	654	0.452632	0.731055
FUTL_3	654	-0.01880	1.135769
INTWO_3	654	0.266174	0.442365
<i>Leverage Variables</i>			
RETA_3	654	-0.32829	1.493576
MVETL_3	654	1.253587	4.159301
TLTA_3	654	0.462704	0.398290
OENEG_3	654	0.062847	0.242911
<i>Firm-Size Variables</i>			
OhlsonSize_3	654	1.468471	1.655274
RelativeSize_3	654	-0.83431	1.174613
<i>Other Firm Characteristics</i>			
Price_3	654	1.305623	1.275785
ExcessReturn_3	654	-0.00563	0.567273
SDReturn_3	654	0.025504	0.027173
FirmAge_3	654	1.200402	1.238483

*Panel**B. Mean and Variability Statistics*

Variable	Observations	Mean	Standard Deviation
Mean_1	327	7.825330	9.150692
SD_1	327	2.210938	3.388707
Mean_2	327	6.310011	8.681792
SD_2	327	1.974335	3.281781
Mean_3	327	4.521087	7.666320
SD_3	327	1.351704	2.782840

C. Rating Categories' Statistics

Variable	Observations	Mean	Standard Deviation
<i>One Year Before</i>			
careeropps_mean_1	327	2.698533	0.873781
careeropps_sd_1	327	0.806543	0.634286
compensation_mean_1	327	3.204054	1.019952
compensation_sd_1	327	0.693903	0.563472
culture_mean_1	327	2.355393	1.496977
culture_sd_1	327	0.742689	0.737253
seniorleadership_mean_1	327	2.615828	1.022545
seniorleadership_sd_1	327	0.896511	0.693012
worklife_mean_1	327	3.160145	1.022187
worklife_sd_1	327	0.855969	0.714880
overall_mean_1	327	3.025266	0.875284
overall_sd_1	327	0.824228	0.599971
<i>Two Years Before</i>			
careeropps_mean_2	327	2.783246	0.870941
careeropps_sd_2	327	0.893290	0.599946
compensation_mean_2	327	3.304863	0.746865
compensation_sd_2	327	0.821785	0.566678
culture_mean_2	327	2.291046	1.435886
culture_sd_2	327	0.882462	0.717721
seniorleadership_mean_2	327	2.609627	0.981455
seniorleadership_sd_2	327	0.962891	0.653779
worklife_mean_2	327	3.114457	0.864257
worklife_sd_2	327	0.889338	0.623642
overall_mean_2	327	3.091540	0.798091
overall_sd_2	327	0.930298	0.589306
<i>Three Years Before</i>			
careeropps_mean_3	327	2.590691	0.847830
careeropps_sd_3	327	0.827568	0.589574
compensation_mean_3	327	3.068088	0.978264
compensation_sd_3	327	0.714481	0.550266
culture_mean_3	327	2.186082	1.338209
culture_sd_3	327	0.722396	0.663047
seniorleadership_mean_3	327	2.441988	0.876578
seniorleadership_sd_3	327	0.862535	0.638214
worklife_mean_3	327	3.008461	0.866640
worklife_sd_3	327	0.844089	0.606252

overall_mean_3	327	2.950795	0.875191
overall_sd_3	327	0.886155	0.582013

*D. Correlations**D1. Correlations between Z-Score and Mean of Ratings One Year Before*

	ZScore_1	Mean_1	SD_1	Career_1	Compensation_1	Culture_1	Leadership_1	Worklife_1	Overall_1
ZScore_1	1								
Mean_1	-0.0405	1							
SD_1	-0.0286	0.6690	1						
Career_1	-0.0385	0.9730	0.6663	1					
Compensation_1	-0.0412	0.9612	0.6560	0.9350	1				
Culture_1	-0.0352	0.8849	0.6094	0.8113	0.8000	1			
Leadership_1	-0.0366	0.9707	0.6272	0.9482	0.9060	0.8301	1		
Worklife_1	-0.0401	0.9691	0.6369	0.9355	0.9394	0.8077	0.9359	1	
Overall_1	-0.0402	0.9746	0.6396	0.9518	0.9212	0.8333	0.9515	0.9310	1

D2. Correlations between Z-Score and Standard Deviation of Ratings One Year Before

	ZScore_1	Mean_1	SD_1	Career_1	Compensation_1	Culture_1	Leadership_1	Worklife_1	Overall_1
ZScore_1	1								
Mean_1	-0.0405	1							
SD_1	-0.0286	0.6690	1						
Career_1	-0.0260	0.6459	0.9515	1					
Compensation_1	-0.0289	0.6203	0.9365	0.8827	1				
Culture_1	-0.0283	0.5912	0.9194	0.8435	0.8355	1			
Leadership_1	-0.0238	0.6586	0.9625	0.9204	0.8662	0.8580	1		
Worklife_1	-0.0269	0.6121	0.9402	0.8405	0.8625	0.8380	0.8989	1	
Overall_1	-0.0286	0.6529	0.9432	0.8925	0.8680	0.8185	0.8866	0.8695	1

Figure 1: Number of Bankruptcies per Year

Figure 1 presents the number of bankruptcies per year from UCLA's LoPucki data over the paper sample from 2008 through 2020.

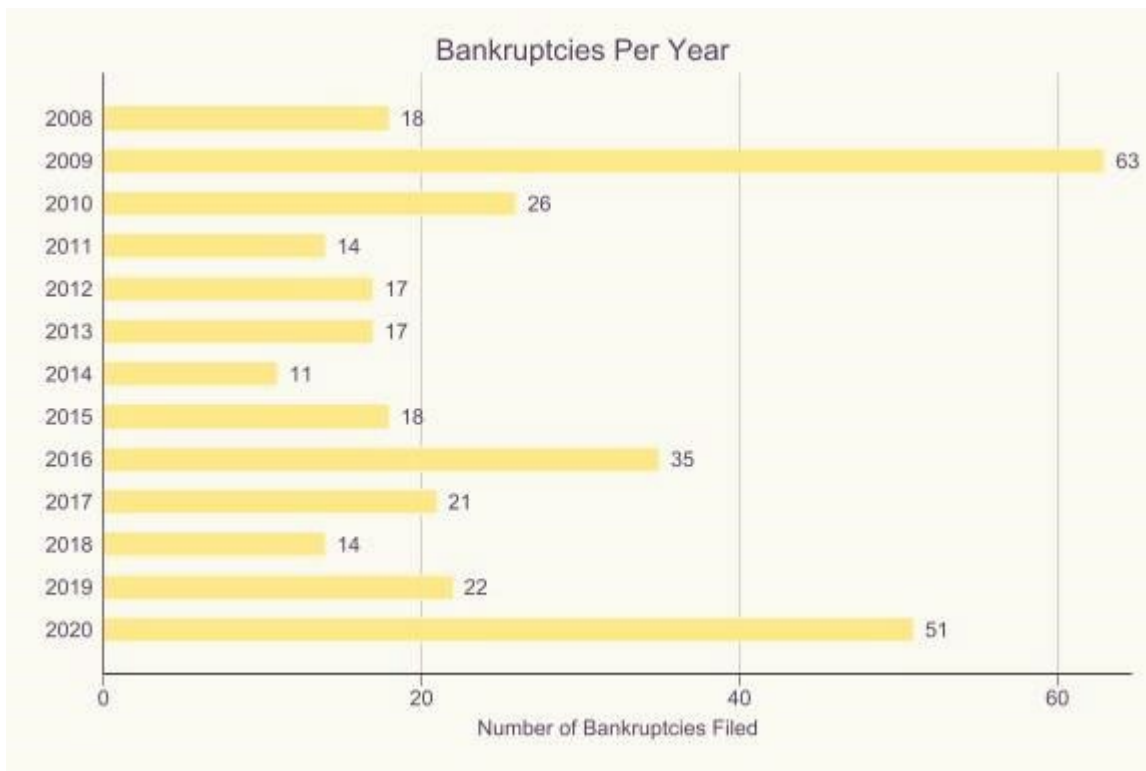
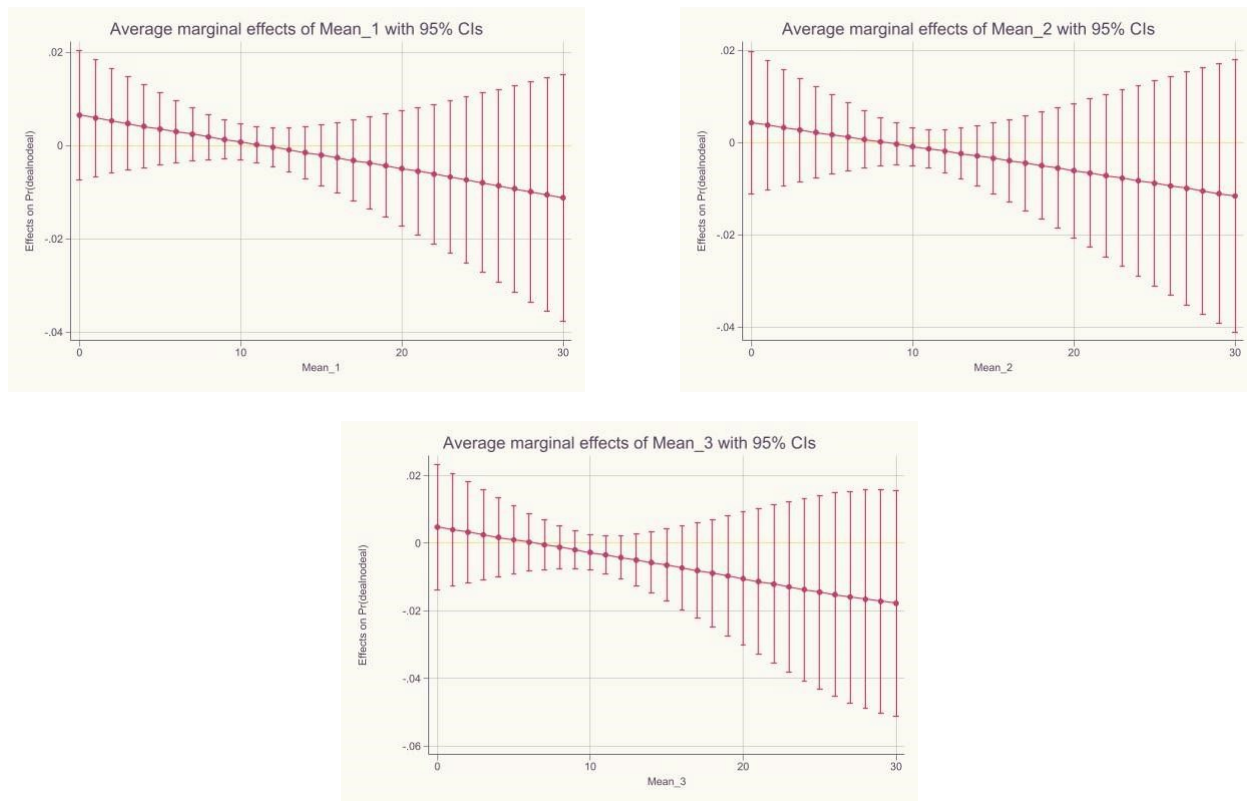


Figure 2: Marginal Effects and Predictive Margins of Mean Ratings

Figure 2 shows a plot of the average marginal effects and the predictive margins with 95% confidence interval of the mean of sum of rating categories one, two, and three years before bankruptcy filings. The figures plot the predicted outcome at various levels of the mean one, two, and three years before. **Panel A.** **Average Marginal Effects**



Panel B. Predictive Margins

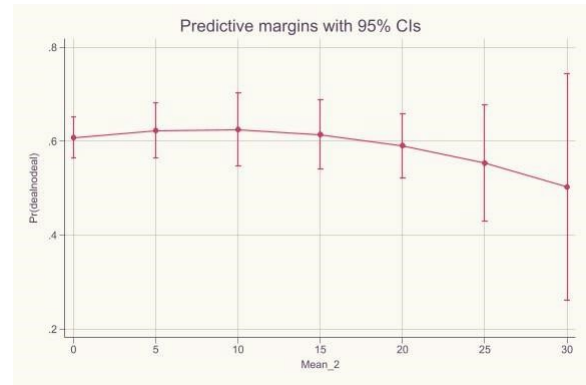


Table 2: Estimation Results for Bankruptcy Models

Table 2 presents the parameter estimates from various bankruptcy prediction models. The sample is from 2008 to 2020 and contains 327 bankruptcies and 327 non-bankrupt firms. Bold font signifies an estimate that is statistically significant. The adjusted Wald Chi-Square statistics from the logistic regression are presented adjacent to each parameter estimate. The Pseudo R-Squared for each model is reported. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<u>Ohlson</u>		<u>Altman</u>		<u>Zmijewski</u>	
	Estimate	Wald-Chi Square	Estimate	Wald-Chi Square	Estimate	Wald-Chi Square
Intercept	0.827***	58.46	1.309***	31.29	0.195	17.66
Mean_1						
EBITTA_1			-1.452 (1.804)	0.648		
Sales_1			0.0000299 (5.68e-05)	0.278		
NITA_1	-0.775 (1.215)	0.407			-0.173 (0.514)	0.113
CHIN_1	-0.00646 (0.00917)	0.496				
WCTA_1	-4.716*** (1.062)	19.712	-1.237* (0.669)	3.416		
CLCA_1	-0.0640 (0.0655)	0.954			0.370*** (0.0918)	16.251
FUTL_1	0.329 (0.238)	1.912				
INTWO_1	0.654* (0.343)	3.637				
RETA_1			0.0424 (0.0573)	0.548		
MVETL_1			-3.753*** (0.951)	15.568		
TLTA_1	-2.137** (0.907)	5.545			-0.138 (0.196)	0.496
OENEG_1	2.095*** (0.616)	11.558				
OhlsonSize_1	0.400*** (0.146)	7.527				

RelativeSize_1

Price_1

ExcessReturn_1

SDReturn_1

FirmAge_1

Log Pseudolikelihood

-275.1780

-174.7317

-347.1652

Pseudo R-Squared

0.2422

0.3914

0.0439

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

	<u>Ratings</u>		<u>Shumway</u>	
	Estimate	Wald-Chi Square	Estimate	Wald-Chi Square
Intercept	0.784	35.66	-0.215	66.82
Mean_1	-0.00630 (0.0391)	0.026		
EBITTA_1	0.848 (0.841)	1.016		
Sales_1				
NITA_1			-0.650 (0.634)	1.054
CHIN_1	-0.224*** (0.0798)	7.882		
WCTA_1	-5.332*** (1.476)	13.056		
CLCA_1				
FUTL_1				
INTWO_1				
RETA_1				
MVETL_1				
OENEG_1				
TLTA_1	0.578 (0.524)	1.217	-1.116 (0.800)	1.945

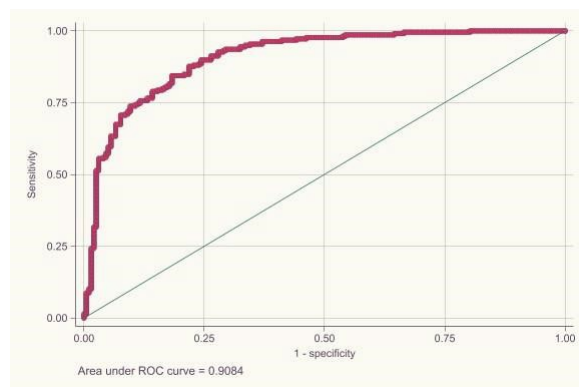
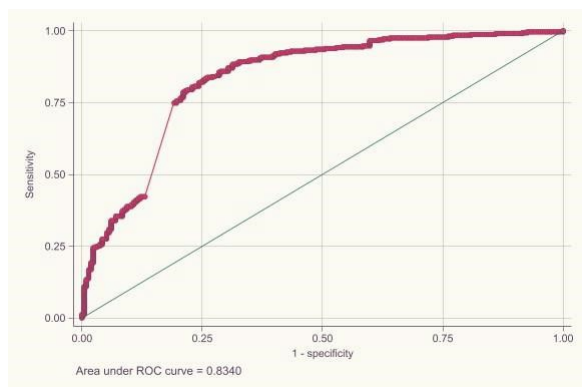
OhlsonSize_1				
RelativeSize_1			-0.669***	11.000
			(0.202)	
Price_1	-1.124***	28.552		
	(0.210)			
ExcessReturn_1	0.0625	0.040	-0.623	0.706
	(0.313)		(0.742)	
SDReturn_1	23.74***	9.814	14.08**	5.076
	(7.577)		(6.249)	
FirmAge_1			-0.237***	9.596
			(0.0765)	
Log Pseudolikelihood	-124.9652		-301.7125	
Pseudo R-Squared	0.2178		0.1691	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Figure 3: ROC Curves for Bankruptcy Models (Ohlson, Altman, Zmijewski, Shumway, and Ratings, respectively from left to right)

Figure 3 plots the Receiver Operating Characteristic (ROC) generated from the various models: Ohlson’s Altman’s, Zmijewski’s, Shumway’s, and Ratings’ models. The ROC for Ohlson’s model is 0.8340, the ROC for Altman’s model is 0.9084, the ROC for Zmijewski’s model is 0.5391, the ROC for Shumway’s model is 0.7920, and the ROC for Ratings’ model is 0.8867.



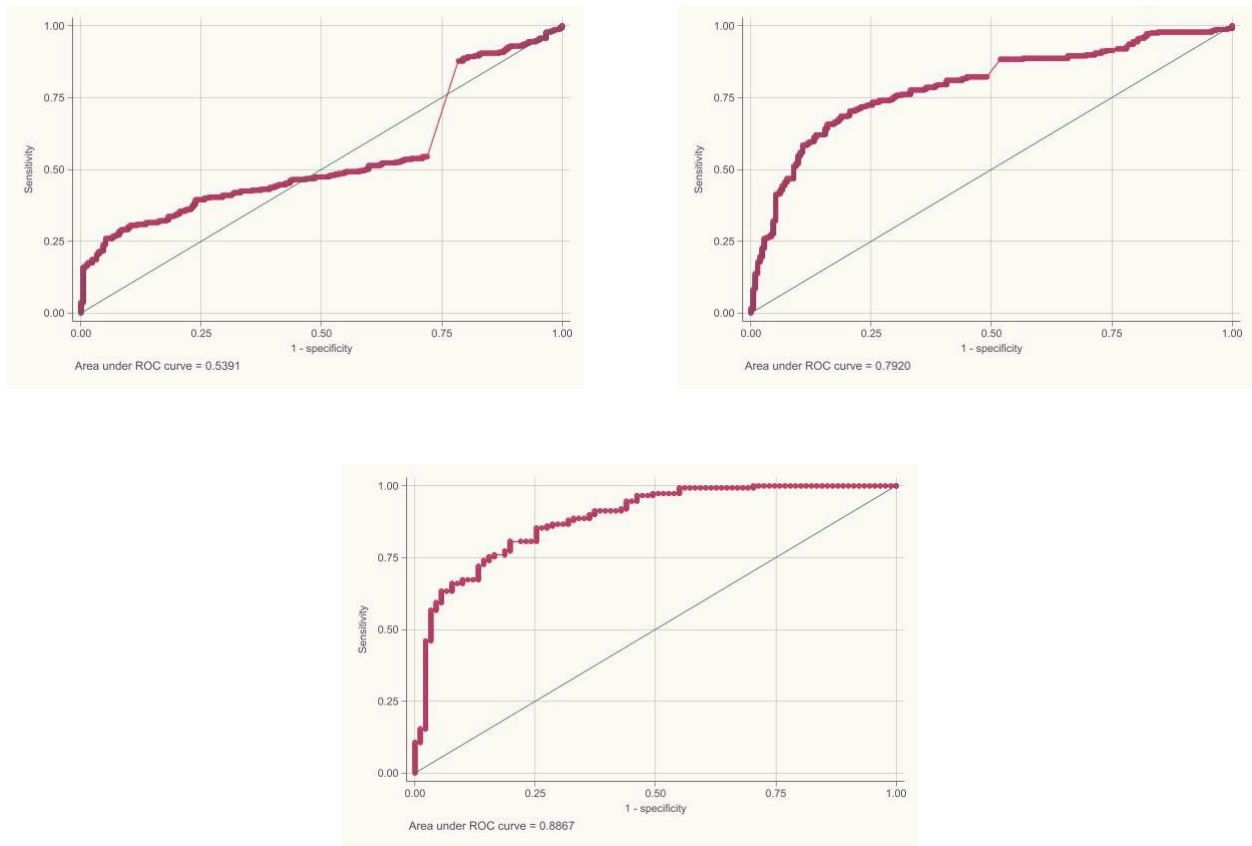
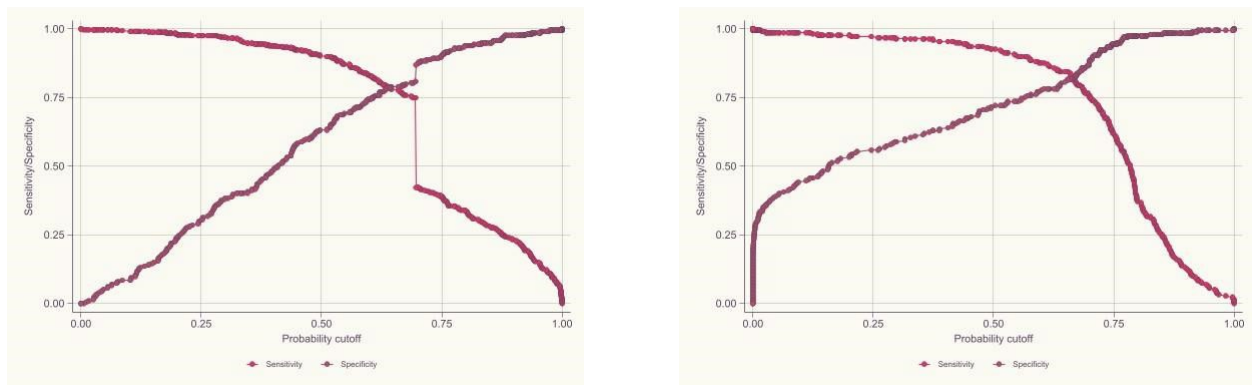


Figure 4: Optimal Cut-Off Points for Bankruptcy Models (Ohlson, Altman, Zmijewski, Shumway, and Ratings, respectively from left to right)

Figure 4 plots the optimal cut-off points generated from the various models: Ohlson's, Altman's, Zmijewski's, Shumway's, and Ratings' models. The horizontal line shows the optimal probability cutoff, while the vertical line shows the sensitivity/specificity of the models.



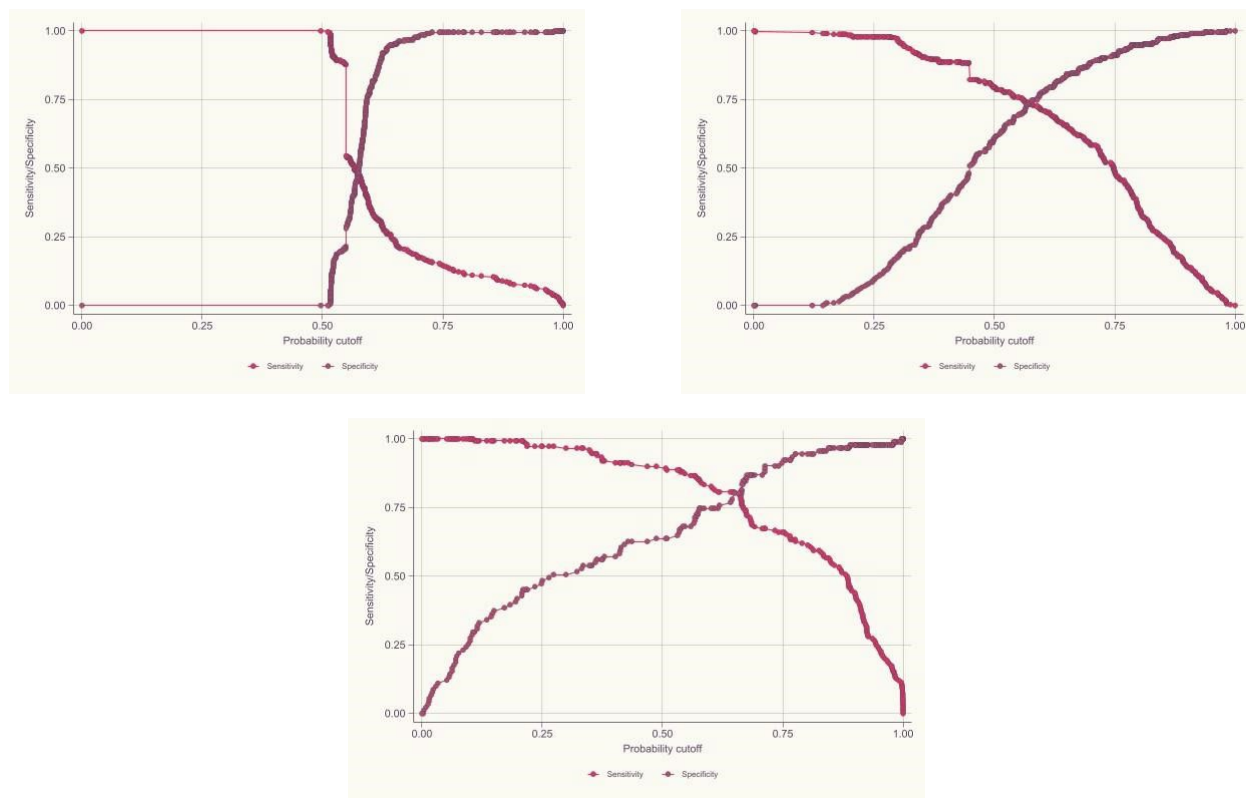


Table 3: Estimation Results for Bankruptcy Models with Ratings

Table 3 presents the parameter estimates from various bankruptcy prediction models with inclusion of mean employee satisfaction. The sample is from 2008 to 2020 and contains 327 bankruptcies and 327 non-bankrupt firms. Bold font signifies an estimate that is statistically significant. The adjusted Wald Chi-Square statistics from the logistic regression are presented adjacent to each parameter estimate. The Pseudo R-Squared for each model is reported. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<u>Ohlson</u>		<u>Altman</u>	
	Estimate	Wald-Chi Square	Estimate	Wald-Chi Square
Intercept	1.330	39.74	1.420	18.20
Mean_1	-0.203 (0.327)	0.385	-0.281 (0.487)	0.333
EBITTA_1			-4.241** (1.903)	4.965
Sales_1			0.000606** (0.000271)	4.989
NITA_1	-2.642** (1.202)	4.828		
CHIN_1	-0.0365 (0.0604)	0.366		

WCTA_1		12.363	-1.765* (0.928)	3.619
	-5.757*** (1.637)			
CLCA_1		0.005		
	0.0125 (0.173)			
FUTL_1	0.405 (0.258)	2.476		
INTWO_1	1.045** (0.506)	4.275		
RETA_1			0.0483 (0.0877)	0.304
MVETL_1			-4.028*** (1.427)	7.968
OENEG_1	1.829** (0.859)	4.530		
TLTA_1	-3.265*** (0.849)	14.777		
OhlsonSize_1	0.639*** (0.166)	14.761		
RelativeSize_1				
Price_1				
ExcessReturn_1				
SDReturn_1				
FirmAge_1				
Log Pseudolikelihood				
	-110.6051		-69.172861	
Pseudo R-Squared	0.3076		0.4696	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

	<u>Zmijewski</u>		<u>Shumway</u>	
	Estimate	Wald-Chi Square	Estimate	Wald-Chi Square
Intercept	0.536	19.24	0.340	33.15
Mean_1	-0.121 (0.306)	0.157	-0.335 (0.385)	0.758

EBITTA_1				
Sales_1				
NITA_1	-0.691 (0.751)	0.847	-1.010 (0.637)	2.516
CHIN_1				
WCTA_1				
CLCA_1	0.644*** (0.162)	15.849		
FUTL_1				
INTWO_1				
RETA_1				
MVETL_1				
TLTA_1	-0.423 (0.354)	1.430	-1.732** (0.688)	6.339
OENEG_1				
OhlsonSize_1				
RelativeSize_1			-0.852*** (0.234)	13.302
Price_1				
ExcessReturn_1			-0.0352 (0.151)	0.054
SDReturn_1			25.09*** (5.848)	18.409
FirmAge_1			-0.101 (0.114)	0.785
Log Pseudolikelihood	-148.2620		-129.5948	
Pseudo R-Squared	0.0719		0.1888	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Figure 5: ROC Curves for Bankruptcy Models with Ratings (Ohlson, Altman, Zmijewski, and Shumway, respectively from left to right)

Figure 5 plots the Receiver Operating Characteristic (ROC) generated from the various models: Ohlson's

Altman's, Zmijewski's, and Shumway's, but including the mean employee satisfaction in those models. The ROC for Ohlson's model is 0.8499, the ROC for Altman's model is 0.9182, the ROC for Zmijewski's model is 0.5987, and the ROC for Shumway's model is 0.7917.

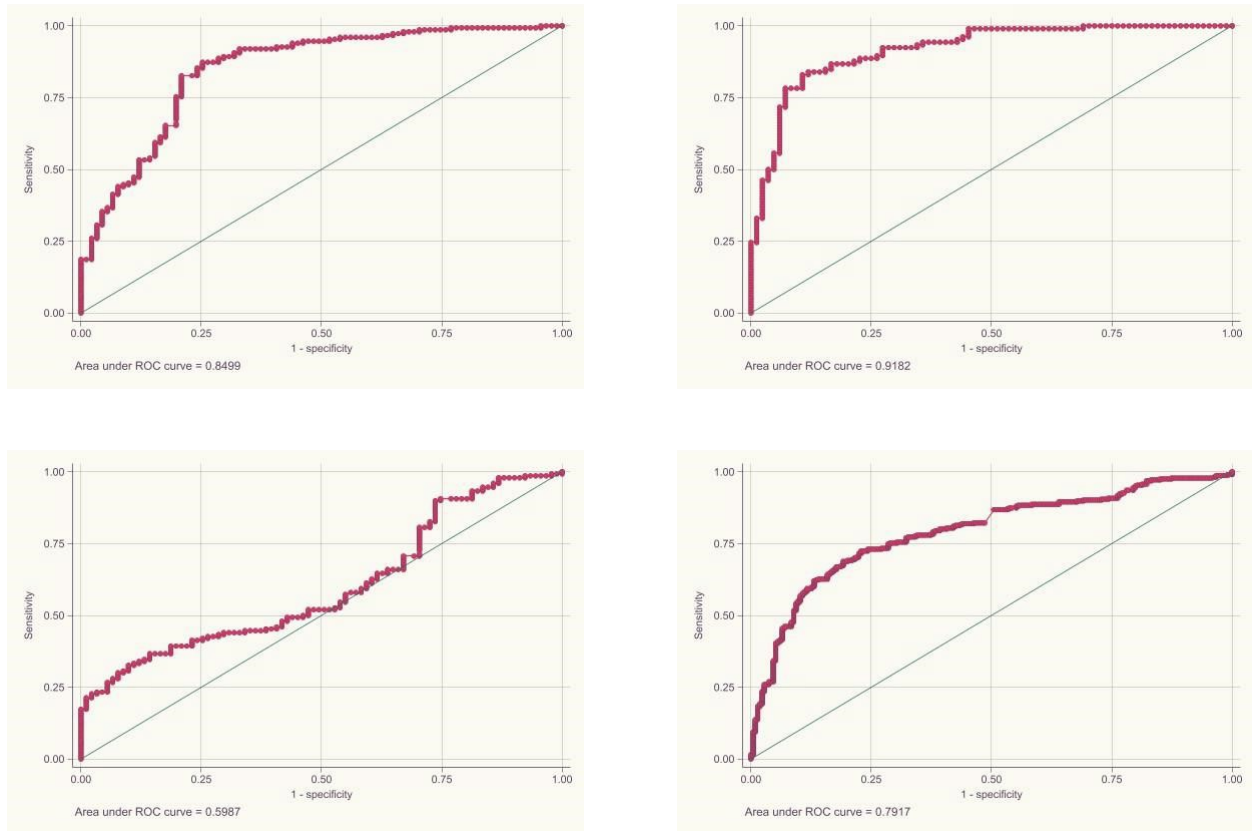


Figure 6: Optimal Cut-Off Points for Bankruptcy Models with Ratings (Ohlson, Altman, Zmijewski, and Shumway, respectively from left to right)

Figure 6 plots the optimal cut-off points generated from the various models: Ohlson's, Altman's, Zmijewski's, and Shumway's, but including the mean employee satisfaction in each model. The horizontal

line shows the optimal probability cutoff, while the vertical line shows the sensitivity/specificity of the models.

