



Handling Imbalanced Data With Weighted Logistic Regression and Propensity Score Matching methods: The Case of P2P Money Transfers


Lavlin Agrawal, North Carolina Agricultural and Technical State University, USA*

 <https://orcid.org/0000-0001-8877-2336>

Pavankumar Mulgund, University of Memphis, USA

 <https://orcid.org/0000-0001-8434-5070>

Raj Sharman, University at Buffalo, USA

 <https://orcid.org/0000-0001-5838-7330>

ABSTRACT

The adoption of empirical methods for secondary data analysis has witnessed a significant surge in IS research. However, the secondary data is often incomplete, skewed, and imbalanced at best. Consequently, there is a growing recognition of the importance of empirical techniques and methodological decisions made to navigate through such issues. However, there is not enough methodological guidance, especially in the form of a worked case study that demonstrates the challenges of imbalanced datasets and offers prescriptive on how to deal with them. Using data on P2P money transfer services, this article presents a running example by analyzing the same dataset using several different methods. It then compares the outcomes of these choices and explicates the rationale behind some decisions such as inclusion and categorization of variables, parameter setting, and model selection. Finally, the article discusses certain regressions models such as weighted logistic regression and propensity matching, and when they should be used.

KEYWORDS

Adoption and Use, Bank-backed P2P, Imbalanced Data, Methodological Decisions, Propensity Match, Rare Event, Weighted Logistic Regression

INTRODUCTION

With the increasing availability of large volumes of publicly available secondary data, the empirical analysis of such data has gained increasing relevance and importance in information systems (IS) research (Black et al., 2020). Secondary data analysis also aligns well with the positivist research paradigm, which is the most dominant research approach within the IS community (Burton-Jones & Lee, 2017). Furthermore, there is an increasing expectation of obtaining data from multiple

DOI: 10.4018/JDM.335888

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

sources to publish research, making the use of secondary data even more relevant. Prior research has also highlighted several benefits of using secondary data, including (a) the reduction of bias that is sometimes introduced in qualitative approaches such as case studies (Choy, 2014); (b) the lack of intrusiveness that is associated with other methods, such as action research and interviews (Rabinovich & Cheon, 2011); (c) the absence of issues, such as survey fatigue (Sinickas, 2007); and (d) efficiency and cost-effectiveness of data procurement and use. With the emergence of reputable and highly credible secondary data sources and improved archival and management processes, the use of secondary data for empirical research is slated to grow even further (Black et al., 2020).

There are some limitations to the use of secondary data. A significant limitation is associated with the imbalanced nature of secondary data, particularly when the research study attempts to explore certain demographic factors or rare events. An imbalanced dataset occurs when the categories for classification are disproportionately represented (Ramayachitra & Manikandan, 2014). For example, in the case of the chosen dataset, if the number of instances of one class (consumer adopting peer-to-peer [P2P] services) is much smaller or larger than the number of instances of the other class (consumer not adopting P2P services), the dataset is said to be imbalanced. Traditional data analysis approaches often fall short when applied to such skewed data, necessitating the adoption of specialized empirical techniques and informed discretion on the part of researchers. Although there is growing recognition of the problem of imbalanced datasets in the IS research community post-COVID-19 pandemic (Dorn et al., 2021), there is insufficient methodological guidance in dealing with the challenge of highly skewed datasets.

We endeavor to address this gap by presenting an example of an empirical analysis of a highly imbalanced dataset. Following prior exemplars that offer methodological guidelines (Gefen et al., 2000; Chua & Storey, 2016), we bring to the fore a series of salient decisions the researchers must make while dealing with imbalanced data, including the selection and categorization of variables, choice of models to use, and parameters to set. Furthermore, we demonstrate how different decisions made during empirical analysis lead to diverse findings. We explore the suitability and use of propensity score matching (PSM) (Rosenbaum & Rubin, 1983) and weighted logistic regression (WLR) techniques (King & Zeng, 2001) to analyze imbalanced data. We also compare the results of the two models and elaborate on when it is appropriate to choose one model over the other.

For illustrative purposes, we make use of secondary data that consists of responses to a survey conducted by one of the top 25 banks in the northeast United States regarding the use of P2P money transfer services. The data are highly skewed, with only 5.4% of customers using bank-based P2P services. We used the responses to this survey in our study to empirically show and explain how methodological decisions impact outcomes. Furthermore, in this study, we use six research questions that are of interest to banks. We focus on demographic factors (age, gender, income, education, and employment status) and trust that can be harnessed for strategic business gain.

The remainder of the paper is organized as follows. The next section provides a methodological background on empirical modeling for imbalanced data with PSM and WLR. This is followed by a section addressing the contextual background of the case of P2P payments. and a section that discusses the research design and demonstrates the development of the research questions, data cleaning, and descriptive statistics of the data. We then provide a model analysis with various methods and decision choices made and follow with a section presenting the results from PSM and WLR combined. The final section discusses the conclusion and limitations of this study.

RELATED WORK

Challenges of Dealing With Imbalanced Data

Handling imbalanced datasets presents multiple challenges, primarily because standard algorithms often favor the majority class to maximize overall accuracy. This bias can result in models that

effectively predict outcomes for the majority class while neglecting the minority class (Mahani & Ali, 2019; Das et al., 2018). Traditional accuracy metrics are not reliable indicators in these scenarios. For example, a model that consistently predicts the majority class might show high accuracy yet fail to capture any insights about the minority class (Krawczyk, 2016). Furthermore, such models are prone to overfitting the majority class, which can compromise their performance on new, unseen data, particularly about the minority class (Spelman & Porkodi, 2018).

Dealing with imbalanced datasets requires careful consideration of several factors. For example, detecting the minority class—such as new technology adoption among people, fraud detection, and disease outbreaks—is crucial, as this is often more important than detecting the majority class. Furthermore, obtaining more data for the minority class can be expensive and time-consuming, and annotation errors exacerbate the imbalance. Such highly skewed data are also found in rare event modeling, such as examining rare diseases or medical conditions.

To address these issues, resampling techniques like oversampling the minority class or undersampling the majority class can also be employed, but they come with their own set of issues, such as overfitting due to replication of minority class instances or loss of potentially valuable information when removing instances from the majority class. Finally, with the rise of big data and data-driven decision-making, the impact of imbalanced data can be more profound. It can even reinforce existing biases if not adequately addressed. Therefore, it is essential to consider the impact of imbalanced data and take appropriate measures to address it.

Empirical Approaches for Handling Imbalanced Data

Many methodologies have been developed to model imbalanced data and are broadly classified into deductive methods, such as WLR (King & Zeng, 2001) and PSM (Rosenbaum & Rubin, 1983), and inductive methods, such as support vector machine (SVM) (Cortes & Vapnik, 1995) and neural network (Raj et al., 2016). Inductive methods are usually used in machine learning approaches, while deductive methods are used in hypothetico-deductive research.

Inductive Methods

Inductive methods (SVM and neural network) represent modern machine learning techniques that are particularly valuable in navigating the complexities of imbalanced data. An SVM is a supervised learning model that classifies objects by learning from examples. It aims to find a hyperplane in an N-dimensional space for clear data point classification, maximizing a specific mathematical function (Noble, 2006). When used for regression, termed a support vector regressor (SVR), it focuses on setting an error threshold (epsilon) adjustable for desired accuracy. However, complexity increases with numerous independent variables. The SVR model, depending on epsilon and variable count, forms complex, non-linear models. Unlike focusing on a single variable, SVM often requires considering combinations of variables. A neural network iteratively learns by evaluating records, predicting outcomes, and adjusting weights for incorrect predictions until prediction accuracy improves or stopping criteria are met. Neural networks are often considered “black boxes” due to their complex nature and high demands for development time and computational power. This complexity can deter researchers from seeking easily interpretable models (Féraud & Clérot, 2002). In summary, SVM can deal with unbalanced data by assigning a larger penalty for misclassification of the minority class during training.

These approaches do not offer a straightforward interpretation of the individual variables in the models but encapsulate a deeper, more intricate understanding. This lack of transparency, however, does not diminish their efficacy. This is a sign of their sophisticated analytical capacity, simultaneously processing multiple variables to reveal patterns and insights that are not immediately apparent through conventional analysis. Nevertheless, one limitation of inductive methods is that they consider the combined effect of multiple variables in the model rather than attempting to explain the specific

influence of individual variables. The complexity and interdependencies of the variables within these models make them less straightforward to interpret compared to more transparent methods.

Deductive Methods

On the other side of the spectrum are deductive approaches. Deductive methods offer the ability to dissect and explain the impact of individual variables on models. Logistic regression is an essential technique for analyzing and classifying binary and proportional response datasets and can be extended to situations involving outcome variables with more than two categories (Wright, 1995; Maalouf, 2011; Hosmer et al., 2013). However, logistic regression results are inconsistent with imbalanced or rare event data (Maalouf, 2011) and tend to be biased toward the majority event (King & Zeng, 2001). Certain corrections must be applied to logistic regression for bias correction. King and Zeng (2001) introduced “prior correlation” and “weighting” for estimate correction. WLR assigns weights to compensate for the difference between sample and population (King & Zeng, 2001). Weighting penalizes the model less for errors made on the majority events and more for errors made on rare events. Winship and Radbill (1994) explained why sampling weights are used and guided how and when to use them. WLR could produce reliable results if the collected sample accurately represents the population. However, if the collected sample suffers from any biases, the weighted logistic results may amplify the bias, as it will be multiplied by the weights and produce unrealistic results. Table 1 provides some of the instances in which WLR was used.

Why Did We Focus on Deductive Methods?

Deductive methods provide a clear, interpretable framework, making them particularly useful for studies in which understanding the role and significance of each factor is crucial. Since the clarity and comprehension of individual variable impacts are paramount in research, we focus our attention on deductive methods. This strategic decision aligns with our goal to not only predict outcomes but also to understand their underlying mechanics while dealing with imbalanced datasets. Deductive approaches serve as ideal tools for this endeavor, allowing us to demonstrate the complexities of dealing with our data comprehensibly and methodically. Specifically, we use WLR and PSM techniques because these methods have been widely adopted as pivotal tools in reducing biases, a utility thoroughly explored and documented in the existing literature (Wang, 2021). The application of PSM has expanded in recent years, solidifying its status as a go-to method for conducting robust analyses (Baser, 2006), while WLR has been recognized for producing dependable results with imbalanced data (Brydon et al., 2019).

PSM matches the treatment and the control groups by balancing the covariates such that both groups respond to intervention in the same manner as in the absence of other confounders. The goal is to reduce the imbalance in the empirical distribution of the pre-treatment confounders between the groups (Stuart, 2010). PSM is beneficial in an observational study where randomization of the experiment is impossible. PSM helps in creating well-matched treated and control groups. By lowering the imbalance, PSM helps improve the parameter estimates (Ho et al., 2007; Iacus et al., 2011). Dehejia and Wahba (2002) used PSM to reduce sample selection bias in a nonexperimental setting. Austin (2008) provided guidance on how to analyze and report the findings that employ PSM. In sampling the control records to match the treatment records, PSM discards many records, leading to information loss. Thus, this information loss may hamper the outcome.

Angrist and Pischke (2008) stated that a lack of standardization in implementing PSM may result in different conclusions, even with the same data. Baser (2006) demonstrated that the estimated cost of illness for asthma patients varies with the selection of propensity-matching algorithms; nearest neighbor, 2 to 1, radius, kernel, and stratified matching algorithms produced similar results. In contrast, Mahalanobis distance (Gorbani, 2019) and Mahalanobis with caliper algorithms (Olmus et al., 2022) produced very different results from those five matching algorithms. Table 2 provides some of the work that used PSM.

