Using R to predict Transaction Fraud

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Allison Spahn, Himanshu Niraj Sethia , Marissa Urbanek , Cyrus Nguyen
Matthew A. Lanham
Purdue University, Mitch Daniels School of Business
alispahn@purdue.edu; hsethia@purdue.edu; murbane@purdue.edu; nguye778@purdue.edu; lanhamm@purdue.edu

ABSTRACT

This study assesses the impact of various financial metrics in predicting transaction fraud. Using the models generated, we formulate a model that helps financial institutions decide which metrics to lookout for inconsistencies that might help prevent transaction fraud.

INTRODUCTION

Credit card fraud remains one of the most pervasive and damaging forms of identity theft in the United States, with far-reaching consequences for individuals, financial institutions, and the broader economy. According to the Federal Trade Commission's 2024 Consumer Sentinel Network report, credit card fraud topped the list of identity theft types, with over 460,000 reports in a single year as shown in Figure 1. This figure represents only a fraction of the actual cases, as many incidents go unreported or undetected, underscoring the scale and seriousness of the problem.

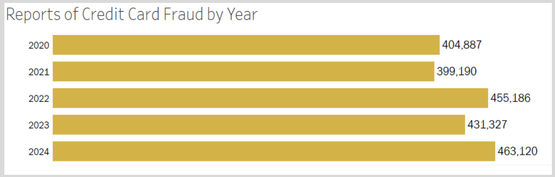

Fig 1. Reports of Credit Card Fraud by Year

For banks and credit card companies, fraud not only results in direct financial losses but also undermines consumer trust and requires heavy investment in fraud prevention infrastructure. In this context, the demand for more intelligent and proactive fraud detection solutions is growing rapidly, as traditional systems struggle to keep pace with the evolving threat landscape.

RESEARCH OBJECTIVES

- Develop a binary classification model to predict if a transaction is fraudulent or not
- Cluster customers based on similarity of certain characteristics, such as age, average transaction amount, etc.

LITERATURE REVIEW

Financial and credit card fraud have increased alongside the rise of e-commerce and digital transactions, making traditional detection systems inadequate. As a result, machine learning (ML) and deep learning (DL) approaches have gained traction for their ability to detect evolving fraud patterns. Supervised learning, including algorithms like Random Forest, Support Vector Machines (SVM), and Logistic Regression, has been widely used, with studies showing good performance in detecting fraudulent transactions. For instance, Baker et al. (2022) used ensemble learning and dimensionality reduction techniques to achieve high accuracy, although recall remained challenging due to the imbalance between legitimate and fraudulent transactions. Alarfaj et al. (2022) further demonstrated that convolutional neural networks (CNNs) outperformed traditional models, achieving a 99.9% accuracy, highlighting the potential of DL for this task. However, the issue of data imbalance, where fraudulent transactions make up a tiny fraction of the total data, persists, and techniques like Synthetic Minority Oversampling Technique (SMOTE) have been employed to address this.

While supervised models dominate the field, unsupervised methods and hybrid models are also gaining interest due to their ability to detect fraud without extensive labeled data. Techniques such as clustering, used by Bekireva et al. (2015), offer flexibility in handling unbalanced datasets. Evaluation metrics, including accuracy, precision, recall, and F1-score, are commonly used, with a focus on minimizing false positives and ensuring reliable fraud detection. Systematic reviews by Ali et al. (2022) and Hernandez Aros et al. (2024) emphasize the importance of using real-world data and adapting models to new fraud strategies.