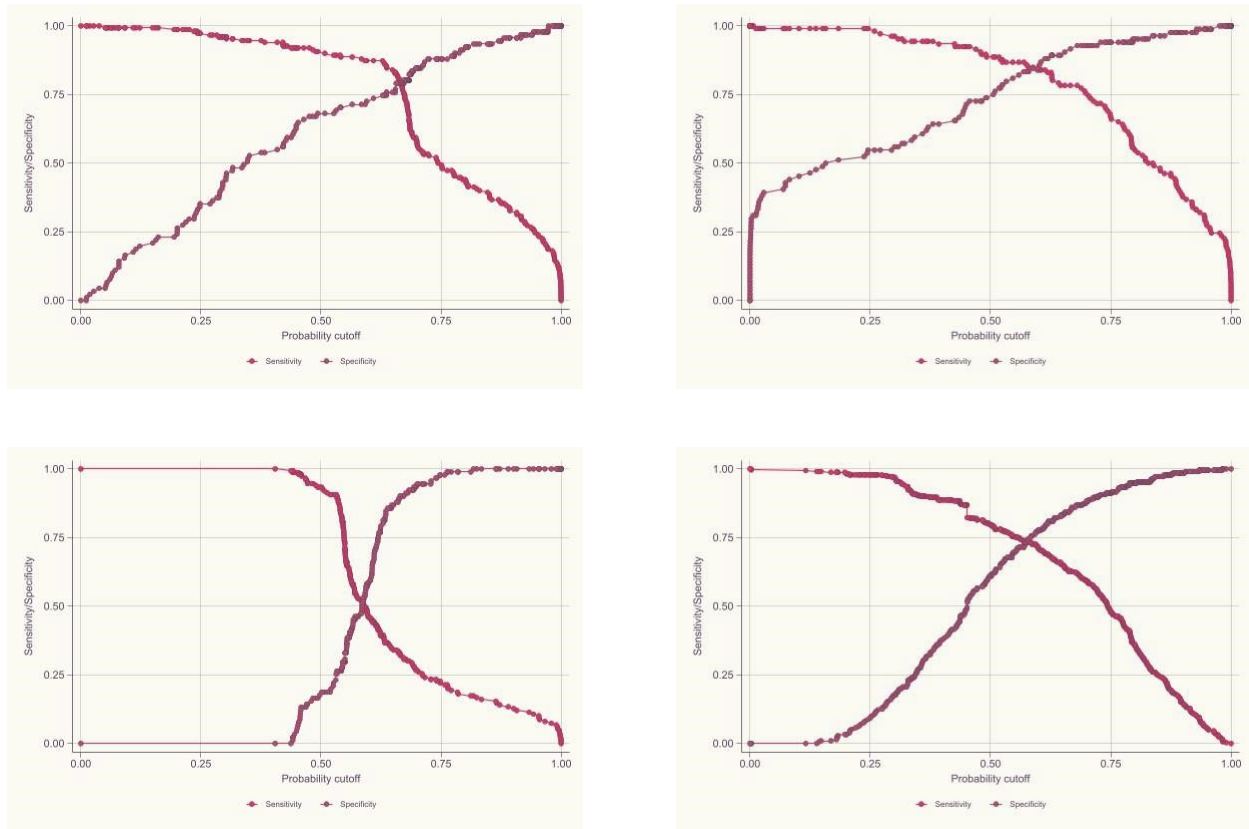


Table 4: Classification Rates for Bankruptcy Models

Table 4 presents the incidence of Type I errors (classifying a bankrupt firm as healthy) and Type II errors (classifying a healthy firm as bankrupt) and Total error (both Type I and Type II errors), according to model scores for Ohlson's, Altman's, Zmijewski's, Ratings', and Shumway's models.

Model Score (Percentile)	<u>Ohlson</u>			<u>Altman</u>			<u>Zmijewski</u>		
	Type I	Type II	Total	Type I	Type II	Total	Type I	Type II	Total
50	0.577982	0.130841	0.708823	0.201835	0.167513	0.369348	0.519878	0.528037	1.047915
70	0.584098	0.121495	0.705593	0.458716	0.030457	0.489172	0.645260	0.214953	0.860213
71	0.596330	0.112150	0.708480	0.477064	0.030457	0.507521	0.654434	0.200935	0.855369
72	0.605505	0.102804	0.708308	0.490826	0.025381	0.516206	0.663609	0.191589	0.855197
73	0.614679	0.093458	0.708137	0.509174	0.025381	0.534555	0.672783	0.182243	0.855026
74	0.626911	0.084112	0.711024	0.532110	0.025381	0.557491	0.678899	0.163551	0.842451
75	0.642202	0.084112	0.726314	0.550459	0.025381	0.575840	0.685015	0.149533	0.834548
76	0.651376	0.070094	0.721470	0.568807	0.025381	0.594188	0.691132	0.130841	0.821973
77	0.660551	0.060748	0.721298	0.587156	0.025381	0.612537	0.694190	0.112150	0.806339
78	0.675841	0.060748	0.736589	0.605505	0.025381	0.630885	0.700306	0.098131	0.798437
79	0.691132	0.056075	0.747206	0.623853	0.025381	0.649234	0.709480	0.084112	0.793592
80	0.703364	0.051402	0.754766	0.642202	0.025381	0.667583	0.721713	0.079439	0.801152
81	0.721713	0.051402	0.773114	0.665138	0.025381	0.690518	0.733945	0.070094	0.804038
82	0.730887	0.042056	0.772943	0.683486	0.025381	0.708867	0.740061	0.056075	0.796136
83	0.746177	0.037383	0.783561	0.697248	0.020305	0.717552	0.755352	0.051402	0.806754
84	0.755352	0.028037	0.783389	0.715596	0.020305	0.735901	0.767584	0.046729	0.814313
85	0.767584	0.023365	0.790949	0.733945	0.020305	0.754250	0.782875	0.046729	0.829604
86	0.785933	0.023365	0.809297	0.752294	0.020305	0.772598	0.795107	0.037383	0.832490
87	0.801223	0.023365	0.824588	0.770642	0.015228	0.785871	0.807339	0.032710	0.840050
88	0.816514	0.018692	0.835205	0.788991	0.015228	0.804219	0.819572	0.023365	0.842936
89	0.831804	0.018692	0.850496	0.807339	0.015228	0.822568	0.828746	0.014019	0.842765
90	0.844037	0.014019	0.858055	0.825688	0.015228	0.840917	0.840979	0.009346	0.850324
91	0.862385	0.014019	0.876404	0.844037	0.015228	0.859265	0.856269	0.004673	0.860942
92	0.874618	0.009346	0.883964	0.862385	0.015228	0.877614	0.871560	0.004673	0.876233
93	0.889908	0.004673	0.894581	0.880734	0.015228	0.895962	0.889908	0.004673	0.894581
94	0.905199	0.004673	0.909872	0.899083	0.010152	0.909235	0.905199	0.004673	0.909872
95	0.920489	0.004673	0.925162	0.912844	0.005076	0.917920	0.920489	0.004673	0.925162
96	0.938838	0.004673	0.943511	0.931193	0.005076	0.936269	0.938838	0.004673	0.943511
97	0.954128	0.004673	0.958801	0.949541	0.005076	0.954617	0.954128	0.004673	0.958801
98	0.972477	0.004673	0.977150	0.967890	0.005076	0.972966	0.969419	0	0.969419
99	0.987768	0.004673	0.992441	0.986239	0.005076	0.991315	0.984710	0	0.984710

Ratings

Shumway

Model Score (Percentile)	Type I	Type II	Total	Type I	Type II	Total
50	0.3730887	0.1448598	0.5179485	0.308869	0.205608	0.514476
70	0.5474006	0.0654206	0.6128212	0.550459	0.070094	0.620552
71	0.5626912	0.0607477	0.6234388	0.565749	0.065421	0.631170
72	0.5749236	0.0560748	0.6309983	0.577982	0.060748	0.638729
73	0.5871559	0.0514019	0.6385578	0.587156	0.051402	0.638558
74	0.6024465	0.0467290	0.6491755	0.605505	0.051402	0.656906
75	0.6177370	0.0467290	0.6644660	0.620795	0.051402	0.672197
76	0.6299694	0.0373832	0.6673526	0.639144	0.051402	0.690546
77	0.6452599	0.0373832	0.6826431	0.654434	0.051402	0.705836
78	0.6605505	0.0373832	0.6979337	0.669725	0.051402	0.721127
79	0.6788991	0.0373832	0.7162823	0.685015	0.046729	0.731744
80	0.6941896	0.0373832	0.7315728	0.700306	0.046729	0.747035
81	0.7125382	0.0373832	0.7499214	0.718655	0.046729	0.765383
82	0.7278287	0.0373832	0.7652119	0.730887	0.042056	0.772943
83	0.7461774	0.0373832	0.7835606	0.740061	0.028037	0.768099
84	0.7614679	0.0373832	0.7988511	0.755352	0.028037	0.783389
85	0.7767584	0.0373832	0.8141416	0.770642	0.028037	0.798680
86	0.7951070	0.0373832	0.8324902	0.785933	0.023365	0.809297
87	0.8073394	0.0327103	0.8400497	0.801223	0.023365	0.824588
88	0.8226300	0.0280374	0.8506674	0.816514	0.018692	0.835205
89	0.8379205	0.0280374	0.8659579	0.828746	0.014019	0.842765
90	0.8501529	0.0233645	0.8735174	0.844037	0.014019	0.858055
91	0.8685015	0.0233645	0.8918660	0.862385	0.014019	0.876404
92	0.8807340	0.0186916	0.8994256	0.874618	0.009346	0.883964
93	0.8960245	0.0140187	0.9100432	0.892966	0.009346	0.902312
94	0.9082569	0.0093458	0.9176027	0.908257	0.009346	0.917603
95	0.9204893	0.0046729	0.9251622	0.920489	0.004673	0.925162
96	0.9388379	0.0046729	0.9435108	0.938838	0.004673	0.943511
97	0.9541284	0.0046729	0.9588013	0.954128	0.004673	0.958801
98	0.9724771	0.0046729	0.9771500	0.972477	0.004673	0.977150
99	0.9847095	0	0.9847095	0.984710	0	0.984710

Figure 7: Classification Rates for Bankruptcy Models (Ohlson, Altman, Zmijewski, Shumway, and Ratings, respectively from left to right)

Figure 7 presents classification rates for the various bankruptcy models – Ohlson’s, Altman’s, Zmijewski’s, Shumway’s, and Ratings’ models – from the tests in Table 4. The horizontal axis plots the percentile model score, while the vertical axis plots the Total error (Type I and Type II errors) for each respective model.

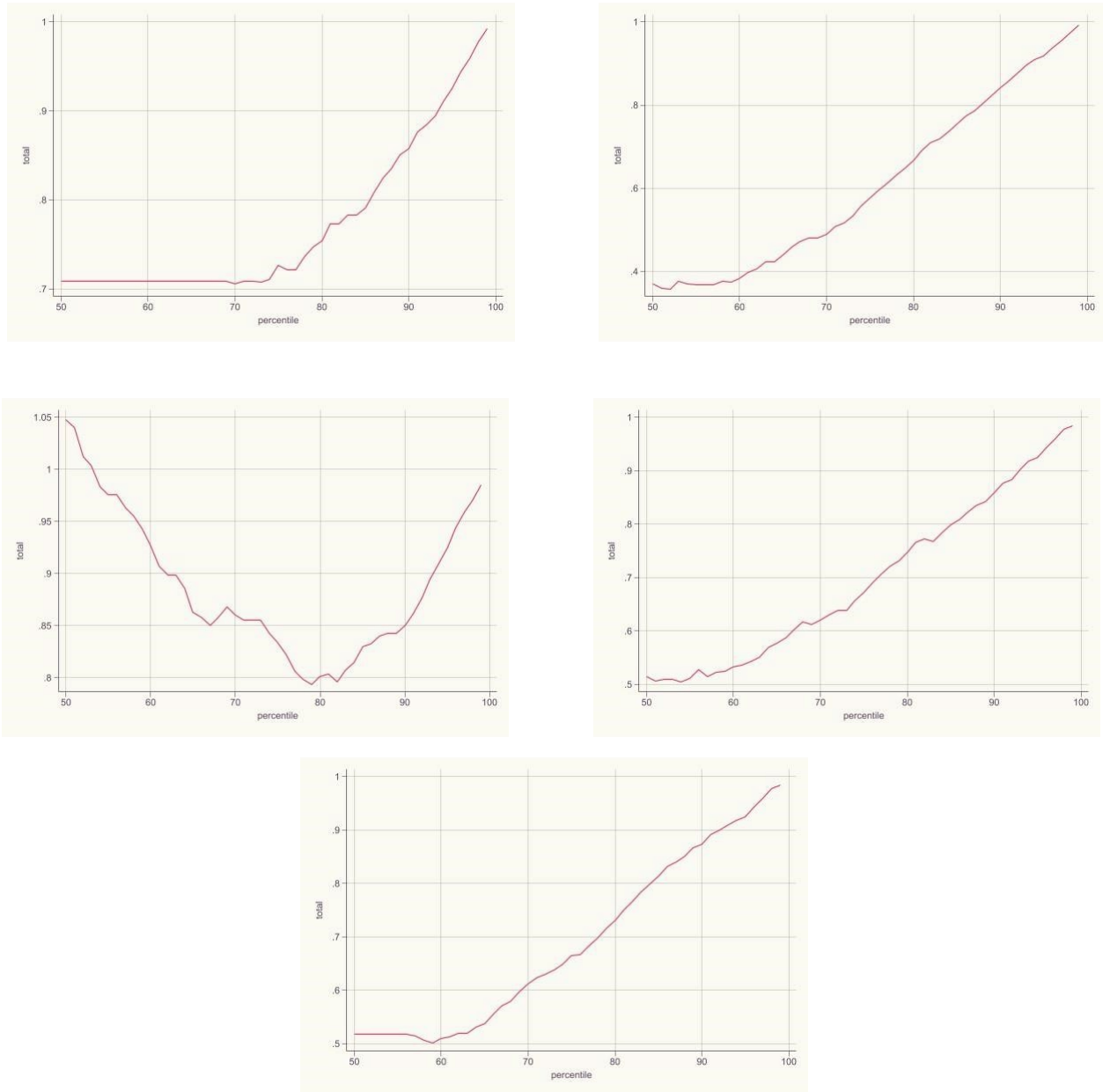


Table 5: Classification Rates for Bankruptcy Models with Ratings

Table 5 presents the incidence of Type I errors (classifying a bankrupt firm as healthy) and Type II errors (classifying a healthy firm as bankrupt) and Total error (both Type I and Type II errors), according to model scores for Ohlson's, Altman's, Zmijewski's, and Shumway's models with inclusion of mean of employee satisfaction.

Model Score (Percentile)	<u>Ohlson</u>			<u>Altman</u>		
	Type I	Type II	Total	Type I	Type II	Total
50	0.299694	0.191589	0.491283	0.174312	0.137056	0.311368
70	0.584098	0.121495	0.705593	0.477064	0.050761	0.527826
71	0.596330	0.112150	0.708480	0.495413	0.050761	0.546174
72	0.605505	0.102804	0.708308	0.509174	0.045685	0.554860
73	0.611621	0.088785	0.700406	0.527523	0.045685	0.573208
74	0.623853	0.079439	0.703293	0.550459	0.045685	0.596144
75	0.636086	0.074766	0.710852	0.568807	0.045685	0.614493
76	0.654434	0.074766	0.729201	0.587156	0.045685	0.632841
77	0.663609	0.065421	0.729029	0.605505	0.045685	0.651190
78	0.675841	0.060748	0.736589	0.623853	0.045685	0.669539
79	0.694190	0.060748	0.754937	0.637615	0.040609	0.678224
80	0.703364	0.051402	0.754766	0.655963	0.040609	0.696572
81	0.721713	0.051402	0.773114	0.674312	0.035533	0.709845
82	0.730887	0.042056	0.772943	0.692661	0.035533	0.728194
83	0.746177	0.037383	0.783561	0.706422	0.030457	0.736879
84	0.755352	0.028037	0.783389	0.724771	0.030457	0.755228
85	0.767584	0.023365	0.790949	0.743119	0.030457	0.773576
86	0.785933	0.023365	0.809297	0.756881	0.025381	0.782262
87	0.801223	0.023365	0.824588	0.779817	0.025381	0.805197
88	0.816514	0.018692	0.835205	0.798165	0.025381	0.823546
89	0.831804	0.018692	0.850496	0.816514	0.025381	0.841895
90	0.844037	0.014019	0.858055	0.834862	0.025381	0.860243
91	0.859327	0.009346	0.868673	0.848624	0.020305	0.868928
92	0.874618	0.009346	0.883964	0.866973	0.020305	0.887277
93	0.889908	0.004673	0.894581	0.885321	0.020305	0.905626
94	0.905199	0.004673	0.909872	0.908257	0.020305	0.928561
95	0.920489	0.004673	0.925162	0.926606	0.020305	0.946910
96	0.938838	0.004673	0.943511	0.944954	0.020305	0.965259
97	0.954128	0.004673	0.958801	0.958716	0.015228	0.973944

98	0.972477	0.004673	0.977150	0.972477	0.010152	0.982629
99	0.987768	0.004673	0.992441	1	0	1

Model Score (Percentile)	<u>Zmijewski</u>			<u>Shumway</u>		
	Type I	Type II	Total	Type I	Type II	Total
50	0.584098	0.467290	1.051388	0.333333	0.242991	0.576324
70	0.663609	0.242991	0.906599	0.571865	0.102804	0.674669
71	0.663609	0.214953	0.878562	0.581040	0.088785	0.669825
72	0.672783	0.205608	0.878390	0.593272	0.084112	0.677384
73	0.672783	0.182243	0.855026	0.602447	0.074766	0.677213
74	0.688073	0.177570	0.865644	0.620795	0.074766	0.695562
75	0.697248	0.168224	0.865472	0.633028	0.070094	0.703121
76	0.709480	0.158879	0.868359	0.645260	0.060748	0.706008
77	0.715596	0.144860	0.860456	0.660551	0.060748	0.721298
78	0.727829	0.140187	0.868016	0.672783	0.056075	0.728858
79	0.737003	0.126168	0.863171	0.685015	0.046729	0.731744
80	0.743119	0.112150	0.855269	0.697248	0.042056	0.739304
81	0.755352	0.102804	0.858155	0.715596	0.042056	0.757652
82	0.758410	0.084112	0.842522	0.727829	0.037383	0.765212
83	0.764526	0.065421	0.829947	0.743119	0.032710	0.775830
84	0.776758	0.060748	0.837506	0.755352	0.028037	0.783389
85	0.785933	0.051402	0.837335	0.770642	0.028037	0.798680
86	0.801223	0.046729	0.847952	0.788991	0.028037	0.817028
87	0.810398	0.037383	0.847781	0.804281	0.028037	0.832319
88	0.822630	0.028037	0.850667	0.819572	0.023365	0.842936
89	0.831804	0.018692	0.850496	0.828746	0.014019	0.842765
90	0.844037	0.014019	0.858055	0.840979	0.009346	0.850324
91	0.859327	0.009346	0.868673	0.856269	0.004673	0.860942
92	0.871560	0.004673	0.876233	0.871560	0.004673	0.876233
93	0.889908	0.004673	0.894581	0.889908	0.004673	0.894581
94	0.905199	0.004673	0.909872	0.905199	0.004673	0.909872
95	0.920489	0.004673	0.925162	0.920489	0.004673	0.925162
96	0.938838	0.004673	0.943511	0.938838	0.004673	0.943511
97	0.954128	0.004673	0.958801	0.954128	0.004673	0.958801
98	0.969419	0	0.969419	0.972477	0.004673	0.977150
99	0.984710	0	0.98471	0.987768	0.004673	0.992441

Figure 8: Classification Rates for Bankruptcy Models with Ratings (Ohlson, Altman, Zmijewski, and Shumway, respectively from left to right)

Figure 8 presents classification rates for the various bankruptcy models – Ohlson’s, Altman’s, Zmijewski’s, and Shumway’s models – from the tests in Table 5. The horizontal axis plots the percentile model score, while the vertical axis plots the Total error (Type I and Type II errors) for each respective model.

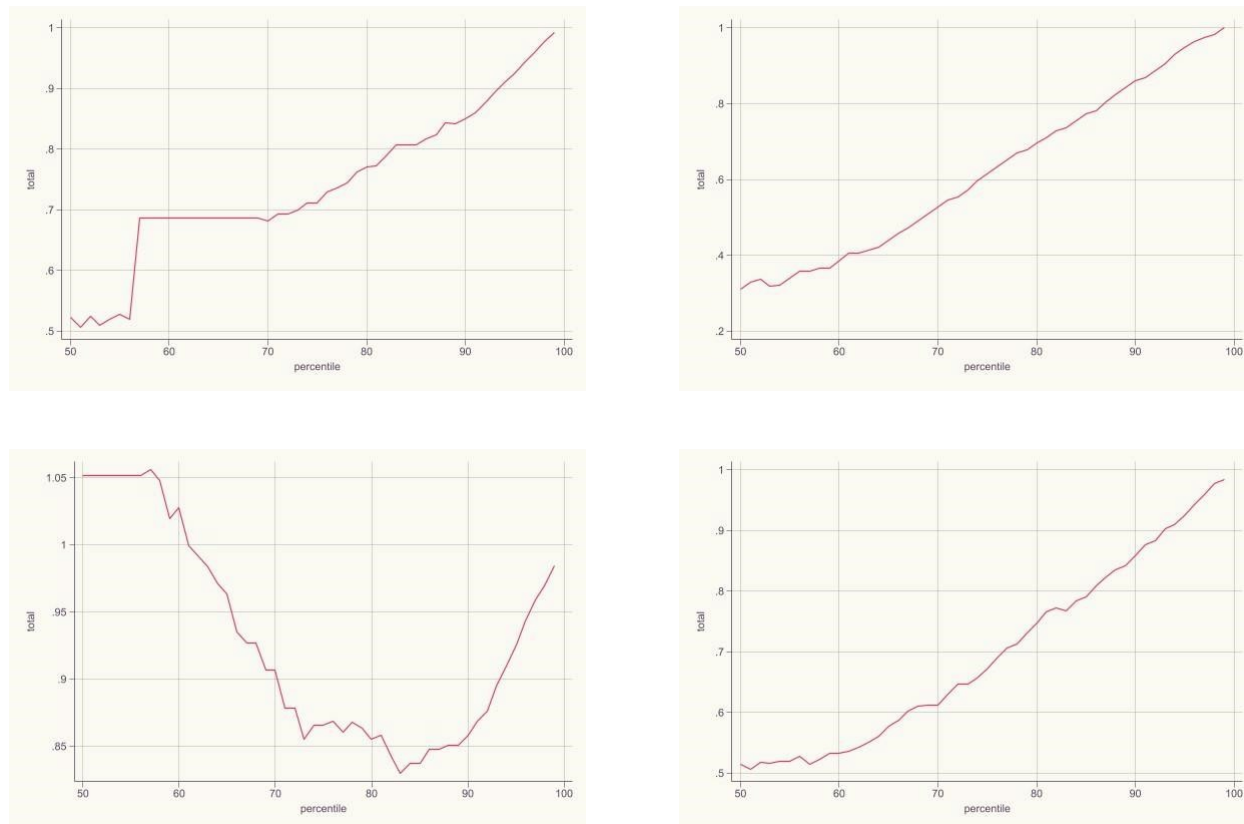


Table 6: Out-of-Sample Forecast Accuracy for Bankruptcy Models

Table 6 reports statistics relating to out-of-sample performance based on the rolling windows method. Observations are classified into deciles based on these bankruptcy probabilities (Decile 1 being the lowest). Actual and predicted probabilities are presented together with Wald-Chi Square statistic. The results for the following models are presented – Ohlson’s, Altman’s, Zmijewski’s, Ratings’, and Shumway’s models.

Ohlson

Altman

Decile	Chi-Square			Chi-Square		
	Actual	Predicted	Stats	Actual	Predicted	Stats
1	5	2.939442	1.444458	1	0.00227	438.5925
2	13	11.30596	0.253828	2	0.646224	2.836032
3	15	18.87824	0.796724	5	7.177679	0.660699
4	22	28.22304	1.372149	13	17.99771	1.387795
5	119	95.85053	5.590977	23	26.08369	0.364563
6	-	-	-	23	28.72941	1.142599
7	15	15.09268	0.000569	33	31.52358	0.069149
8	42	44.13479	0.103259	40	32.35796	1.804834
9	44	50.51492	0.84023	40	35.31721	0.620902
<u>10</u>	<u>52</u>	<u>53.65069</u>	<u>0.050787</u>	<u>38</u>	<u>38.16425</u>	<u>0.000707</u>

Zmijewski

Ratings

Decile	Chi-Square			Chi-Square		
	Actual	Predicted	Stats	Actual	Predicted	Stats
1	29	28.43557	0.011204	5	2.124327	3.892759
2	120	83.91169	15.52068	7	7.020329	0.0000589
3	-	-	-	23	16.44145	2.616228
4	2	4.407028	1.314669	22	27.70281	1.173961
5	19	30.41686	4.285279	65	58.7226	0.6710498
6	17	31.31273	6.542204	11	14.56417	0.8722303
7	24	31.94927	1.977851	46	45.55432	0.0043603
8	25	33.09774	1.981206	48	49.36272	0.0376199
9	39	35.80448	0.285198	51	52.12202	0.0241535
10	52	47.66462	0.394329	49	53.81192	0.4302869

Shumway

Decile	Actual	Predicted	Chi-Square Stats
1	14	12.85991	0.101074

2	23	18.86195	0.90783
3	21	24.50963	0.502558
4	16	24.57253	2.990669
5	27	30.98908	0.513496
6	35	35.56945	0.009117
7	44	39.78366	0.446856
8	49	42.8446	0.884336
9	47	46.26814	0.011577
10	<u>51</u>	<u>50.74106</u>	<u>0.001322</u>

Table 7: Out-of-Sample Forecast Accuracy for Bankruptcy Models with Ratings

Table 7 reports statistics relating to out-of-sample performance based on the rolling windows method. Observations are classified into deciles based on these bankruptcy probabilities (Decile 1 being the lowest). Actual and predicted probabilities are presented together with Wald-Chi Square statistic. The results for the following models are presented – Ohlson’s, Altman’s, Zmijewski’s, and Shumway’s models with the inclusion of mean of employee satisfaction.

Ohlson

Altman

Decile	Actual	Predicted	Chi-Square Stats	Actual	Predicted	Chi-Square Stats
1	5	6.060679	0.185629	1	0.000797	1252.523
2	13	14.78499	0.215503	2	0.496633	4.550865
3	13	22.86834	4.258467	6	7.084383	0.165983
4	26	29.6532	0.450065	9	17.34269	4.013246
5	41	35.35978	0.899668	20	25.60863	1.228366
6	92	73.92711	4.418264	29	30.59115	0.082761
7	1	1.406342	0.117407	37	34.33963	0.206104
8	39	41.60372	0.162951	39	36.56068	0.162751
9	46	48.33648	0.112941	39	40.21312	0.036597
10	51	52.99935	0.075424	36	40.90007	0.587057

Decile	<u>Zmijewski</u>			<u>Shumway</u>		
	Actual	Predicted	Chi-Square Stats	Actual	Predicted	Chi-Square Stats
1	24	27.1957	0.37552	20	15.20559	1.511703
2	39	29.32692	3.190536	10	24.02962	8.191154
3	33	30.27457	0.245353	28	29.15312	0.045611
4	24	32.43731	2.19464	29	32.06665	0.293276
5	71	55.41735	4.381645	22	36.14635	5.536362
6	9	12.70476	1.080324	37	40.35836	0.279462
7	17	35.53437	9.667337	41	44.36263	0.254883
8	26	37.63573	3.597387	41	48.00068	1.021017
9	33	41.4557	1.724704	47	50.19291	0.20311
10	51	51.34185	0.002276	52	52.63543	0.007671

Table 8: Classification Rates Using Altman’s Z-Score and Ratings’ Model One, Two, and Three Years Before

Table 8 reports classification rates for Ratings’ and Altman’s Z-Score model which shows sensitivity, specificity, positive predictive value, negative predictive value, false positives, and false negatives. The results are reported for each model one, two, and three years before merger announcement for Ratings’ and Altman’s Z-Score models.

Panel A. Classification Rates for Ratings' Model

Classified + if predicted Pr(D) $\geq .5$		<u>One</u>	<u>Two</u>	<u>Three</u>
		<u>Year</u>	<u>Years</u>	<u>Years</u>
True D defined as dealnodeal != 0		<u>Before</u>	<u>Before</u>	<u>Before</u>
Sensitivity	Pr(+ D)	86.54%	81.65%	89.91%
Specificity	Pr(~D)	62.15%	68.69%	43.46%
Positive predictive value	Pr(D +)	77.75%	79.94%	70.84%
Negative predictive value	Pr(~D -)	75.14%	71.01%	73.81%
False + rate for true ~D	Pr(+~D)	37.85%	31.31%	56.54%
False - rate for true D	Pr(- D)	13.46%	18.35%	10.09%
False + rate for classified	Pr(~D +)	22.25%	20.06%	29.16%
False - rate for classified	Pr(D -)	24.86%	28.99%	26.19%
Correctly classified		76.89%	76.52%	71.53%

Panel B. Classification Rates for Altman's Z-Score Model

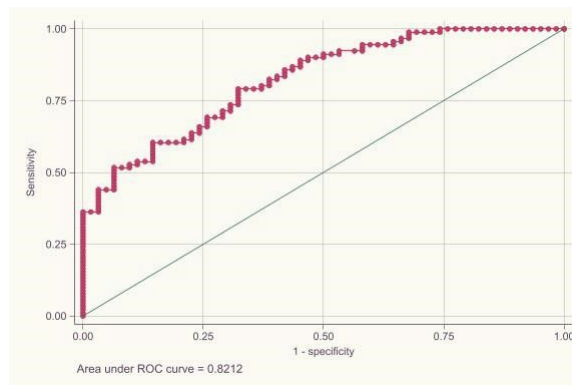
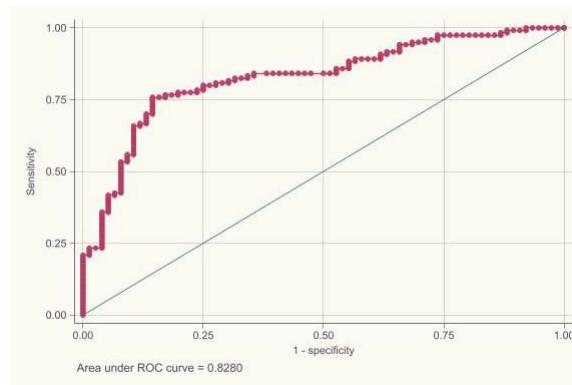
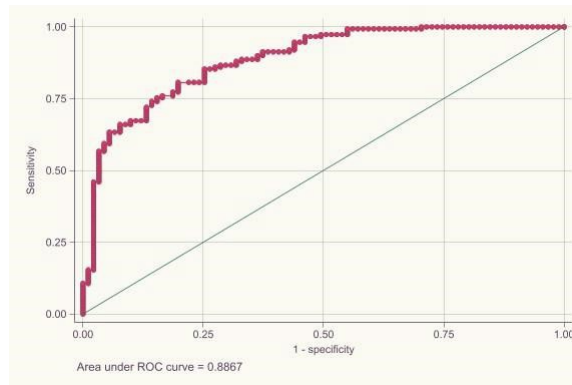
Classified + if predicted Pr(D) $\geq .5$		<u>One</u>	<u>Two</u>	<u>Three</u>
		<u>Year</u>	<u>Years</u>	<u>Years</u>
True D defined as dealnodeal != 0		<u>Before</u>	<u>Before</u>	<u>Before</u>
Sensitivity	Pr(+ D)	98.67%	99.69%	99.08%
Specificity	Pr(~D)	24.18%	1.87%	8.88%
Positive predictive value	Pr(D +)	68.20%	60.82%	62.43%
Negative predictive value	Pr(~D -)	91.67 %	80.00 %	86.36 %
False + rate for true ~D	Pr(+~D)	75.82%	98.13%	91.12%
False - rate for true D	Pr(- D)	1.33%	0.31%	0.92%
False + rate for classified +	Pr(~D +)	31.80%	39.18%	37.57%
False - rate for classified -	Pr(D -)	8.33%	20.00%	13.64%
Correctly classified		70.54%	61.00%	63.40%

Figure 9: ROC Curves for Ratings' Model and Altman's Model One, Two, and Three Years Before, Respectively Figure 9 plots the Receiver Operating Characteristic (ROC) generated from Ratings' and

Ohlson's Altman's models. The ROC for Ratings' model one year before merger announcement is 0.8867, the ROC for

Ratings' model two years before is 0.8280, the ROC for Ratings' model three years before is 0.8212, while the ROC for Altman's model one year before merger announcement is 0.9084, the ROC for Altman's model two years before merger announcement is 0.7808, and the ROC for Altman's model three years before is 0.7808.