Table 1. Weighted logistic regression in research

Research	Objective	Comments
Laitinen (1999)	Predict a corporate credit analyst's risk estimate	This work used WLR and linear regression on the filed information from a credit information agency to predict the corporate analyst's risk estimate of a company.
Hsu et al. (2007)	Estimate the recurrence rate for a colorectal polyp prevention trial in which a participant might have a variable follow-up	This paper compared the logistic, weighted logistic, and Kaplan–Meier estimator. WLR outperformed the two.
Hu et al. (2007)	Predict the odds of head, face, or neck injuries during rollover crashes	This paper used WLR to provide evidence that unbelted occupants have significantly higher injury risks than belted occupants.
Zare et al. (2013)	Determine the risk factors for female breast cancer in a low socioeconomic population in Iran	Using the weighting method in “relogit,” this study found factors such as a positive history of ovarian cancer, hormone therapy, positive history of breast cancer in first relatives, and no history of OCP use to be significant predictors of breast cancer.
Guillen et al. (2018)	Predict the decision to purchase full coverage of motor insurance versus a basic insurance product	Using publicly available data sets, this work showcased that WLR is performing better in predicting the decisions to purchase full motor insurance coverage.

Table 2. Propensity score matching in research

Research	Objective	Comments
Czajka et al. (1992)	Demonstrate the benefit of propensity modeling on advanced data collected by the Internal Revenue Service	This paper showcased how using propensity modeling on advanced data can help in producing estimates very close to that of final data.
Dehejia and Wahba (2002)	Use PSM to accurately estimate the treatment effect in nonexperimental settings	This is one of the early research projects that implemented PSM in economics. This work used observational data from the National Supported Work Demonstration to estimate the effects of a training program.
Baser (2006)	Demonstrate how different matching algorithms may produce different results	This study compared the seven different algorithms to estimate the cost of asthma and confirmed that different matching techniques produce different results.
Caliendo and Kopeinig (2008)	Discuss the implementation issues and guide researchers on how to use PSM for evaluation purposes	This paper discussed the basic steps in implementing the PSM. It further guided the selection of the appropriate matching algorithm.
Austin (2008)	Provide five guidelines for analyzing and reporting studies that employ propensity-score matching	This work reviewed the 47 research studies in medical literature and provided recommendations for the design, analysis, and reporting of results while using PSM.
Shipman (2017)	Address the benefits and limitations of PSM and discuss the design choices available in PSM	A review of 86 articles in accounting journals demonstrated a substantial rise in the use of PSM in accounting.

CASE CONTEXT

Background

We demonstrate the methodological nuances of handling imbalanced data using P2P money transfer data. Earlier (even today), people transacted money among themselves using cash or checks. However, with the emergence of new technologies, customers can transact electronically. Electronic P2P payment service is an emerging area within the fintech domain that will likely impact how customers

transact (Agarwal & Zhang, 2020). An individual makes P2P transactions to transfer money to another entity using an online intermediary application, generally through a mobile device (Lara-Rubio et al., 2021). Earlier, only nonbank financial technology organizations (PayPal, Google, and Apple) offered electronic P2P money transfer services. However, in June 2017, an ally of traditional banks launched Zelle, an electronic P2P money transfer service. Zelle has emerged as a leading P2P application with support from more than 100 financial institutes and has changed the dynamics of P2P money transfer services.

According to a 2017 eMarketer forecast, the transaction value of all U.S. P2P payments will exceed \$156 billion in 2018 – rising to more than \$244 billion by 2021 (Van Dyke, 2022). A recent survey by the Aite Group (Groenfeldt, 2019) found that 57% of the respondents had made at least one P2P money transfer in the past year. Furthermore, the COVID-19 pandemic has created an environment where contactless payment systems are more of a necessity to prevent the spread of infections. Thus, banks are trying to promote the use of P2P systems; they are safe from a community health perspective and cost-effective from the banks' perspective (Humphrey et al., 2003). In this work, we study specific demographic factors that impact the use of P2P payment services, which are more important from a business perspective when devising tactical investment strategies.

Research Questions of Interest

Previous research has extensively studied the importance of ease of use, usefulness, self-efficacy, risk, social influence, and facilitation conditions (Fishbein & Ajzen, 1975; Ajzen, 1991; Davis et al., 1989; Venkatesh & Davis, 2000; Venkatesh et al., 2003; Venkatesh et al., 2012; Shaikh & Karjaluo, 2015). Today's P2P money transfer applications are easy to use and come with robust security. Consequently, fintech businesses are leveraging market demographics to formulate better investment strategies and attract more customers to use their P2P services. With this objective, one of the top 25 banks in the northeast United States conducted a survey in 2019. This survey was developed with a business and market orientation and did not include traditional variables that are part of the technology acceptance model (Davis, 1989) or the unified theory of acceptance and use of technology (UTAUT) (Venkatesh et al., 2003) models. The survey emphasized understanding demographic factors, such as gender, age, household income, number of children, educational attainment, and amount and type of investments made. Furthermore, as banks compete with fintech companies in P2P money transfer services, "trust in the banks" was included as a key question in this survey (Appendix A). This survey was conducted to answer the following research questions (RQs) so that better advertisement and promotion efforts could be established targeting a specific segment of customers.

- RQ1: How does age impact the use of P2P payment services?
- RQ2: How does gender impact the use of P2P payment services?
- RQ3: How does income impact the use of P2P payment services?
- RQ4: How does the extent of education impact the use of P2P payment services?
- RQ5: How does employment status impact the use of P2P payment services?
- RQ6: How does trust in a bank impact the use of its P2P payment services?

Age and gender have been extensively considered moderators in the adoption of technology. The UTAUT and UTAUT2 models depict how gender and age moderate the relationship between technology usage and its predictor variables. Studies have noted that younger people are more comfortable accepting technology than older people (Morris & Venkatesh, 2000; Venkatesh et al., 2003). On the other hand, some predictors of technology adoption, such as usefulness, have strong interactions with the male population, while ease of use and subjective norms have strong interactions with the female population (Venkatesh & Morris, 2000; Venkatesh et al., 2003). Laukkanen and Pasanen (2008) found age and gender to be differentiating factors between mobile banking and other

forms of banking. Age has been found to negatively influence the adoption and use of online banking (Polatoglu et al., 2001; Karjaluoto et al., 2002).

Katz et al. (2016) mentioned that technology adoption is informed and constrained by limited discretionary income. Polatoglu et al. (2001) and Karjaluoto et al. (2002) found that higher income was favorable to online banking use. Higher education leads to the adoption and use of online banking (Polatoglu et al., 2001). The findings of Karjaluoto et al. (2002) further reinforce this notion. However, research in China suggests that education level has no impact on the usage of online and mobile banking (Laforet & Li, 2005).

Karjaluoto et al. (2002) highlighted that occupation is one of the critical predictors of online banking usage, with white-collar employees more likely to adopt the technology. However, to our knowledge, no study has used “employment status (ranging from student to retired)” as the predictor variable.

Trust is crucial in a risky, uncertain, and interdependent environment (Mayer et al., 1995). The literature on e-commerce has pointed to trust as a major obstacle to its growth and adoption (Gefen et al., 2003). Yousafzai et al. (2005) mentioned that a high level of trust in banks does not translate to a bank’s digital services and how trust in digital services can be enhanced so that more customers will adopt these services. Alalwan et al. (2017) confirmed that trust is important in determining users’ likelihood of adopting mobile technologies. McKnight et al. (1998) categorized trust as an institution, personal and environmental, and defined “institution-based trust” as a user’s belief that impersonal structures support the user’s likelihood of success in a given situation. Institution-based trust is characterized by a firm’s size, capability, integrity, role in the market, benevolence, reputation, or brand (Oliveira et al., 2014). We used the responses from this survey as data for our work, with the main objective of explaining how to handle imbalanced data and the impact of such empirical decisions on the outcomes.

METHODOLOGICAL CONSIDERATIONS

Empirical Methods

From a methodological perspective, prior literature has also discussed two specific approaches: oversampling and undersampling. Undersampling is a widely used method for imbalanced data in which a rebalance between majority and minority classes is made by deleting records from majority observations (Gazzah et al., 2015). Several prior studies have utilized undersampling to address the problem of data imbalance (Lin et al., 2022; Koziarski, 2020; Hoyos-Osorio et al., 2021). However, two main questions arise while performing undersampling: (a) Which records should be preserved? and (b) How many records should be preserved? (Xie et al., 2021). Deleted records from the majority class can result in significant information loss in a model (Xie et al., 2021). Oversampling methods rely on generating synthetic minority records and adding them to the minority class to balance the majority and minority classes (Gazzah et al., 2015). Oversampling methods have been applied in prior research (Zhu et al., 2020; Xie et al., 2020; Wang et al., 2021). However, adding synthetic records to the minority class leads to the sink of the original minority record among the synthetic ones. This can potentially distort the classification results (Barua et al., 2012).

To overcome these issues, we analyze the data using two deductive methods for handling imbalanced data: (a) WLR (King & Zeng, 2001; Zare et al., 2013; Maalouf et al., 2018) and (b) PSM (Rosenbaum & Rubin, 1983; Ho et al., 2007; Iacus et al., 2011). Both methods are widely used in situations that generate imbalanced data, such as rare and extreme events, particularly in healthcare research. However, a lack of a standardized approach to implementing these methods may result in different conclusions, even with the same data (Angrist & Pischke, 2008). The impact of such empirical decisions within the umbrella of these two methods remains underexplored. Baser (2006) attempted to address these issues to some extent by demonstrating how the estimated cost of illness for asthma patients varies with the selection of propensity-matching algorithms.

Several decision choices are available for both methodologies. We compare the outcomes of different decisions and attempt to provide the rationale behind the choices and corresponding outcomes. Figure 1 provides a summary of the methodological decisions we made in this work. From this point on, the paper’s structure flows in the same sequence as decisions in a typical research process.

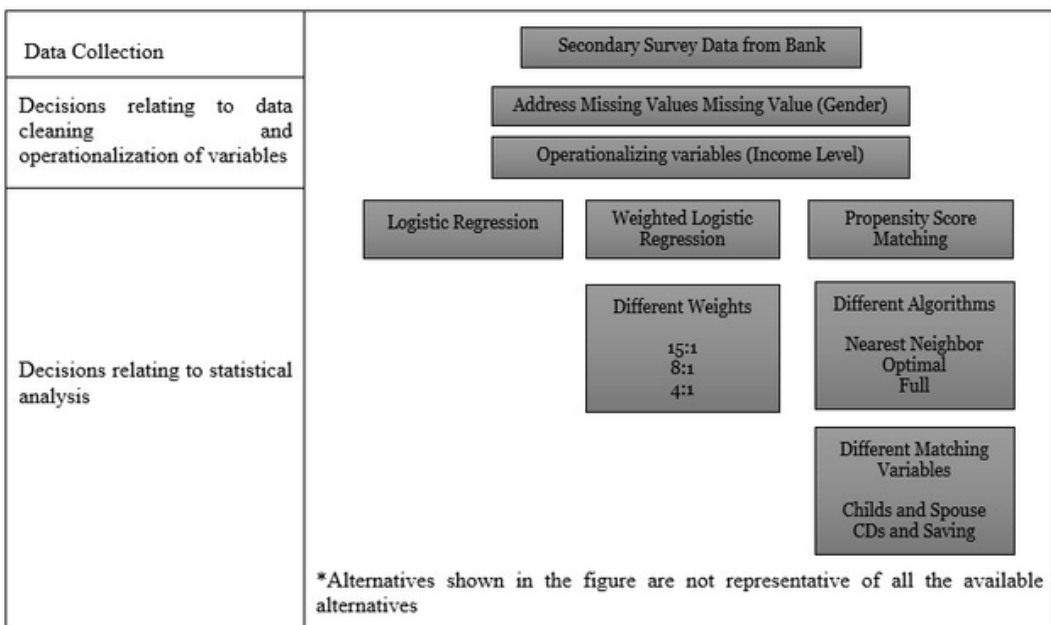
Data

This paper utilizes secondary data from a survey conducted in 2019 by one of the top 25 banks in the United States in terms of total assets and market capitalization regarding the use of P2P money transfer services. The data consist of 2,250 respondents who completed the survey (906 respondents were men with an average age of 54 years, and 1,338 were women with an average age of 49; 6 persons did not disclose their gender). Data are very skewed or imbalanced in terms of bank-based P2P, with 138 respondents using bank-backed P2P and 2,112 not using bank-backed P2P.

