# 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*

iD https://orcid.org/0000-0001-8877-2336

Pavankumar Mulgund, University of Memphis, USA

iD https://orcid.org/0000-0001-8434-5070

Raj Sharman, University at Buffalo, USA

iD 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

*Corresponding Author

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 (Ramyachitra & 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

35 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/handling-imbalanced-data-with-weighted-logistic-regression-and-propensity-score-matching-methods/335888

## Related Content

Implementing an Object-Oriented Deductive Database Using Temporal Reasoning
Nihan Kesimand Marek Sergot (1996). *Journal of Database Management (pp. 21-34).*
www.irma-international.org/article/implementing-object-oriented-deductive-database/51170

Applying Database Techniques to the Semantic Web
Maria del Mar Roldán-García, Ismael Navas-Delgadoand José F. Aldana-Montes (2005). *Encyclopedia of Database Technologies and Applications (pp. 18-22).*
www.irma-international.org/chapter/applying-database-techniques-semantic-web/11116

Dynamic Integration in Multidatabase Systems
Wen-Syan Liand Chris Clifton (1996). *Journal of Database Management (pp. 28-40).*
www.irma-international.org/article/dynamic-integration-multidatabase-systems/51160

Designing Graph Databases With GRAPHED
Gustavo Cordeiro Galvão Van Erven, Rommel Novaes Carvalho, Waldeyr Mendes Cordeiro da Silva, Sergio Lifschitz, Harley Vera-Oliveraand Maristela Holanda (2019). *Journal of Database Management (pp. 41-60).*
www.irma-international.org/article/designing-graph-databases-with-graphed/230294

Towards Autonomic Workload Management in DBMSs
Baoning Niu, Patrick Martinand Wendy Powley (2011). *Theoretical and Practical Advances in Information Systems Development: Emerging Trends and Approaches (pp. 154-173).*
www.irma-international.org/chapter/towards-autonomic-workload-management-dbmss/52956