Panel A. Ratings' Model



Panel B. Altman’s Model

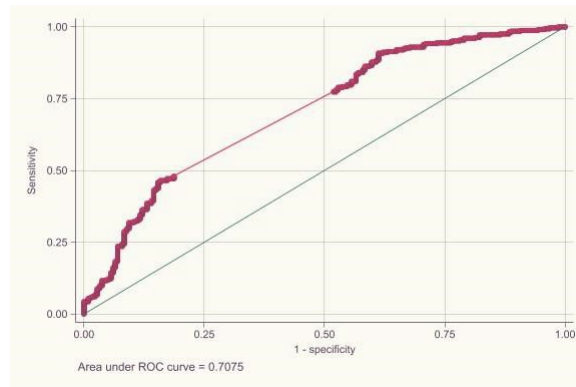
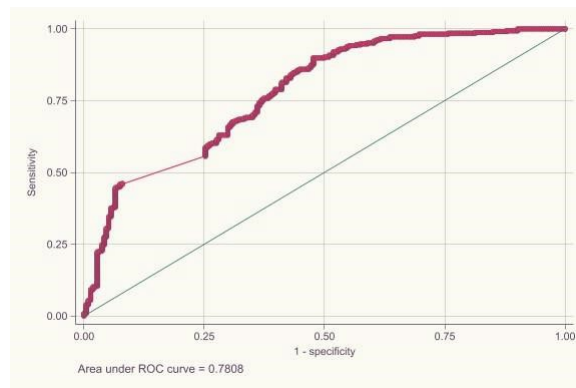
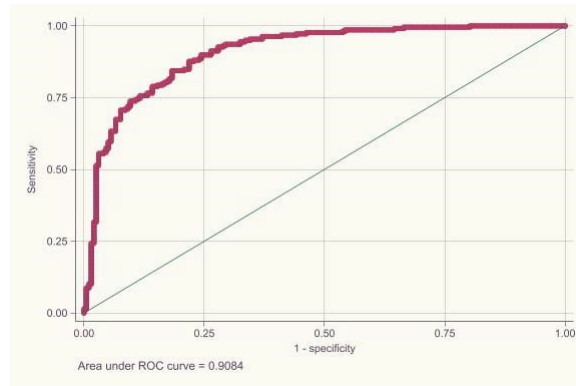


Table 9: Classification Rates Using Rating Categories’ Model One, Two, and Three Years Before

Table 9 reports classification rates for breakdown of ratings’ which shows sensitivity, specificity, positive predictive value, negative predictive value, false positives, and false negatives. The results are reported for each model one, two, and three years before merger announcement for Ratings’ model using breakdown of rating categories.

Classified + if predicted

Career_ Comp_ Culture_ Senior_ Worklife_ Overall_

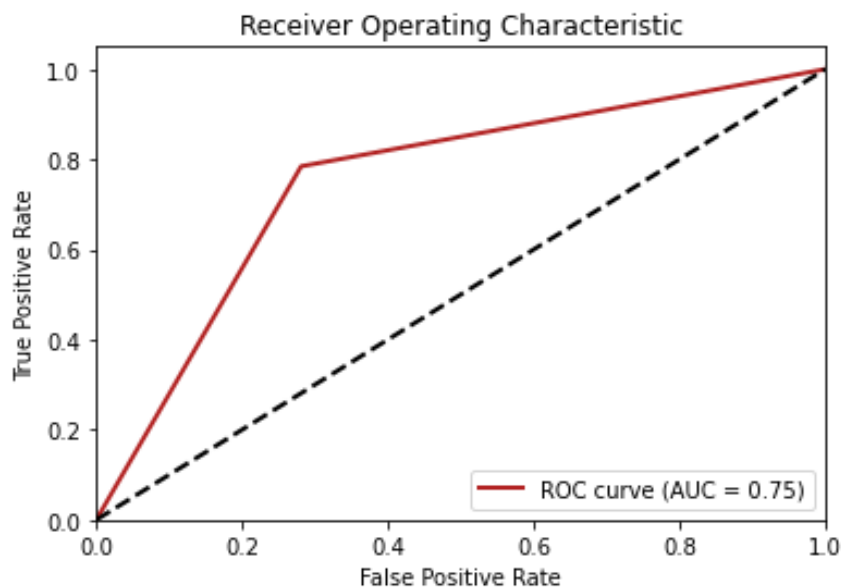
Pr(D)	$\geq .5$	Op ps 1	Benefits_1	Values_1	Lead_1	Balance_1	Rating_1
True D defined as dealnodeal !=	0						
Sensitivity	Pr(+ D)	86.85%	86.85%	86.85%	86.85%	86.85%	86.85%
Specificity	Pr(~D)	65.42%	65.42%	65.89%	66.36%	66.36%	66.36%
Positive predictive value	Pr(D +)	79.33%	79.33%	79.55%	79.78%	79.78%	79.78%
Negative predictive value	Pr(~D -)	76.50%	76.50%	76.63%	76.76%	76.76%	76.76%
False + rate for true ~D	Pr(+~D)	34.58%	34.58%	34.11%	33.64%	33.64%	33.64%
False - rate for true D	Pr(- D)	13.15%	13.15%	13.15%	13.15%	13.15%	13.15%
False + rate for classified +	Pr(~D +)	20.67%	20.67%	20.45%	20.22%	20.22%	20.22%
False - rate for classified -	Pr(D -)	<u>23.50%</u>	<u>23.50%</u>	<u>23.37%</u>	<u>23.24%</u>	<u>23.24%</u>	<u>23.24%</u>
<u>Correctly classified</u>		<u>78.37%</u>	<u>78.37%</u>	<u>78.56%</u>	<u>78.74%</u>	<u>78.74%</u>	<u>78.74%</u>

Classified + if predicted Pr(D)	$\geq .5$	Career_ Opps 2	Comp_ Benefits 2	Culture_ Values 2	Senior_ Lead 2	Worklife_ Balance 2	Overall_ Rating 2
True D defined as dealnodeal !=	0						
Sensitivity	Pr(+ D)	81.96%	81.35%	81.65%	81.96%	81.96%	81.96%
Specificity	Pr(~D)	62.15%	59.81%	61.68%	62.15%	61.21%	61.68%
Positive predictive value	Pr(D +)	76.79%	75.57%	76.50%	76.79%	76.35%	76.57%
Negative predictive value	Pr(~D -)	69.27%	67.72%	68.75%	69.27%	68.95%	69.11%
False + rate for true ~D	Pr(+~D)	37.85%	40.19%	38.32%	37.85%	38.79%	38.32%
False - rate for true D	Pr(- D)	18.04%	18.65%	18.35%	18.04%	18.04%	18.04%
False + rate for classified +	Pr(~D +)	23.21%	24.43%	23.50%	23.21%	23.65%	23.43%
False - rate for classified -	Pr(D -)	<u>30.73%</u>	<u>32.28%</u>	<u>31.25%</u>	<u>30.73%</u>	<u>31.05%</u>	<u>30.89%</u>
<u>Correctly classified</u>		<u>74.12%</u>	<u>72.83%</u>	<u>73.75%</u>	<u>74.12%</u>	<u>73.75%</u>	<u>73.94%</u>

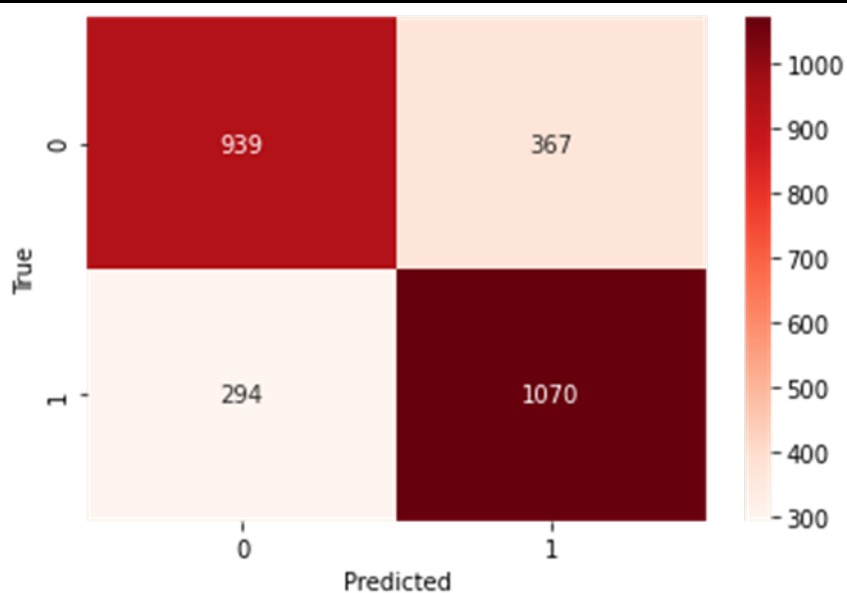
Classified + if predicted Pr(D)	$\geq .5$	Career_ Opps 3	Comp_ Benefits 3	Culture_ Values 3	Senior_ Lead 3	Worklife_ Balance 3	Overall_ Rating 3
True D defined as dealnodeal !=	0						
Sensitivity	Pr(+ D)	81.96%	91.74%	89.91%	91.13%	92.05%	90.21%
Specificity	Pr(~D)	61.68%	31.78%	35.51%	31.78%	32.24%	32.71%
Positive predictive value	Pr(D +)	76.57%	67.26%	68.06%	67.12%	67.49%	67.20%
Negative predictive value	Pr(~D -)	69.11%	71.58%	69.72%	70.10%	72.63%	68.63%

False + rate for true ~D	Pr(+~D)	38.32%	68.22%	64.49%	68.22%	67.76%	67.29%
False - rate for true D	Pr(- D)	18.04%	8.26%	10.09%	8.87%	7.95%	9.79%
False + rate for classified +	Pr(~D +)	23.43%	32.74%	31.94%	32.88%	32.51%	32.80%
<u>False - rate for classified -</u>	<u>Pr(D -)</u>	<u>30.89%</u>	28.42%	30.28%	29.90%	27.37%	31.37%
<hr/>							
<u>Correctly classified</u>		<u>73.94%</u>	<u>68.02%</u>	<u>68.39%</u>	<u>67.65%</u>	<u>68.39%</u>	<u>67.47%</u>

Table and Figure 10: Neural Network Learning from Textual Reviews' Data (ROC Curve and Accuracy – based on 6,906 actual failed company reviews and 6,443 pseudo company reviews)



X_train.shape	(10679, 1500)
y_train.shape	(10679,)
X_test.shape	(2670, 1500)
y_test.shape	(2670,)
Accuracy	73.37%
Standard Deviation	1.16%

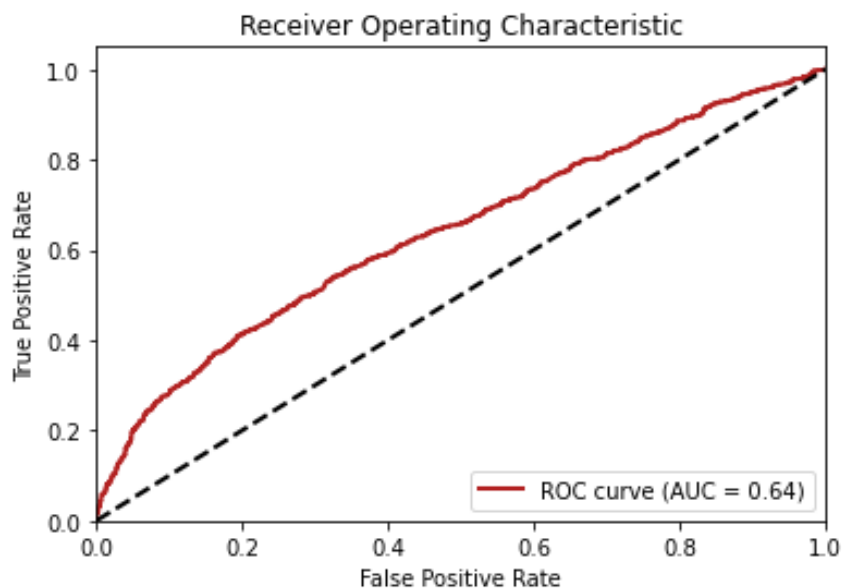


True/False Positives/Negatives

True Negatives (TN)	939
False Positives (FP)	367
False Negatives (FN)	294

True Positives (TP)	1070
---------------------	------

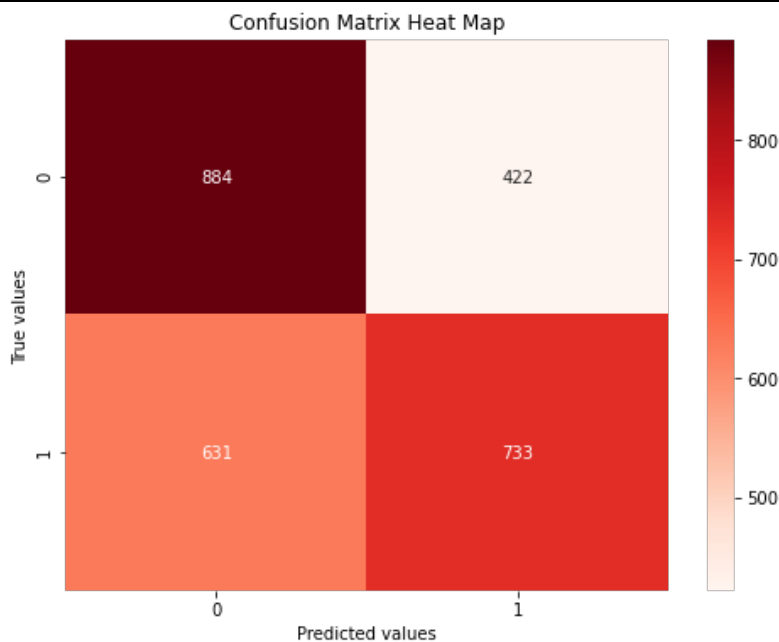
Table and Figure 11: Neural Network Learning from Ratings' Data (ROC Curve and Accuracy – based on 6,906 actual failed company ratings and 6,443 pseudo company ratings)



X_train.shape	(10679, 7)
y_train.shape	(10679,)
X_test.shape	(2670, 7)
y_test.shape	(2670,)
Accuracy	61.66%

Standard Deviation

2.37%



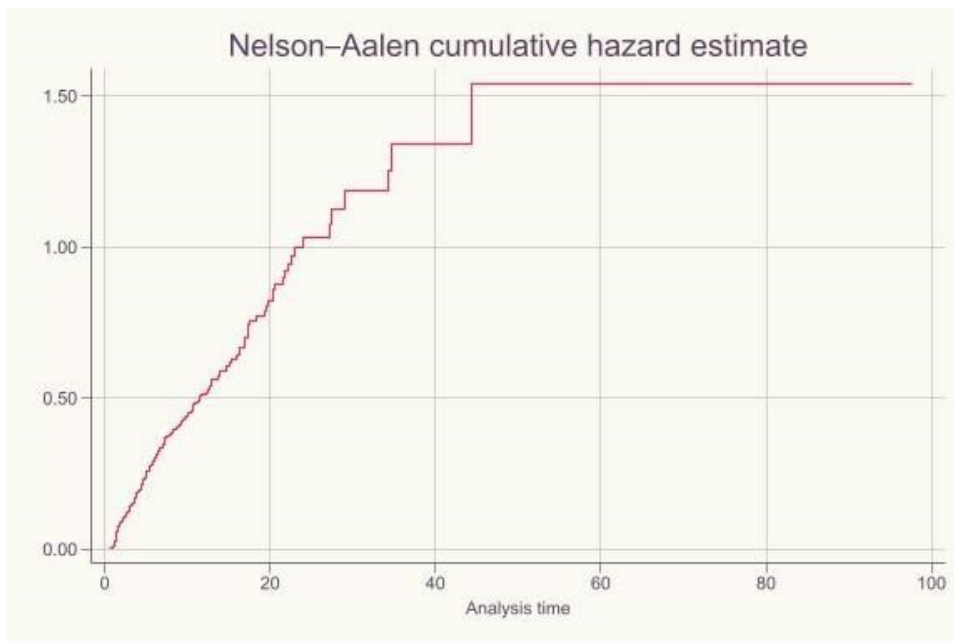
True/False Positives/Negatives

True Negatives (TN)	884
False Positives (FP)	422
False Negatives (FN)	631
True Positives (TP)	733

Figure 12: Cumulative Hazard Curve

Figure 10 presents Nelson-Aalen cumulative hazard estimate (Panel A) and Kaplan-Meier survival estimates (Panel B). Nelson-Aalen cumulative hazard estimate figure plots the analysis time (time from bankruptcy filing to time of emergence) on the horizontal axis, while the Kaplan-Meier survival estimates figure plots the analysis time (time from bankruptcy filing to time of emergence) on the horizontal axis.

Panel A. Nelson-Aalen Cumulative Hazard Estimate



Panel B. Kaplan-Meier Survival Estimates

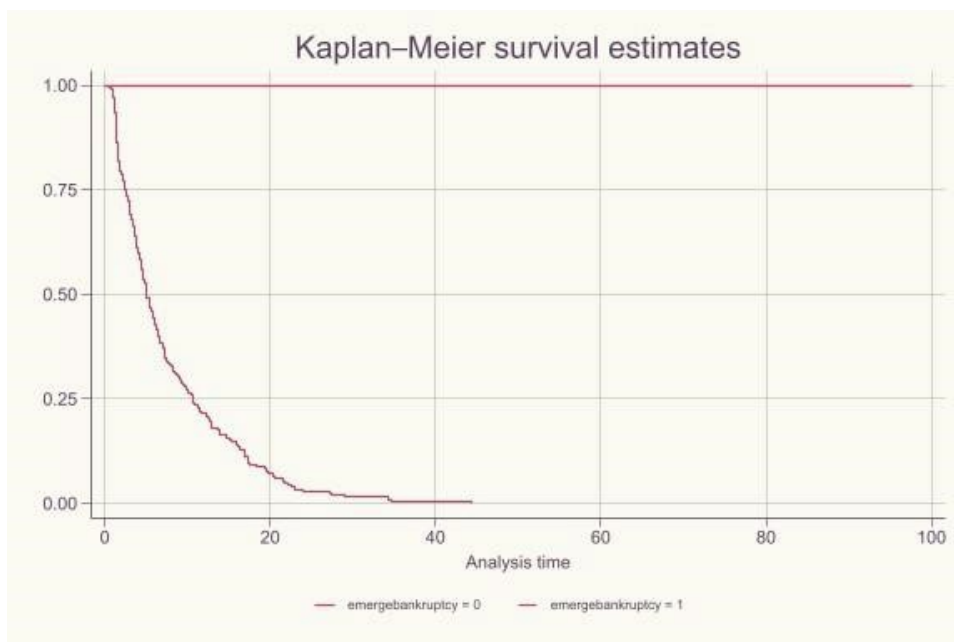
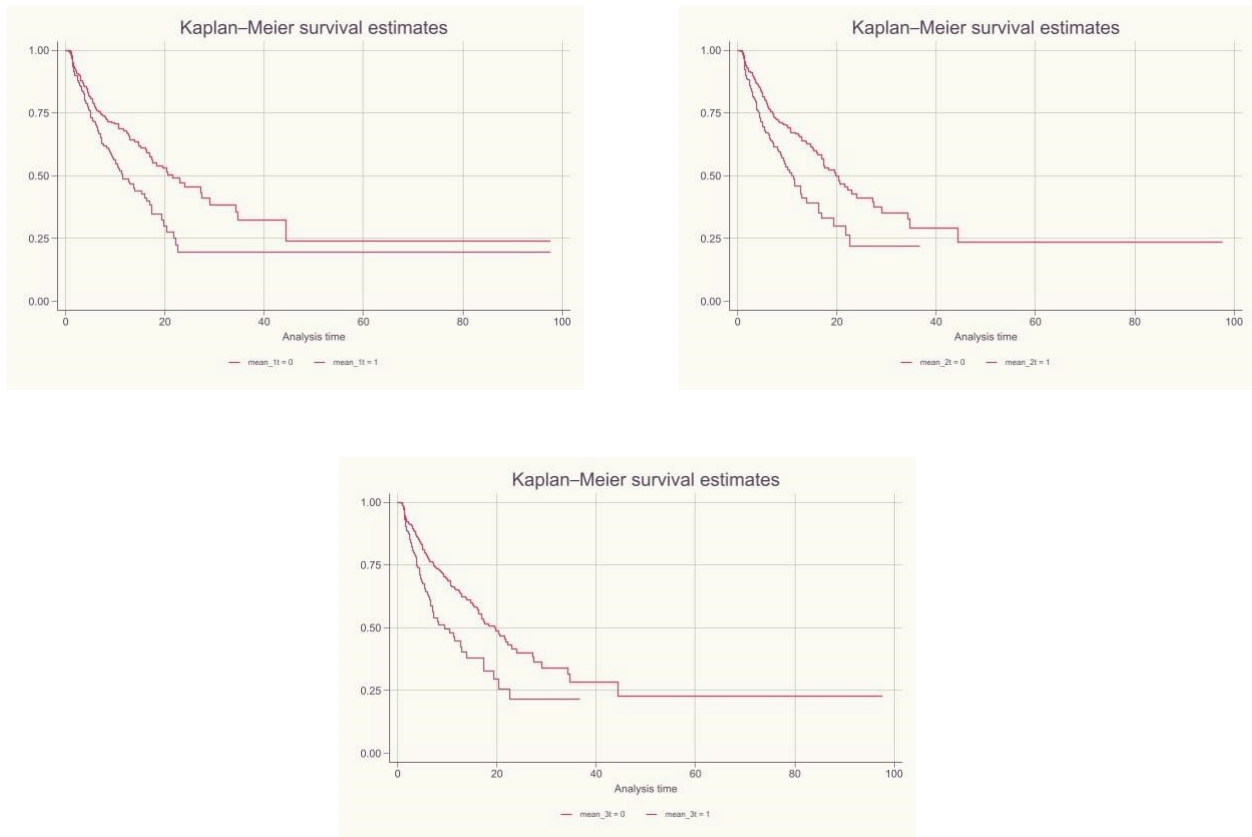


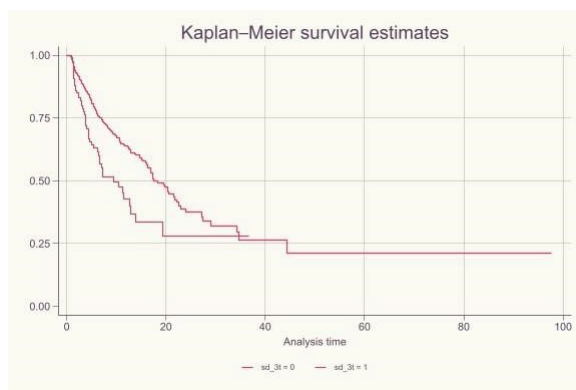
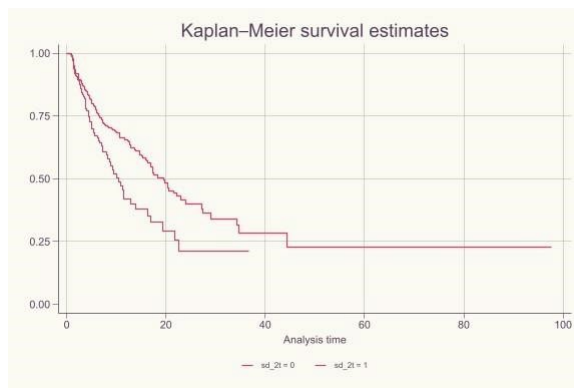
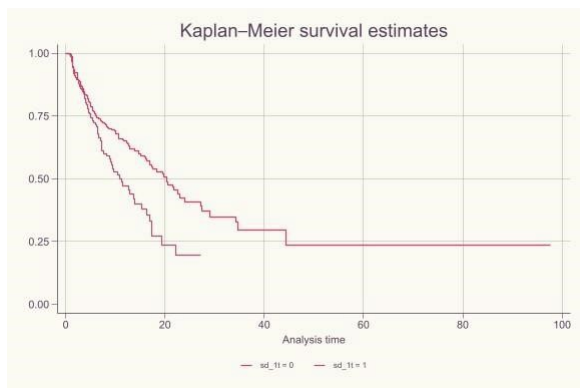
Figure 13: Univariate Analyses

Figure 11 presents Kaplan-Meier survival estimates from univariate analyses with mean employee satisfaction one, two, and three years before merger announcement (Panel A), standard deviation of employee satisfaction one, two, and three years before merger announcement (Panel B), and Z-Score one, two, and three years before merger announcement (Panel C).

Panel A. Kaplan-Meier Survival Estimates for Models with Mean of Employee Satisfaction



Panel B. Kaplan-Meier Survival Estimates for Models with Standard Deviation of Employee Satisfaction



Panel C. Kaplan-Meier Survival Estimates for Models with Z-Score

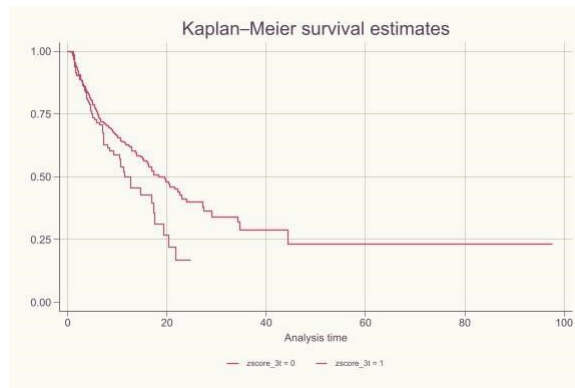
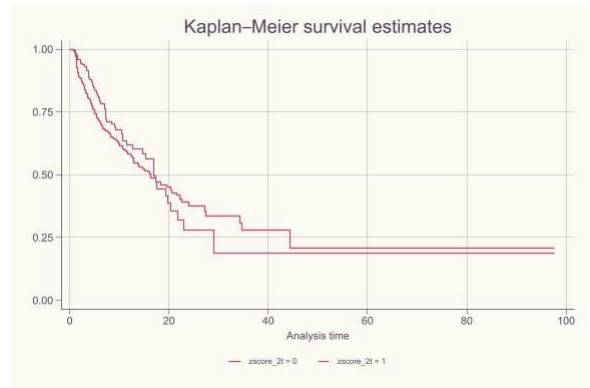
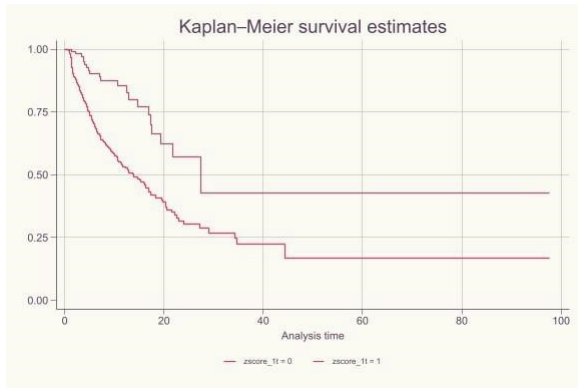


Table 12: Survival Analyses with, and Z-Score Mean of Ratings, Standard Deviation of Ratings

Table 12 presents survival analyses with Cox regression for bankruptcy emergence with mean of employee satisfaction (Panel A), standard deviation of employee satisfaction (Panel B), and Z-Score (Panel C) one, two, and three years before bankruptcy filings. Firm controls (ROA, Sale Growth, Cash, Book Leverage, R&D, and BTM) are also included in the model. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A. Cox Regression for Bankruptcy Emergence with Mean of Employee Satisfaction

Cox Regression for Bankruptcy Emergence with Mean						
<i>Main Effects</i>						
Mean_1	0.490***	(3.37)				
Mean_2			0.346*	(2.19)		
Mean_3					0.273	(1.62)
<i>Controls</i>						
ROA_1	0.309	(1.31)	1.635**	(2.62)	1.393**	(3.05)
Sale_Growth_1	-0.0509	(-1.10)	-0.0233	(-0.22)	0.274	(1.91)
Cash_1	-3.569**	(-3.15)	-1.300	(-1.48)	-1.987*	(-2.15)
Book_Leverage_1	0.705***	(4.62)	1.156***	(5.31)	1.049***	(4.12)
R&D_1	-1.160	(-0.56)	2.784	(1.64)	2.716	(0.58)
BTM_1	0.393**	(2.60)	0.746**	(3.15)	-0.0657	(-0.34)
Observations	301		301		301	
T-statistics in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"						

Panel B. Cox Regression for Bankruptcy Emergence with Standard Deviation of Employee Satisfaction

Cox Regression for Bankruptcy Emergence with Standard Deviation

<i>Main Effects</i>						
SD_1	0.470**	(3.07)				
SD_2			0.159	(0.99)		
SD_3					0.269	(1.54)
<i>Controls</i>						
ROA_1	0.296	(1.25)	1.676**	(2.73)	1.343**	(2.89)
Sale_Growth_1	-0.0517	(-1.14)	0.0283	(0.28)	0.308*	(2.22)
Cash_1	-3.563**	(-3.10)	-1.205	(-1.36)	-2.003*	(-2.12)
Book_Leverage_1	0.723***	(4.68)	1.153***	(5.32)	1.033***	(4.03)
R&D_1	-1.344	(-0.64)	2.564	(1.50)	2.858	(0.60)
BTM_1	0.394**	(2.59)	0.771**	(3.24)	-0.0528	(-0.27)
Observations	301		301		301	
T-statistics in parentheses						
="* p<0.05						
** p<0.01						
*** p<0.001"						

Panel C. Cox Regression for Bankruptcy Emergence with Z-Score

Cox Regression for Bankruptcy Emergence with Z-Score						
<i>Main Effects</i>						
ZScore_1	-0.251	(-1.06)				
ZScore_2			0.143	(0.89)		
ZScore_3					0.0857	(0.51)
<i>Controls</i>						
ROA_1	0.360	(1.54)	1.662**	(2.72)	1.326**	(2.90)
Sale_Growth_1	-0.0473	(-1.10)	0.0574	(0.59)	0.322*	(2.33)
Cash_1	-3.486**	(-3.05)	-1.280	(-1.44)	-2.043*	(-2.19)
Book_Leverage_1	0.703***	(4.67)	1.202***	(5.44)	1.106***	(4.40)
R&D_1	-0.607	(-0.30)	2.404	(1.41)	3.823	(0.82)
BTM_1	0.438**	(2.89)	0.798***	(3.33)	-0.0589	(-0.30)
Observations	301		301		301	
T-statistics in parentheses						
="* p<0.05						
** p<0.01						
*** p<0.001"						