This dataset aligns perfectly with our research aims due to its inherent imbalances. This imbalanced nature reflects a real-world scenario wherein a relatively small segment of consumers fully adopts and utilizes P2P money transfer services. At the same time, a more significant majority may not engage with these services as extensively. Such disparity in user behavior is not only common in emerging financial technologies but is also a critical element that influences the applicability of relevant methods. This natural imbalance in the dataset presents an excellent opportunity to showcase the effectiveness of our methodological considerations. Our approach is designed explicitly to address and correct for an uneven distribution in the data.

The richness and granularity of the data are exceptional, with an extensive array of variables that facilitate the application of sophisticated methods. These data provide fertile ground for robust methodological exploration, enabling a detailed and comprehensive analysis that can yield rich, insightful conclusions. The chosen P2P service data are of high quality and reliability, with a large sample size of consumers with verified bank accounts from diverse demographic profiles of the typical American population. This reliable data is necessary for our statistical techniques, which increase our findings’ generalizability.

Figure 1. Decisions taken for data analysis



Although data collection was completed by the bank in late 2019, data were only made available to researchers in 2021 due to procedural prerequisites and due diligence activities. As service adoption is imbalanced, therefore, our selected dataset is a fitting testbed for our methodological scrutiny, which serves as an excellent use case through which we can examine and understand the methodological elements of dealing with imbalanced data.

Addressing Missing Data

The first methodological decision often relates to addressing the issue of missing values. Given that six records were missing gender values, the researcher must decide to either impute the missing values or delete their records. Prior research informs us that dropping records may hamper the statistical power of the model (Cohen, 1992). There are many imputation techniques, including imputing with the mean, minimum, maximum, highest frequency, and value from a regression model. However, each imputation technique may yield different results for the same data (Engels et al., 2003).

We found that all six persons with missing gender values did not use bank-backed P2P. Given the skewness and size of the dataset, we decided to drop those six records from the analysis. First, the impact of the deletion of these six records was minimal. Second, the imputing of gender could not be rationalized.

Further, we deleted 21 records of all non-Zelle users who did not disclose either education, employment, or relationship status (four did not disclose education, fifteen did not disclose employment status, and seven did not disclose relationship status). Finally, we included 2,223 records. Table 3 provides a descriptive statistic. It shows that the responses in the dataset for analysis consist of 138 bank-based P2P users and 2,085 not using bank-based P2P services, making it highly skewed for regular analysis. We ran the multicollinearity test before proceeding further (please refer to Appendix B).

Impact of Data Operationalization Decisions

Before a statistical analysis of the data, a researcher makes several assumptions relating to the operationalization of the data. The operationalization of input variables impacts the outcomes (Ada et al., 2012). We used “income” and “education” to illustrate the impact of operationalization. We used a simple model with few predictors and ran a logistic regression.

Operationalization of Income. In the survey, income has seven categories ranging from “none” to “more than \$200,000” based on the bank’s segregation rationale. Hence, we planned to classify these seven categories into three groups: “Low,” “Medium,” and “High,” as used in previous research. Pew Research has defined middle household income as ranging from about \$45,200–\$135,600 in 2016 (Bennett et al., 2020). This classification does not naturally align with the income categories in the survey. Hence, we classify the categories “none,” “less than \$15,000,” “\$15,000 to under \$35,000,” and “\$35,000 to under \$50,000” in the “Low” category, “\$50,000 to under \$100,000” in “Medium,” and “\$200,000 or higher” in the “High” category. However, we faced a dilemma about whether to place the category “\$100,000 to under \$200,000” in “Medium” or “High.” Hence, we tried both classification schemes. In Classification 1, we placed “\$100,000 to under \$200,000” in “Medium,” whereas in Classification 2, we placed “\$100,000 to under \$200,000” in “High.” Table 4 illustrates the sample size of these two classification schemes and the bank’s classification.

We used these three classification schemes and ran logistic regression to model the effect of age, gender, and income level on the likelihood of using bank-based P2P services. Table 5 provides the results.

Observations. Table 5 shows that the likelihood of using bank-backed P2P increases with income level in all three cases. However, with the classification used by the bank, none of the income levels were significant. Whereas in new classifications 1 and 2, the low-income level (< \$50,000) became significant. Additionally, the medium-income level ($\geq 50,000$ and < \$100,000) became significant for classification 2. Hence, we can see that income was not initially significant. However, with different operationalizations, income up to \$50,000 became significant in both cases, while income up to

Table 3. Descriptive statistics

			Use bank-based P2P	Do not use bank-based P2P
Sample Size (N)			138	2085
Gender	gender_M	Male	45	854
	gender_F	Female	93	1231
Age (years)		Minimum	18	18
	Age	Mean	45.4	51.46
		Maximum	76	85
Income (dollar)	income_0	None	0	6
	income_1	< \$15,000	9	139
	income_2	\$15,000–\$35,000	17	394
	income_3	\$35,000–\$50,000	19	322
	income_4	\$50,000–\$100,000	45	728
	income_5	\$100,000–\$200,000	40	409
	income_6	> \$200,000	8	87
Education	education_1	Less than a high school grad	0	30
	education_2	High school diploma or equivalent (e.g., GED)	13	406
	education_3	Some college/technical school/associate's degree	45	667
	education_4	4-year college degree/bachelor's degree	49	627
	education_5	Graduate or professional degree (e.g., JD, MA, MBA, MD)	31	355
Employment Status	employment_1	Employed full-time by someone else (30+ hours per week)	70	850
	employment_2	Employed part-time by someone else (less than 30 hours per week)	15	202
	employment_3	Self-employed	8	93
	employment_4	Business owner	0	21
	employment_5	Not employed but looking for work	5	81
	employment_6	Not employed and not looking for work	8	58
	employment_7	Full-time student	4	33
	employment_8	Part-time student and employed	1	9
	employment_9	Homemaker	7	136
	employment_10	Retired	20	602
Trust in Bank	trust_Yes	Yes	99	1510
	trust_neutral	Neutral	26	409
	trust_No	No	13	166
In Relationship	relationship_Y	Yes	93	1316
	relationship_N	No	45	769
Have Children	child_Y	Yes	60	571
	child_N	No	78	1514
Have CDs	cds_Y	Yes	25	369
	cds_N	No	113	1716
Have Saving Account	saving_Y	Yes	106	1554
	saving_N	No	32	531

Table 4. Operationalization of income

Income Range	Bank's Classification	Size	Classification 1	Size	Classification 2	Size
I don't have any income	0	6	Low	906	Low	906
Less than \$15,000	1	148				
\$15,000 to under \$35,000	2	411				
\$35,000 to under \$50,000	3	341	Medium	1222	Medium	773
\$50,000 to under \$100,000	4	773				
\$100,000 to under \$200,000	5	449				
\$200,000 or higher	6	95	High	95	High	544

Table 5. Parameter estimates for logistic regression

Bank's Classification		With Classification 1		With Classification 2	
Variables	Estimates (Significance)	Variables	Estimates (Significance)	Variables	Estimates (Significance)
Intercept	-14.563098	Intercept	-1.024705 (*)	Intercept	-0.986897 (**)
Age	-0.024026 (***)	Age	-0.023567 (***)	Age	-0.023882 (***)
Gender_M	-0.326135 (.)	Gender_M	-0.288751	Gender_M	-0.321631 (.)
Income_1	12.920216	Income_L	-0.742985 (.)	Income_L	-0.758229 (***)
Income_2	12.645734	Income_M	-0.293840	Income_M	-0.517507 (*)
Income_3	12.991948				
Income_4	13.066711				
Income_5	13.586589				
Income_6	13.576942				
AIC	1018.7		1017.7		1012.6

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 . 1

\$100,000 became significant in one case. It makes the case that the operationalization of a variable may lead to different outcomes.

Operationalization of Education. In the survey, education has five categories, as shown in Table 6. We classified these five categories into three categories: “Up to High School,” “Undergraduate,” and “Graduate.”

Using these two classification schemes, we ran logistic regression to model the effect of age, gender, and education level on the likelihood of using bank-based P2P services, as shown in Table 7.

Observations. Table 7 shows that the significance of educational attainment significantly changed between the two cases. With the new classification, as education level increases, the likelihood of using bank-backed P2P increases as the magnitude of beta coefficients surges, in addition to the statistical significance levels showing that outcomes vary with the operationalization of variables.

Analysis. Increasing the granularity of variables results in more information. However, the formation of more categories decreases the variance among the categories and makes it harder for a model to detect the impact of a particular category, especially in rare event modeling. That said, specific groupings increase the variance in the category, allowing for the manifestation of significance. However, categorization, and re-categorization are vital, as they allow us to understand at what

Table 6. Operationalization of education

Education Level	Bank's Classification	Size	New Classification	Size
Less than a high school grad	1	30	Up to High School	449
High school diploma or equivalent (e.g., GED)	2	419		
Some college/technical school/associate's degree	3	712	Undergraduate	1,388
4-year college degree/bachelor's degree	4	676		
Graduate or professional degree (e.g., JD, MA, MBA, MD)	5	386	Graduate	386

Table 7. Parameter estimates for logistic regression

Bank's Classification		With New Classification	
Variables	Estimates (Significance)	Variables	Estimates (Significance)
Intercept	-15.573087	Intercept	-2.333784 (***)
Age	-0.023178 (***)	Age	-0.023016 (***)
Gender_M	-0.280149	Gender_M	-0.279411
Education_2	13.335117	Education_UnderGrad	0.915176 (**)
Education_3	14.106299	Education_Graduate	1.182014 (***)
Education_4	14.216290		
Education_5	14.429474		
AIC	1011.1		1009.6

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 ' ' 1

granularity level interventions or advertising efforts should be targeted to impact the outcome variable, which, in our case, is the use of bank-based P2P payment services.

Operationalization for the Rest of the Work. With the bank's consent, for the rest of the work, we continued with the operationalization of income, employment, and education, as shown in Table 8. Table 9 presents the descriptive statistics of the new categories.

We ran the multicollinearity test and found the variance inflation factor well below 5 (please refer to Appendix B). Often, age is considered highly correlated with income and education. We also visualize age with various predictor variables for a robustness check, as in Figure 2. We noticed no issue of high correlation with the other predictor variables in this dataset.

REGRESSION METHODS FOR ANALYSIS

Given that the outcome variable is binary, the natural choice of regression is logistic regression, which can be implemented in many ways, such as "Simple Logistic," "Weighted Logistics," "Rare Event Logistics," "Rare Events Weighted Logistic," and "Simple Logistics with Propensity Score Matching." These algorithms can easily be implemented in "R" using packages "glm," "relogit," "rewlr," and "matchit." In this work, we use R and focus mainly on "Weighted Logistic" and "Simple Logistic with Propensity Score Matching" to showcase how different decision choices lead to different outcomes.

Table 8. Operationalization of variables

Variable	Original Categories	New Categories
Income	None	Low
	< \$15,000	
	\$15,000–\$35,000	
	\$35,000–\$50,000	Medium
	\$50,000–\$100,000	
	\$100,000–\$200,000	High
	> \$200,000	
Education	Less than a high school grad	Up to High School
	High school diploma or equivalent (e.g., GED)	Undergraduate
	Some college/technical school/associate’s degree	
	4-year college degree/bachelor’s degree	Graduate
	Graduate or professional degree (e.g., JD, MA, MBA, MD)	
Employment Status	Employed full-time by someone else (30+ hours per week)	Employed
	Employed part-time by someone else (less than 30 hours per week)	
	Self-employed	Self Employed
	Business owner	
	Not employed but looking for work	Not Employed
	Full-time student	Student
	Part-time student and employed	
	Not employed and not looking for work	Not Working
	Homemaker	
	Retired	

Comparison of Logistic Regression and Weighted Logistic Regression

Logistic regression is best suited for a binary classification problem (use vs. not use) (Press & Wilson, 1978; Wright, 1995). However, it is ineffective for imbalanced or skewed data, such as ours. The use of bank-based P2P systems is relatively rare in the dataset, making it highly skewed. Hence, we opted for WLR and compared it against simple logistic regression. King and Zeng (2001) suggested weighting as a procedure that weighs the data to compensate for differences in sample and population. Table 10 compares logistic regression and WLR, with weights equal to the inverse of the class distribution (i.e., weight 15:1).