METHODOLOGY

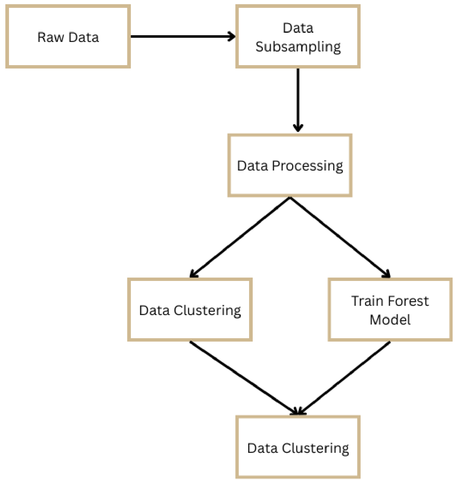

Fig 2. Methodology Diagram

STATISTICAL RESULTS

The Logistic Regression predictive model provided an AUC of 0.7193, with a 95% confidence intervals falling between (0.7078, 0.7307).

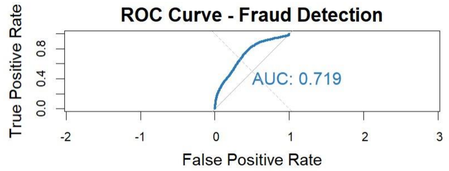

Fig 3. ROC Curve of Logistic Regression Model

The Random Forest model provided an accuracy of 85.45%, with a 95% confidence interval falling between (84.63%, 86.24%).

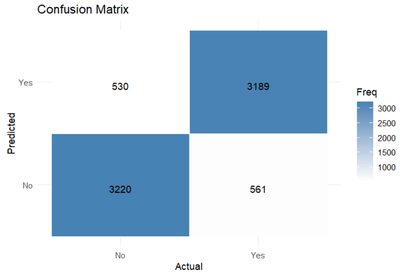

Fig 4. Confusion Matrix of Random Forest Model

Figure 5 showcases the weighting of each predictor measured by the decrease in Gini index. Throughout this we were able to find the most crucial features in classifying a transaction correctly.

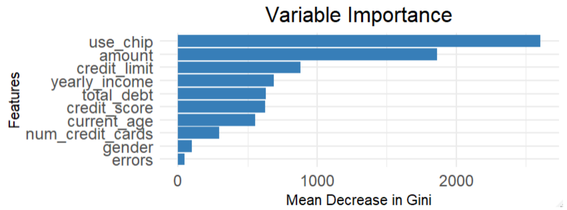

Fig 5. Variable Importance Plot of Random Forest Model

EXPECTED IMPACT

The fraud detection model, which achieved over 80% accuracy, demonstrates potential for real-world impact if implemented by a financial institution. Compared to a baseline 75% model, a 5% improvement could lead to a 20% increase in fraud prevention, 25% in cost savings, and a 15% increase in review efficiency, as shown in Figure 6.

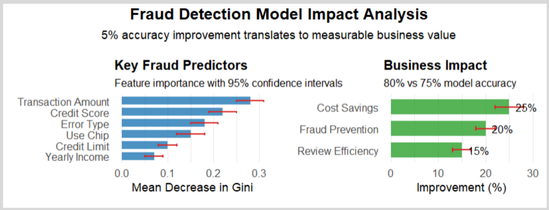

Fig 6. Predicted Business Impact of Improved Fraud Detection Accuracy

This translates to fewer financial losses, reduced operational strain, and improved customer trust. The model also identifies key fraud indicators, such as transaction amount and credit score, with confidence intervals to support their influence on predictions. Overall, even small improvements in accuracy can generate significant business value at scale. Additionally, this model could help compliance teams prioritize high-risk transactions more efficiently, freeing up resources for more strategic work.

CONCLUSIONS

When dealing with financial fraud, human review alone is not scalable enough to catch evolving threats. With the increasing volume of digital transactions, it is essential for financial institutions to utilize automated detection methods powered by machine learning.

Through our research, we have been able to answer the following:

- Supervised models like Random Forest can effectively classify fraudulent transactions
- Features such as chip usage, transaction amount, and credit limit were key predictors of fraud
- Clustering helped segment customers and uncover suspicious patterns

Given more real-time and balanced data, future models could improve recall and reduce false positives, enhancing fraud detection systems even further.

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham and our industry partner for this opportunity, their guidance, and support on this project.

https://github.com/murbane/Shiny_App