Table 13: Survival Analyses with Mean and Standard Deviation of Rating Categories

Table 13 presents survival analyses with Cox regression for bankruptcy emergence with mean of employee satisfaction and standard deviation of employee satisfaction rating categories (Career Opportunities, Compensation Benefits, Culture Values, Senior Leadership, Work-Life Balance, and Overall Rating) one, two, and three years before bankruptcy filings. Firm controls (ROA, Sale Growth, Cash, Book Leverage, R&D, and BTM) are also included in the model. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Cox Regression for Bankruptcy Emergence with Mean Rating Categories

<i>Main Effects</i>								
careeropps_mean_1	0.282	(1.79)						
compensation_mean_1			0.404*	(2.55)				
culture_mean_1					0.339*	(2.21)		
seniorleadership_mean_1							0.129	(0.78)
<i>Controls</i>								
ROA_1	0.381	(1.63)	0.395	(1.69)	0.302	(1.29)	0.363	(1.56)
Sale_Growth_1	-0.0497	(-1.18)	-0.0502	(-1.17)	-0.0491	(-1.16)	-0.0501	(-1.20)
Cash_1	-3.550**	(-3.13)	-3.488**	(-3.12)	-3.440**	(-3.00)	-3.538**	(-3.10)
Book_Leverage_1	0.725***	(4.89)	0.728***	(4.88)	0.692***	(4.57)	0.720***	(4.83)
R&D_1	-0.637	(-0.32)	-1.236	(-0.59)	-0.265	(-0.13)	-0.586	(-0.29)
BTM_1	0.401**	(2.68)	0.387**	(2.58)	0.405**	(2.69)	0.408**	(2.72)
Observations	301		301		301		301	
T-statistics in parentheses								
="* p<0.05								
** p<0.01								
*** p<0.001"								

Cox Regression for Bankruptcy Emergence with Mean Rating Categories

<i>Main Effects</i>				
worklife_mean_1	0.275	(1.71)		
overall_mean_1			0.298	(1.75)
<i>Controls</i>				
ROA_1	0.370	(1.59)	0.356	(1.54)
Sale_Growth_1	-0.0492	(-1.18)	-0.0494	(-1.18)
Cash_1	-3.593**	(-3.18)	-3.483**	(-3.07)
Book_Leverage_1	0.719***	(4.83)	0.720***	(4.86)
R&D_1	-0.874	(-0.43)	-0.390	(-0.20)
BTM_1	0.406**	(2.71)	0.392**	(2.62)
Observations	301		301	
	T-statistics in parentheses			
	="* p<0.05			
	** p<0.01			
	*** p<0.001"			

Cox Regression for Bankruptcy Emergence with Mean Rating Categories

<i>Main Effects</i>								
careeropps_mean_2	0.685***	(3.76)						
compensation_mean_2			0.688***	(3.72)				
culture_mean_2					0.325	(1.92)		
seniorleadership_mean_2							0.723***	(3.97)
<i>Controls</i>								
ROA_2	1.706**	(2.78)	1.670**	(2.74)	1.636**	(2.66)	1.683**	(2.76)
Sale_Growth_2	0.0133	(0.13)	0.00970	(0.10)	0.0201	(0.20)	0.0165	(0.16)
Cash_2	-1.137	(-1.27)	-1.187	(-1.33)	-1.120	(-1.26)	-1.100	(-1.23)
Book_Leverage_2	1.181***	(5.36)	1.195***	(5.56)	1.151***	(5.27)	1.164***	(5.35)
R&D_2	2.823	(1.67)	2.830	(1.67)	2.637	(1.55)	2.789	(1.65)
BTM_2	0.778**	(3.27)	0.722**	(3.08)	0.769**	(3.23)	0.773**	(3.24)
Observations	301		301		301		301	
T-statistics in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"								

Cox Regression for Bankruptcy Emergence with Mean Rating Categories

<i>Main Effects</i>				
worklife_mean_2		0.700***	(3.78)	
overall_mean_2				0.723*** (3.89)
<i>Controls</i>				
ROA_2		1.641**	(2.68)	1.654** (2.71)
Sale_Growth_2		0.0110	(0.11)	0.0158 (0.16)
Cash_2		-1.484	(-1.74)	-1.201 (-1.35)
Book_Leverage_2		1.175***	(5.48)	1.193*** (5.51)
R&D_2		2.938	(1.76)	2.800 (1.65)
BTM_2		0.758**	(3.23)	0.698** (2.96)
Observations		301		301
T-statistics in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"				

Cox Regression for Bankruptcy Emergence with Mean Rating Categories

<i>Main Effects</i>								
careerops_mean_3	0.302	(1.51)						
compensation_mean_3			0.424*	(2.12)				
culture_mean_3					0.129	(0.69)		
seniorleadership_mean_3							0.388*	(1.97)
<i>Controls</i>								
ROA_3	1.324**	(2.90)	1.420**	(3.20)	1.393**	(3.07)	1.430**	(3.23)
Sale_Growth_3	0.285*	(2.05)	0.261	(1.87)	0.319*	(2.32)	0.288*	(2.07)
Cash_3	-2.092*	(-2.21)	-2.051*	(-2.25)	-1.928*	(-2.07)	-2.013*	(-2.19)
Book_Leverage_3	1.068***	(4.20)	1.096***	(4.35)	1.082***	(4.29)	1.129***	(4.52)
R&D_3	3.083	(0.66)	2.650	(0.57)	3.625	(0.78)	3.489	(0.74)
BTM_3	-0.0322	(-0.17)	-0.0506	(-0.26)	-0.0603	(-0.31)	-0.0910	(-0.47)
Observations	301		301		301		301	
T-statistics in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"								

Cox Regression for Bankruptcy Emergence with Mean Rating Categories

<i>Main Effects</i>				
worklife_mean_3		0.293	(1.41)	
overall_mean_3				0.563** (2.94)
<i>Controls</i>				
ROA_3		1.394**	(3.14)	1.407** (3.25)
Sale_Growth_3		0.312*	(2.27)	0.266 (1.89)
Cash_3		-2.032*	(-2.16)	-2.217* (-2.37)
Book_Leverage_3		1.101***	(4.39)	1.121*** (4.47)
R&D_3		2.850	(0.60)	3.395 (0.70)
BTM_3		-0.0630	(-0.33)	-0.0813 (-0.42)

Observations	301	301
	T-statistics in parentheses	
	="* p<0.05	
	** p<0.01	
	*** p<0.001"	

Cox Regression for Bankruptcy Emergence with Standard Deviation Rating Categories

<i>Main Effects</i>								
careerops_sd_1	0.375*	(2.39)						
compensation_sd_1			0.451**	(2.87)				
culture_sd_1					0.458**	(2.80)		
seniorleadership_sd_1							0.388*	(2.48)
<i>Controls</i>								
ROA_1	0.320	(1.36)	0.293	(1.23)	0.306	(1.30)	0.318	(1.34)
Sale_Growth_1	-0.0506	(-1.16)	-0.0513	(-1.15)	-0.0498	(-1.16)	-0.0509	(-1.16)
Cash_1	-3.533**	(-3.09)	-3.486**	(-3.04)	-3.329**	(-2.91)	-3.493**	(-3.03)
Book_Leverage_1	0.725***	(4.76)	0.709***	(4.59)	0.714***	(4.66)	0.719***	(4.70)
R&D_1	-0.990	(-0.46)	-1.083	(-0.50)	-0.323	(-0.16)	-0.975	(-0.46)
BTM_1	0.389**	(2.58)	0.409**	(2.68)	0.383*	(2.53)	0.403**	(2.66)

Observations	301	301	301	301
		T-statistics in parentheses		
		="* p<0.05		
		** p<0.01		
		*** p<0.001"		

Cox Regression for Bankruptcy Emergence with Standard Deviation Rating Categories

<i>Main Effects</i>				
worklife_sd_1		0.336*	(2.12)	
overall_sd_1				0.393* (2.50)
<i>Controls</i>				
ROA_1		0.317	(1.34)	0.313 (1.33)
Sale_Growth_1		-0.0514	(-1.17)	-0.0517 (-1.16)
Cash_1		-3.572**	(-3.09)	-3.559** (-3.07)

Book_Leverage_1	0.717***	(4.70)	0.725***	(4.74)
R&D_1	-0.435	(-0.21)	-0.839	(-0.41)
BTM_1	0.394**	(2.60)	0.400**	(2.64)
Observations	301		301	

T-statistics in parentheses
 ="* p<0.05
 ** p<0.01
 *** p<0.001"

Cox Regression for Bankruptcy Emergence with Standard Deviation Rating Categories

<i>Main Effects</i>								
careeropps_sd_2	0.151	(0.90)	-0.00912	(-0.05)				
compensation_sd_2								
culture_sd_2				0.249	(1.43)			
seniorleadership_sd_2						0.150	(0.89)	
<i>Controls</i>								
ROA_2	1.674**	(2.74)	1.722**	(2.85)	1.637**	(2.66)	1.664**	(2.71)
Sale_Growth_2	0.0308	(0.30)	0.0551	(0.55)	0.0244	(0.24)	0.0349	(0.35)
Cash_2	-1.230	(-1.38)	-1.252	(-1.41)	-1.121	(-1.26)	-1.206	(-1.36)
Book_Leverage_2	1.152***	(5.32)	1.162***	(5.40)	1.141***	(5.25)	1.150***	(5.30)
R&D_2	2.528	(1.48)	2.341	(1.36)	2.542	(1.49)	2.496	(1.46)
BTM_2	0.777**	(3.27)	0.793***	(3.34)	0.755**	(3.18)	0.782***	(3.29)
Observations	301		301		301		301	

T-statistics in parentheses
 ="* p<0.05
 ** p<0.01
 *** p<0.001"

Cox Regression for Bankruptcy Emergence with Standard Deviation Rating Categories

<i>Main Effects</i>				
worklife_sd_2	0.181	(1.06)		
overall_sd_2			0.107	(0.63)

<i>Controls</i>				
ROA_2	1.663**	(2.71)	1.683**	(2.75)
Sale_Growth_2	0.0310	(0.31)	0.0410	(0.41)
Cash_2	-1.198	(-1.35)	-1.241	(-1.40)
Book_Leverage_2	1.145***	(5.27)	1.157***	(5.36)
R&D_2	2.512	(1.47)	2.456	(1.44)
BTM_2	0.768**	(3.23)	0.775**	(3.25)
Observations	301		301	

T-statistics in parentheses
 ="* p<0.05
 ** p<0.01
 *** p<0.001"

Cox Regression for Bankruptcy Emergence with Standard Deviation Rating Categories

<i>Main Effects</i>								
careeropps_sd_3	0.317	(1.74)						
compensation_sd_3			0.200	(1.07)				
culture_sd_3					0.335	(1.68)		
seniorleadership_sd_3							0.168	(0.91)
<i>Controls</i>								
ROA_3	1.303**	(2.81)	1.347**	(2.93)	1.319**	(2.82)	1.357**	(2.95)
Sale_Growth_3	0.332*	(2.39)	0.318*	(2.32)	0.330*	(2.40)	0.327*	(2.39)
Cash_3	-2.015*	(-2.12)	-1.997*	(-2.13)	-1.922*	(-2.05)	-1.973*	(-2.12)
Book_Leverage_3	1.008***	(3.88)	1.052***	(4.11)	0.991***	(3.78)	1.058***	(4.14)
R&D_3	3.305	(0.69)	3.027	(0.64)	3.581	(0.76)	3.377	(0.73)
BTM_3	-0.0605	(-0.31)	-0.0488	(-0.25)	-0.0490	(-0.25)	-0.0554	(-0.29)
Observations	301		301		301		301	

T-statistics in parentheses
 ="* p<0.05
 ** p<0.01
 *** p<0.001"

Cox Regression for Bankruptcy Emergence with Standard Deviation Rating Categories

<i>Main Effects</i>				
worklife_sd_3	0.200	(1.07)		
overall_sd_3			0.318	(1.72)
<i>Controls</i>				
ROA_3	1.351**	(2.94)	1.345**	(2.87)
Sale_Growth_3	0.314*	(2.29)	0.315*	(2.30)
Cash_3	-2.011*	(-2.15)	-2.015*	(-2.14)
Book_Leverage_3	1.060***	(4.17)	1.009***	(3.88)
R&D_3	3.410	(0.73)	2.979	(0.63)
BTM_3	-0.0452	(-0.24)	-0.0459	(-0.24)
Observations	301		301	
	T-statistics in parentheses			
	="* p<0.05			
	** p<0.01			
	*** p<0.001"			

Online Appendix

for

Predicting Bankruptcy: Ask the Employees

By John Knopf and Kristina Lalova

Table A1: Correlation Matrix A1.1. Correlations between Z-Score and Mean of Ratings Two Years Before

	ZScore_2	Mean_2	SD_2	Career_2	Compensation_2	Culture_2	Leadership_2	Worklife_2	Overall_2
ZScore_2	1								
Mean_2	-0.0107	1							
SD_2	-0.0476	0.7552	1						
Career_2	-0.0119	0.9738	0.7347	1					
Compensation_2	0.0005	0.9774	0.7285	0.9532	1				
Culture_2	-0.0492	0.8980	0.7624	0.8409	0.8437	1			
Leadership_2	0.0057	0.9707	0.7133	0.9501	0.9344	0.8373	1		
Worklife_2	-0.0041	0.9744	0.7103	0.9412	0.9536	0.8288	0.9424	1	
Overall_2	-0.0045	0.9746	0.7122	0.9381	0.9503	0.8396	0.9405	0.9498	1

. Correlations between Z-

<i>A1.2</i>	<i>Score and Standard Deviation of Ratings Two Years Before</i>								
	ZScore_2	Mean_2	SD_2	Career_2	Compensation_2	Culture_2	Leadership_2	Worklife_2	Overall_2
ZScore_2	1								
Mean_2	-0.0107	1							
SD_2	-0.0476	0.7552	1						
Career_2	-0.0465	0.7306	0.9595	1					
Compensation_2	-0.0453	0.6919	0.9468	0.9019	1				
Culture_2	-0.0421	0.6953	0.9460	0.8751	0.8783	1			
Leadership_2	-0.0462	0.7449	0.9660	0.9217	0.8794	0.9076	1		
Worklife_2	-0.0462	0.7276	0.9613	0.9055	0.8911	0.8790	0.9182	1	
Overall_2	-0.0472	0.7451	0.9637	0.9088	0.8981	0.8863	0.9148	0.9274	

<i>A1.3</i>	<i>Score and Mean of Ratings Three Years Before</i>							
	ZScore_3	Mean_3	SD_3	Career_3	Compensation_3	Culture_3	Leadership_3	Worklife_3
ZScore_3	1							
Mean_3	0.0622	1						
SD_3	0.0714	0.7774	1					
Career_3	0.0573	0.9776	0.7745	1				
Compensation_3	0.0838	0.9731	0.7328	0.9482	1			
Culture_3	0.0684	0.9221	0.7474	0.8724	0.8613	1		
Leadership_3	0.0361	0.9830	0.7602	0.9598	0.9417	0.9038	1	
Worklife_3	0.0511	0.9748	0.7597	0.9493	0.9578	0.8527	0.9491	1
Overall_3	0.0609	0.9748	0.7434	0.9473	0.9327	0.8765	0.9574	0.9428

A1.4

Score and Standard Deviation of Ratings Three Years Before

	ZScore_3	Mean_3	SD_3	Career_3	Compensation_3	Culture_3	Leadership_3	Worklife_3	Overall_3
ZScore_3	1								
Mean_3	0.0622	1							
SD_3	0.0714	0.7774	1						
Career_3	0.0646	0.7683	0.9641	1					
Compensation_3	0.0685	0.7257	0.9484	0.9014	1				
Culture_3	0.0754	0.7073	0.9330	0.8741	0.8523	1			
Leadership_3	0.0199	0.7479	0.9722	0.9331	0.8999	0.8999	1		
Worklife_3	0.0891	0.7484	0.9725	0.9208	0.9277	0.8795	0.9332	1	
Overall_3	0.0951	0.7666	0.9514	0.9071	0.8736	0.8455	0.9113	0.9244	

Table A2: Linear Discriminant Analysis with Unstandardized and Standardized Coefficients for Altman's Model (equivalent to Tables 2 and 3)

Predictors	Unstandardized Coefficients (1)	Standardized Coefficients (2)	Unstandardized Coefficients (3)	Standardized Coefficients (4)
Mean_1		-		0.00466
WCTA_1	1.476932		0.550663	-1.4739
RETA_t1	-0.07409		-0.13922	0.072564
EBITTA_t1	-0.16224		-0.0527	0.171106

MVETL_t1	0.204272	0.501703	-0.20474	-0.50285
STA_t1	0.698869	0.60159	-0.69903	-0.60173
_cons	-0.72771	-	0.692398	-

Table A3: Error Rates for Altman’s and Ratings’ Models in a Discriminant Analysis Setting with Variable Information One, Two, and Three Years Before Bankruptcy Filing

Ratings' Model	Bankruptcy?								
	<u>One Year Before</u>			<u>Two Years Before</u>			<u>Three Years Before</u>		
Error rate	0	1	Total	0	1	Total	0	1	Total
Stratified	0.205768	0.268095	0.236931	0.315487	0.279221	0.297354	0.292589	0.41908	0.355835
Unstratified	0.291163	0.183687	0.237425	0.375824	0.220211	0.298017	0.331577	0.380586	0.356081
Priors	0.5	0.5		0.5	0.5		0.5	0.5	

Altman's Model	Bankruptcy?								
	<u>One Year Before</u>			<u>Two Years Before</u>			<u>Three Years Before</u>		
Error rate	0	1	Total	0	1	Total	0	1	Total
Stratified	0.371231	0.152863	0.262047	0.464176	0.226945	0.34556	0.536242	0.225025	0.380634

Unstratified	0.445255	0.082962	0.264108	0.513077	0.187525	0.350301	0.57001	0.198978	0.384494
Priors	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

The Evolution of the Financial Technology Network

Corresponding author: Alisher Mansurov, Ph.D.

Assistant Professor

School of Business

Nipissing University

Office H129, 100 College Drive, Box 5002, North Bay, ON P1B 8L7

Phone: (705) 474-3450 ext. 4033

Email: alisherm@nipissingu.ca

Co-author: Marc Pilon, PhD

Assistant Professor

Department of Accounting

HEC Montréal

3000 Côte-Sainte-Catherine Road, Montréal, Québec H3T 2A7

Phone: (514) 340-6510

Email: marc.pilon@hec.ca

ABSTRACT

The widespread adoption of financial technologies (FinTech) in the realm of financial services has garnered global attention. Yet, there is little known on the involvement of various types of organizations in the advancement of FinTech. In this study, we apply a network analysis on 23,000 FinTech news articles from 2008-2022 to illustrate the network of FinTech organizations and examine its characteristics. The results of the network analysis show that incumbent financial services providers are central in the FinTech network, while the prominence of technology giants, regulatory bodies, and FinTech start-ups has grown. Moreover, ties between similar organizations are stronger, although these connections have weakened over time.

Keywords: *network analysis, FinTech, collaboration, news articles, financial services*

Introduction

FinTech, as the name implies, arises from merging finance and technology. Broadly speaking, FinTech refers to a set of technological innovations in the financial field. Although the interaction between finance and technological innovation is not a new phenomenon, the rapid emergence and application of high-tech solutions in financial services is what defines FinTech. Early start-ups as well as established tech companies are responsible for much of this disruption which is modernizing many financial and accounting functions using innovative software, algorithms, and applications.

The success of FinTech firms is attributed to not only to their ability to provide fast and convenient financial services using sophisticated technologies (Puschmann, 2017; Boot et al., 2021) but also their expanded financial inclusion (Salampasis & Mention, 2018) that are generally not offered by incumbent financial intermediaries. This has pushed incumbents, which are often thought to be less adaptable to new changes (Hornuf, 2021), to come up with more appealing financial solutions to compete with those offered by new entrants (Kohtamäki et al., 2019). Incumbents' reaction to FinTech innovations includes developing own FinTech solutions in-house, collaborating with FinTech start-ups, and obtaining access to FinTech through acquisitions (Chemmanur et al., 2020; Murinde et al., 2022; KPMG, 2022). Incumbents are also addressing the threat of new technology by forming strategic alliances with other established players within the industry as developing FinTech internally can be expensive, slow, and complex (EY, 2019).

Despite many benefits of collaboration in FinTech, traditional players may be reluctant to cooperate on hi-tech solutions. Part of the reason may be incumbents' cybersecurity concerns related to working with new technological applications (Najaf et al., 2021). Traditional players

may also be less agile and slower to adopt innovative technologies due to their costly legacy infrastructure (Brandl & Hornuf, 2020). Moreover, regulations and compliance requirements are often shown as another reason for incumbents' passive role in FinTech (Hornuf et al., 2020). Vendrell-Herrero et al. (2017) point to innovations in digital services as a discouraging factor for incumbents' willingness to come up with their own distribution paths. Overall, the importance of traditional financial services firms in FinTech is blurred.

In this paper, we first identify the main players in global FinTech and examine how they interact. We show how traditional banks, credit card companies, FinTech start-ups, financial regulators, and other types of organizations co-exist the FinTech network over time. More precisely, we answer the following two main research questions:

RQ1. What is a FinTech network and how has it evolved over time?

RQ2. Who are the central players of the FinTech network and what are some collaboration patterns between them?

Collaboration in FinTech is still an underexplored area (Mention, 2019; Hornuf et al., 2021; Murinde et al., 2022). Our study contributes to the emerging literature by describing the global FinTech network and shedding light into the interaction patterns of its key members. The study also offers a methodological contribution by applying a network analysis on FinTech news over the last 15 years. The rest of the document is structured as follows. Section 2 reviews literatures on FinTech collaboration and network science. The following section presents the data and methodology used in the study. Next, results are presented. The final section discusses the results and concludes the study.

Literature Review

This section reviews the existing literature as it relates to Fintech collaboration and network analysis.

Collaboration in FinTech

Collaboration can be defined as “the act of working with another person or group of people to create or produce something” (Oxford English Dictionary, 2019:online). In the business context, collaboration refers to company’s cooperation with other network participants to achieve business targets and create stakeholder value (Batt & Purchase, 2004).

Although a growing number of empirical business papers have studied FinTech, literature on FinTech collaboration is scarce. Brandl & Hornuf (2020) apply a mixed-methods approach using involving network analysis to learn how FinTech companies and banks interact in the German market. By representing companies as nodes and connections as links, the authors present a graphical illustration of how banks, FinTechs, and FinTech banks are related in terms of investments, partnerships, and spin-offs. Their findings indicate that banks are more likely to form strategic partnerships to advance new technologies than directly integrate FinTech. The authors argue that the characteristics of new digital solutions and delayed decisions to upgrade legacy systems explain banks’ unwillingness to fully support new digital solutions.

FinTech-bank interactions are also the focus in Hornuf et al. (2021) who extend the findings of Brandl & Hornuf (2020) by investigating particular bank characteristics (e.g., product-related, investment) associated with a bank–fintech alliance. Using manually collected data on bank alliances that cover the major banks of Canada, France, Germany, and the UK, the study answers the questions “which banks typically collaborate with fintechs, how intensely they do so, and

which form of alliance they prefer” (p. 1505, Hornuf et al., 2020). The authors perform regression analyses and provide evidence for banks’ alliances with FinTechs based on product-related causes, and investment purposes to a lesser extent.

Boot et al. (2021) forecast the future of banks by reviewing established trends and new developments in FinTech. More specifically, they argue that new technologies will transform banks away from their traditional business model, both horizontally (i.e., operations enabled by bank’s balance sheet) and vertically (i.e., informational and communication synergies). They note the following factors that may speed up this transformation: specialized competitors offering horizontally integrated financial services, easing regulations on certain bank operations, and digital platforms offered by large technology firms (BigTech firm). Boot et al. (2021) argue that BigTech firms are capable of developing “ecosystems to rival banks’ information and communication advantages” (p. 53). The authors also discuss the important and challenging role of regulators in creating a promoting environment for FinTech.

We use a holistic approach to study collaboration and trends in FinTech. More specifically, unlike Brandl & Hornuf (2020) and Hornuf et al. (2020), our study addresses how a wide range of FinTech players (e.g., FinTech startups, traditional banks, payment processors, technology firms, regulators) come together in the network of FinTech organizations. Furthermore, we analyze how FinTech, its players, and its network evolved over time to propose our view of what the future holds.

[Network Science/Analysis](#)

A network is a structure that can be graphically represented by its points (or nodes) and linkages (or links) that connect the nodes. The concepts of a network, nodes, and linkages come from

graph theory which uses formal math, statistics, and computer science to explain information contained in the vertices (i.e., nodes) and edges (i.e., links). Mathematically, a network can be expressed in the form of a square matrix that assigns all points and linkages its rows and columns. This matrix is called as an adjacency matrix since its elements express how close or adjacent any given pair is in the network. The representation of a network in the form of an adjacency matrix allows us to understand various network properties about the structure of the network, the importance of each point in the network, and other relationships.

In social sciences, network analysis methods have been developed over several decades by coupling mathematical and statistical foundations with theoretical and empirical research in the domain. Wasserman & Faust (1994) define a social network as a social environment of a finite set or sets of members and relationships among them. A social network expresses “dependency relationships resulting from the differential possession of scarce resources at the nodes and the structured allocation of these resources at the ties.” (Wellman, 1983, p. 157). Thus, a social network is instrumental in distributing ideas, information, and other organizational resources among their members (Kempe et al., 2003; Lea et al., 2006).

Since it borrows from multiple disciplines, social network analysis is somewhat a sophisticated research area. However, at its core, social network analysis entails analyzing relationships among interacting entities in a social system. It is a broad intellectual approach to study patterns of ties linking organizations, individuals, or groups (Wellman, 1983). Patterns and relationships in a network can be analyzed through various analytical and visual techniques, as described below.

The widespread use of computers and programming by academic researchers has accelerated the application of social network analysis in social science disciplines. In the management and

business research, researchers have used network analysis in project management, supply chain management, and strategy research. Chinowski et al. (2010) analyze survey data collected from full-service engineering companies, that exhibit a high level of collaboration, to understand the characteristics of high-performance teams. To accomplish this goal, the authors obtain density and centrality measures (see APPENDIX A for definitions) for discipline managers that are tasked with the implementation of the strategic vision of the company. Based on calculated measures, they identify the level of interaction within each team and the importance each discipline manager to lead and communicate in high-performance environments. The study demonstrates that networks analysis can effectively be used to examine the role of trust, communication, and knowledge transfer in enhancing project teams' performance.

In the financial services industry, Minoiu and Reyes (2013) perform a network analysis of geographical linkages in the global banking system. Participants (or actors) in this study include 184 countries that were responsible for cross-border banking and capital flows in the period of 1978-2010. The authors used such measures as network density, node centrality, and connectedness to describe the structural properties and dynamics of the global banking network before and after the global financial crisis of 2008. In addition, the analysis ranks the participants (countries) based on their importance to the global network. This helped monitor the relative position (or stability) of leaders (important nodes) around financial crises. As a major direction for future research, Minoiu and Reyes (2013) indicate the need for studies on network typology that cover more asset classes and indicators.

Data and Methodology

Data

This study uses data from Finextra - a prominent and independent source of information for the global FinTech community (Finextra Research, 2022). It provides news, press releases, and research related to financial technology through its website finextra.com.

As a major financial technology newswire, Finextra has been providing news reports covering FinTech and other firms around the world since 2003. Each news article includes details regarding related firms, news release date, channels, and keywords, and date of news release.

APPENDIX B provides a sample news article from Finextra. The number of related firms, which are organizations mentioned in the article, ranges from one to several.

This study uses FinTech news articles, published between 2008 and 2022. The financial crisis of 2008 marks the start of FinTech 3.0 that saw an unprecedented growth in the use of technology such as smartphones, applications, and application programming interfaces (APIs) in financial services (Giglio, 2021; Paul and Sadath, 2021). 23,266 FinTech news articles have been published during the study period. The number of the news articles, organization mentions, unique organizations, unique channels, and unique keywords are provided in APPENDIX C.

Organizations mentioned in the FinTech news articles include payment processing companies such as Visa and Mastercard, commercial banks such as Barclays and Citi, financial regulators such as Financial Conduct Authority (FCA) and European Central Bank (ECB), FinTech startups such as Monzo and Revolut, and other entities. APPENDIX D lists top 50 organizations by the number of mentions in the FinTech news throughout the study period. Channels in the

FinTech news refer to the nature of business conducted by the mentioned organizations (e.g., wholesale banking, retail banking, cryptocurrency). Keywords provide more specific information regarding the main topic of the FinTech news articles (e.g., cards, trade execution, blockchain, artificial intelligence). We perform a clustering analysis¹ based on the channels and keywords included in the FinTech news (see APPENDIX E). This segmentation of FinTech organizations provides high-level descriptions of FinTech groups for the 15-year study period. Almost half of the news articles are in Segment 1 (49%) which can be characterized by such channels and keywords as wholesale banking, regulation & compliance, trade execution, dealing rooms, mergers & acquisitions, blockchain. Segment 2 (28%) and Segment 3 (23%) are similar in that the main channel in both is retail banking. However, the focus in Segment 2 is start-ups, mobile & online, artificial intelligence, open APIs, whereas the main keywords in Segment 3 are payments, security, mobile & online, cards, start-ups, and mobile.

Methodology

The methodology entails the investigation of the FinTech network and the evolution of its properties in different periods using network science. We start by constructing the FinTech network based on the adjacency matrix of organizations mentioned in the FinTech news for each year in the period from 2008-2022. Then, we explore the features and shape of the FinTech network for various periods. These network attributes include nodes (organizations), edges (a connection between organizations), isolates (organization not connected to the network), and degree (average number of connections per organization) of the network. APPENDIX A provides the definitions of measures used in the study. Next, we examine some of the most connected organizations in the network of FinTech firms to provide insights on connections

¹ k-means clustering using principal components identified from channels and keywords in new articles

among organizations. Finally, in the analysis of FinTech network properties we will show the most important players based on their various centrality or popularity measures.

Results

FinTech Network from 2008 to 2022: Overall Properties and Isolates

TABLE 1 provides information on the counts of nodes (or organizations), edges (or connections), and isolates over time. Overall, there is an increasing trend in each of the three metrics. This is an indication of a growing number of FinTech organizations and connections.

The table also presents the percentage of isolates and the network degree in the study period.

[Insert TABLE 1 about here]

Figure 1 (below), which graphically represents the information in TABLE 1, shows that the proportion of unconnected organizations has been increasing. This is one of the factors that has led to lower levels of connectedness within the network. While some isolate organizations are disappearing from the FinTech network, others are becoming more entrenched in the network of FinTech firms. For example, there were 185 isolate organizations in total in the three years from 2010 to 2012. Ten years after, in the three-year period 2020-2022, 100 (54.1%) of the isolate organizations disappeared from the network and 85 (45.9%) remained in the network. Of these remaining isolate organizations, none stayed as isolates, 31 (36.5%) had one connection, and 54 (63.5%) had two or more connections. Most of the deeply connected 54 organizations, that were

not previously connected with other organizations in the network, were start-up firms² founded from the mid-2000s to early 2010s.

[Insert FIGURE 1 about here]

Most Connected Organizations and Their Similarities

We analyze main players and connections between them in the network of FinTech organizations. FIGURE 2 visualizes how organizations with more than 20 connections are linked together. The more the connections an organization has, the more central its position in the network. The width of each link is determined by the strength of connection between two organizations connected by the link.

[Insert FIGURE 2 about here]

Panel A and B of Figure 2 present a pictorial representation of the FinTech network in 2010 and 2020, respectively. Both networks are dominated by large incumbent financial institutions - commercial banks, wholesale banks, and payment processors. In the networks, one can observe that Citi, Visa, MasterCard, Deutsche Bank, Barclays Bank, JPMorgan Chase, HSBC, Goldman Sachs, UBS, and others are central figures in each network. One can also see the presence of European financial regulators such as the Financial Conduct Authority (FCA), in both 2010 and 2022, and the European Central Bank (ECB), in 2022. However, one of the distinguishing features between the two networks is the involvement of tech giants such as Google, Facebook,

² Examples include Adyen, eToro, GoCardless, Jumio, Nutmeg, Payoneer, SumUp, Wise, Zopa.