Observations

While “Age” and “Education Up to HS” are significant in logistic regression, more variables become significant and parameter estimates change slightly in the WLR. WLR advocates that using bank-based P2P is negatively associated with age, education level, and income. Males are less likely to use bank-based P2P than females. Trust and employment status are not significant. “Employment_not_emp” changed the direction, although it is not significant. With weighted regression (15:1), the area under the curve (AUC) increases from 0.500 to 0.552.

Table 9. Descriptive statistics with new categories

		Use bank-based P2P	Do not use bank-based P2P
Sample Size (N)		138	2,085
Gender	Male	45	854
	Female	93	1,231
Age (years)	Minimum	18	18
	Mean	45.4	51.47
	Maximum	76	85
Income (dollar)	Low	45	861
	Medium	85	1,137
	High	8	87
Education	Graduate	31	355
	Undergraduate	94	1,294
	Up to High School	13	436
Employment Status	Employed	85	1,052
	Not Employed	5	81
	Not Working	35	796
	Self Employed	8	114
	Student	5	42
Trust in Bank	Yes	99	1,510
	Neutral	26	409
	No	13	166
In Relationship	Yes	93	1,316
	No	45	769
Have Children	Yes	60	571
	No	78	1,514
Have CDs	Yes	25	369
	No	113	1,716
Have Saving Account	Yes	106	1,554
	No	32	531

Analysis

We see that with WLR, several factors, such as age, income, and education, become highly significant. This is because weighting increases the number of responses relating to bank-based P2P payment service responses. This prevents the drowning-out effect of the control group. Weighting penalizes the model less for errors made on majority events and more for errors made on rare events. This helps reduce model bias toward the majority event. Our analysis supports the arguments of King and Zeng (2001), explaining how statistical models, such as logistic regression, tend to bias toward the majority event and underestimate the rare event.

Figure 2. Relationship between age and other variables



Table 10. Logistic regression versus weighted logistic regression

	Logistic Regression	Weighted Logistic Regression (Use:Not Use – 15:1)
N	2223	2223
Variables	Estimates (significance)	Estimates (significance)
Intercept	-0.9686741 (.)	2.008201 (***)
Age	-0.0221855 (**)	-0.025809 (***)
Gender_M	-0.3072995	-0.337708 (***)
Income_Low	-0.4608820	-0.622431 (***)
Income_Medium	-0.2255489	-0.308998 (*)
Trust_No	0.0911125	0.076932
Trust_Yes	0.0475557	0.080549
Education_Under Grad	-0.1947449	-0.215887 (*)
Education Up to HS	-1.0086506 (**)	-1.035474 (***)
Employment_not_emp	0.0018259	-0.053754
Employment_not_working	-0.0929406	-0.036490
Employment_self_emp	0.0001106	0.086217
Employment_student	0.0451278	0.057957
AIC	1023.3	5456.9
AUC	0.500	0.552

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 ' ' 1

AUC = area under the curve (calculated for the entire 2223 observations)

Impact of Utilizing Different Weights in Weighted Logistic Regression

Weights in WLR generally represent the proportion of events in the sample and the population. However, it is not always feasible to have this ratio. In such cases, the researcher may need to use different weights. Table 11 presents the results from WLR with different weights.

Observation

The results point to some robustness regarding the change in the significance level of key predictor variables and their likelihood estimates (beta-coefficients). For significant variables, we notice that the significance level increases, and the parameter estimate decreases as the weight decreases. “Income_Medium” is only significant with a weight of 15:1, whereas “Education_Under Grad” became non-significant with a weight of 4:1. The AUC reduces as we move from weights (15:1) to (4:1).

Analysis

In a sense, weighting artificially amplifies the distribution (i.e., the impact of the mean and variance of the smaller group) to bring the effect of the change in the predictor variable in the minority group on the outcome. As weights increase, the internal number of control records increases, and the model can better calculate parameter estimates with an increased significance level. Weighting could also help answer the question of how the results would change if the data collection process included more data from one of the groups (the critical assumption being that the mean and variance are the same). It assumes that the additional data collected will have the same profile as the data from the smaller

Table 11. Weighted logistic regression with different weights

Weights	Use:Not Use 15:1	Use:Not Use 8:1	Use:Not Use 4:1
N	2223	2223	2223
Variables	Estimates (significance)	Estimates (significance)	Estimates (significance)
Intercept	2.008201 (***)	1.274592 (***)	0.498965
Age	-0.025809 (***)	-0.024326 (***)	-0.023223 (***)
Gender_M	-0.337708 (***)	-0.325904 (***)	-0.316694 (**)
Income_Low	-0.622431 (***)	-0.567740 (**)	-0.517700 (*)
Income_Medium	-0.308998 (*)	-0.285205	-0.259083
Trust_No	0.076932	0.086542	0.090954
Trust_Yes	0.080549	0.071220	0.061085
Education_Under Grad	-0.215887 (*)	-0.205766 (*)	-0.199283
Education_Up to HS	-1.035474 (***)	-1.021359 (***)	-1.013610 (***)
Employment_not_emp	-0.053754	-0.028566	-0.011269
Employment_not_working	-0.036490	-0.059680	-0.076740
Employment_self_emp	0.086217	0.054028	0.027311
Employment_student	0.057957	0.064460	0.060839
AIC	5456.9	3921.2	2603.6
AUC for entire dataset	0.552	0.501	0.500

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 ' ' 1

AUC = area under the curve (calculated for the entire 2223 observations)

group. The probability of that assumption holding may be small but depends on the remainder of the data distribution of the population from which the sample survey was selected.

Comparison of Logistic Regression and Logistic Regression With Propensity Score Matching

Like WLR, PSM (Cepeda et al., 2003) is often used to address highly skewed data. However, PSM techniques reduce the number of observations of majority events to a level that matches the rare event to balance the number of responses. In PSM, the control and treatment groups have the same probability of reacting to the treatment if exposed (Austin, 2011; Abadie & Imbens, 2016). The critical question is: How does the researcher weed out responses? At a high level, variables are used to select records from the larger group that match the records from the smaller group. The selection process is a methodological choice made by researchers for justifiable reasons. However, alternate reasons may hold equal appeal. We used “Have Child” and “Relationship Status” as matching variables and nearest neighbor (with ratio 1) as a matching algorithm for PSM and ran logistic regression. The results are presented in Table 12.

Observation

We note that “Age” and “Education Up to HS” became nonsignificant. In comparison, gender became significant with PSM. Further, parameter estimates changed slightly between logistics and logistics with PSM. The AUC for PSM is 0.556, whereas for weighted logistics, it was only 0.500.

Table 12. Logistic regression with and without PSM

	Logistic Regression	Logistic Regression after PSM
n	2223	276
Variables	Estimates (significance)	Estimates (significance)
Intercept	-0.9686741 (.)	0.700619
Age	-0.0221855 (**)	-0.006589
Gender_M	-0.3072995	-0.681189 (*)
Income_Low	-0.4608820	0.638174
Income_Medium	-0.2255489	0.697547
Trust_No	0.0911125	-0.249053
Trust_Yes	0.0475557	-0.136901
Education_Under Grad	-0.1947449	-0.511303
Education_Up to HS	-1.0086506 (**)	-1.294111
Employment_not_emp	0.0018259	0.229658
Employment_not_working	-0.0929406	-0.357194
Employment_self_emp	0.0001106	-0.215604
Employment_student	0.0451278	15.159302
AIC	1023.3	380.63
AUC	0.500	0.556

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 . 1

AUC = area under the curve (calculated for the entire 2223 observations)

Analysis

When the groups are matched based on a specific variable, the PSM algorithm selects observations from larger groups that are very similar to the smaller group. New groups created have different variances and means for predicting variables than the original groups. This leads to different parameter estimates.

Comparison of Propensity Score Matching Methods Using Different Matching Algorithms

PSM allows researchers to use different methods to match the two groups. The various matching algorithms, as discussed by Stuart et al. (2011), are briefly explained in Table 13.

From a methodological perspective, several variables or a combination of variables could be used for matching purposes. To isolate the impact of the different matching variables, we utilize “Have Child” and “Relationship Status” as matching variables for all the matching algorithms we executed. Table 14 presents the results of the nearest neighbor (with ratios 1 and 2) and optimal (with ratios 1 and 2).

Observations

We notice a change not only in likelihood estimates but also in the significance level of variables. “Age” consistently shows a negative relationship with the use of bank-backed P2P; however, this relationship is only significant in nearest neighbor (with ratio 2). “Education_upto HS” is significant for each algorithm except nearest neighbor (with ratio 1). The direction of the relationship among low-income, trust, and employment status changes with the algorithm used. In addition, we can observe that the AUC varies among the algorithms and reduces as we increase the ratio.

Analysis

The algorithm used to match the control and treatment groups utilizes different procedures, which essentially changes how the data are processed; hence, different control groups are formed, even for the same treatment group. In other words, each algorithm induces a balance in baseline covariates, reflecting the variance–bias tradeoff (Austin, 2014). It is unclear which method has superior performance, and further investigation is a matter for future research.

Table 13. Matching algorithms in PSM

Matching Algorithm	Underlying Mechanism
Exact	It selects the exact match from the control group for each individual in the treatment group. If a match is not found, the individual from the treatment is dropped.
Subclassification	When there are more variables to be used for the matching algorithm, subclassification forms subclasses, such that treated and control groups are as similar as possible in covariate distribution.
Nearest Neighbor	It selects the best control match for each individual in the treatment group.
Optimal	Similar to the nearest neighbor, it selects the control with the smallest average absolute distance across all the matched pairs.
Full	For each treated individual, it fetches all the matching control individuals.
Genetic	It automates the process of finding a good matching solution.
Coarsened	The best match is found using ax ante rather than using the original data.

Table 14. Logistic regression with different PSM algorithms

PSM Algorithm	Nearest (Ratio = 1)	Nearest (Ratio = 2)	Optimal (Ratio = 1)	Optimal (Ratio = 2)
n	276	414	276	414
Variables	Estimates (Significance)	Estimates (Significance)	Estimates (Significance)	Estimates (Significance)
Intercept	0.700619	0.963278	1.40924 (.)	0.645315
Age	-0.006589	-0.016031 (.)	-0.01082	-0.008303
Gender_M	-0.681189 (*)	-0.519339 (*)	-0.24016	-0.284651
Income_Low	0.638174	0.02715	0.00199	-0.533654
Income_Medium	0.697547	0.099921	-0.06243	-0.288819
Trust_No	-0.249053	-0.016511	-0.15732	0.003031
Trust_Yes	-0.136901	-0.180783	0.10640	-0.010182
Education_Under Grad	-0.511303	-0.494889 (.)	-0.59209	-0.233479
Education_upto HS	-1.294111	-1.297311 (*)	-1.93310 (***)	-1.249909 (**)
Employment_not_emp	0.229658	0.422127	-0.18838	-0.115539
Employment_not_working	-0.357194	-0.399566	-0.34608	-0.307862
Employment_self_emp	-0.215604	0.049897	-0.19186	-0.343021
Employment_student	15.159302	1.891102 (.)	0.13918	-0.086327
AIC	380.63	514.97	383.23	529
AUC	0.556	0.513	0.555	0.500

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 . 0.1 ' ' 1
AUC = area under the curve (calculated for the entire 2,223 observations)

Comparison of Propensity Score Matching Methods Using Different Matching Variables

The choice of variables used for matching significantly impacts the outcomes. When a variable or a combination of variables is chosen for matching purposes, the PSM method essentially picks out responses from the major class that match the responses from the minor class, eliminating the variance in the responses for these variables in the final dataset. We used the following combination of variables for propensity matching: (a) “Have Child” and “Relationship Status,” (b) “Have CDs” or “Have a Saving account,” and (c) “Have Child,” “Relationship Status,” “Have CDs,” or “Have Saving account.” Table 15 presents the logistic regression results after PSM with the nearest neighbor (with ratio 1) algorithm.

Observations

“Gender_M” consistently shows a negative relationship with using bank-backed P2P. Undergraduate customers are less likely to use bank-backed P2P than graduate customers. However, this relationship is not consistently significant across the three iterations. Although insignificant, the “Income_Medium,” “Trust_No,” and “Employment_not_emp” categories changed the direction among the three iterations. We observe that the AUC varies among the algorithms and reduces as we increase the ratio. We can see that the AUC varies as we change the variables for matching.

Table 15. PSM algorithms with different matching variables

PSM Matching Variables	Child and Relationship	CDs and Saving	Child, Relationship, CDs, and Saving
n	276	276	276
Variables	Estimates (Significance)	Estimates (Significance)	Estimates (Significance)
Intercept	0.700619	1.728937 (*)	0.945939
Age	-0.006589	-0.024849 (*)	-0.007009
Gender_M	-0.681189 (*)	-0.757216 (**)	-0.654654 (*)
Income_Low	0.638174	0.425943	0.494483
Income_Medium	0.697547	0.636961	0.544445
Trust_No	-0.249053	-0.4693	-0.193018
Trust_Yes	-0.136901	-0.185822	-0.177379
Education_Under Grad	-0.511303	-0.546706	-0.671439 (.)
Education_upto HS	-1.294111	-1.573731 (**)	-1.495153 (*)
Employment_not_emp	0.229658	0.580929	0.482946
Employment_not_working	-0.357194	0.009727	-0.287977
Employment_self_emp	-0.215604	0.355421	0.320305
Employment_student	15.159302	1.046359	15.257323
AIC	380.63	379.03	380.31
AUC	0.556	0.573	0.543

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 . 1
 AUC = area under the curve (calculated for the entire 2223 observations)

Analysis

When groups are matched based on a specific variable, PSM algorithms select responses from the larger group that are similar to those from the smaller group. Essentially, the difference in mean and variance values for these variables between the groups becomes similar, these variables from becoming predictor variables. From a statistical perspective, the population characteristics of the major and minor groups vary significantly depending on the choice of the matching variable used. These differences significantly impact the results, and one must choose which set of results is preferable to report. In a sense, using statistical analysis results to report outcomes is quite flawed. The choice of PSM must find justification in the application domain.

Comparison of Propensity Score Matching With the Sort Order of Input Data

The formation of the control and treatment groups depends on the mechanism of the algorithm selected and how that algorithm treats the input data. Hence, the same input data but different sorting orders of variables may generate completely different control and treatment groups and may lead to different outcomes. Table 16 demonstrates the impact of sorting the order of input data with “Have Child” and “Relationship Status” as matching variables and nearest neighbor (with ratio 1) as the matching algorithm. The first column presents the result with data as it is, the second column with input data sorted based on income in ascending order, and the third with input data sorted based on income in descending order.

Table 16. PSM algorithms with sorted input data

Sorting order based on	None	Ascending Income	Descending Income
n	276	276	276
Variables	Estimates (Significance)	Estimates (Significance)	Estimates (Significance)
Intercept	0.700619	22.03428	-3.20193 (***)
Age	-0.006589	-0.04015 (**)	0.03144 (*)
Gender_M	-0.681189 (*)	-0.01920	-0.52599
Income_Low	0.638174	- 20.89115	21.79992
Income_Medium	0.697547	-0.12119	3.21504 (***)
Trust_No	-0.249053	-0.20274	-0.18542
Trust_Yes	-0.136901	-0.06641	-0.45416
Education_Under Grad	-0.511303	0.03793	-0.18772
Education_upto HS	-1.294111	-1.07241	-1.22442 (.)
Employment_not_emp	0.229658	0.46667	15.61159
Employment_not_working	-0.357194	-0.40176	-0.68857
Employment_self_emp	-0.215604	-0.37196	-0.8855
Employment_student	15.159302	1.22271	0.14391
AIC	380.63	199.97	252.5
AUC	0.556	0.551	0.503

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 . 1
AUC = area under the curve (calculated for 2,223 observations)

Observations

We observe that even with the same input data, same matching variables, and same matching algorithm, results vary just by sorting the input data before processing it through PSM. “Age” is not significant in the first column and becomes significant when we sort the data. “Income_Medium” becomes significant when we sort the data in descending order by income. In addition, parameter estimates differ between the three iterations.

Analysis

We observed that the results varied based on the order of the input data. The treatment or control group formation depends on how the algorithm is designed to process the input data. For example, the nearest neighbor algorithm selects the best control match for each individual in the treatment group. In skewed data, there is a good chance that for a given record in the treatment group, multiple records exist with the same propensity score. Hence, the algorithm selects the first record it encounters in the treatment group with the same propensity score. As the ordering of input data changes, a different control record is selected for the same treatment record; thus, the covariates’ variance changes, leading to different outcomes.

PSM AND WEIGHTED LOGISTIC REGRESSION COMBINED

This section provides the results when PSM and WLR are combined. First, we ran PSM techniques with a ratio of 2 that reduced the number of observations of majority events to twice that of minority events. We then used this reduced dataset and modeled it with WLR with a weight of 1:2. We used “Have Child” and “Relationship Status” as matching variables. Table 17 shows the results.

Table 17. PSM algorithms with sorted input data

	PSM (Nearest)	PSM (Nearest) + Weighted Logistic (1:2)	PSM (Optimal)	PSM (Optimal) + Weighted Logistic (1:2)
n	414	414	414	414
Variables	Estimates (significance)	Estimates (significance)	Estimates (significance)	Estimates (significance)
Intercept	0.963278	1.770703 (**)	0.645315	1.412374
Age	-0.016031 (.)	-0.016734 (*)	-0.008303	-0.00967 (*)
Gender_M	-0.519339 (*)	-0.529274 (**)	-0.284651	-0.29829
Income_Low	0.02715	-0.045674	-0.533654	-0.54937
Income_Medium	0.099921	0.055924	-0.288819	-0.28808
Trust_No	-0.016511	-0.068124	0.003031	-0.00985
Trust_Yes	-0.180783	-0.190757	-0.010182	-0.00737
Education_Under Grad	-0.494889 (.)	-0.516766 (*)	-0.233479	-0.24252
Education Up to HS	-1.297311 (*)	-1.311982 (***)	-1.249909 (**)	-1.26017 (***)
Employment_not_emp	0.422127	0.411186	-0.115539	-0.12645
Employment_not_working	-0.399566	-0.385234 (.)	-0.307862	-0.28354
Employment_self_emp	0.049897	0.052972	-0.343021	-0.3144
Employment_student	1.891102 (.)	1.851161 (.)	-0.086327	-0.13046
AIC	514.97	734.27	529	753.94
AUC	0.513	0.574	0.500	0.541

Significance level: 0 **** 0.001 *** 0.01 ** 0.05 * 0.1 ' ' 1
 AUC = area under the curve (calculated for the entire 2223 observations)

Observations

We observe that the AUC increases as we use weighted logistics on the dataset generated using PSM. We can see that “Education_upto HS” remains significant across all four scenarios. “Age” is also significant across all three scenarios. Furthermore, parameter estimates differ among the four iterations, and some estimates change the direction of some of the variables.

Analysis

By combining the two methods, we can achieve a better result in terms of identifying the rare event in the entire sample. However, it may produce more false negatives and identify the non-important event as an event of importance.

DISCUSSION AND CONCLUSION

Table 18 summarizes the comparison of the various methods used in our analysis.

The results for the six research questions varied considerably based on empirical decisions. “Employment status” and “Trust in a bank” are not significant in any of the iterations, whereas “Age,” “Gender,” and “Education up to HS” are significant in most of the iterations. On the other hand, “Income Level (Low)” is significant, mainly in WLR.

Furthermore, the significance levels and parameter estimates greatly vary among the iterations. However, it should not lead us to have less confidence in them. The results from the most appropriate

Table 18. Comparison of different methods

	RQ1		RQ2		RQ3		RQ4		RQ5			RQ6	
	Age	Gender	Low	Medium	Under Grad	Up to HS	Not Emp	Not Working	Self Emp	Student	No	Yes	(Trust in a bank)
	(.)	(**)				(**)							
Logistic Regression													
Weights Logistic Regression with Different Weights (Use Zelle: Do not Use Zelle)													
15:1	(***)	(***)	(***)	(*)	(*)	(***)							
8:1	(***)	(***)	(**)		(*)	(***)							
4:1	(***)	(**)	(*)			(***)							
Propensity Score Matching with Different Algorithms (Matching Variables Child and Relationship Status)													
Nearest (Ratio =1)		(*)											
Nearest (Ratio =2)		(.)	(*)		(.)	(*)					(.)		
Optimal (Ratio =1)	(.)					(***)							
Optimal (Ratio =2)						(**)							
Propensity Score Matching with Different Matching Variables (Nearest ratio =1)													
“Have Child” and “Relationship Status”		(*)											
“Have CDs” and “Have Saving”	(*)	(**)				(**)							
“Child,” “Relationship,” “CDs” and “Saving”		(*)			(.)	(*)							
Propensity Score Matching with Different Sort Order (Nearest Ratio =1 and Matching Variables Child and Relationship Status)													
None		(*)											
Ascending Income		(**)											
Descending Income	(***)	(*)		(***)		(.)							
Significance level: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘.’ 1													
Blue cells represent a positive relationship							Red cells represent a negative relationship						

choices based on domain consideration related to variable selection in PSM using equal weights should guide our decision-making. The different choices of matching variables provide a different view of the data, and the results from different choices are not necessarily incorrect. While our prior experience with “all things being equal (*Citrus Paribus*)” points to equal weighting as the ideal choice for a researcher, domain consideration for such decisions is critical. Based on the initial analyses, we propose some general guidelines and technique-specific guidelines in Tables 19 and 20, respectively.

Contribution to IS Literature

Addressing the issue of data imbalance is a critical aspect of data analysis and modeling within the realm of IS research (Gao et al., 2018). Data imbalance refers to an asymmetrical distribution of data across different classes or categories within a dataset, whereby certain classes are underrepresented compared to others (He & Garcia, 2009). This issue can manifest in diverse contexts, including financial fraud detection (Xia & Zhang, 2023; Al-Shabi, 2019), medical diagnosis (Bridge et al., 2020), sentiment analysis (Al Shamsi & Abdallah, 2022), text analysis (Li et al., 2023), image classification (Tian & Han, 2022), and recommendation systems (Zhang et al., 2019).

In such scenarios, addressing the challenge of unbalanced data in the IS literature holds considerable significance for several compelling reasons. First, it aims to enhance the performance of quantitative models by rectifying the inherent bias introduced by unbalanced data. This bias often skews the models’ performance in favor of the majority class, consequently leading to suboptimal outcomes (Krawczyk, 2016). Researchers have proposed various techniques to mitigate this issue, including resampling methods, algorithmic modifications with weighting, and ensemble methods (Fernández et al., 2018). We contribute to the IS literature by demonstrating the use of three specific methods that can be very effective in dealing with imbalanced data. Furthermore, we point out several key empirical decisions and critical tradeoffs that researchers need to make.