Amazon, and Microsoft in the network of FinTech firms in 2022. Consistent with Boot et al. (2021), this is indicative of the increasing involvement of large information technology (IT) companies in FinTech, and their alliance with the established players in financial services.

Connections between organizations that are in the same nature of business seem to be strong, although with some caveats. For example, the width of the link between Visa and MasterCard is the strongest in the 2010 network, although this connection weakened noticeable in 2022. Panel B of FIGURE 2 also highlights strong connections within the commercial banks and the wholesale banks in the two networks, as it can be observed the thicker links between this type of organizations.

Next, we examine organization's overall connections (e.g., Visa has - footnote) to quantify the similarity between some of the main players of the FinTech network. We also observe how these similarities have changed over time. TABLE 2 presents Jaccard similarity scores between the 10 most connected organizations in 2010 and 2022.

[Insert TABLE 2 about here]

Jaccard similarity scores in the bottom left part of TABLE 2 shows how similar each pair of organizations in terms of their own networks in 2010. All values are color coded – i.e., the higher the similarity/Jaccard score the higher the color intensity. Some of the most similar pairs are HSBC & Barclays Bank (0.37), Visa & MasterCard (0.31), JPMorgan Chase & Swift (0.31), Citi & Goldman Sachs (0.31), and JPMorgan Chase and Barclays Bank (0.22). The top right part of the figure displays similarity scores in 2022. Here the most similar pairs are HSBC and NatWest (0.25) and Citi and Goldman Sachs (0.21). You can observe that the similarity values are lower compared to the earlier period. In fact, the average similarity score in 2010 is 0.14, whereas it is

0.10 in 2022.

Central Organizations in the Network of FinTech: Centrality Measures

In this section we examine node centrality measures to analyze how important an organization is within the network of FinTech organizations. Four panels in APPENDIX F present top 30 organizations based on their degree, eigenvector, betweenness, and closeness centrality measures for the years 2008, 2015, and 2022. These seven-years intervals are chosen to observe the popularity, importance, and influence of organizations over time.

Based on degree centrality (APPENDIX F1), most connected organizations are incumbent financial institutions such as MasterCard, Citi, Visa, JPMorgan Chase, Barclays Bank, HSBC, etc. APPENDIX F2 shows most influential organizations as measured by eigenvalue centrality. Citi is on the top of the list in all three periods, while most of the list is comprised of established incumbent financial institutions in each period. Unlike the lists for degree centrality, the lists for eigenvalue centrality show fewer start-up firms. This is somewhat expected as it might be less likely for relatively newer firms to create connections with the core nodes of the network.

APPENDIX F3 ranks organizations by closeness centrality which quantifies the closeness of an organization to many other organizations in the network. Similar to the previous centrality lists, this list also is dominated by the incumbent payment processors and banks, with a few exceptions. Technology firms Google and Amazon are among the top 10 organizations in 2022. This implies that an increasing number of FinTech firms are collaborating with IT firms. This is also true for betweenness centrality (APPENDIX F4) which measures the number of times an organization is between two other organizations - technology giants Microsoft and Google among the most central organizations in 2022. Moreover, start-up FinTech firms are more

represented in APPENDIX F4, especially in 2022. This points to the important influence of firms like Klarna, Monzo, Starling Bank, Coinbase, Revolut in the flow of information in the network.

A few trends can be easily observed from the provided centrality measures. For example, stock exchanges such as Nasdaq, NYSE, London Stock Exchange have become less central over time. Instead, cryptocurrency exchanges such as Coinbase and Binance appear in the network of FinTech organizations in 2022. Another trend is related to technology giants such as Microsoft, Google, Facebook, and Amazon, which have become some of the most central figures. Similarly, the role of supervisory authorities such as the Financial Conduct Authority (FCA) and the Bank for International Settlements (BIS) has risen in the network.

Discussion and Concluding Remarks

As demonstrated in the previous section, our study defines the FinTech network as a dynamic group of connected organizations that collaborate to advance their technological and innovative capabilities. The co-occurrence of organizations in the FinTech news articles explains the strength of connections between organizations in this network. Overall, the last 15 years have seen expanding participation in the network from a variety of organizations. However, the level of connectedness between the organizations has been declining, while the number of organizations with no connections (isolates) has been growing. This is primarily attributed to new members joining the network as the isolates of the prior years become progressively more integrated into the network over years. Our findings also indicate that organization in the same nature of business not only rely on or cooperate with each other more intensely but also have similar collaboration patterns. However, the degree of cooperation among traditional players has subsided over time, while their engagement in partnerships with other types of organizations has

intensified. This is consistent with the view that incumbents are benefiting from improving regulatory conditions (Daud et al., 2022; Sun et al., 2022) while compensating for being late in the game of financial inclusion and digitization by forming alliances with organizations outside their traditional networks (Brandl & Hornuf, 2020; Feyen et al., 2021).

Although FinTech firms pose a threat to incumbents' market share (Nienaber, 2016; Goldstein, 2019), they also induce incumbents to be more involved in the FinTech network (Klus et al., 2019; Feyen et al., 2021; Murinde et al., 2022). Greater involvement of incumbent financial intermediaries in the FinTech network promotes their operational efficiency (Hendershott et al., 2021) and financial stability (Fung et al., 2020; Daud et al., 2023). Therefore, traditional financial service providers, who constantly seek collaborations opportunities, are entrenched in the FinTech network. Besides cooperating among themselves, they are establishing strong ties with large software and technology firms. This is in line with the growing application of big tech solutions (e.g., APIs, ecosystems, and cloud computing) in financial services as incumbents shift to online environment (Huang, 2018, Boot et al., 2021). Another major group of players in the FinTech network are FinTech start-ups which are leveraging incumbents' reputation and resources (Klus et al., 2019) in exchange for their technological disruptions and innovative solutions (Chemmanur et al., 2020).

Our network centrality analysis highlights the increased prominence of regulatory agencies and cryptocurrency exchanges. This coincides with the emergence of digital currencies and crypto assets. This domain of FinTech have attracted the attention of policymakers who are responsible for the creation of new policies concerning the activities of cryptocurrency exchanges. In more general terms, the higher importance of regulatory bodies in the network confirms the increasing role of regulation which creates challenges and opportunities for FinTech (Sun et al., 2022).

Incumbents' profound involvement in FinTech can also be observed in the payment processing field. Visa and Mastercard are among the most connected organizations, while their collaboration with each other one of the strong. They also are top two companies linking other organizations based on betweenness centrality, indicating a great level of involvement in FinTech developments.

In this paper, we offered a novel way of defining and examining the FinTech network over time. Our findings add to the extant literatures on business network analysis and collaboration in FinTech. Besides, our insights can benefit many types organizations, regardless of their participation level in the FinTech network, in crafting collaboration strategies. However, despite a broad coverage of FinTech news internationally, our dataset of news articles is subject to the constraints of Finextra Research. Future research may address this limitation by aggregating FinTech news from multiple sources and reliably removing redundant news publications. Another possible limitation of the study is its selective approach to examining collaboration patterns between FinTech participants. Future studies can extend our analysis by performing the investigation of connections and collaboration elements in the FinTech network with more breadth.

References

- Batt, P.J. and Purchase, S. (2004), “Managing collaboration within networks and relationships”, *Industrial Marketing Management*, Vol. 33, No. 3, pp. 169–174.
- Boot, A., Hoffmann, P., Laeven, L., & Ratnovski, L. (2021). Fintech: what’s old, what’s new?. *Journal of Financial Stability*, 53, 100836.
- Brandl, B., & Hornuf, L. (2020). Where did FinTechs come from, and where do they go? The transformation of the financial industry in Germany after digitalization. *Frontiers in Artificial Intelligence*, 3, 8.
- Chemmanur, T. J., Imerman, M. B., Rajaiya, H., & Yu, Q. (2020). Recent developments in the fintech industry. *Journal of Financial Management, Markets and Institutions*, 8(01), 2040002.
- Cheng, M., & Qu, Y. (2020). Does bank FinTech reduce credit risk? Evidence from China. *Pacific-Basin Finance Journal*, 63, 101398.
- Chinowsky, P. S., Diekmann, J., & O’Brien, J. (2010). Project organizations as social networks. *Journal of construction engineering and management*, 136(4), 452-458.
- Daud, S. N. M., Khalid, A., & Azman-Saini, W. N. W. (2022). FinTech and financial stability: Threat or opportunity?. *Finance Research Letters*, 47, 102667.
- Feyen, E., Frost, J., Gambacorta, L., Natarajan, H., & Saal, M. (2021). Fintech and the digital transformation of financial services: implications for market structure and public policy. *BIS Papers*.
- Fu, J., & Mishra, M. (2022). Fintech in the time of COVID– 19: Technological adoption during crises. *Journal of Financial Intermediation*, 50, 100945.
- Fung, D. W., Lee, W. Y., Yeh, J. J., & Yuen, F. L. (2020). Friend or foe: The divergent effects of FinTech on financial stability. *Emerging Markets Review*, 45, 100727.
- Giglio, F. (2021). Fintech: A literature review. *European Research Studies Journal*, 24(2B), 600627.
- Goldstein, I., Jiang, W., & Karolyi, G. A. (2019). To FinTech and beyond. *The Review of Financial Studies*, 32(5), 1647-1661.
- Hendershott, T., Zhang, X., Zhao, J. L., & Zheng, Z. (2021). FinTech as a game changer: Overview of research frontiers. *Information Systems Research*, 32(1), 1-17.
- Hendrikse, R., Bassens, D., & Van Meeteren, M. (2018). The Appleization of finance: Charting incumbent finance’s embrace of FinTech. *Finance and Society*, 4(2), 159-80.

- Hornuf, L., Klus, M. F., Lohwasser, T. S., & Schwienbacher, A. (2021). How do banks interact with fintech startups?. *Small Business Economics*, 57, 1505-1526.
- Huang, R. H. (2018). Online P2P lending and regulatory responses in China: Opportunities and challenges. *European Business Organization Law Review*, 19, 63-92.
- Index, G. F. A. (2019). Ernst & Young, 2019. 44 p. URL: <https://www.EY-Global-FinTechAdoption-Index-2019.pdf>.
- Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 137-146): ACM.
- Klus, M. F., Lohwasser, T. S., Holotiuk, F., & Moormann, J. (2019). Strategic alliances between banks and fintechs for digital innovation: Motives to collaborate and types of interaction. *The Journal of Entrepreneurial Finance (JEF)*, 21(1), 1-23.
- Kohtamäki, M., Parida, V., Oghazi, P., Gebauer, H., & Baines, T. (2019). Digital servitization business models in ecosystems: A theory of the firm. *Journal of Business Research*, 104, 380-392.
- KPMG Survey 2022 - <https://info.kpmg.us/news-perspectives/industry-insights-research/kpmg2022-banking-survey.html>
- Lea, B. R., Yu, W. B., Maguluru, N., & Nichols, M. (2006). Enhancing business networks using social network based virtual communities. *Industrial Management & Data Systems*, 106(1), 121-138.
- Mention, A. L. (2019). The future of fintech. *Research-Technology Management*, 62(4), 59-63.
- Minoiu, C., & Reyes, J. A. (2013). A network analysis of global banking: 1978–2010. *Journal of Financial Stability*, 9(2), 168-184.
- Murinde, V., Rizopoulos, E., & Zachariadis, M. (2022). The impact of the FinTech revolution on the future of banking: Opportunities and risks. *International Review of Financial Analysis*, 81, 102103.
- Najaf, K., Mostafiz, M. I., & Najaf, R. (2021). Fintech firms and banks sustainability: why cybersecurity risk matters?. *International Journal of Financial Engineering*, 8(02), 2150019.
- Nienaber, R. (2016). Banks need to think collaboration rather than competition. *The FinTech book: The financial technology handbook for investors, entrepreneurs and visionaries*, 20-21.

- Paul, L. R., & Sadath, L. (2021, February). A systematic analysis on fintech and its applications. In *2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)* (pp. 131-136). IEEE.
- Puschmann, T. Fintech. *Bus Inf Syst Eng* 59, 69–76 (2017). <https://doi.org/10.1007/s12599-0170464-6>
- Salampasis, D., & Mention, A. L. (2018). FinTech: Harnessing innovation for financial inclusion. In *Handbook of Blockchain, Digital Finance, and Inclusion, Volume 2* (pp. 451-461). Academic Press.
- Sun, Y., Li, S., & Wang, R. (2022). Fintech: From budding to explosion-an overview of the current state of research. *Review of Managerial Science*, 17(3), 715-755.
- Vendrell-Herrero, F., Bustinza, O. F., Parry, G., & Georgantzis, N. (2017). Servitization, digitization and supply chain interdependency. *Industrial Marketing Management*, 60, 69-81.
- Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.
- Wellman, B. (1983). Network analysis: Some basic principles. *Sociological theory*, 155-200.

FIGURES

FIGURE 1: Network Degree over Time

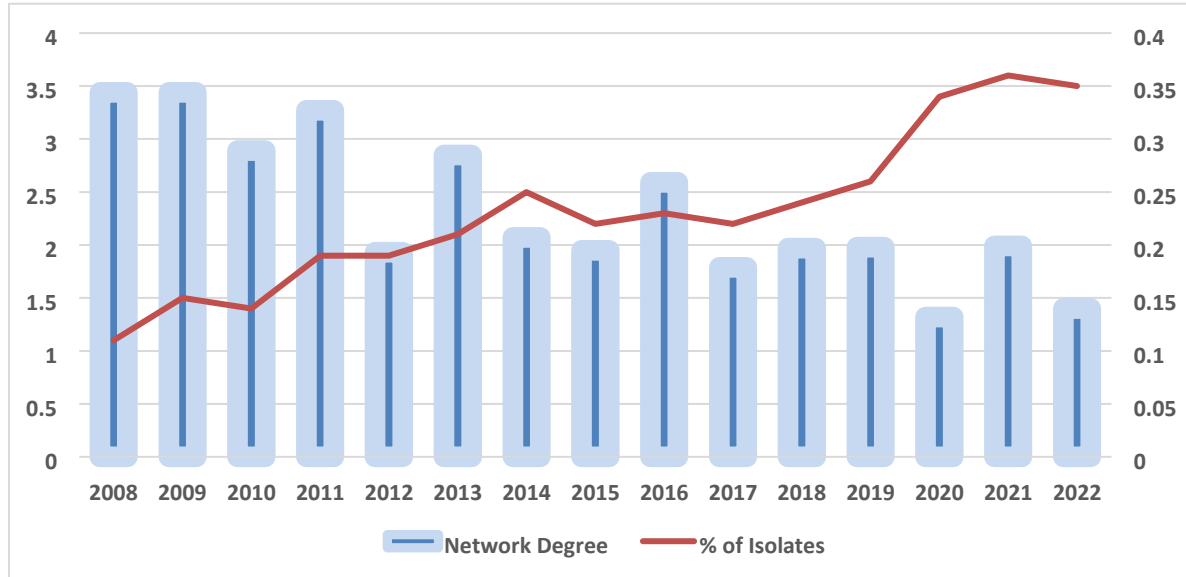
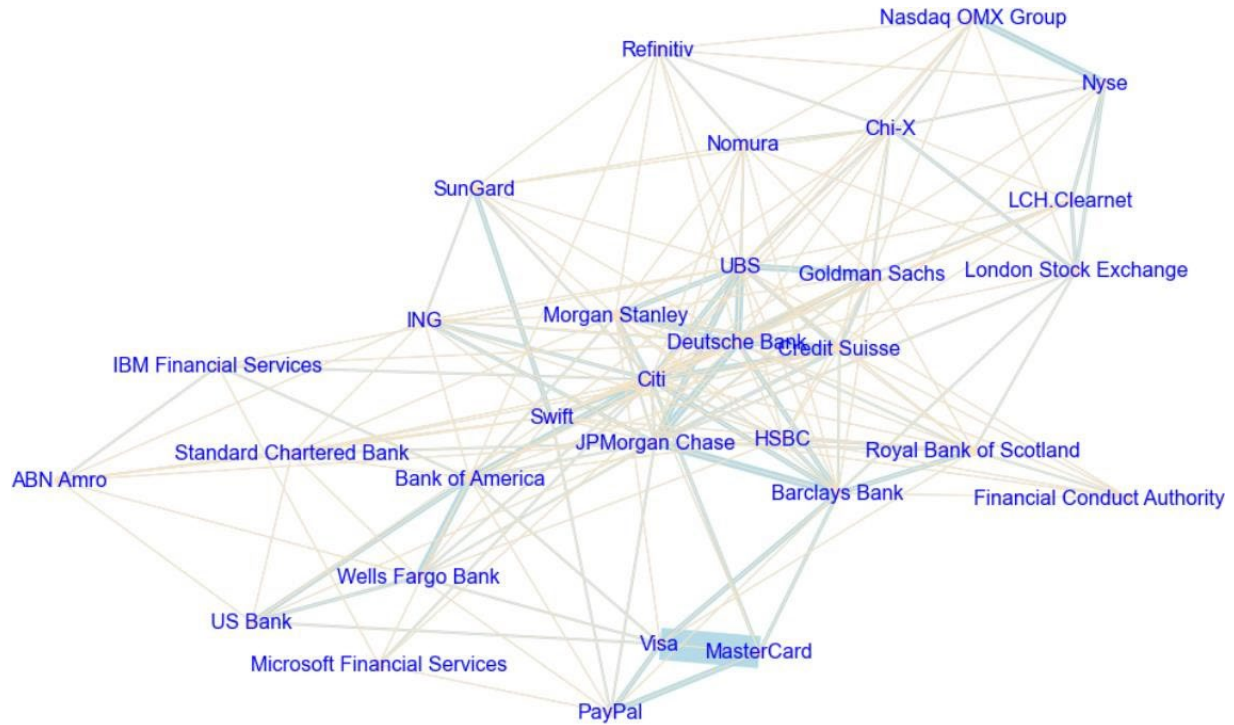
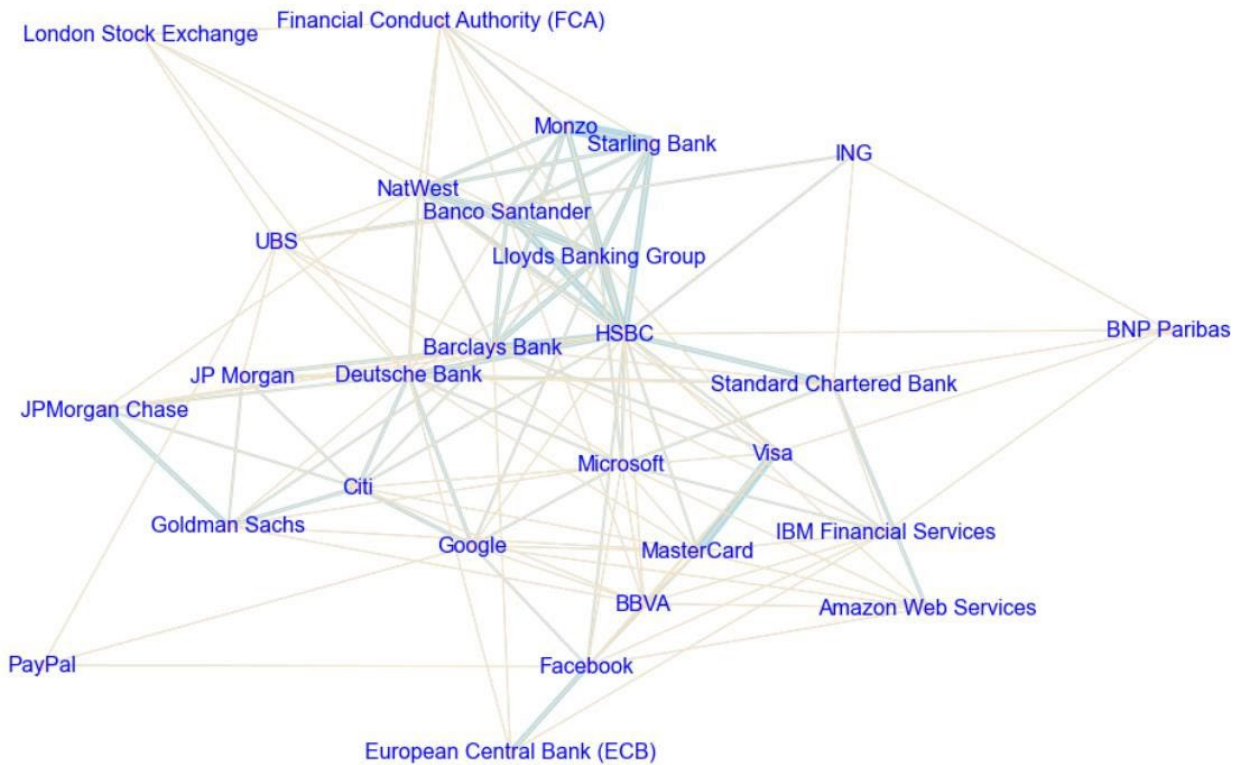


FIGURE 2: The Network of Most Connected Organizations in 2010 versus 2022 Panel

A: The Network in 2010



Panel B: The Network in 2020



TABLES

TABLE 1: FinTech Network Characteristics (2008-2022)

Year	# of Nodes	# of Edges	# of Isolates	% of Isolates	Network Degree
2008	568	1,952	65	0.11	3.44
2009	496	1,704	72	0.15	3.44
2010	449	1,297	65	0.14	2.89
2011	380	1,244	71	0.19	3.27
2012	376	725	73	0.19	1.93
2013	402	1,145	84	0.21	2.85
2014	364	754	90	0.25	2.07
2015	370	720	82	0.22	1.95
2016	396	1,024	92	0.23	2.59
2017	486	872	108	0.22	1.79
2018	702	1,380	170	0.24	1.97
2019	858	1,702	220	0.26	1.98
2020	1,076	1,422	369	0.34	1.32
2021	1,219	2,423	436	0.36	1.99

2022	1,056	1,477	367	0.35	1.4
------	-------	-------	-----	------	-----

TABLE 2: Jaccard Similarity Heatmap 2010 (bottom left) versus 2022 (top right)

	Mastercard	Citi	Visa	JPMorgan Chase	Barclays Bank	HSBC	NatWest	FCA	Swift	Goldman Sachs
Mastercard	1.00	0.10	0.14	0.04	0.09	0.17	0.04	0.05	0.09	0.09
Citi	0.05	1.00	0.07	0.16	0.14	0.14	0.09	0.03	0.08	0.21
Visa	0.31	0.08	1.00	0.03	0.05	0.11	0.04	0.07	0.04	0.07
JPMorgan Chase	0.05	0.20	0.09	1.00	0.14	0.08	0.08	0.05	0.11	0.16
Barclays Bank	0.13	0.15	0.12	0.22	1.00	0.08	0.08	0.05	0.11	0.16
HSBC	0.07	0.17	0.10	0.22	0.37	1.00	0.25	0.13	0.08	0.10
NatWest	0.03	0.07	0.04	0.03	0.21	0.22	1.00	0.19	0.07	0.07
FCA	0.02	0.16	0.05	0.13	0.16	0.20	0.12	1.00	0.09	0.06
Swift	0.04	0.22	0.07	0.31	0.12	0.17	0.00	0.06	1.00	0.05
Goldman Sachs	0.05	0.31	0.07	0.19	0.17	0.19	0.05	0.19	0.16	1.00

APPENDIX

APPENDIX A: Measures and Definitions

Variable/Measure	Definition
<i>Adjacency Matrix</i>	A square matrix representing a FinTech network into a mathematical form which shows any connections between all organizations in the network.
<i>Centrality</i>	An importance measure of an organization in a FinTech network. It is a popularity measure.

<i>Degree Centrality</i>	A centrality measure that shows the number of connections to the organization. The higher the degree, more central the organization is.
<i>Eigenvector Centrality</i>	A centrality measure that considers the value of its immediate neighbors to show how influential an organization is in the FinTech network.
<i>Betweenness Centrality</i>	A centrality measure based on the number of times an organization is between two other organizations.
<i>Closeness Centrality</i>	A centrality measure that quantifies the closeness of an organization to many other organizations in the Fintech network.
<i>Node</i>	An organization mentioned in a FinTech news article. It is an object of interest in a FinTech network.
<i>Edge</i>	A connection between two organizations based on their co-occurrence a FinTech news article.
<i>Isolate</i>	An organization which is not connected with other organizations in a FinTech network.
<i>Network Degree</i>	The average number of connections per organization in a FinTech network.
<i>Jaccard Similarity</i>	A similarity measure between connected components of two organizations.

B: Sample News Article Published on Finextra

Related Companies

MasterCard >

Coinbase >

Lead Channel

Payments >

Channels

Cryptocurrency >

Retail Banking >

Keywords

Cards >

Editorial | what does this mean?

Mastercard and Coinbase partner to let people buy NFTs with cards

18 January 2022



7



7



0



Mastercard is joining forces with Coinbase to enable people to buy NFTs with their debit and credit cards.

Coinbase recently joined the NFT party, outlining plans to build a decentralised marketplace where users can mint, collect and trade non-fungible tokens.

However, the firm acknowledges that the process of buying an NFT can be daunting for novices, requiring them to open a crypto wallet, buy crypto, then use it to purchase the NFT in an online marketplace.

The firm is working with Mastercard to classify NFTs as "digital goods" and unlock a new way to pay using cards.

Raj Dhamodharan, EVP, digital asset and blockchain products and partnerships, Mastercard, says: "We're working to make NFTs more accessible because we believe tech should be inclusive.

"When more people are included in new technologies, it spurs innovation, helps economies grow and expands choices for consumers."

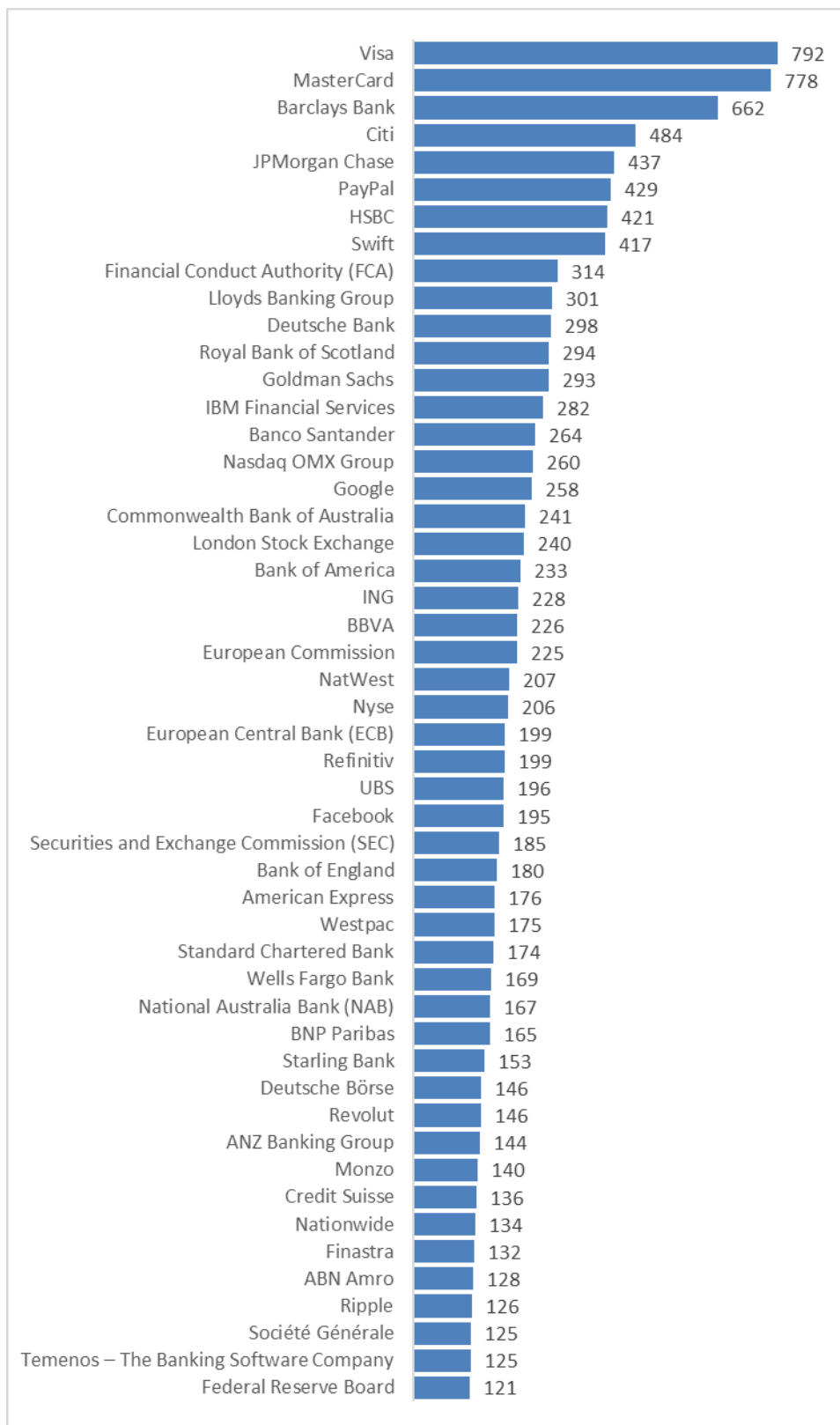
Source: Finextra Research, 2022

APPENDIX C: Details of News Articles by Year (2008-2022)

Year	News articles	Organization mentions	Unique organizations	Unique channels	Unique keywords
2008	1,502	2,561	568	9	70
2009	1,391	2,172	496	9	66
2010	1,208	1,826	449	9	66
2011	1,139	1,713	380	10	66
2012	1,112	1,497	376	11	62
2013	1,156	1,451	402	11	70
2014	1,252	1,346	364	11	63
2015	1,409	1,476	370	13	64
2016	1,634	1,847	396	14	69
2017	1,537	1,867	486	15	72
2018	1,635	2,264	702	15	74
2019	1,826	2,597	858	17	83
2020	2,175	3,016	1,076	19	85
2021	2,222	3,230	1,219	19	93
2022	2,068	2,810	1,056	19	100
Total	23,266	31,673	3,328*	19*	115*

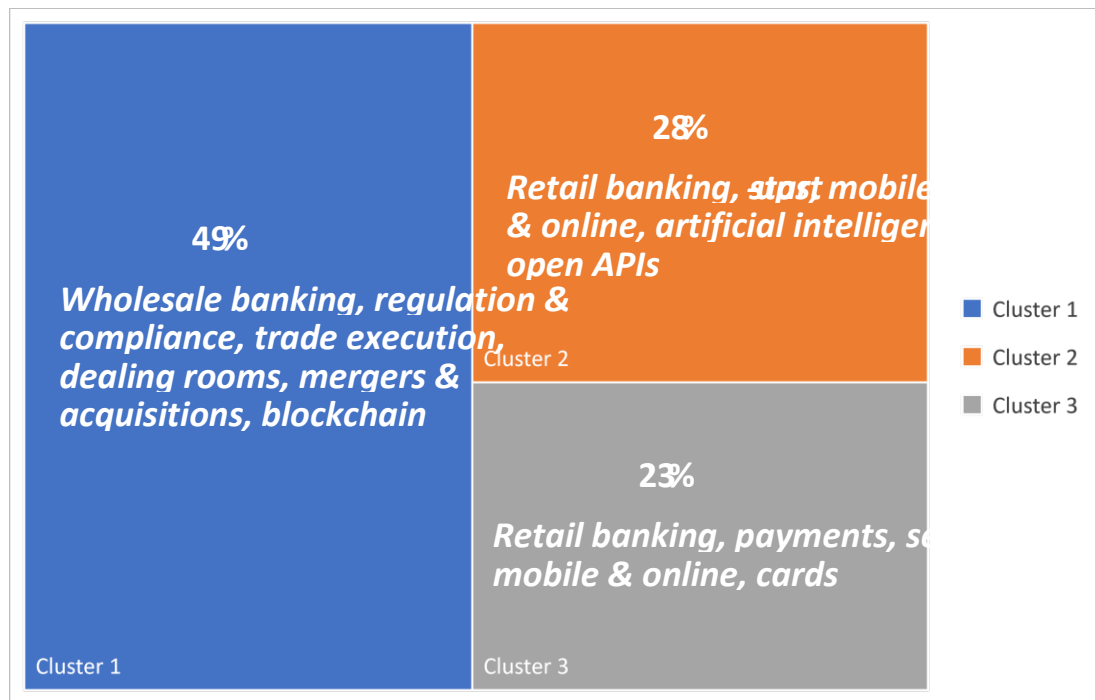
Notes: *The totals represent unique sums for unique organizations, unique channels, and unique keywords.