Second, addressing the data imbalance aims to improve decision-making processes that rely heavily on empirical models (Kubat & Matwin, 1997). Incorrect predictions of rare events can have profound consequences within IS, potentially leading to significant negative outcomes (Galar et al., 2012). By tackling class imbalance, researchers endeavor to improve the accuracy of predictive

Table 19. General guideline for methodological decisions

<p>1. Determine the research objective Clearly define the research objective and the problem you are trying to address. This will help guide the selection of appropriate methods and techniques.</p>
<p>2. Understand the data Gain a thorough understanding of the dataset, including the nature of the variables, their distributions, and any potential data quality issues. Assess the extent of the data imbalance between important (minority) observations and other (majority) observations.</p>
<p>3. Perform data analysis Follow the technique-specific guidelines provided in the table below (refer to Table 20) to perform a robust data analysis</p>
<p>4. Evaluate and interpret results Carefully analyze the results obtained from the chosen methods. Assess the model’s performance, interpret the estimated parameters, and draw meaningful conclusions in relation to the research objective.</p>
<p>5. Sensitivity analysis Conduct sensitivity analyses to test the robustness of the results. Explore variations in the weights, covariate selection, and matching algorithms to evaluate the stability of the findings.</p>
<p>6. Reporting and documentation Document all steps taken, including the rationale behind the choices made at each stage. Clearly report the methods employed, results obtained, and limitations encountered during the research process.</p>

Table 20. Guidelines specific to the chosen technique

WLR
<p style="text-align: center;">1. Assign weights based on sample-to-population ratio</p> <p>When using WLR, determine the weights to be assigned to observations based on their importance and the extent of the data imbalance. Consider using the inverse ratio of the number of important observations to the number of other observations as a starting point.</p>
<p style="text-align: center;">2. Consider domain expertise</p> <p>If possible, choose weights based on domain expertise. Expert knowledge can provide valuable insights into the importance of different observations or groups within the data.</p>
<p style="text-align: center;">3. Evaluate different weights</p> <p>If domain expertise is not available or inconclusive, try different weights and assess their impact on the performance of the logistic regression model. Experimentation can help identify the weights that yield a more accurate classification of records.</p>
Logistic Regression With PSM
<p style="text-align: center;">1. Ground matching variables on prior theories or domain knowledge</p> <p>When applying PSM, select matching variables for the control and treatment groups based on prior theories or domain knowledge. This ensures that the matching process is based on relevant factors and increases the validity of the analysis.</p>
<p style="text-align: center;">2. Explore multiple matching algorithms</p> <p>PSM involves various permutations and combinations. Given that the choice of matching algorithm affects parameter estimates and model performance, try multiple matching algorithms. By comparing the results, you can identify the algorithm that provides the most reliable and consistent outcomes.</p>
<p style="text-align: center;">3. Consider the order of the input data</p> <p>Be aware that the order of the input data can influence the results obtained from PSM. Pay attention to the ordering of the data when performing PSM to minimize any potential bias.</p>
WLR With PSM
<p style="text-align: center;">1. Opt for a large control group using the ratio option</p> <p>In WLR with PSM, select a control group as large as possible using the ratio option. This helps maintain a more substantial control group and reduces information loss. Using weights in logistic regression compensates for any imbalance between the control and treatment groups.</p>
<p style="text-align: center;">2. Minimize the impact of sort order</p> <p>Choosing a larger control group, as suggested above, also reduces the impact of the sort order of the data. This reduces the potential bias introduced by the order of the observations.</p>
<p style="text-align: center;">3. Experiment with different PSM algorithms and the weights</p> <p>To determine the most accurate option for classifying records, try different PSM algorithms with various weights. The weight should depend on the majority/minority ratio used in PSM. However, you may have to try various weights. By evaluating their performance, you can identify the combination of the PSM algorithm and the weights that yield the best classification accuracy.</p>

models by reducing the occurrence of false negatives (i.e., missed occurrences of the minority class) and false positives (i.e., incorrect predictions of the minority class) (Provost & Fawcett, 2001). Consequently, this improvement enhances decision-making processes guided by predictions generated by IS (Zliobaite, 2010).

Third, we note that imbalanced data can lead to biased models that perform well on the majority class but exhibit inferior performance on the minority class (Kamishima et al., 2012). Such bias can result in an inadequate representation of certain individuals or groups, perpetuating social and ethical concerns (Dwork et al., 2012). Prior literature has emphasized the need for inclusive and responsible research within the IS field (Pushkarna et al., 2022). We contribute to the emerging stream of IS literature by presenting a worked example of methods that yield equitable outcomes across different populations or contextual settings (Kleinberg et al., 2016). Therefore, our work aligns with prior literature on fairness and generalizability within IS (Kamiran & Calders, 2012; Zafar et al., 2017).

In summary, our exploration of handling unbalanced data within IS research serves to provide a playbook for future researchers to improve model performance, enhance decision-making processes, ensure fairness, and promote generalizability.

Business Implications

Unbalanced secondary data can yield numerous managerial implications and provide practitioners with several advantages. We briefly highlight some of these implications below.

- **Data analysis and empirical tradeoffs:** The performance of well-known models are susceptible to imbalanced data (Das et al., 2018). In particular, models with imbalanced data perform poorly on minority-class examples (Weiss & Provost, 2001). The techniques, strategies, and empirical tradeoffs presented in this paper provide a playbook that practitioners can use to arrive at accurate conclusions through robust analysis, leading to more reliable and valid findings.
- **Decision-making:** Managers and analysts must recognize the presence of data imbalance and acknowledge its implications when analyzing imbalanced data and interpreting the findings. Neglecting data imbalance may result in biased outcomes and inaccurate conclusions.
- **Bias detection and correction:** Unbalanced secondary data can uncover biases in the data collection process or underlying systems. For instance, imbalanced data introduce bias between racial groups in a deep learning model (Puyol-Antón et al., 2021). By examining patterns and distributions within the data, practitioners can identify potential biases and take corrective measures. This may involve adjusting sampling methods, collecting additional data, or employing statistical techniques outlined in this study to balance the dataset.
- **Performance evaluation:** Unbalanced secondary data can impact the performance evaluation of models or systems trained on such data. For instance, a quantitative model trained on imbalanced data may exhibit high accuracy but struggle to predict minority classes accurately (Thabtah et al., 2020). Practitioners must exercise caution when evaluating the effectiveness of models or systems using imbalanced data and consider metrics that account for data imbalance.
- **Risk assessment:** Unbalanced secondary data may lead to an erroneous assessment of risks. For example, in fraud detection, if the majority of transactions are non-fraudulent, a model trained on imbalanced data might inadequately identify fraudulent transactions. It is vital to comprehend the limitations of imbalanced data for precise risk assessment and the implementation of appropriate risk mitigation strategies.
- **Targeted interventions:** Unbalanced secondary data can assist practitioners in identifying specific areas or segments that necessitate targeted interventions. By analyzing minority classes or underrepresented categories, managers can gain insights into potential opportunities or challenges (Puyol-Antón et al., 2021). These insights can inform decision-making, resource allocation, and strategies to address specific needs or capture untapped markets.
- **Data collection and augmentation:** Unbalanced secondary data can reveal gaps in data collection efforts, particularly regarding the representation of minority classes or underrepresented groups. Practitioners can utilize this knowledge to guide future data collection initiatives and ensure a more balanced representation of all relevant classes. Data augmentation techniques can also be employed to artificially balance the dataset by generating synthetic instances of minority classes. This approach enhances model performance and generalization.

LIMITATIONS

This paper has several limitations. It uses secondary data from a recent survey conducted by a bank. Certain key variables that may interest a general IS audience were not included. Second, we have not provided deeper explanations or a more intrusive analysis in this work. Third, many other algorithms

are available for PSM, but we have only demonstrated our work with a few of them. Future studies could address many of the limitations and shortcomings of this paper. A researcher or practitioner has many options for studying a phenomenon, and different options may lead to different outcomes. Hence, we recommend that researchers try various options and choose the best for them.

ACKNOWLEDGMENTS

The authors are grateful to Paris Roselli, Michaele Walser, and Bank for providing an opportunity to work with them and share the survey response. This work was completed without any funding or financial support.

REFERENCES

- Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*(2), 781–807. doi:10.3982/ECTA11293
- Ada, S., Sharman, R., & Balkundi, P. (2012). Impact of meta-analytic decisions on the conclusions drawn on the business value of information technology. *Decision Support Systems*, *54*(1), 521–533. doi:10.1016/j.dss.2012.07.001
- Agarwal, S., & Zhang, J. (2020). FinTech, lending and payment innovation: A review. *Asia-Pacific Journal of Financial Studies*, *49*(3), 353–367. doi:10.1111/ajfs.12294
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. doi:10.1016/0749-5978(91)90020-T
- Al Shamsi, A., & Abdallah, S. (2022). Sentiment analysis of Emirati dialect. *Big Data and Cognitive Computing*, *6*(2), 57. doi:10.3390/bdcc6020057
- Alalwan, A. A., Dwivedi, Y. K., & Rana, N. P. (2017). Factors influencing adoption of mobile banking by Jordanian bank customers: Extending UTAUT2 with trust. *International Journal of Information Management*, *37*(3), 99–110. doi:10.1016/j.ijinfomgt.2017.01.002
- Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374. doi:10.1002/wics.84
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. doi:10.2307/j.ctvc4j72
- Anong, S. T., & Kunovskaya, I. (2013). M-finance and consumer redress for the unbanked in South Africa. *International Journal of Consumer Studies*, *37*(4), 453–464. doi:10.1111/ijcs.12014
- Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049. doi:10.1002/sim.3150 PMID:18038446
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. doi:10.1080/00273171.2011.568786 PMID:21818162
- Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057–1069. doi:10.1002/sim.6004 PMID:24123228
- Austin, P. C., Mamdani, M. M., Stukel, T. A., Anderson, G. M., & Tu, J. V. (2005). The use of the propensity score for estimating treatment effects: Administrative versus clinical data. *Statistics in Medicine*, *24*(10), 1563–1578. doi:10.1002/sim.2053 PMID:15706581
- Barua, S., Islam, M. M., Yao, X., & Murase, K. (2012). MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, *26*(2), 405–425. doi:10.1109/TKDE.2012.232
- Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*, *9*(6), 377–385. doi:10.1111/j.1524-4733.2006.00130.x PMID:17076868
- Bennett, J., Fry, R., & Kochhar, R. (2020, July 23). *Are you in the American middle class? Find out with our income calculator*. Pew Research Center. <https://www.pewresearch.org/fact-tank/2018/09/06/are-you-in-the-american-middle-class/>
- Black, S., Davern, M. J., Maynard, S., & Naseer, H. (2020). Opportunity or threat: Board perspectives on the secondary use of data. *24th Pacific Asia Conference on Information Systems*, (pp. 1–8). IEEE.
- Bridge, J., Meng, Y., Zhao, Y., Du, Y., Zhao, M., Sun, R., & Zheng, Y. (2020). Introducing the GEV activation function for highly unbalanced data to develop COVID-19 diagnostic models. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2776–2786. doi:10.1109/JBHI.2020.3012383 PMID:32750973

- Burton-Jones, A., & Lee, A. S. (2017). Thinking about measures and measurement in positivist research: A proposal for refocusing on fundamentals. *Information Systems Research*, 28(3), 451–467. doi:10.1287/isre.2017.0704
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31–72. doi:10.1111/j.1467-6419.2007.00527.x
- Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158(3), 280–287. doi:10.1093/aje/kwg115 PMID:12882951
- Chao, C. M. (2019). Factors determining the behavioral intention to use mobile learning: An application and extension of the UTAUT model. *Frontiers in Psychology*, 10, 1652. doi:10.3389/fpsyg.2019.01652 PMID:31379679
- Choy, L. T. (2014). The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches. *IOSR Journal of Humanities and Social Science*, 19(4), 99–104. doi:10.9790/0837-194399104
- Chua, C. E., & Storey, V. C. (2016). Dealing with dangerous data: Part-whole validation for low incident, high risk data. [JDM]. *Journal of Database Management*, 27(1), 2–57. doi:10.4018/JDM.2016010102
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1(3), 98–101. doi:10.1111/1467-8721.ep10768783
- Cortes, C., & Vapnik, V. (1995). Support vector machine. *Machine Learning*, 20(3), 273–297. doi:10.1007/BF00994018
- Czajka, J. L., Hirabayashi, S. M., Little, R. J., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, 10(2), 117–131.
- Das, S., Datta, S., & Chaudhuri, B. B. (2018). Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81, 674–693. doi:10.1016/j.patcog.2018.03.008
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly*, 13(3), 319–340. doi:10.2307/249008
- Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1), 151–161. doi:10.1162/003465302317331982
- Dorn, M., Grisci, B. I., Narloch, P. H., Feltes, B. C., Avila, E., Kahmann, A., & Alho, C. S. (2021). Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets. *PeerJ. Computer Science*, 7, e670. doi:10.7717/peerj-cs.670 PMID:34458574
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, (pp. 214–226). ACM. doi:10.1145/2090236.2090255
- Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56(10), 968–976. doi:10.1016/S0895-4356(03)00170-7 PMID:14568628
- Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1), 92–107. doi:10.2307/1937887
- Féraud, R., & Clérot, F. (2002). A methodology to explain neural network classification. *Neural Networks*, 15(2), 237–246. doi:10.1016/S0893-6080(01)00127-7 PMID:12022511
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer., doi:10.1007/978-3-319-98074-4
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.