D: 50 Most Commonly Mentioned Organizations in News Articles (2008-2022)



APPENDIX

E: Clusters of Organizations based on News Channels and Keywords



F1: Top 30 Organizations by Degree Centrality in 2008, 2015, and 2022

	2008	2015	2022
1	Citi	Visa	Mastercard
2	HSBC	Citi	Citi
3	Barclays Bank	Deutsche Bank	Visa
4	MasterCard	JPMorgan Chase	JPMorgan Chase
5	Deutsche Bank	Barclays Bank	Barclays Bank
6	NYSE	MasterCard	HSBC
7	Credit Suisse	HSBC	NatWest
8	JPMorgan Chase	UBS	FCA
9	Nasdaq OMX Group	Bank of America	Swift
10	Visa	Bank of New York Mellon	Goldman Sachs

11	Refinitiv	BBVA	Lloyds Banking Group
12	Royal Bank of Scotland	Nasdaq OMX Group	BNP Paribas
13	LCH.Clearnet	PayPal	Coinbase
14	London Stock Exchange	Accenture Financial Services	Google
15	Bank of America Merrill Lynch	Credit Suisse	Klarna
16	Bank of America	Royal Bank of Scotland	UBS
17	IHS Markit	American Express	Bank of America
18	Temenos	Goldman Sachs	BNY Mellon
19	Goldman Sachs	Lloyds Banking Group	BIS
20	Lehman Brothers	Toronto Dominion Bank	Wells Fargo
21	BNP Paribas	Morgan Stanley	Gemini
22	Lloyds Banking Group	US Bank	Microsoft
23	FIS Global	Fiserv, Inc.	Robinhood
24	UBS	Royal Bank of Canada (RBC)	US Bank
25	ACI Worldwide	Banco Santander	Amazon Web Services
26	Morgan Stanley	Refinitiv	Deutsche Bank
27	Chi-X	Commerzbank	Monzo
28	Instinet	ING	Starling Bank
29	Swift	Swift	Zero Hash
30	Turquoise	CME Group	NAB

Notes:

FCA – Financial Conduct Authority; BIS – Bank for International Settlement; NAB – National Australia Bank

F2: Top 30 Organizations by Eigenvector Centrality in 2008, 2015, and 2022

	2008	2015	2022
1	Citi	Citi	Citi
2	HSBC	Deutsche Bank	Mastercard
3	Deutsche Bank	Bank of America	JPMorgan Chase
4	Barclays Bank	Bank of New York Mellon	Wells Fargo
5	Credit Suisse	JPMorgan Chase	BNY Mellon
6	JPMorgan Chase	HSBC	US Bank
7	Royal Bank of Scotland	Barclays Bank	PNC Financial Services Group

8	Goldman Sachs	Morgan Stanley	Truist
9	Bank of America Merrill Lynch	UBS	HSBC
10	IHS Markit	Visa	Swift
11	UBS	Toronto Dominion Bank	Amazon Web Services
12	Bank of America	Goldman Sachs	Federal Reserve Board
13	Morgan Stanley	Société Générale	SETL
14	Lehman Brothers	Royal Bank of Canada (RBC)	Digital Asset Holdings
15	BNP Paribas	Credit Suisse	Barclays Bank
16	NYSE	Royal Bank of Scotland	TD Bank
17	London Stock Exchange	NAB	Bank of America
18	Refinitiv	Commerzbank	Goldman Sachs
19	Chi-X	SEB	NatWest
20	Instinet	MasterCard	Visa
21	MasterCard	Refinitiv	Capital One
22	Visa	US Bank	Lloyds Banking Group
23	LCH.Clearnet	Bloomberg	Google
24	Société Générale	American Express	Morgan Stanley
25	CME Group	PayPal	BNP Paribas
26	Turquoise	BlackRock	UBS
27	Lloyds Banking Group	Standard Chartered Bank	Deutsche Bank
28	ABN Amro	Nasdaq OMX Group	Standard Chartered Bank
29	Tradeweb	BNP Paribas	Early Warning Services
30	HBOS	BBVA	FCA

Notes:

NAB – National Australia Bank; FCA – Financial Conduct Authority

APPENDIX F3: Top 30 Organizations by Closeness Centrality in 2008, 2015, and 2022

	2008	2015	2022
1	Citi	Deutsche Bank	Mastercard
2	Deutsche Bank	Barclays Bank	Citi
3	HSBC	Citi	JPMorgan Chase
4	Barclays Bank	UBS	Visa
5	JPMorgan Chase	JPMorgan Chase	Barclays Bank
6	Credit Suisse	Visa	NatWest
7	Royal Bank of Scotland	HSBC	Google
8	Refinitiv	Nasdaq OMX Group	Lloyds Banking Group
9	Bank of America	Royal Bank of Scotland	FCA
10	IHS Markit	Bank of America	Amazon Web Services
11	MasterCard	BBVA	Goldman Sachs
12	Nyse	Bank of New York Mellon	Klarna
13	BNP Paribas	Lloyds Banking Group	BNP Paribas
14	Bank of America Merrill Lynch	Credit Suisse	HSBC
15	Morgan Stanley	BNP Paribas	Swift
16	UBS	Morgan Stanley	Wells Fargo
17	Lehman Brothers	MasterCard	Payment Systems Regulator
18	LCH.Clearnet	Standard Chartered Bank	Standard Chartered Bank
19	Goldman Sachs	PayPal	US Bank
20	London Stock Exchange	Goldman Sachs	Deutsche Bank
21	Visa	Banco Santander	BNY Mellon
22	CME Group	Refinitiv	UBS
23	Société Générale	Accenture Financial Services	Morgan Stanley
24	ING	Fiserv, Inc.	NAB
25	SAP	Rabobank	Starling Bank
26	Nasdaq OMX Group	Swift	MasterCard
27	Swift	CME Group	PNC Financial Services Group
28	Chi-X	Capital One	Truist
29	Instinet	American Express	Capital One

30	IBM Financial Services	Nationwide	CMA
----	------------------------	------------	-----

FCA – Financial Conduct Authority; NAB – National Australia Bank

APPENDIX F4: Top 30 Organizations by Betweenness Centrality in 2008, 2015, and 2022

	2008	2015	2022
1	Citi	Visa	Mastercard
2	MasterCard	Deutsche Bank	Visa
3	HSBC	Barclays Bank	Citi
4	Nasdaq OMX Group	Citi	JPMorgan Chase
5	Refinitiv	JPMorgan Chase	Barclays Bank
6	Nyse	UBS	NatWest
7	Deutsche Bank	Nasdaq OMX Group	Lloyds Banking Group
8	Visa	Swift	Microsoft
9	Barclays Bank	MasterCard	Swift
10	IBM Financial Services	Accenture Financial Services	Google
11	Credit Suisse	Banco Santander	HSBC
12	JPMorgan Chase	BBVA	FCA
13	Temenos	HSBC	Klarna
14	FIS Global	PayPal	BNP Paribas
15	SAP	Euroclear	UBS
16	IHS Markit	ECB	NAB
17	BNP Paribas	ING	Goldman Sachs
18	LCH.Clearnet	Fiserv, Inc.	BIS
19	Swift	Ripple	Deutsche Bank
20	Bank of America	ANZ Banking Group	Monzo
21	Bank of America Merrill Lynch	Lloyds Banking Group	MasterCard
22	PayPal	Royal Bank of Scotland	Money20/20
23	London Stock Exchange	Six Group AG	Starling Bank
24	Microsoft Financial Services	EBA Clearing	Coinbase
25	Lehman Brothers	CME Group	Amazon
26	CME Group	Google	Nexi

27	Société Générale	IHS Markit	FTX
28	FCA	Bank of America	ING
29	Royal Bank of Scotland	Refinitiv	Binance
30	Nordea	Goldman Sachs	Revolut

ECB - European Central Bank; FCA – Financial Conduct Authority; BIS – Bank for International Settlement; NAB – National Australia Bank

Optimal Sustainable Pest and Soil Management for Farmers: A Dynamic Bioeconomic Modeling Framework^{1,2}

Michael A. Meneses,
Clare L. Casteel, Miguel I. Gómez, David R. Just,
Ravi Kanbur, David R. Lee, C.-Y. Cynthia Lin Lawell

Abstract

Organic farming -- wherein farmers do not use synthetic fertilizers, pesticides, herbicides, or fungicides to grow their produce -- is widely considered to be a far more sustainable alternative to conventional food production. In this paper, we review, synthesize, and discuss the economics literature on organic farming; review our research that combines insights from economics and the natural sciences to study and inform farmer transitions from conventional to organic management; and present a framework for dynamic bioeconomic modeling of a farmer's pest and soil management decisions. Our research and framework aims to help farmers improve decisionmaking around synthetic compound use and organic production, with the ultimate goal of improving soil bacteria stewardship, crop yields, farmer profits, agricultural sustainability, greenhouse gas mitigation, biodiversity, resilience of the organic farming system, the protection of water and other resources, the provision of ecosystem services, and public and environmental health.

¹ Meneses: Cornell University; mam789@cornell.edu. Casteel: Cornell University; clc269@cornell.edu. Gómez: Cornell University; mig7@cornell.edu. Just: Cornell University; drj3@cornell.edu. Kanbur: Cornell University; sk145@cornell.edu. Lee: Cornell University; drl5@cornell.edu. Lin Lawell: Cornell University; clinlawell@cornell.edu.

² We thank Shadi Atallah, Apurva Borar, Nano Barahona, Peter Blair Henry, Jacob Bradt, Stephanie Brockmann, Vincenzina Caputo, Bingyan Dai, Will Davis, Eric Denemark, Avanti Dey, Victor Simoes Dornelas, Yurou He, John Hubbard, Jenny Ifft, Kelsey Jack, Ted Jaenicke, Kaleb Javier, Joseph Kuvshnikov, Tongzhe Li, Francesco Lissoni, Tong Liu, Trevon Logan, Khyati Malik, Conrad Miller, Rosina Moreno, José Nuño-Ledesma, Jess Rudder, Michelle Segovia, Michael Toman, Harold van Es, Joakim Weill, and Lawrence White for detailed and helpful

comments and discussions. We benefited from comments from participants at the 2024 International Business Analytics Conference (IBAC); the Ph.D. Excellence Initiative Summer Research Workshop at New York University; an Association of Environmental and Resource Economists (AERE) Session at the Eastern Economics Association (EEA) Annual Conference (AERE@EEA); the Heartland Environmental and Resource Economics Workshop; the American Economic Association (AEA) Summer Mentoring Pipeline Conference (SMPC); the Agricultural and Applied Economics Association (AAEA) Annual Meeting; the California Department of Pesticide Regulation (DPR) Pesticide Use Report (PUR) Analysis Workgroup; the Southern Economic Association (SEA) Annual Meeting; the Interdisciplinary PhD Workshop in Sustainable Development (IPWSD) at Columbia University; the Economics of Science and Innovation Summer School at the University of Barcelona; the Ph.D. Applied Macroeconomics guest lecture at Cornell University; and the Voices of Quebec Labrador Foundation panel ‘The Rise of Young Professionals’. We received funding from the U.S. Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), Cornell Graduate School Conference Travel Grants, Cornell TREESPEAR Research Grants, and the Cornell DEEP-GREEN-RADAR Khaled H. Kheiravar Memorial Scholarship. Gómez, Just, Kanbur, Lee, and Lin Lawell are Faculty Fellows at the Cornell Atkinson Center for Sustainability. All errors are our own.

1. Introduction

Conventional farming has been heavily criticized for causing biodiversity loss, soil erosion, and increased water pollution due to the rampant usage of synthetic fertilizers and pesticides. Organic farming -- wherein farmers do not use synthetic fertilizers, pesticides, herbicides, or fungicides to grow their produce -- is widely considered to be a far more sustainable alternative to conventional food production (Varanasi, 2019).

In this paper, we review, synthesize, and discuss the economics literature on organic farming; review our research in Meneses et al. (2024) that combines insights from economics and the natural sciences to study and inform farmer transitions from conventional to organic management; and present a framework for dynamic bioeconomic modeling of a farmer's pest and soil management decisions. Our research and framework aims to help farmers improve decision-making around synthetic compound use and organic production, with the ultimate goal of improving soil bacteria stewardship, crop yields, farmer profits, agricultural sustainability, greenhouse gas mitigation, biodiversity, resilience of the organic farming system, the protection of water and other resources, the provision of ecosystem services, and public and environmental health.

2. The Economics of Organic Farming

Transitioning to organic farming entails the discontinuation of pesticide use, a change that may impact farm profits. The relationship between pesticide use and farm profit has been the subject of many studies. Chambers et al. (2010) shows pesticide use as increasing returns to quasi-fixed factors of production like capital and land. In contrast, Jacquet et al. (2011) use a mathematical programming model to determine whether pesticide use can be reduced without affecting farmer income and find that a up to a 30 percent reduction is possible.

Multiple studies have applied the dynamic optimization and programming toolkits to the study of optimal agricultural management practices. Jaenicke (2000) develops a dynamic data envelopment analysis (DEA) model of crop production to investigate the role soil capital plays in observed productivity growth and the crop rotation effect. Yeh et al. (2024) develop a novel dynamic bioeconomic analysis framework that combines numerical dynamic optimization and dynamic structural econometric estimation, and apply it to analyze the optimal management strategy for Spotted Wing Drosophila, a pest affecting soft-skinned fruits. Wu (2000) develops a dynamic model and solves for the optimal time path for herbicide application. Dynamic models have also been developed to study agricultural productivity (Carroll et al., 2019), agricultural groundwater management (Sears et al., 2019, 2024a, 2024b, 2024c), agricultural disease control (Carroll et al., 2024a), pollination input decisions by apple farmers (Wilcox et al., 2024), supply chain externalities (Carroll et al., 2024b), optimal bamboo forest management (Wu et al., 2024), fisheries management (Conrad et al., 2024; Shin et al., 2024), and grapes (Sambucci et al., 2024).

Delbridge and King (2016) use dynamic programming to address the question of why so few farmers choose to transition to organic farming. They model the decision to transition to organic production as a dynamic programming problem where the transition involves sunk costs, and find the slow uptake of organic farming may be partially driven by the option value generated by the

sunk costs associated with organic transition. Other studies have sought to incorporate transition dynamics, such as the empirically documented initial decrease in crop yields associated with conventional to organic transitions, into profitability assessments of organic farming. Dabbert and Madden (1986) find in their multi-year simulation of a 117-hectare crop-livestock farm that the initial decrease in crop yields during an organic transition results in a 30 percentage point decrease in income in the first year of transition. The biological underpinnings of this initial decrease in productivity, and their response to farmer control variables are not made explicit.

3. Adding Insights from the Natural Sciences

Soil microbes benefit agricultural production by enhancing crop nutrient use, stress tolerance, and pest resistance (Lori et al., 2017). New insights from soil science show that the use of synthetic fertilizers and pesticides can be harmful to these beneficial soil microbes (Hussain et al., 2009; Lo, 2010; Kalia et al., 2011, Lori et al., 2017; Blundell et al., 2020). For example, Blundell et al. (2020) find that organic management is associated with decreased pest pressure on tomato plants. This effect is driven by an accumulation of salicylic acid in plant tissue, and is likely mediated by soil microbe communities. Similarly, Lori et al. (2017) find that organic management is associated with increased microbial abundance and activity.

Thus, while using synthetic fertilizers and pesticides may have the initial effect of increasing crop yields, over time these synthetic compounds exert an indirect negative effect on crop yields through their negative effects on soil health. This insight has implications for a farmer's optimal synthetic fertilizer and pesticide strategy, and for whether and how a farmer should transition from conventional to organic farming.

In the long run, pesticide use may even negatively affect profits due to their effects on soil productivity through soil health. Sexton et al. (2007) acknowledge the effect that pesticide use can have on soil health through its impact on soil microbiomes. Kalia and Gosal (2011) also document the damaging effects that the application of pesticides in conventional farming has on soil microorganisms that benefit plant productivity. Jaenicke and Lengnick (1999) estimate a soil quality index consistent with the notion of technical efficiency. Murphy et al. (2020) find that farmers in developing countries usually do not have sufficient information about their soil nutrient levels to make profit-maximizing decisions about fertilizer usage, and that there can be potentially large net benefits to providing farmers with soil information.

Owing to intertwined feedback links between biological and economic systems, bioeconomic modeling is challenging, and there is a considerable need for studies that couple economic models of decision-making with biophysical models to provide policy-relevant implications (Kling et al., 2017). Stevens (2018) argues that optimal control models may be well suited for studying the economics of soil management. In Meneses et al. (2024), we argue further that dynamic optimization and dynamic programming may help shed light on the optimal rate of transition from conventional to organic farming, by allowing us to better capture the countervailing and dynamic effects that pesticide use has on profits through its effect on pest pressure and soil health.

In our research in Meneses et al. (2024), we develop a dynamic bioeconomic model of a farmer's decisions regarding the use of synthetic compounds (e.g., synthetic fertilizers and

pesticides) and the transition from conventional to organic management. Our crop production model accounts for the newly documented interrelationships among synthetic compound use, soil health, and crop yields. We characterize and solve for a “fully informed” farmer’s optimal synthetic compound use strategy, and for whether and how a farmer should transition from conventional to organic farming. These solutions are compared to those from a “misinformed” model in which the farmer is not aware of the interactions between synthetic compound use, soil health, and crop yields, allowing us to assess how gaining knowledge of these interactions might be expected to change farmers’ synthetic compound use strategies and, ultimately, their decisions around adopting organic management. We identify and discuss agricultural and economic conditions under which farmers can be expected to voluntarily reduce their reliance on synthetic compounds, and possibly even adopt organic management, upon learning of the benefits associated with stewardship of their soil’s microbiome. We apply our model to farmer-level pesticide-use panel data to estimate parameters governing farmers’ current understanding of the interrelations between soil microbes, pesticides, and crop yields, and to examine possible effects of extension programs targeting farmers’ understanding of soil microbes.

4. Dynamic Bioeconomic Modeling Framework

The dynamic optimization problem faced by the farmer is to choose a pesticide and fertilizer input trajectory $c(t)$ to maximize the present discounted value (PDV) of their entire stream of profits from crop production:

$$\max_{\{c(t)\}} \int_0^{\infty} \left((P_{con} \cdot (1 - I_{org}) + P_{org} \cdot I_{org}) \cdot f(c(t), b(t)) - c(t) \right) e^{-\rho t} dt, \quad (1)$$

where P_{con} is the conventional crop price; P_{org} is the organic crop price; I_{org} is a dummy variable that equals 1 if the farm is organic and 0 otherwise; $f(c(t), b(t))$ is the crop production function as a function of pesticide and fertilizer input use $c(t)$ and soil microbes $b(t)$; and ρ is the interest rate.

The choice of the functional form and/or parameter values for the crop production function $f(c(t), b(t))$ is best informed by relevant scientific information from biology, plant sciences, and agronomy. For example, one can calibrate and parameterize the crop production function $f(c(t), b(t))$ with estimates from agronomic and other relevant literatures for the relevant crops. Other model parameters -- including organic crop prices P_{org} , conventional crop prices P_{con} , and the interest rate ρ -- can be calibrated using best available literature estimates, market data, and data and information from stakeholders.

In our research in Meneses et al. (2024), we incorporate newly documented interrelationships among synthetic compound use, soil health, and crop yield from soil science via a biological production for soil microbes $b(t)$ that depends on pesticide and fertilizer input use $c(t)$ as well as the total stock of synthetic compounds in the soil. Our model in Meneses et al. (2024) enables us to analyze the dynamics of soils and pests for better management, and to determine and document the effects of integrating soil and pest management on soil health and fertility, greenhouse gas

mitigation, enhanced biodiversity, resilience of the organic farming system, water and other resources, and other ecosystem services.

If the solution to the dynamic optimization problem (1) yields an optimal strategy that differs from farmers' actual decisions, then, to the extent that some of the differences between actual behavior and optimal strategy reflect possible sub-optimal behavior on the part of farmers, the model may suggest ways to improve farmers' soil and pest management. Our research and framework aims to help farmers improve decision-making around synthetic compound use and organic production, with the ultimate goal of improving soil bacteria stewardship, crop yields, farmer profits, and agricultural sustainability.

References

- Blundell, R., J.E. Schmidt, A. Igwe, A.L. Cheung, R.L. Vannette, A.C. Gaudin, and C.L. Casteel. (2020). Organic management promotes natural pest control through altered plant resistance to insects. *Nature Plants*, 6(5):483–491.
- Carroll, C.L., C.A. Carter, R.E. Goodhue, and C.-Y.C. Lin Lawell. (2019). Crop disease and agricultural productivity: Evidence from a dynamic structural model of verticillium wilt management. In W. Schlenker, editor, *Agricultural Productivity and Producer Behavior*, pages 217–249. University of Chicago Press, Chicago.
- Carroll, C.L., C.A. Carter, R.E. Goodhue, and C.-Y.C. Lin Lawell. (2024a). The economics of decision-making for crop disease control. Working paper, Cornell University. URL: http://clinlawell.dyson.cornell.edu/Vwilt_short_long_paper.pdf
- Carroll, C.L., C.A. Carter, R.E. Goodhue, and C.-Y.C. Lin Lawell. (2024b). Supply chain externalities and agricultural disease. Working paper, Cornell University. URL: http://clinlawell.dyson.cornell.edu/spinach_seed_paper.pdf
- Chambers, R.G., G. Karagiannis, and V. Tzouvelekas. Another look at pesticide productivity and pest damage. *American Journal of Agricultural Economics*, 92(5):1401–1419, 2010.
- Conrad, J.M., C.-Y.C. Lin Lawell, and B.B. Shin. (2024). On the optimal moratorium in an overfished fishery: The case of Canadian cod. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/fishery_UI_paper.pdf
- Dabbert, S., and P. Madden. (1986). The transition to organic agriculture: A multi-year simulation model of a Pennsylvania farm. *American Journal of Alternative Agriculture*, pages 99–107.
- Delbridge, T.A., and R.P. King. (2016). Transitioning to organic crop production: A dynamic programming approach. *Journal of Agricultural and Resource Economics*, pages 481–498.
- Hussain, S., T. Siddique, M. Saleem, M. Arshad, and A. Khalid. (2009). Chapter 5: Impact of pesticides on soil microbial diversity, enzymes, and biochemical reactions. *Advances in Agronomy*, 102:159–200.
- Jacquet, F., J.-P. Butault, and L. Guichard. (2011). An economic analysis of the possibility of reducing pesticides in French field crops. *Ecological Economics*, 70(9):1638–1648.
- Jaenicke, E. (2000). Testing for intermediate outputs in dynamic DEA models: Accounting for soil capital in rotational crop production and productivity measures. *Journal of Productivity Analysis*, 14(4):247–266.
- Jaenicke, E., and L. Lengnick. (1999). A soil-quality index and its relationship to efficiency and productivity growth measures: Two decompositions. *American Journal of Agricultural Economics*, 81(4):881–893.
- Kalia, A., and S. Gosal. (2011). Effect of pesticide application on soil microorganisms. *Archives of Agronomy and Soil Science*, 57(6):569–596.
- Kling, C.L., R.W. Arritt, G. Calhoun, and D.A. Keiser. (2017). Integrated assessment models of the food, energy, and water nexus: a review and an outline of research needs. *Annual Review of Resource Economics*, 9(1):143–163.
- C.-C. Lo. (2010). Effect of pesticides on soil microbial community. *Journal of Environmental Science and Health Part B*, 45(5):348–359.
- Lori, M., S. Symnackzik, P. Mäder, G. De Deyn, and A. Gattinger. (2017). Organic farming enhances soil microbial abundance and activity^{aa} meta-analysis and meta-regression. *PLOS ONE*, 12(7):e0180442.
- Meneses, M.A., C.L. Casteel, M.I. Gómez, D.R. Just, R. Kanbur, D.R. Lee, and C.-Y.C. Lin Lawell. (2024). Organic farming transitions: A dynamic bioeconomic model. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/organic_transitions_paper.pdf
- Murphy, D.M., D. Roobroeck, D.R. Lee, and J. Thies (2020). Underground knowledge: Estimating the impacts of soil information transfers through experimental auctions. *American Journal of Agricultural Economics*, 102(5):1468–1493.
- Sambucci, O., C.-Y.C. Lin Lawell, and T.J. Lybbert (2024). Pesticide spraying and disease forecasts: A dynamic structural econometric model of grape growers in California. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/PMI_dynamic_paper.pdf
- Sears, L., D. Lim, and C.-Y.C. Lin Lawell. (2019). Spatial groundwater management: A dynamic game framework and application to California. *Water Economics and Policy*, 5(1):1850019.
- Sears, L., C.-Y.C. Lin Lawell, G. Torres, and M.T. Walter. (2024a). Adjudicated groundwater property rights: A structural model of the dynamic game among groundwater users in California. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/property_rights_struc_paper.pdf

- Sears, L., C.-Y.C. Lin Lawell, G. Torres, and M.T. Walter. (2024b). Moment-based Markov equilibrium estimation of high-dimension dynamic games: An application to groundwater management in California. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/mme_groundwater_paper.pdf
- Sears, L., C.-Y.C. Lin Lawell, and M.T. Walter. (2024c). Groundwater under open access: A structural model of the dynamic common pool extraction game. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/open_access_structural_paper.pdf
- Sexton, S.E., Z. Lei, and D. Zilberman. (2007). The economics of pesticides and pest control. *International Review of Environmental and Resource Economics*, 1(3):271–326.
- Shin, B.B., J.M. Conrad, and C.-Y.C. Lin Lawell. (2024). On the optimality of a fishery moratorium. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/moratorium_paper.pdf
- Stevens, A.W. (2018). The economics of soil health. *Food Policy*, 80:1–9.
- Varanasi, A. (2019) Is organic food really better for the environment? *State of the Planet*. URL <https://news.climate.columbia.edu/2019/10/22/organic-food-better-environment/>
- Wilcox, S.W., D.R. Just, C.-Y.C. Lin Lawell, M.I. Gómez, and H. Grab. (2024). To (rent) bees or not to (rent) bees?: An examination of the farmer's question. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/rent_bees_paper.pdf
- Wu, J. (2000). Optimal weed control under static and dynamic decision rules. *Agricultural Economics*, 25 (1):119–130.
- Wu, T., D.R. Just, C.-Y.C. Lin Lawell, J. Zhao, Z. Fei, A. Ortiz-Bobea and Q. Wei. (2024). Optimal forest management for interdependent products: A nested dynamic bioeconomic model and application to bamboo. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/Moso_bamboo_dynamic_paper.pdf
- Yeh, D.A., M.I. Gómez, and C.-Y.C. Lin Lawell. (2024). Sustainable pest management, beliefs, and behavior: A dynamic bioeconomic analysis. Working paper, Cornell University. URL: https://clinlawell.dyson.cornell.edu/Maine_SWD_dynamic_bioeconomic_paper.pdf

The heterogeneous effects of COVID-19 on young consumer purchasing patterns: An examination using causal random forest

Youngran Choi
Assistant Professor
David B. O'Maley College of Business
Embry-Riddle Aeronautical University
1 Aerospace Boulevard
Daytona Beach, FL 32114
CHOIY20@erau.edu
ORCID: 0000-0002-6868-2381

Abstract

The coronavirus (COVID-19) pandemic prompted governments worldwide to implement lockdown measures that resulted in an increase in online purchases. Young consumers, who are familiar with browsing online platforms, may respond differently than older consumers to the pandemic in terms of online purchases. The generalized random forest (GRF) was employed to identify the heterogeneous causal effects of COVID-19 on the online shopping habits of young consumers. The attributes the GRF analysis identified as sensitive to COVID-19 in inducing online purchases include: gender based on employment, race, television ownership, and specific types of grocery products (i.e., dry foods and frozen foods). Understanding the heterogeneity in the effects of COVID-19 that induce young consumers to shop online can provide practitioners with insights on how to prioritize and target high-impact groups with specific characteristics while improving management and policy strategies for future pandemics or other extended periods of disruption of in-store businesses.

Keywords: young consumer, consumption pattern, COVID-19, random forest, machine learning
JEL Codes: D12, L81, R20

1 Introduction

The coronavirus (COVID-19) was declared a pandemic in March 2020 by the World Health Organization (WHO) which prompted governments to implement a number of measures such as lockdowns, social distancing, and off-line store closures to prevent wide transmission. These preventive methods restricted people from traveling and shopping as usual, triggering consumers to turn to online shopping options to reduce the risk of infection (Ellison et al., 2020; Shaw et al., 2022). The resultant changes in shopping behavior patterns are evident in the statistics gathered by the Annual Retail Trade Survey issued by U.S. Census Bureau. Sales of e-commerce in 2020, the first year of the pandemic reached US\$815.4 billion, an increase of 43 percent from US\$571.2 billion in 2019 (Brewster, 2022). The share of e-commerce in the U.S. increases dramatically, reaching 14.6 percent of total sales in 2020, up from 10.6 in 2019; the ecommerce share continues to be significant, maintaining a level of 14.7 percent of total sales in 2021 (U.S. Census Bureau, 2022).

As online shopping has increased, studies that seek to associate new shopping patterns with COVID-19 have also increased. The growth of e-commerce during the pandemic has prompted questions about the changes in consumer habits and the implications of those changes, which we seek to understand. Understanding is particularly important for businesses as firms optimize their models and marketing strategies. The changes in shopping behaviors are evident as firms respond to the disruptions in behavioral patterns catalyzed by efforts to slow the spread of COVID-19 (Mehta et al., 2020). Young adult consumers, who are one segment among consumer groups, are more prone to explore and experiment, and therefore are potentially more adaptable to change. This adaptability would be revealed as young consumers adopt online buying practices that reflect their technological savvy while using digital media (Heaney, 2007; Dannenberg, 2020). However, young consumers are not a homogeneous subgroup, and variations exist across preferences and shopping habits. Therefore, understanding intragenerational differences in online shopping can help businesses develop strategies to leverage the market (Johnson et al., 2022; Debevec et al., 2013).

For firms developing integrated business models with e-commerce platforms, understanding this segment of consumers has the potential to lead to significant growth for retail businesses. Researching the factors associated with “increased online purchases” by young consumers in response to the restrictions related to COVID-19 can improve our understanding of the consequences of disruptive events in industries. A growing body of literature exists on consumer behavioral changes during COVID-19, however many of these studies focus on general consumers. Studies have found that the number of confirmed cases of COVID-19 are positively associated with food purchases made online in China, especially for young consumers living in large cities (Gao et al., 2020). Other research has also found that the perception of enjoyment and utilitarian aspects (Sumi and Ahmed, 2022) as well as convenience (Shaw et al., 2022) affect younger consumers’ choice to shop online.

This paper focuses on identifying the characteristics of young consumers who were more likely to purchase groceries online in response to COVID-19 restrictions and using our findings to offer insights on marketing to this subgroup. Identifying the heterogeneous effects of COVID-19 across attributes offers insights to not only researchers but also to business practitioners building marketing strategies to effectively target specific subgroup populations of young consumers. This paper adopts a recently developed machine learning technique, generalized random forest, to select the most important of more

than a hundred attributes. This approach has the added benefit of not restricting variable selection based on a prior model structure.

This paper is organized as follows: Section 2 presents literature related to studies on young consumer behavior during COVID-19 restrictions. Data and empirical methods are explained in Section 3. Section 4 reports results and discussions. Conclusions are presented in Section 5.

2 Related literature

Existing research mostly focuses on the behavioral changes of consumer groups in general during the period of the pandemic, and relatively few studies explored whether or how young consumers' online purchasing patterns changed in specific. During the COVID-19 pandemic, consumers' behavior was transformed, which resulted in changes in market structure and dynamics (Mehta et al., 2020). Ellison et al. (2020) present empirical evidence of increased online grocery shopping and reduced food-away-from-home expenses for U.S. households during the pandemic. Similar findings were presented by Truong and Truong (2022), who applied the theory of fear, stating that consumers altered their shopping behavior by choosing curbside pick-up and online methods instead of shopping in physical stores due to fear of being exposed to COVID-19 infection. Using the Nielsen COVID-19 Shopper Behavior Survey, Lo et al. (2021) presented similar findings of increased online shopping by U.S. households, and their intention to continue shopping online post-COVID-19 due to fear of risk of infection. Grashuis et al. (2020) conducted a choice experiment method that they adapted to determine consumer preferences for online grocery shopping; they found that consumer willingness to grocery shop online is positively associated with COVID-19 spreading at an increasing rate. East (2023) predicts that the pattern of online grocery ordering in the U.K. will decrease short after the pandemic ends but then steadily grow as it did prior to the pandemic.

Showrav et al. (2021) and Yan et al. (2021) investigated key determinants of consumers moving toward online shopping platforms in Bangladesh and China, respectively. They found that the significant benefits included contactless transactions, convenience, and technological support for online shopping. In addition, Watanabe and Omori (2020) stated that reduction in transportation costs and search costs are some of the benefits of online shopping, as internet platforms offer extensive lists of products and services that are easily accessible and comparable. However, the costs associated with adopting online shopping, such as computer installation costs and the investment of learning to browse websites, can hinder some consumers from adopting online shopping platforms, such that online consumption varies across countries and regions. From the viewpoint of consumers, loss of privacy is another potential cost of online shopping (Watanabe and Omori, 2020). Hoseini and Valizadeh (2021) employed a difference-in-difference approach to analyze how the consumption pattern and online-offline substitution occurred during the COVID-19 lockdown in Iran. They found that online consumption spiked during the lockdown and the preference for online purchases persisted even after the lifting of lockdown measures. Hamermesh (2020) used survey data to show how "with whom and how" time spent affects life satisfaction and simulated the impacts of the pandemic on life satisfaction.

Young consumers in particular exhibited increased online shopping behavior during COVID-19. Hadler et al. (2021) conducted a survey of US young consumers on their perspectives on and experiences with online grocery shopping during the pandemic. The researchers found that young consumers had positive perceptions of their online grocery shopping experiences and widely utilized this shopping method during

the pandemic with a tendency to purchase more snacks and non-perishable foods. Shaw et al. (2020) showed that younger Germans preferred the convenience of online shopping more than older Germans. The study also found that younger consumers in Germany found it easier to browse through and compare products on different websites. The data from Germany revealed that young female adults presented a high level of hedonic shopping motives compared to young males, and were also more active in using social media during the time of COVID-19 shutdowns (Koch et al., 2020). Increased online shopping behavior was also evident in China. Gao et al (2020) found that young consumers living in large cities perceived online purchases as lower in risk and were more likely to purchase food online compared to the elderly. Moon et al. (2021) presented similar findings using data from Korea, where younger consumers, along with single females, tended to shop more online during COVID-19. The qualitative study by Amatulli et al (2023) examined the consumption patterns of Generation Z from the perspective of how they coped with the pandemic. The authors found that during the lockdown the younger generation's consumption was linked to experiences that pleased them, however, their shopping behaviors correlated with their social and personal welfare.

Several studies explored the effects of COVID-19 on consumer consumption patterns with applications of machine learning techniques. German et al. (2022) adopted a random forest classifier and artificial neural network to identify factors affecting consumers' preferences in selecting third-party logistics such as Amazon. They showed that the two machine learning algorithms reliably revealed that consumers' attitudes, satisfaction, perception of values, the trust level of third-party logistics, and perception of environmental friendliness to be significant attributes of concern in choosing third-party logistics during COVID-19. Using the latent Dirichlet allocation (LDA) tool, a type of machine learning technique used in analyzing text data, Meena and Kumar (2022) explored social media data to examine the performance of online food delivery services and consumer satisfaction in India during the pandemic. Schlegelmilch et al. (2022) employed supervised machine learning and artificial neural network in analyzing consumer sentiments about COVID-19 in six countries for constructing marketing implications using text-mining tools.

3 Data and empirical method

We collected the data from the Neilson Survey Data, which recorded detailed long-form information on grocery purchases by individual consumers for the years 2019 and 2020 in the U.S. The data includes spending records for individuals, including product information such as price and unit, as well as the demographics of the individuals and the households that the individuals belong to. The study data covers purchase information for young consumers within the age range of 18 to 25 years. This data set on young consumers contains 446,782 observations. We began with 110 variables that were mostly binary to indicate specific attributes and used empirical analysis to reduce the dimensions of these potential covariates.

The variables are generally related to two factors: items purchased and demographics. The information about items purchased includes method of payment, type of item (whether dairy, dry grocery, frozen foods, health care and beauty, non-food grocery, fresh produce, etc.), whether the item was purchased with a coupon or a discount deal, price per unit, and total spent for the purchase. The demographics include information on household income and size, occupation, race, ethnicity (Hispanic or not), gender, marital status, level of education, whether the household has an internet connection, type of residence, and region. Variables that are selected as important using the generalized random forest, and the

COVID-19 period as a treatment, are summarized in Table 1. The mean and mean difference of each variable by treatment is also presented. The correlation table is shown in Table 2.

3.1 Generalized causal forest

The empirical analysis involved a high dimensional dataset with many variables as candidates for explaining the heterogeneous effects of COVID-19 on the purchasing patterns of young consumers. We employed the generalized random forest (GRF) to identify the attributes that explain the heterogeneous effects of COVID-19 in inducing more online consumption by young consumers. The attributes that are most important can be identified as those that are most closely associated with changing online shopping patterns. This method also allows researchers to identify the particular subgroups of young consumers who altered their online shopping behavior the most during COVID-19.

To examine the heterogeneous effects of COVID-19, we defined y as the outcome variable for an individual i , where $y_i = 1$ if an individual purchased an item online or 0 if it was purchased in an offline shop. The vector of covariates is denoted as X_i . The treatment variable represents the period of the COVID-19 pandemic with a binary variable where a value 1 represents the COVID-19 period, in this case the year 2020, and 0 represents the pre-pandemic period. For a consumer who purchased items online, the outcome variable is denoted as $y_i = 1$ while $y_i = 0$ denotes items purchased in offline stores. The factual treatment effects cannot be examined since only one of two outcomes is observed in the dataset (Holland & Rubin, 1987). Therefore, conditional average treatment effects were identified following the difference between the expected potential outcomes conditional at X_i .

Eq. (1)

The heterogeneous treatment effects were estimated using a semi-parametric approach. The effects are defined as $\tau(X_i)$. The term \hat{y}_i represents the out-of-sample prediction of a conditional outcome y_i . The conditional probability of being exposed to COVID-19 is denoted as p_i for X_i . The GRF approach estimates \hat{y}_i and p_i separately. Then the GRF computes the residuals of treatment and the outcome y_i . The ATE is estimated based on the following equation.

Eq. (2)

We assumed that unconfoundedness was satisfied as exposure to COVID-19 occurred unexpectedly and free of dependence on potential outcomes, Since each young consumer had a probability of being exposed to COVID-19 that was greater than 0 and less than 1, that is, $0 < p_i < 1$, we assumed that the overlap assumption was also satisfied. In addition, although lockdown was implemented during the pandemic stage, which may have prevented offline consumption, young consumers were not selected based on particular covariates (Rosenbaum & Rubin, 1983).

3.2 Robustness - comparison of mean outcomes

To support the robustness of the results produced with GRF, we employed a quasi-difference-indifference method to estimate the treatment effects of COVID-19 on young consumers' online purchases during the pandemic. The coefficients of logistic regression were compared to the results of conditional average

treatment effects (CATEs) from the GRF for heterogeneous directions and statistical significance. This is to compare the mean outcome among different treatment group. The probability that groceries were purchased online is denoted as y_i for individual i . The logistic regression model was defined as:

$$\text{Eq. (3)}$$

where t indicates the time period during which a young consumer experienced the pandemic as a treatment. X_i denotes a characteristic selected under the GRF algorithm as an important conditional variable. The selected attribute interacts with the COVID time period, and the coefficient of the interaction, β , is the DID estimator for the effect that measures the difference between the treatment effects when $t=1$ and $t=0$. The interaction coefficient captures the heterogeneous effects of COVID-19 conditional on a specific attribute based on the following equation where the attribute is denoted as X_i .

$$\text{Eq. (4)}$$

The term ϵ_i in Eq. (3) represents the remaining selected characteristics, excluding the particular attribute interacting with the COVID-19 treatment. In addition, a location code—the first three digits of the zip code, denoted as LOC_i —is added to capture fixed effects for controlling unobserved characteristics of young consumers. The error term ϵ_i is expected to have a zero mean.

4 Results and discussion

The empirical method of employing a causal random forest, a type of machine learning technique, enabled us to detect heterogeneity in the effects the COVID-19 pandemic had on increasing young consumers' online purchases. Table 3 presents the CATEs of important selected attributes used to identify the heterogeneity in the effects of COVID-19. The variables were selected in order based on how frequently each attribute was selected when splitting data to build a tree (Athey et al., 2019). The variables selected as important suggest that young consumers with these attributes tended to shift more swiftly to online shopping during the pandemic. The identification of heterogeneous effects is also demonstrated in Figure 1 with an associated 95% confidence interval.

The GRF method was able to identify specific demographic characteristics associated with increasing online shopping behavior during COVID-19. Household income and size contributed to explaining the heterogeneous effects of the pandemic in inducing young consumers to choose to shop online. Both the below-median and above-median income groups experienced the positive effects of the COVID-19 pandemic and turned to online shopping more during lockdown. However, young consumers from the below-median income group showed greater effects of the pandemic in moving toward online shopping compared to the above-median income group. Duffy et al. (2022) stated that for the general consumer, those of lower income in their sample who were also experiencing food insecurity transitioned to online grocery shopping at a rate of 36%. The authors also explained that this lower income group was also more likely to benefit from the Supplemental Nutrition Assistance Program (SNAP), as the United States Department of Agriculture started allowing online grocery outlets to accept SNAP during the COVID-19 pandemic. Our findings also show that young consumers from large households were more likely to shop online during the pandemic, which is consistent with the findings of Duffy et al. (2022), who found that those with children and those experiencing food insecurity were more likely to shop online for groceries during the pandemic period.

Our findings also indicated that the number of televisions that young households owned affected the likelihood of purchasing groceries online. Young households with no television purchased more groceries online than those with one or more televisions. It is likely that members of young households without televisions use a computer or a laptop as the main means of media, thereby making them more familiar with the processes of e-commerce purchases. In addition, shopping websites are optimized for computer browsers and screens along with keyboard and mouse control (Wagner et al., 2017), so that those using a computer or a laptop as their primary device for tasks may find online shopping using the device more convenient.

Other demographic characteristics that were identified as attributes associated with purchasing groceries online during COVID-19 include gender based on employment status. It is interesting to observe that employed females responded to COVID-19 with the highest rates of online purchasing, followed by employed males. Unemployed females responded negatively with a drop in online purchasing, while both unemployed and employed males responded positively. Unemployed males experienced lower effects with respect to online purchasing compared to employed males. Kock et al. (2020) found that women and Generation Z, defined as those born after 1997, presented higher motivation to shop online compared to older, male groups. Bardey et al. (2023) also uncovered gender disparities in the shopping pattern of beauty and fashion products post-COVID-19. Our findings contribute to the literature by confirming that the tendency to make online purchases differs not only across gender but also across employment status.

Racial disparities were also observed in a heterogeneous tendency to purchase online while being affected by COVID-19. We found that young black and Asian consumers were more likely to increase online shopping compared to those who are white. Similar to our findings, Sze et al. (2020) also found that certain racial groups, such as blacks and Asians, are at a higher risk of being infected by COVID-19 compared to whites. Martin et al (2020) offered explanations in their study, noting that these minorities are more likely to live in households of larger sizes, thereby reducing the effectiveness of lockdown and social distancing for those individuals. Hawkins (2020) explains that these minority groups are also more likely to be employed in sectors with close proximity to others and therefore a higher risk of exposure to COVID-19. Thus, young consumers of these minority groups, if aware of risks associated with both occupation and higher household size, may be inclined to purchase more online and avoid face-to-face contact to reduce the risks of infection.

It is interesting to observe how the GRF method allows the detection of heterogeneity in the effects of COVID-19 on which types of products are more likely to be consumed online. The estimated treatment effects, the effect of COVID-19 in enhancing the consumption of a specific product type, is higher for frozen foods. Chenarides et al. (2021) surveyed consumers in the U.S. and found that due to store service closures and lockdown, consumers purchased more than normal during the service times. When stores visits are constrained, foods that are not perishable or that can be stored for a long time are preferred. The empirical evidence of Chang et al. (2021) also revealed that frozen foods were among the food items most in demand during COVID-19, based on data from Taiwan.

Consistent with research on how coupons and deal-flags induce consumers to make online purchases (Ren et al., 2021), we also found that more online purchases occurred during COVID-19 for items with dealflags or discounts compared to regular-priced items. Our results provide evidence that young consumers, a particular population group by age strata, were sensitive to when products were offered with price deals while purchasing online items during COVID-19.

4.1 Difference-in-difference estimators for further robustness

The generalized random forest (GRF) provides researchers with the advantage of handling data by reducing its dimensionality when the data has many covariates without a pre-identifying model structure. For the robustness of the results produced with random forest algorithm, we estimated the treatment effects using the difference-in-difference approach. Table 4 presents the marginal effect estimators generated by interacting the treatment with the attributes selected as important in GRF. Overall, the coefficients from the logistic regression show signs that are consistent compared to the CATEs estimated using GRF. Those young adults with below-median incomes who own no televisions and/or are from large households were likely to make more online grocery purchases compared to those with above-median incomes who own more than one television and/or are from small households, during the pandemic. Employed females also had a higher likelihood of purchasing online groceries compared to others. The negative signs associated with employed males indicate that employed males were less likely to purchase groceries online compared to others, including both females and unemployed males. Price deal flags showed a negative sign, which is inconsistent with the CATEs estimated in GRF, however this result is not statistically significant. Dummies for different races also showed consistent signs that in general young black and Asian consumers were more likely to purchase groceries online compared to other races, while white young adults were less likely to during the pandemic. Although discrepancies exist in the magnitudes of the effects, the signs are consistent in general, and the statistical significance of the variables indicate that GRF is indeed effective in selecting the important variables when data is high dimensional. In addition, the estimation using GRF was able to effectively identify the heterogeneous effects of COVID-19.

5 Conclusion

In this paper, we identify a set of characteristics that contribute to heterogeneity in the effects of COVID19 in moving young consumers toward online shopping. We adopt a recently developed machine learning method, a generalized random forest, which affords us the benefit of being able to analyze data in a nonparametric approach that is not also bounded by prior model specification. Among many candidate variables, certain demographic characteristics were associated with higher heterogeneous effects of COVID-19. Young Asian and black consumers were more likely to purchase products online than those who were white. Young consumers from larger-sized households and/or with lower incomes were also more prone to shop online compared to those from single-family households and/or with higher incomes. Specific product types, such as frozen foods, were also associated with higher purchase rates by young consumers during COVID-19.

In addition, the application of a recently developed machine learning tool on a large-scale dataset with many variables enables researchers to obtain empirical evidence related to causal factors without prior model specification as well as with a low computational burden (Dorie et al., 2019; Wendling et al., 2018). The current study demonstrates that the generalized causal forest tool can effectively determine the attributes of consumer subgroup characteristics that are sensitive with respect to responses to the pandemic, that is, attributes that contribute to an increase in online grocery shopping.

Our findings offer business practitioners, particularly those in marketing, insights about methods of finding specific characteristics of consumers being affected by treatment, thereby facilitating the implementation of effective personalized promotions and services targeting specific subgroups. Understanding the variables mostly likely to be useful in analyzing the effects of a particular event or treatment can also provide insights for policymakers on factors underlying the heterogeneous effects without restricting the analysis to a limited number of candidate variables. In addition, implementing the generalized random forest model captures the treatment effects in high dimensional data with a low computation burden, which results in timely analysis of causal factors. Prompt identification of the factors that cause changes in consumption patterns is critical in implementing effective business and marketing strategies. The technique used in this paper can screen initial data and identify the characteristics to be prioritized in personalized marketing and consumer communication. This paper's findings can also be used to develop ways to improve nutritional support and a food safety net for young consumers. Regulations to ensure the safety of groceries, including frozen foods, through disclosure of information and labelling can support consumers by allowing them to make informed purchasing decisions.

While we provide insights on how young consumers were affected by COVID-19 when making online purchases during lockdown, based on results obtained by investigating the heterogeneity of the effects of various attributes, our results are limited in terms of offering guidelines on whether the effects continue to persist during the post-COVID-19 period. This limitation is due to a lack of data on which to conduct analyses to determine post-COVID-19 behavior. Post-COVID-19 shopping behavior can be investigated in future research. In addition, to obtain more consistent and timely results, it is recommended that practitioners employ the most up-to-date data in applications. We also acknowledge that bias may exist as a result of omitted variables of other events and policies that can influence consumption choices of young consumers during the study period.

Table 1. Summary statistics and mean comparisons by pre-COVID and COVID periods

	Full sample			(1) Pre-COVID		(2) COVID		Difference	Chisquare test
	Obs	Mean	SD	Mean	SD	Mean	SD		
COVID-19 (treatment)	446,782	0.47	0.5						
Online shopping	446,782	0.01	0.12	0.01	0.10	0.02	0.14	-0.01***	866.09
Household income (below median)	446,782	0.61	0.49	0.61	0.49	0.62	0.49	-0.01***	55.82
Household size (above median)	446,782	0.42	0.49	0.41	0.49	0.42	0.49	-0.01***	35.38
TV owned, none	446,782	0.2	0.40	0.23	0.42	0.18	0.38	0.05***	1,900.00
Female, employed	446,782	0.75	0.43	0.75	0.43	0.75	0.43	0.01***	16.02
Male, employed	446,782	0.96	0.21	0.98	0.15	0.93	0.25	0.04***	5,200.00
Dry grocery	446,782	0.43	0.50	0.43	0.49	0.43	0.50	-0.01***	18.05
Non-food grocery	446,782	0.07	0.25	0.07	0.25	0.07	0.25	-0.00**	2.92
Deal flag	446,782	0.09	0.29	0.11	0.31	0.07	0.26	0.04***	1,700.00
Price per unit (below median)	446,782	0.62	0.49	0.64	0.48	0.60	0.49	0.04***	652.64
White/Caucasian	446,782	0.68	0.46	0.66	0.47	0.71	0.45	-0.05***	1,400.00
Black/African American	446,782	0.15	0.36	0.18	0.38	0.12	0.32	0.06***	3,600.00
Asian	446,782	0.06	0.23	0.06	0.23	0.05	0.23	0.00***	17.49
Other races	446,782	0.11	0.31	0.10	0.31	0.12	0.32	-0.01***	212.41
Observations				236,388		210,394		446,782	

Note: The minimum and maximum values are 0 and 1 for all variables as variables included in our analysis are binary.

Table 2. Correlation matrix of selected important variables

Variables	(1)	(2)	(3)	(4)	(7)	(5)	(6)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) Online shopping	1.00														
(2) COVID period	0.04	1.00													
(3) Household income (below median)	0.02	0.01	1.00												
(4) Household size (above median)	0.03	0.01	-0.01	1.00											
(7) TV owned none	0.04	-0.06	0.10	-0.03	1.00										
(5) Female, employed	0.02	-0.01	-0.13	-0.23	0.01	1.00									
(6) Male, employed	-0.03	-0.11	-0.04	-0.01	0.10	-0.02	1.00								
(8) Deal flag	-0.01	-0.06	-0.04	-0.01	0.04	0.06	0.00	1.00							
(9) Price per unit (below median)	-0.06	-0.04	0.04	-0.01	0.03	-0.01	0.01	0.04	1.00						
(10) White/Caucasian		-0.06	0.06	0.05	-0.04	0.03	0.02	-0.05	-0.03	0.01	1.00				
(11) Black/African American		0.08	-0.09	0.06	0.08	-0.05	-0.06	0.02	0.03	0.00	-0.62	1.00			
(12) Asian	0.00	-0.01	-0.06	0.00	0.07	0.00	0.05	0.01	-0.03	-0.36	-0.10	1.00			
(13) Other races	0.00	0.02	-0.09	-0.03	-0.04	0.04	0.01	0.01	-0.01	-0.52	-0.15	-0.09	1.00		
(14) Dry grocery	-0.02	0.01	0.03	0.04	0.00	-0.02	-0.02	0.00	0.26	0.00	0.02	-0.03	0.00	1.00	
(15) Non-food grocery	-0.23	1.00	0.01	0.00	0.02	0.01	-0.01	-0.02	0.01	0.01	-0.09	-0.01	0.02	-0.01	0.00

Table 3. CATEs estimated using subgroup by binary levels of selected important variables

Variable	Level	CATE	SE	Lower bound	Upper bound
Household income	Below median	12.62	1.98	8.75	16.49
	Above median	5.84	2.25	1.42	10.25
Household size	Below median	4.00	1.85	0.38	7.62
	Above median	16.33	2.36	11.70	20.96
Female	Employed	13.91	1.70	10.57	17.25
	Unemployed	-3.59	2.99	-9.45	2.27
Male	Employed	10.21	1.58	7.11	13.31
	Unemployed	5.77	3.73	-1.54	13.08
TV owned	None	21.27	3.21	14.97	27.57
	One or more	5.80	1.65	2.56	9.04
Price deal flag	Yes	23.83	6.39	11.30	36.35
	No	8.53	1.51	5.57	11.49
Price per unit	Below median	13.55	2.18	9.27	17.84
	Above median	6.49	2.03	2.51	10.47
Race	White/Caucasian	3.78	1.82	0.22	7.35
	Black/African American	26.17	3.21	19.88	32.46
	Asian	16.58	6.76	3.32	29.83
	Other	-1.33	4.36	-9.88	7.21
Product type	Dry grocery	9.14	2.19	4.85	13.43
	Non-food grocery	0.20	5.16	-9.92	10.32

Figure 1. CATE estimated separately over subsamples (by binary level)

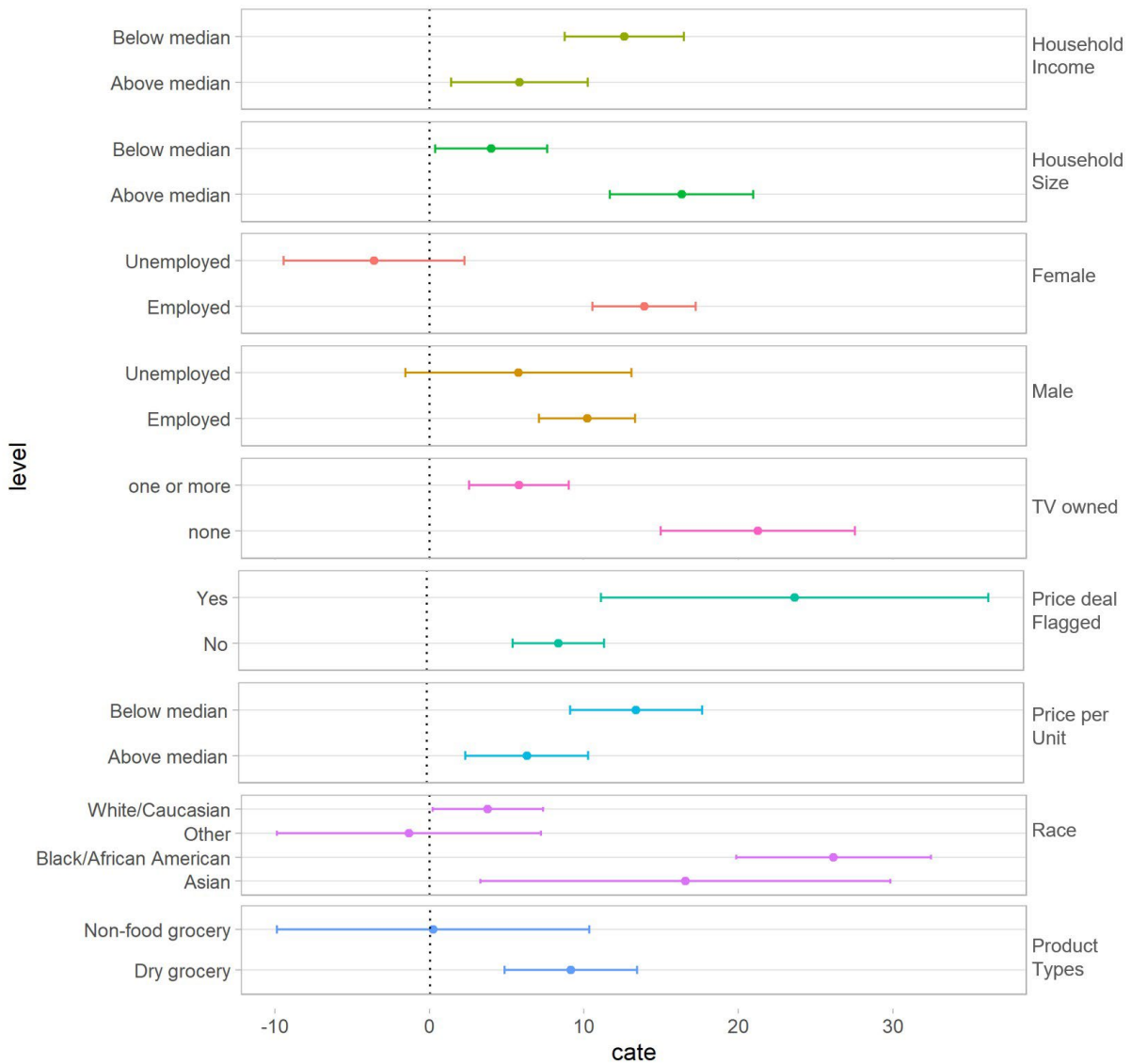


Table 4. Summary of difference-in-difference (DID) marginal effects estimates comparing pre-and during-COVID-19

			Pre-COVID (%)	During-COVID (%)	Effects* (During-Pre) (%)	zscore	pvalue
Household income	Above median	0	1.07	1.36			
	Below median	1	0.86	2.49			
Household size	Diff.		-0.21	1.14	1.35	15.30	0.00
	Below median	0	0.76	1.48			
	Above median	1	1.20	2.94			
	Diff.		0.44	1.46	1.02	4.68	0.00

Female	Unemployed	0	1.08	0.71			
	Employed	1	0.88	2.64			
	Diff.		-0.20	1.93	2.13	22.76	0.00
Male	Unemployed	0	1.04	4.70			
	Employed	1	0.93	1.95			
	Diff.		-0.11	-2.74	-2.63	-6.60	0.00
TV owned	One or more	0	0.92	1.51			
	None	1	1.06	4.77			
	Diff.		0.15	3.26	3.11	17.27	0.00
Price deal flag	No	0	0.98	2.15			
	Yes	1	0.63	1.58			
	Diff.		-0.35	-0.57	-0.22	1.26	0.21
Price per unit	Above median	0	1.67	3.05			
	Below median	1	0.50	1.50			
	Diff.		-1.17	-1.55	-0.38	8.96	0.00
Race	White	0	1.02	2.65			
		1	0.88	1.63			
		Diff.	-0.15	-1.02	-0.88	-6.57	0.00
	Black	0	0.84	1.38			
		1	1.50	5.37			
		Diff.	0.66	3.99	3.33	15.02	0.00
Asian	0	0.95	2.07				
	1	0.76	2.17				
	Diff.	-0.18	0.10	0.29	2.15	0.03	
Product type	Dry grocery	0	1.01	2.05			
		1	0.82	2.12			
		Diff.	-0.19	0.07	0.26	4.49	0.00
	Non-food grocery	0	0.90	2.07			
		1	1.34	2.09			
		Diff.	0.44	0.02	-0.42	-4.51	0.00

Note: * the effects indicate estimated coefficients for DID estimators.

Statements & Declarations

Conflict of interest: The author declare that no funds, grants, or other support were received during the preparation of this manuscript.

Financial statement: The author has no relevant financial or non-financial interests to disclose.

Data: Researcher's own analyses calculated based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher's and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. The code to analyze the data can be provided upon request.