- Galar, D., Gustafson, A., Tormos Martínez, B. V., & Berges, L. (2012). Maintenance decision making based on different types of data fusion. *Eksplotacja i niezawodność—Maintenance and Reliability*, 14(2), 135–144.
- Gao, Y., Bu, X., Hu, Y., Shen, H., Bai, T., Li, X., & Wen, S. (2018). Solution for large-scale hierarchical object detection datasets with incomplete annotation and data imbalance. *arXiv preprint arXiv:1810.06208*. <https://doi.org/10.1810.0620810.48550>
- Gazzah, S., Hechkel, A., & Amara, N. E. B. (2015, March). A hybrid sampling method for imbalanced data. In 2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15) (pp. 1–6). IEEE. doi:10.1109/SSD.2015.7348093
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *Management Information Systems Quarterly*, 27(1), 51–90. doi:10.2307/30036519
- Gefen, D., Straub, D., & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems*, 4(1), 7. doi:10.17705/1CAIS.00407
- Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583–595. Facta Universitatis. 10.22190/FUMI1903583G
- Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, 84(11), 2809–2815. doi:10.1890/02-3114
- Groenfeldt, T. (2019, February 8). Real-time person-to-person payments are on the rise in the U.S.—Aité. *Forbes Magazine*. <https://www.forbes.com/sites/tomgroenfeldt/2019/02/08/real-time-person-to-person-payments-are-on-the-rise-in-the-u-s-aite/#76a541fc609d>
- Guillen, M., & Pesantez-Narvaez, J. (2018). Machine learning and predictive modeling for automobile insurance pricing. *Anales del Instituto de Actuarios Españoles*, 24, 123–147. doi:10.26360/2018_6
- Guo, T., & Li, G. Y. (2008). Neural data mining for credit card fraud detection. *2008 International Conference on Machine Learning and Cybernetics*. IEEE. doi:10.1109/ICMLC.2008.4621035
- Hair, J. F. Jr, Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). Macmillan.
- Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, 19(2), 139–152. doi:10.2753/MTP1069-6679190202
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. doi:10.1109/TKDE.2008.239
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199–236. doi:10.1093/pan/mpj013
- Hong, X., Chen, S., & Harris, C. J. (2007). A kernel-based two-class classifier for imbalanced data sets. *IEEE Transactions on Neural Networks*, 18(1), 28–41. doi:10.1109/TNN.2006.882812 PMID:17278459
- Hosmer, D. W. Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons., doi:10.1002/9781118548387
- Hoyos-Osorio, J., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021). Relevant information undersampling to support imbalanced data classification. *Neurocomputing*, 436, 136–146. doi:10.1016/j.neucom.2021.01.033
- Hsu, C. H., Green, S. B., & He, Y. (2007). A weighted logistic regression model for estimation of recurrence of adenomas. *Statistics in Medicine*, 26(7), 1567–1578. doi:10.1002/sim.2648 PMID:16850435
- Hu, J., Chou, C. C., Yang, K. H., & King, A. I. (2007). A weighted logistic regression analysis for predicting the odds of head/face and neck injuries during rollover crashes. *Annual Proceedings - Association for the Advancement of Automotive Medicine. Association for the Advancement of Automotive Medicine*, 51, 363–379. PMID:18184502
- Hughes, N., & Lonie, S. (2007). M-PESA: Mobile money for the “unbanked” turning cellphones into 24-hour tellers in Kenya. *Innovations: Technology, Governance, Globalization*, 2(1-2), 63–81. doi:10.1162/itgg.2007.2.1-2.63

- Humphrey, D., Willeson, M., Lindblom, T., & Bergendahl, G. (2003). What does it cost to make a payment? *Review of Network Economics*, 2(2). doi:10.2202/1446-9022.1024
- Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493), 345–361. doi:10.1198/jasa.2011.tm09599
- John Hopkins University & Medicine. (2020). *Mortality analyses*. John Hopkins University & Medicine Coronavirus Resource Center. <https://coronavirus.jhu.edu/data/mortality>
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33. doi:10.1007/s10115-011-0463-8
- Karjaluoto, H., Mattila, M., & Pento, T. (2002). Factors underlying attitude formation towards online banking in Finland. *International Journal of Bank Marketing*, 20(6), 261–272. doi:10.1108/02652320210446724
- Katz, V. S., & Gonzalez, C. (2016). Community variations in low-income Latino families' technology adoption and integration. *The American Behavioral Scientist*, 60(1), 59–80. doi:10.1177/0002764215601712
- Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(1), 51. doi:10.1186/1472-6947-11-51 PMID:21801360
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163. doi:10.1093/oxfordjournals.pan.a004868
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent tradeoffs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*. <https://doi.org/10.1101/090507>
- Koziarski, M. (2020). Radial-based undersampling for imbalanced data classification. *Pattern Recognition*, 102, 107262. doi:10.1016/j.patcog.2020.107262
- Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232. doi:10.1007/s13748-016-0094-0
- Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. *14th International Conference on Machine Learning*, 97(1), 179–186.
- Laforet, S., & Li, X. (2005). Consumers' attitudes towards online and mobile banking in China. *International Journal of Bank Marketing*, 23(5), 362–380. doi:10.1108/02652320510629250
- Laitinen, E. K. (1999). Predicting a corporate credit analyst's risk estimate by logistic and linear models. *International Review of Financial Analysis*, 8(2), 97–121. doi:10.1016/S1057-5219(99)00012-5
- Lara-Rubio, J., Villarejo-Ramos, A. F., & Liébana-Cabanillas, F. (2021). Explanatory and predictive model of the adoption of P2P payment systems. *Behaviour & Information Technology*, 40(6), 528–541. doi:10.1080/0144929X.2019.1706637
- Laukkanen, T., & Pasanen, M. (2008). Mobile banking innovators and early adopters: How they differ from other online users? *Journal of Financial Services Marketing*, 13(2), 86–94. doi:10.1057/palgrave.fsm.4760077
- Li, R., Chen, S., Zhao, F., & Qiu, X. (2023). Text detection model for historical documents using CNN and MSER. *Journal of Database Management (JDM)*, 34(1), 1–23. doi:10.4018/JDM.322086
- Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3), 281–299. doi:10.1504/IJDATS.2011.041335
- Maalouf, M., Homouz, D., & Trafalis, T. B. (2018). Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence*, 34(1), 161–174. doi:10.1111/coin.12123
- Macharia, J., & Okunoye, A. (2013). Mobile banking influence on wealth creation for the unbanked. *2013 IST-Africa Conference and Exhibition*. IST.
- Mahani, A., & Ali, A. R. B. (Eds.). (2019). *Classification problem in imbalanced datasets*. IntechOpen. doi:10.5772/intechopen.89603

- Mattila, M., Karjaluoto, H., & Pentto, T. (2003). Internet banking adoption among mature customers: Early majority or laggards? *Journal of Services Marketing*, 17(5), 514–528. doi:10.1108/08876040310486294
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. doi:10.2307/258792
- McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, 23(3), 473–490. doi:10.2307/259290
- Merritt, C. (2011). Mobile money transfer services: The next phase in the evolution of person-to-person payments. *Journal of Payments Strategy & Systems*, 5(2), 143–160.
- Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253–267. doi:10.1080/09720502.2010.10700699
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). John Wiley & Sons.
- Morris, M. G., & Venkatesh, V. (2000). Age differences in technology adoption decisions: Implications for a changing work force. *Personnel Psychology*, 53(2), 375–403. doi:10.1111/j.1744-6570.2000.tb00206.x
- Mukherjee, A., & Nath, P. (2003). A model of trust in online relationship banking. *International Journal of Bank Marketing*, 21(1), 5–15. doi:10.1108/02652320310457767
- Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, 24(12), 1565–1567. doi:10.1038/nbt1206-1565 PMID:17160063
- Oliveira, T., Faria, M., Thomas, M. A., & Popovič, A. (2014). Extending the understanding of mobile banking adoption: When UTAUT meets TTF and ITM. *International Journal of Information Management*, 34(5), 689–703. doi:10.1016/j.ijinfomgt.2014.06.004
- Olmuş, H., Beşpınar, E., & Nazman, E. (2022). Performance evaluation of some propensity score matching methods by using binary logistic regression model. *Communications in Statistics. Simulation and Computation*, 51(4), 1647–1660. doi:10.1080/03610918.2019.1679181
- Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8, 761–773.
- Patzer, G. L. (1995). *Using secondary data in marketing research: United States and worldwide*. Greenwood Publishing Group.
- Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. *IASRI, New Delhi*, 1(1), 58–65. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=300f806f39ea84cfe999cb52903726a2e8f0458>
- Pikkarainen, T., Pikkarainen, K., Karjaluoto, H., & Pahlila, S. (2004). Consumer acceptance of online banking: An extension of the technology acceptance model. *Internet Research*, 14(3), 57–63. doi:10.1108/10662240410542652
- Polatoglu, V. N., & Ekin, S. (2001). An empirical investigation of the Turkish consumers' acceptance of Internet banking services. *International Journal of Bank Marketing*, 19(4), 156–165. doi:10.1108/02652320110392527
- Potter, F., Grau, E., Williams, S., Diaz-Tena, N., & Carlson, B. L. (2006). An application of propensity modeling: Comparing unweighted and weighted logistic regression models for nonresponse adjustments. In *Proceedings of the Survey Research Methods Section*. American Statistical Association.
- Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), 699–705. doi:10.1080/01621459.1978.10480080
- Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3), 203–231. doi:10.1023/A:1007601015854
- Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, (pp. 1776–1826). ACM. doi:10.1145/3531146.3533231

- Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., & King, A. P. (2021). Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. *24th International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer. doi:10.1007/978-3-030-87199-4_39
- Rabinovich, E., & Cheon, S. (2011). Expanding horizons and deepening understanding via the use of secondary data sources. *Journal of Business Logistics*, 32(4), 303–316. doi:10.1111/j.0000-0000.2011.01026.x
- Raj, V., Magg, S., & Wermter, S. (2016). Towards effective classification of imbalanced data with convolutional neural networks. In F. Schwenker, H. Abbas, N. El Gayar, & E. Trentin (Eds.), *Artificial Neural Networks in Pattern Recognition, ANNPR 2016* (pp. 150–162). Springer. doi:10.1007/978-3-319-46182-3_13
- Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: A review. [IJCBR]. *International Journal of Computing and Business Research*, 5(4), 1–29.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. doi:10.1093/biomet/70.1.41
- Sahin, Y., & Duman, E. (2011, June). Detecting credit card fraud by ANN and logistic regression. In *2011 International Symposium on Innovations in Intelligent Systems and Applications* (pp. 315–319). IEEE. doi:10.1109/INISTA.2011.5946108
- Senso, N. C., & Venkatakrishnan, V. (2013). Challenges of mobile-phone money transfer services' market penetration and expansion in Singida District, Tanzania. *International Journal of Research in Management & Technology*, 3(6), 205–215.
- Shaikh, A. A., & Karjaluoto, H. (2015). Mobile banking adoption: A literature review. *Telematics and Informatics*, 32(1), 129–142. doi:10.1016/j.tele.2014.05.003
- Shen, A., Tong, R., & Deng, Y. (2007, June). Application of classification models on credit card fraud detection. In *2007 International conference on service systems and service management* (pp. 1–4). IEEE. doi:10.1109/ICSSM.2007.4280163
- Shipman, J. E., Swanquist, Q. T., & Whited, R. L. (2017). Propensity score matching in accounting research. *The Accounting Review*, 92(1), 213–244. doi:10.2308/accr-51449
- Shishehbor, M. H., Baker, D. W., Blackstone, E. H., & Lauer, M. S. (2002). Association of educational status with heart rate recovery: A population-based propensity analysis. *The American Journal of Medicine*, 113(8), 643–649. doi:10.1016/S0002-9343(02)01324-4 PMID:12505114
- Singh, S., Srivastava, V., & Srivastava, R. K. (2010). Customer acceptance of mobile banking: A conceptual framework. *Sies journal of management*, 7(1).
- Sinickas, A. (2007). Finding a cure for survey fatigue. *Strategic Communication Management*, 11(2), 11.
- Spelmen, V. S., & Porkodi, R. (2018, March). A review on handling imbalanced data. In *2018 international conference on current trends towards converging technologies (ICCTCT)* (pp. 1–11). IEEE.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1), 1.
- Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*.
- Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, 429–441. doi:10.1016/j.ins.2019.11.004
- Tian, X., & Han, H. (2022). Deep convolutional neural networks with transfer learning for automobile damage image classification. [JDM]. *Journal of Database Management*, 33(3), 1–16. doi:10.4018/JDM.309738
- Van Dyke, D. (2022). Mobile banking is a major driver of bank switching in the US — here are the 3 top features customers crave. *Insider Intelligence*. <https://www.insiderintelligence.com/insights/top-mobile-banking-features/>
- Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186–204. doi:10.1287/mnsc.46.2.186.11926

- Venkatesh, V., & Morris, M. G. (2000). Why don't men ever stop to ask for directions? gender, social influence, and their role in technology acceptance and usage behavior. *Management Information Systems Quarterly*, 24(1), 115–139. doi:10.2307/3250981
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly*, 27(3), 425–478. doi:10.2307/30036540
- Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *Management Information Systems Quarterly*, 36(1), 157–178. doi:10.2307/41410412
- Wang, X., Xu, J., Zeng, T., & Jing, L. (2021). Local distribution-based adaptive minority oversampling for imbalanced data classification. *Neurocomputing*, 422, 200–213. doi:10.1016/j.neucom.2020.05.030
- Weiss, G. M., & Provost, F. (2001). *The effect of class distribution on classifier learning: An empirical study* (Technical report ML-TR-44). Rutgers University. 10.7282/t3-vpfw-sf95
- Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2), 230–257. doi:10.1177/0049124194023002004
- Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association.
- Xia, H., An, W., & Zhang, Z. J. (2023). Credit risk models for financial fraud detection: A new outlier feature analysis method of XGBoost with SMOTE. [JDM]. *Journal of Database Management*, 34(1), 1–20. doi:10.4018/JDM.321739
- Xie, Y., Qiu, M., Zhang, H., Peng, L., & Chen, Z. (2020). Gaussian distribution-based oversampling for imbalanced data classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(2), 667–679. doi:10.1109/TKDE.2020.2985965
- Yang, A. S. (2009). Exploring adoption difficulties in mobile banking services. *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration*, 26(2), 136–149. 10.1002/cjas.102
- Yousafzai, S. Y., Pallister, J. G., & Foxall, G. R. (2005). Strategies for building and communicating trust in electronic banking: A field experiment. *Psychology and Marketing*, 22(2), 181–201. doi:10.1002/mar.20054
- Yu, C. S. (2012). Factors affecting individuals to adopt mobile banking: Empirical evidence from the UTAUT model. *Journal of Electronic Commerce Research*, 13(2), 104–121.
- Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR*, 54, (pp. 962–970). doi:10.48550/arXiv.1507.05259
- Zare, N., Haem, E., Lankarani, K. B., Heydari, S. T., & Barooti, E. (2013). Breast cancer risk factors in a defined population: Weighted logistic regression approach for rare events. *Journal of Breast Cancer*, 16(2), 214–219. doi:10.4048/jbc.2013.16.2.214 PMID:23843856
- Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), 1–38. doi:10.1145/3158369
- Zhao, Y., Wong, Z. S. Y., & Tsui, K. L. (2018). A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering*, 2018, 1–11. Advance online publication. doi:10.1155/2018/6275435 PMID:29951182
- Zhu, T., Lin, Y., & Liu, Y. (2020). Improving interpolation-based oversampling for imbalanced data learning. *Knowledge-Based Systems*, 187, 104826. doi:10.1016/j.knosys.2019.06.034
- Žliobaitė, I. (2010). Learning under concept drift: An overview. *arXiv preprint arXiv:1010.4784*.

APPENDIX A

Survey Questionnaire as in Table 21

Table 21. Survey questionnaire in brief

Are you...	
<input type="checkbox"/> Male	<input type="checkbox"/> Female
<input type="checkbox"/> Other	<input type="checkbox"/> Prefer not to say
What is your age?	
<input type="checkbox"/> _____	
What is your household's total annual income from all sources, including retirement and investment income?	
<input type="checkbox"/> Less than \$15,000	<input type="checkbox"/> \$15,000 to under \$35,000
<input type="checkbox"/> \$35,000 to under \$50,000	<input type="checkbox"/> \$50,000 to under \$100,000
<input type="checkbox"/> \$100,000 to under \$200,000	<input type="checkbox"/> \$200,000 or higher
<input type="checkbox"/> I don't have any income	
Do you have any children under the age of 18 living at least part-time in your household?	
<input type="checkbox"/> Yes	<input type="checkbox"/> No
Which of the following types of accounts do you have in your name or have for your household? (Select all that apply)	
<input type="checkbox"/> Checking	<input type="checkbox"/> Mortgage
<input type="checkbox"/> Savings / Money market	<input type="checkbox"/> Home equity loan or line of credit
<input type="checkbox"/> CDs	<input type="checkbox"/> IRA (Individual Retirement Account)
<input type="checkbox"/> Auto loan	<input type="checkbox"/> Investments (mutual funds, stocks/bonds, etc.)
<input type="checkbox"/> None of these	
Which P2P have you used in the past 12 months?	
<input type="checkbox"/> _____	
What is your relationship status?	
<input type="checkbox"/> Single	<input type="checkbox"/> Divorced
<input type="checkbox"/> Married	<input type="checkbox"/> Separated
<input type="checkbox"/> In a relationship similar to marriage	<input type="checkbox"/> Widowed
<input type="checkbox"/> Prefer not to answer	
What is the highest level of education you have achieved?	
<input type="checkbox"/> Less than a high school grad	<input type="checkbox"/> 4-year college degree/Bachelor's degree
<input type="checkbox"/> High school diploma or equivalent	<input type="checkbox"/> Graduate or professional degree
<input type="checkbox"/> Some college/Associate degree	<input type="checkbox"/> Prefer not to answer
Are you currently ...? (Select one)	
<input type="checkbox"/> Employed full-time	<input type="checkbox"/> Not employed and not looking for work
<input type="checkbox"/> Employed part-time	<input type="checkbox"/> Full-time student
<input type="checkbox"/> Self-employed	<input type="checkbox"/> Part-time student and employed
<input type="checkbox"/> Business owner	<input type="checkbox"/> Homemaker
<input type="checkbox"/> Not employed but looking for work	<input type="checkbox"/> Retired

APPENDIX B

Multicollinearity

Before proceeding further with regression analysis, we did a multicollinearity test. Multicollinearity refers to the linear relationship between two or more predictor variables (Farrar et al., 1967; Alin, 2010). The more multicollinear the data is, the less reliable the estimates are (Graham, 2003; Alin, 2010). We used the Variance Inflation Factor (VIF) test to detect if data is suffering from any multicollinearity (as used by Midi et al., 2010). Hair et al., (1995), Paul (2006), and Montgomery et al. (2012) suggest VIF over 10 as the strong indicator of multicollinearity. However, other researchers suggest that VIF less than 5 is appropriate (Hair et al., 2011). Table 22 provides VIF statistics for the original data we received from the bank. VIF for income_4 is slightly above 6 rest are under 5.

Table 23 provides VIF statistics with the categories that we used for our work.

Table 22. Variance inflation factor for original data

Variable	VIF	Variable	VIF	Variable	VIF
age	2.22	child_Y	1.31	employment_1	2.73
gender_M	1.11	saving_Y	1.18	employment_2	1.53
relationship_Y	1.30	cds_Y	1.10	employment_3	1.26
income_0	1.10	trust_No	1.32	employment_4	1.07
income_1	3.05	trust_Yes	1.32	employment_5	1.31
income_2	5.18	education_1	1.16	employment_6	1.22
income_3	4.29	education_2	2.12	employment_7	1.31
income_4	6.23	education_3	2.26	employment_8	1.08
income_5	4.60	education_4	1.99	employment_9	1.53

Table 23. VIF with new categories

Variable	VIF	Variable	VIF
age	1.85	employment_NotEmp	1.09
gender_M1	1.05	employment_NotWorking	1.64
income_Medium	1.53	employment_SelfEmp	1.06
income_Low	1.20	employment_Student	1.11
trust_No	1.32	relationship_Y	1.23
trust_Yes	1.32	saving_Y	1.14
education_UnderGrad	1.71	cds_Y	1.09
education_Grad	1.84	child_Y	1.27

Lavlin Agrawal completed his Ph.D. at the University at Buffalo in the Department of management science and Systems. His research interests include a variety of avenues spanning from Financial Technologies, Information Security and Assurance, Healthcare, and Data Analytics. Mr. Agrawal is a Presidential Scholar and holds a master's degree in Management Information Systems from the University at Buffalo. Another area of Mr. Agrawal's expertise is the Management of IT Projects. He holds several professional licenses and expert-level certifications in Project Management and Agile Methodologies such as PMP and CSM.

Pavankumar Mulgund is a Assistant Professor at the University of Memphis. He received his Ph.D. in management information system from the SUNY Buffalo. He has more than 12 years' corporate and consulting experience. He extensively researches digital health interventions and their impact on various stakeholders (e.g., patients, providers, payers, and public health organizations), and his interests include technology strategy, the business value of IT, and the application of novel technologies (e.g., AI, IoT, and blockchain) in the context of health information technology. Mr. Mulgund has published several papers in leading academic and industry journals, is a frequent speaker at IS conferences, and has consulted for several organizations. He has developed and taught graduate-level IS courses in database management systems, systems analysis and design, data visualization with Tableau, and experiential IT projects. Before joining SUNY Buffalo, Mr. Mulgund was leading product and delivery teams for a primary contractor of the US Centers for Medicare & Medicaid Services. He has also worked for IBM and Mindtree, among others. Mr. Mulgund holds several professional licenses and expert-level certifications in project management, Agile methodologies, design thinking, and digital business.

Raj Sharman is a Professor in the Management Science and Systems Department at The State University of New York at Buffalo. His expertise is in the areas of Artificial Intelligence, Patient Safety and HealthIT, Information Assurance and the use of biologically inspired computer security models, and Disaster Response Management. He has published widely in national and international journals and is the recipient of several grants from university and external agencies, including the National Science Foundation. He also served as the Director of MS in MIS program and as the Ph.D. Faculty Adviser for the Department of Management Science and Systems from 2005 to 2016. He received his Ph.D. in Computer Science and a Master of Science degree in Industrial Engineering from the Louisiana State University. He received his Bachelor's degree in Engineering and Master's Degree in Management from the Indian Institute of Technology, Bombay, India.