References

- Amatulli, C., Peluso, A. M., Sestino, A., Guido, G., & Belk, R. (2023). The influence of a lockdown on consumption: An exploratory study on generation Z's consumers. *Journal of Retailing and Consumer Services*, 73, 103358. doi:10.1016/j.jretconser.2023.103358
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148-1178. doi:10.1214/18-AOS1709
- Ben Hassen, T., El Bilali, H., & Allahyari, M. S. (2020). Impact of COVID-19 on food behavior and consumption in qatar. *Sustainability*, 12(17) doi:10.3390/su12176973
- Bardey, A., Daniel, A. B., Rogaten, J., Mala, A., & Khadaroo, A. Exploring the post-COVID lipstick effect: A short report. *International Journal of Market Research*, 14707853231201856. doi:10.1177/14707853231201856
- Brewster, M. (2022). Annual retail trade survey shows impact of online shopping on retail sales during COVID-19 pandemic. Retrieved from <https://www.census.gov/library/stories/2022/04/ecommerce-sales-surged-duringpandemic.html>
- Chang, H., & Meyerhoefer, C. D. (2021). COVID 19 and the demand for online food shopping services: Empirical evidence from Taiwan. *American Journal of Agricultural Economics*, 103(2), 448-465. doi:10.1111/ajae.12170
- Chauhan, V., & Shah, M. (2020). *Purakala (UGC care journal) an empirical analysis into sentiments, media consumption habits, and consumer behaviour during the coronavirus (COVID-19) outbreak* doi:10.13140/RG.2.2.32269.15846
- Chen, Y., Guang-Wen Zheng, Abu Bakkar Siddik Qian Li Dong, & Chang, D. (2021). Factors affecting the consumers online shopping during the COVID-19 pandemic in China. *Revista*

Argentina De Clínica Psicológica, 30(1), 853. doi:10.24205/03276716.2020.2081

Chenarides, L., Grebitus, C., Lusk, J. L., & Printezis, I. (2021). Food consumption behavior during the COVID 19 pandemic. *Agribusiness (New York, N.Y.)*, 37(1), 44-81. doi:10.1002/agr.21679

Dannenberg, P., Fuchs, M., Riedler, T., & Wiedemann, C. (2020a). *Digital transition by COVID 19 pandemic? the German food online retail* Wiley. doi:10.1111/tesg.12453

Dannenberg, P., Fuchs, M., Riedler, T., & Wiedemann, C. (2020b). Digital transition by COVID-19 pandemic? the german food online retail. *Tijdschrift Voor Economische En Sociale Geografie*, 111(3), 543-560. doi:10.1111/tesg.12453

Debevec, K., Schewe, C. D., Madden, T. J., & Diamond, W. D. (2013). Are today's millennials splintering into a new generational cohort? maybe! *Journal of Consumer Behaviour*, 12(1), 20-31. doi:10.1002/cb.1400

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1), 43. doi:10.1214/18-STS667

Duffy, E. W., Lo, A., Hall, M. G., Taillie, L. S., & Ng, S. W. (2022). Prevalence and demographic correlates of online grocery shopping: Results from a nationally representative survey during the COVID-19 pandemic. *Public Health Nutrition*, 25(11), 3079-3085. doi:10.1017/S1368980022001756

East, R. (2022). Online grocery sales after the pandemic. *International Journal of Market Research*, 64(1), 13-18. doi:10.1177/14707853211055047

- Ellison, B., McFadden, B., Rickard, B. J., & Wilson, N. L. W. (2021). Examining food purchase behavior and food values during the COVID-19 pandemic. *Applied Economic Perspectives and Policy*, 43(1), 58-72. doi:10.1002/aep.13118
- Gao, X., Shi, X., Guo, H., & Liu, Y. (2020). To buy or not buy food online: The impact of the COVID-19 epidemic on the adoption of e-commerce in China. *PloS One*, 15(8), e0237900. doi:10.1371/journal.pone.0237900
- German, J. D., Ong, A. K. S., Redi, A. A. N. P., & Robas, K. P. E. (2022). Predicting factors affecting the intention to use a 3PL during the COVID-19 pandemic: A machine learning ensemble approach. *Heliyon*, 8(11)
- Grashuis, J., Skevas, T., & Segovia, M. S. (2020). Grocery shopping preferences during the COVID-19 pandemic. *Sustainability*, 12(13) doi:10.3390/su12135369
- Guthrie, C., Fosso-Wamba, S., & Arnaud, J. B. (2021). Online consumer resilience during a pandemic: An exploratory study of e-commerce behavior before, during and after a COVID-19 lockdown. *Journal of Retailing and Consumer Services*, 61, 102570. doi:10.1016/j.jretconser.2021.102570
- Hadler, N., Locher, I., Waselewski, M., Hoffs, C., Chang, T., & Brandt, E. (2021). Youth perspectives and experiences with in-person and online grocery shopping during the COVID-19 pandemic. *Current Developments in Nutrition*, 5, 224. doi:10.1093/cdn/nzab029_025
- Hamermesh, D. S. (2020). Life satisfaction, loneliness and togetherness, with an application to covid-19 lock-downs. *Review of Economics of the Household*, 18(4), 983-1000. doi:10.1007/s11150-020-09495-x

- Hawkins, D. (2020). Differential occupational risk for COVID 19 and other infection exposure according to race and ethnicity. *American Journal of Industrial Medicine*, 63(9), 817-820.
doi:10.1002/ajim.23145
- Heaney, J. (2007). Generations X and Y's internet banking usage in Australia. *Journal of Financial Services Marketing*, 11(3), 196-210. doi:10.1057/palgrave.fsm.4760052
- Holland, P. W., & Rubin, D. B. (1987). Causal inference in retrospective studies. *ETS Research Report Series*, 1987(1), 203-231. doi:10.1002/j.2330-8516.1987.tb00211.x
- Hoseini, M., & Valizadeh, A. (2021). The effect of COVID-19 lockdown and the subsequent reopening on consumption in Iran. *Review of Economics of the Household*, 19(2), 373-397.
doi:10.1007/s11150-021-09557-8
- Johnson, O., Cho, H., & Patwary, S. (2023). Online shopping orientations of Latino millennial generational cohorts. *Journal of Consumer Behaviour*, 22(1), 14-26. doi:10.1002/cb.2107
- Kim, H. E., Cho, Y., & Youn, N. (2021). COVID-19 uncertainty and temporal framing in advertising for online experiential consumption. *Journal of Advertising*, 50(3), 280-289.
doi:10.1080/00913367.2021.1927915
- Koch, J., Frommeyer, B., & Schewe, G. (2020). Online shopping motives during the COVID-19 pandemic – lessons from the crisis. *Sustainability (Basel, Switzerland)*, 12(24), 10247.
doi:10.3390/SU122410247
- Lo, A., Duffy, E., & Ng, S. W. (2021). Who's grocery shopping online and why: Cross-sectional analysis of a nationally-representative sample since the pandemic. *Current Developments in Nutrition*, 5, 231. doi:10.1093/cdn/nzab029_032

- Martin, C. A., Jenkins, D. R., Minhas, J. S., Gray, L. J., Tang, J., Williams, C., . . . Pareek, M. (2020). Socio-demographic heterogeneity in the prevalence of COVID-19 during lockdown is associated with ethnicity and household size: Results from an observational cohort study. *EClinicalMedicine*, 25, 100466. doi:10.1016/j.eclinm.2020.100466
- Mehta, S., Saxena, T., & Purohit, N. (2020). The new consumer behaviour paradigm amid COVID-19: Permanent or transient? *Journal of Health Management*, 22(2), 291-301. doi:10.1177/0972063420940834
- Meena, P., & Kumar, G. (2022). Online food delivery companies' performance and consumers expectations during covid-19: An investigation using machine learning approach. *Journal of Retailing and Consumer Services*, 68, 103052. doi:10.1016/j.jretconser.2022.103052
- Moon, J., Choe, Y., & Song, H. (2021). *Determinants of consumers' online/offline shopping behaviours during the COVID-19 pandemic* MDPI AG. doi:10.3390/ijerph18041593
- Morales, D. X., Morales, S. A., & Beltran, T. F. (2021). Racial/ethnic disparities in household food insecurity during the COVID-19 pandemic: A nationally representative study. *Journal of Racial and Ethnic Health Disparities*, 8(5), 1300-1314. doi:10.1007/s40615-020-00892-7
- Punj, G. (2012). Income effects on relative importance of two online purchase goals: Saving time versus saving money? *Journal of Business Research*, 65(5), 634-640. doi:10.1016/j.jbusres.2011.03.003
- Ren, X., Cao, J., Xu, X., & Gong, Y. (. (2021). A two-stage model for forecasting consumers' intention to purchase with e-coupons. *Journal of Retailing and Consumer Services*, 59(59), 102289. doi:10.1016/j.jretconser.2020.102289
- Rosenbaum, P. R. (2006). The central role of the propensity score in observational studies for causal effects. *Matched sampling for causal effects* (pp. 170-184) Cambridge University

Press. doi:10.1017/CBO9780511810725.016 Retrieved
from <http://dx.doi.org/10.1017/CBO9780511810725.016> <http://dx.doi.org/10.1017/CBO9780511810725.016>

- Schlegelmilch, B. B., Sharma, K., & Garg, S. (2022). Employing machine learning for capturing COVID-19 consumer sentiments from six countries: A methodological illustration. Bradford, England] : *MCB University Press*. doi:10.1108/IMR-06-2021-0194
- Shaw, N., Eschenbrenner, B., & Baier, D. (2022). Online shopping continuance after COVID-19: A comparison of canada, germany and the united states. *Journal of Retailing and Consumer Services*, 69, 103100. doi:10.1016/j.jretconser.2022.103100
- Showrav, D. G. Y., Hassan, A., Anam, S., & Chakrabarty, A. K. (2021). Factors influencing the rapid growth of online shopping during covid-19 pandemic time in Dhaka city, Bangladesh. *Academy of Strategic Management Journal, Special Issue 2*, 20, 1-13.
- Sumi, R. S., & Ahmed, M. (2022). Investigating young consumers' online buying behavior in COVID-19 pandemic: Perspective of Bangladesh. *IIM Ranchi Journal of Management Studies*, 1(2), 108-123. doi:10.1108/IRJMS-09-2021-0127
- Sze, S., Pan, D., Nevill, C. R., Gray, L. J., Martin, C. A., Nazareth, J., . . . Pareek, M. (2020). Ethnicity and clinical outcomes in COVID-19: A systematic review and metaanalysis. *EClinicalMedicine*, 29-30, 100630. doi:10.1016/j.eclinm.2020.100630
- Taylor, C. R. (2020). The urgent need for more research on influencer marketing. *International Journal of Advertising*, 39(7), 889-891. doi:10.1080/02650487.2020.1822104
- United States Census Bureau. (2022). Annual retail trade survey 2020. Retrieved from <https://www.census.gov/data/tables/2020/econ/arts/annual-report.html> Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects

using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242. doi:10.1080/01621459.2017.1319839

Wagner, G., Schramm-Klein, H., & Steinmann, S. (2017). Consumers' attitudes and intentions toward internet-enabled TV shopping. *Journal of Retailing and Consumer Services*, 34, 278-286. doi:10.1016/j.jretconser.2016.01.010

Watanabe, T., & Omori, Y. (2020). Online consumption during the COVID-19 crisis: Evidence from Japan. *The Impact of COVID-19 on E-Commerce*, doi:10.51432/978-1-8381524-8-2_4

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37(23), 3309-3324. doi:10.1002/sim.7820

Missing Data Imputation – A Two-Phase Model

Di Wu, Lehman College, CUNY
Hsien Tseng Wang, Lehman College, CUNY

May 2024

1 Abstract

In the domain of business analytics, data science, and information science, the challenge of addressing missing data is ever-present and often complicated to both scholars and practitioners. This paper introduces a two-phase model designed to navigate this challenge with precision and clarity. In phase 1, a supervised approach is deployed to comprehensively understand the unique characteristics of the specific dataset. This initial phase conducts an exhaustive evaluation of existing missing data imputation methods, including both traditional statistical techniques and advanced machine learning-based approaches. These methods are rigorously assessed to determine their applicability to the dataset's specific attributes. In phase 2, the selected methods from the previous phase are applied to the dataset for the actual missing data imputation. The contribution of this two-phase model lies in its practicality and its ability to resolve the common dilemma faced by practitioners when confronted with a myriad of imputation methods. By adhering to a structured and data-driven approach, this model minimizes the uncertainty and complexity that often surround the method selection process. The outcome is the confident and informed selection of the most effective imputation method fitted to the unique characteristics of the given dataset. In doing so, this approach contributes not only to enhanced clarity but also to increased effectiveness in the domain of missing data imputation, thus strengthening the foundations of sound data science and rigorous research practices.

2 Introduction

Missing data is an issue in research across various fields, ranging from epidemiology and clinical trials to social sciences and finance. The presence of missing values in datasets can compromise the validity and reliability of statistical analyses, leading to biased estimates and conclusions. Addressing missing data effectively is therefore essential for producing accurate and robust research findings. In recent years, there has been growing interest in developing advanced methodologies for handling missing data, with a focus on automating the selection of appropriate imputation methods to streamline the data preprocessing pipeline. This paper introduces a two-phase model for automating the selection of optimal imputation methods, aimed at addressing the challenges associated with missing data and promoting evidence-based approaches to data analysis.

Our paper makes several contributions to the field of missing data imputation. First, we systematically outline the nature of missing data and the challenges associated with handling incomplete datasets in research. By discussing the different types of missing data and their implications, we provide a comprehensive understanding of the complexity of the problem. Second, we introduce a two-phase model for automating the selection of optimal imputation methods. This model involves systematically exploring the space of possible imputation methods and hyperparameters in the first phase, followed by the application of the identified best-performing imputation method to the dataset in the second phase. Third, we shed light on the automation of the imputation process, highlighting the potential benefits of removing human bias and promoting evidence-based approaches to method selection. Finally, our paper lays the groundwork for future experiments to evaluate the effectiveness and efficiency of the proposed two-phase model in real-world research settings.

Our paper is organized into four main sections. We begin by discussing the nature of missing data, including the various types of missingness and their implications for statistical analyses. Next, we review existing imputation methods, ranging from traditional approaches such as mean imputation and regression imputation to more advanced techniques such as multiple imputation and deep learning-based methods. By systematically evaluating

the strengths and limitations of each method, we provide insights into the landscape of missing data imputation techniques. Following this, we introduce our proposed twophase model for automating the selection of optimal imputation methods. This model aims to improve the reliability of statistical analyses by leveraging automated model selection algorithms and evidence-based approaches. Finally, we conclude the paper by discussing directions for future research, including experiments to evaluate the effectiveness of the proposed model in real-world research settings and potential extensions to further enhance missing data imputation methodologies.

3 Missing Data

Missing data refers to the absence of values in a dataset that were expected to be present. It occurs when no data value is stored for the variable in a particular observation. There are all kinds of reasons and scenarios that missing data occurs. We summarize the types of scenarios of missing, the types of data missing, and the types of dependencies of missing data and existing data in this section.

3.1 Types of Missingness

Missing Completely at Random (MCAR) MCAR is a fundamental concept in the study of missing data, characterized by the absence of any systematic relationship between the missingness of data and observed or unobserved variables within the dataset. In other words, the probability of data being missing is unrelated to both the observed values and the unobserved values of the variables under study. This means that any missingness in the data occurs purely by chance, without any discernible pattern or reason [Rubin, 1976]. MCAR is often considered the most favorable scenario for handling missing data because it simplifies the assumptions underlying imputation methods and statistical analyses. Under the MCAR assumption, the observed data are still representative of the population from which they were sampled, even in the presence of missing values. Thus, complete case analysis (CCA) or listwise deletion, where cases with any missing values are excluded from the analysis, can yield unbiased estimates if data are MCAR [Allison, 2003]. Despite its theoretical simplicity and attractiveness for statistical analysis, verifying whether data are truly missing completely at random can be challenging in practice. Researchers often resort to statistical tests, such as Little's MCAR test [Little, 1988], to assess the plausibility of the MCAR assumption. Additionally, sensitivity analyses and robustness checks are recommended to evaluate the robustness of findings under different missing data assumptions, including MCAR.

Missing at Random (MAR) MAR is another critical concept in the analysis of missing data, where the probability of missingness depends on observed variables within the dataset but not on the missing values themselves. In other words, the missingness mechanism is related to the observed data but not to the unobserved data. MAR assumes that conditional on the observed variables, the probability of data being missing is constant across different levels of the missing variable [Rubin, 1976]. MAR is a common and less restrictive assumption compared to MCAR, as it allows for systematic patterns in missingness that can be accounted for in statistical analyses. Under the MAR assumption, the missing data are not a random sample of all data, but the observed data remain representative of the population conditional on the variables used to predict missingness. This makes MAR amenable to various imputation techniques, such as regression imputation and multiple imputation, which utilize observed information to impute missing values. However, it's essential to carefully consider the assumptions underlying MAR and assess their validity in the context of the specific dataset and research question to ensure unbiased and accurate results.

Missing Not at Random (MNAR) MNAR represents a more complex and challenging scenario compared to MCAR and MAR, where the probability of missingness is related to the unobserved data itself, even after conditioning on observed variables. In MNAR, the missingness mechanism depends on information that is not included in the dataset, making it non-ignorable and potentially biasing statistical analyses [Rubin, 1976]. Handling MNAR data requires careful consideration and often involves the use of specialized imputation techniques that incorporate information about the missingness mechanism. However, accurately identifying and modeling the missingness

mechanism in MNAR data can be challenging, as the information needed to fully characterize the missingness is typically unavailable.

3.2 Types of Data Missing

Numerical Data Missing numerical data refers to situations where numeric values are absent or incomplete in a dataset. This type of missing data can arise for various reasons, such as equipment malfunction, human error in data entry, or intentional non-response by survey participants.

Categorical Data Missing categorical data refers to instances where categorical values are absent or incomplete in a dataset. This can occur due to various reasons such as survey non-response, data entry errors, or intentional omission.

Ordinal Data Missing ordinal data refers to situations where values within ordinal variables are absent or incomplete in a dataset. Ordinal data represent categories with a natural ordering, such as levels of education or Likert scale responses.

3.3 Dependencies between Missing and Existing Data

Understanding dependencies between missing and existing data is crucial for the effective handling of missing data in statistical analysis. Missing data mechanisms can be influenced by the values of other variables in the dataset, leading to dependencies between the presence of missing values and observed data. This phenomenon has been extensively studied in the literature, with various methods proposed to account for dependencies in missing data imputation [Graham et al., 2007, Sterne et al., 2009, Azur et al., 2011].

4 Imputation Methods

4.1 Statistical Methods

Mean/Median Imputation Mean/Median Imputation is a commonly used method for handling missing data, particularly for numerical variables. In this approach, missing values are replaced with the mean or median of the observed values for that variable. Mean imputation involves replacing missing values with the arithmetic average of the observed values, while median imputation replaces missing values with the middle value of the observed values when arranged in ascending order. Mean/Median Imputation is straightforward to implement and can help preserve the overall distribution of the data. However, it may lead to biased estimates, particularly if missingness is related to the underlying distribution of the variable. Despite its limitations, Mean/Median Imputation remains a widely used method due to its simplicity and computational efficiency [Graham et al., 2007].

Mode Imputation Mode Imputation is a straightforward method for handling missing categorical data by replacing missing values with the mode, or most frequent category, of the observed values for that variable. This approach is commonly used when dealing with categorical variables where there is a clear majority category. Mode imputation is simple to implement and preserves the categorical structure of the data. However, it may not be suitable for variables with multiple equally frequent categories or when the mode does not accurately represent the underlying distribution of the data. Despite its limitations, Mode Imputation remains a popular choice for handling missing categorical data due to its simplicity and ease of implementation [Azur et al., 2011].

Hot Deck Imputation Hot deck imputation is a method commonly used for handling missing data, particularly in survey research, where missing values are replaced with observed values from similar cases or "donors." In hot deck imputation, a "donor" is selected from the cases with complete data based on similarity measures, such as propensity scores or Mahalanobis distances, to the case with the missing value. The observed value from the selected donor is then imputed for the missing value. This approach aims to preserve the distribution and

relationships present in the observed data while filling in missing values. Hot deck imputation can be particularly useful when there is a considerable amount of missing data and when the missingness mechanism is related to the observed data. However, it requires careful consideration of how to define similarity between cases and how to select donors to ensure the validity of imputed values

[Little, 1988].

Cold Deck Imputation Cold deck imputation, a method used to handle missing data, involves replacing missing values with predetermined or externally sourced values. Unlike hot deck imputation, where donors are selected from the existing dataset, cold deck imputation relies on values from external sources or historical data to fill in missing values. This approach is commonly used when no similar cases are available in the dataset or when the structure of the data prevents the use of hot deck imputation. While cold deck imputation can provide a quick solution for handling missing data, it may introduce bias if the externally sourced values do not accurately represent the missing values in the current dataset. Careful consideration of the appropriateness and relevance of the external data source is necessary to ensure the validity of imputed values [Sterne et al., 2009].

4.2 Machine Learning Methods

Regression Imputation Regression imputation is a widely used method for handling missing data, particularly when the missing values are believed to be related to other variables in the dataset. In this approach, a regression model is constructed using the observed data, with the variable containing missing values treated as the dependent variable and other relevant variables serving as predictors. The model is then used to predict the missing values based on the observed values of the predictor variables. Regression imputation leverages the relationships between variables to estimate missing values, thereby preserving the structure and patterns present in the observed data. While regression imputation is commonly employed due to its flexibility and ability to handle complex missing data patterns, it relies on the assumption that the relationship between the variables used in the regression model accurately represents the missingness mechanism [Allison, 2003].

Multiple Imputation Multiple imputation is a powerful technique for handling missing data that involves generating multiple plausible values for each missing observation based on the observed data and the assumed underlying distribution. Each imputed dataset is analyzed separately, and the results are combined using specific rules to produce valid parameter estimates and standard errors that appropriately account for the uncertainty introduced by imputation. Multiple imputation offers several advantages over single imputation methods, including the ability to incorporate variability in the imputed values, thereby providing more accurate estimates of uncertainty. It also allows for the inclusion of auxiliary variables to improve imputation accuracy and flexibility in handling complex missing data patterns. While multiple imputation requires careful consideration of the imputation model and convergence diagnostics, it is widely regarded as a flexible and robust approach for handling missing data in practice [Rubin and Schenker, 1986, Rubin and Schenker, 1991, Rubin, 2018].

K-nearest Neighbors Imputation K-nearest neighbors (KNN) imputation is a non-parametric method used for handling missing data by imputing missing values based on the values of the nearest neighbors in the dataset. In this approach, the missing value is replaced with the average of the values of its k-nearest neighbors, where similarity between data points is typically measured using distance metrics such as Euclidean distance or cosine similarity. KNN imputation is particularly useful when the dataset contains both numerical and categorical variables, as it can handle mixed data types effectively. This method offers simplicity and flexibility in handling missing data without making strong distributional assumptions, making it applicable to a wide range of datasets. However, the choice of the value of k and the distance metric can significantly impact imputation accuracy, and careful selection and tuning of these parameters are necessary for optimal performance [Troyanskaya et al., 2001, Schwender, 2012].

Expectation-Maximization (EM) Algorithm The Expectation-Maximization (EM) algorithm is a powerful iterative method commonly used for handling missing data by estimating the parameters of a statistical model in the presence of incomplete data. The algorithm alternates between two main steps: the expectation (E) step, where the missing

data are imputed based on the current estimates of the model parameters, and the maximization (M) step, where the model parameters are updated based on the observed and imputed data. This iterative process continues until convergence is reached, resulting in maximum likelihood estimates of the model parameters. The EM algorithm offers a flexible and computationally efficient approach for handling missing data in a wide range of statistical models, including mixture models and latent variable models [Dempster et al., 1977]. However, the EM algorithm relies on the assumption that the data are missing at random, and its performance can be sensitive to the initial parameter values and convergence criteria, necessitating careful implementation and validation [Enders, 2001, Arminger and Sobel, 1990].

Bayesian Imputation Bayesian imputation is a principled approach for handling missing data that leverages Bayesian inference to estimate the missing values based on the observed data and prior information about the underlying data distribution. In this method, a Bayesian model is specified to represent the joint distribution of the observed and missing data, along with any relevant covariates or auxiliary variables. The missing values are then imputed by drawing samples from the posterior distribution of the missing data given the observed data and the specified model. Bayesian imputation offers several advantages, including the ability to incorporate uncertainty in the imputed values and the flexibility to incorporate prior knowledge about the data generating process. Additionally, Bayesian imputation allows for the propagation of uncertainty from the imputed values to subsequent analyses, leading to more accurate inference [Schafer, 1997]. However, Bayesian imputation requires careful specification of the prior distributions and model assumptions, and computational challenges may arise when dealing with complex models or large datasets [Ma and Chen, 2018, Kong et al., 1994].

Decision-Trees-Based Imputation Decision-trees-based imputation is an innovative approach for handling missing data that utilizes decision-tree algorithms to predict missing values based on the observed data. In this method, a decision tree is constructed using the observed data as input features, and the missing values are imputed by traversing the decision tree to predict the missing values based on the observed values of other variables. Decision-trees-based imputation offers several advantages, including the ability to handle both numerical and categorical variables, as well as interactions between variables, without making strong distributional assumptions [Little and Rubin, 2019]. Additionally, decision-trees-based imputation can handle non-linear relationships and complex data structures, making it suitable for a wide range of datasets. However, careful consideration should be given to the selection of appropriate decision tree algorithms and tuning of parameters to ensure optimal imputation accuracy and generalization performance [Rahman and Islam, 2011, Rahman and Islam, 2013].

4.3 Deep Learning Methods

Deep learning methods for missing data imputation have emerged as powerful techniques for handling missing values in complex datasets. These methods leverage the capacity of deep neural networks to learn intricate patterns and relationships from the observed data to impute missing values effectively. Various deep learning architectures, such as autoencoders, generative adversarial networks (GANs), and recurrent neural networks (RNNs), have been adapted for missing data imputation tasks [Lee et al., 2019, Chakraborty et al., 2021]. Deep learning methods for missing data imputation offer several advantages, including the ability to handle high-dimensional data, nonlinear relationships, and missing patterns, without relying on strong distributional assumptions [Che et al., 2018]. However, these methods may require large amounts of data and computational resources for training, and careful validation is necessary to ensure robust performance in diverse datasets.

5 Two-Phase Model

5.1 Concept and Rational

Choosing the appropriate methods for missing data imputation for specific datasets can be challenging due to the diverse nature of missing data patterns and the complexity of statistical models. The difficulty lies in selecting

methods that not only account for the missing data mechanism but also preserve the underlying structure and relationships present in the data. Several factors contribute to this challenge, including the type of missing data, the distributional properties of the data, and the presence of interactions or dependencies between variables. Moreover, the performance of imputation methods can vary depending on the characteristics of the dataset, such as the amount of missingness, the presence of outliers, and the dimensionality of the data. Researchers must carefully evaluate the strengths and limitations of different imputation methods and consider the specific characteristics of their dataset to make informed decisions [Sterne et al., 2009, Azur et al., 2011].

The automation of finding the most appropriate imputation methods for specific datasets has garnered significant attention in recent years, given the increasing complexity and size of datasets encountered in various fields. Machine learning techniques, such as automated model selection algorithms and ensemble methods, have been proposed to streamline the process of selecting optimal imputation methods. These approaches leverage the computational power of algorithms to systematically evaluate and compare the performance of different imputation methods across diverse datasets, considering factors such as missing data patterns, data distribution, and model complexity [Che et al., 2018].

Following this trend, we would like to construct a two-phase model that automatically explores the space of possible imputation methods and hyperparameters so we can identify and then apply the best-performing imputation method.

5.2 Phase One: Identifying

Phase one involves exploring the space of possible imputation methods and hyperparameters to identify the best-performing model for handling missing data. To understand the nature of missing values, we use the attributes that have missing values as the dependent attributes for a supervised study. This process typically entails systematically evaluating a range of imputation techniques, such as mean imputation, regression imputation, and multiple imputation, along with various hyperparameters that govern the behavior of these methods. Automated model selection algorithms, such as grid search and Bayesian optimization, are commonly employed to efficiently search through the space of possible combinations and identify the optimal model configuration based on predefined performance metrics, such as imputation accuracy or predictive performance [Che et al., 2018, Van Buuren and Oudshoorn, 2000]. By systematically exploring the space of imputation methods and hyperparameters, researchers can identify the most effective approach for imputing missing data in their specific dataset, ensuring robust and reliable results.

5.3 Phase Two: Implementing

Phase Two involves applying the identified best-performing imputation method, determined in Phase One, to the dataset containing missing values. Once the optimal imputation method and corresponding hyperparameters have been selected through systematic exploration of the imputation space, the chosen method is applied to impute missing values in the dataset. This step ensures that the missing data are accurately and effectively handled according to the selected approach, thus preparing the dataset for subsequent analyses. By implementing the best-performing imputation method identified in Phase One, researchers can mitigate bias and uncertainty associated with missing data, ultimately improving the validity and reliability of study results.

6 Discussion

This two-phase model involves automating the exploration of the space of possible imputation methods and hyperparameters to identify the best-performing model for handling missing data. This automated process significantly lowers the barrier of making the correct choice by systematically evaluating a range of imputation techniques and hyperparameters without requiring extensive manual intervention. By removing human bias from the decision-making process, this approach ensures an objective and evidence-based selection of the optimal imputation method. Automated model selection algorithms, such as grid search and Bayesian optimization, efficiently search through the space of possible combinations to identify the most effective approach for imputing

missing data. By adopting this evidence-based approach, researchers can rely on empirical data and performance metrics to guide their decision-making, ultimately improving the validity and reliability of study results.

While the two-phase model for handling missing data offers several advantages, it also presents certain disadvantages. One notable drawback is the potential time-consuming nature of the process, particularly during Phase One, where systematic exploration of the imputation space and hyperparameter tuning may require significant computational resources and time. Additionally, the computational complexity of identifying the best-performing imputation method and hyperparameters can pose challenges, especially for large and highdimensional datasets. Furthermore, the implementation of the identified imputation method in Phase Two may also incur computational costs, particularly for computationally intensive techniques such as deep learning-based imputation methods [Che et al., 2018, Van Buuren and Oudshoorn, 2000]. Thus, while the two-phase model offers a robust and systematic approach to handling missing data, researchers must consider the potential drawbacks, including time constraints and computational requirements.

A possible experiment to evaluate the effectiveness of the proposed twophase model for automating missing data imputation could involve the following steps. Firstly, various datasets could be generated based on different types of missingness and data characteristics. These datasets would simulate real-world scenarios with varying degrees of missing data. Secondly, hypotheses could be developed based on the characteristics of the datasets and existing literature on imputation methods. These hypotheses would predict the performance of different imputation methods under different missing data scenarios. Thirdly, the two-phase model would be applied to the generated datasets to systematically explore the space of possible imputation methods and hyperparameters. The model would identify the optimal imputation method for each dataset based on predefined performance metrics, such as imputation accuracy or predictive performance. Finally, the "best method" as determined by theory and the "optimal method" identified by the two-phase model would be compared to evaluate the effectiveness of the proposed approach. This comparison would provide insights into the ability of the two-phase model to automate the selection of optimal imputation methods and improve the reliability of missing data imputation in real-world research settings.

7 Conclusion

In conclusion, the two-phase model outlined in this paper provides a systematic framework for automating the selection of optimal imputation methods, thereby lowering the barrier to making the correct choice and removing human bias. By leveraging automated model selection algorithms and evidence-based approaches, researchers can improve the validity and reliability of their analyses, ultimately enhancing the robustness of study results. Future research efforts may focus on developing more efficient algorithms and strategies for automating the selection of imputation methods, thereby further streamlining the process and enhancing the accessibility of advanced imputation techniques for researchers across diverse domains. Through continued advancements in missing data imputation methodologies, researchers can overcome the limitations posed by incomplete datasets and produce more reliable and insightful analyses in various fields.

Disclaimer

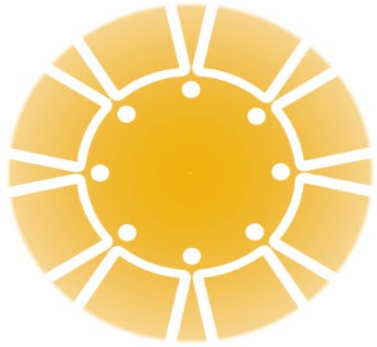
Please note that the authors utilized ChatGPT to polish the writing style to improve the readability of the article.

References

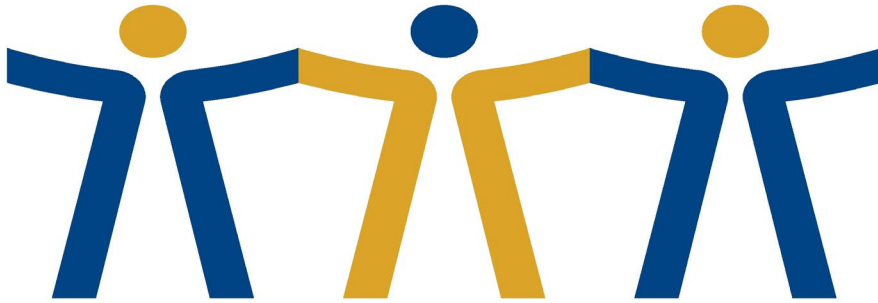
- [Allison, 2003] Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology*, 112(4):545.
- [Arminger and Sobel, 1990] Arminger, G. and Sobel, M. E. (1990). Pseudomaximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, 85(409):195–203.

- [Azur et al., 2011] Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.
- [Chakraborty et al., 2021] Chakraborty, H., Samanta, P., and Zhao, L. (2021). Sequential data imputation with evolving generative adversarial networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- [Che et al., 2018] Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.
- [Enders, 2001] Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8(1):128–141.
- [Graham et al., 2007] Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, 8:206–213.
- [Kong et al., 1994] Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288.
- [Lee et al., 2019] Lee, D., Kim, J., Moon, W.-J., and Ye, J. C. (2019). Collagan: Collaborative gan for missing image data imputation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2487–2496.
- [Little and Rubin, 2019] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- [Little, 1988] Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.
- [Ma and Chen, 2018] Ma, Z. and Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, 47:297–313.
- [Rahman and Islam, 2011] Rahman, M. G. and Islam, M. Z. (2011). A decision tree-based missing value imputation technique for data pre-processing. In *The 9th Australasian Data Mining Conference: AusDM 2011*, pages 41–50. Australian Computer Society Inc.
- [Rahman and Islam, 2013] Rahman, M. G. and Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowledge-Based Systems*, 53:51–65.
- [Rubin, 1976] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- [Rubin, 2018] Rubin, D. B. (2018). Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC.
- [Rubin and Schenker, 1986] Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American statistical Association*, 81(394):366–374.
- [Rubin and Schenker, 1991] Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in medicine*, 10(4):585–598.
- [Schafer, 1997] Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

- [Schwender, 2012] Schwender, H. (2012). Imputing missing genotypes with weighted k nearest neighbors. *Journal of Toxicology and Environmental Health, Part A*, 75(8-10):438-446.
- [Sterne et al., 2009] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338.
- [Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520-525.
- [Van Buuren and Oudshoorn, 2000] Van Buuren, S. and Oudshoorn, C. G. (2000). Multivariate imputation by chained equations.



NORTHERN CHAUTAUQUA
**COMMUNITY
FOUNDATION**



*Today.
Tomorrow.
Together.*

The campaign for northern Chautauqua

NORTHERN CHAUTAUQUA COMMUNITY FOUNDATION



FREDONIA

STATE UNIVERSITY OF NEW YORK

