Mya Bass

Mbass83@wgu.edu

**Fraud Detection and Prediction on PaySim data**

011006868

WGU

1/26/2025

1. Statement of Problem and Hypothesis

Financial fraud is an issue that is exponentially increasing. Banks, financial institutions, insurance companies, and more need fraud analytics. Cybercriminals can easily hack credit card machines, ATMs, and online banks to siphon funds. Though, with the help of data analytics fighting against fraudulent activity is not only possible but essential. Fraud analytics tools can filter through large datasets of transactions to spot inconsistencies and fraud patterns. Allowing machine learning models to process the synthetic dataset I am using (PaySim) will produce an effective fraud detection and prediction tool. Models like Random Forests aid in diagnostic analysis and can regulate imbalanced datasets. In conclusion, applying data analysis to fraud detection using machine learning increases the accuracy of fraud prediction and provides a more scalable and adaptive solution. It allows financial institutions to detect fraudulent activities in real time, minimize financial losses, and enhance the overall security of digital payment systems.

This analysis aims to improve the understanding of the factors contributing to fraudulent transactions. The research question guiding this study is: To what extent do transaction amount, transaction type, and account balance changes affect the classification of fraudulent transactions? The insights gained from this analysis can help financial institutions detect fraud early, preventing financial losses and protecting users. Machine learning models can automate the detection of potential fraud in large volumes of transactions (Nanda, 2024).

The alternative hypothesis for this study is that transaction amount, transaction type, and account balance changes significantly affect the classification of fraudulent transactions, as measured by performance metrics such as precision, recall, and F1-score.

1. Summary of Data Analysis Process

The analysis includes a synthetic dataset from PaySim, a mobile provider that mimics real-world financial transactions. The data contains one month of financial logs in Africa. The dataset was chosen because it was previously studied for its similarity to real financial data. In that research, it was decided to be just as effective without creating potential risk for real customers. Once the data is loaded, the data preprocessing steps include handling missing values, feature selection, feature renaming, encoding categorical variables, and feature scaling. Following that, SMOTE (Synthetic Minority Oversampling Technique) was applied to attempt to balance fraudulent and authentic transactions. Feature engineering was also essential to enhancing the model’s predictive power in identifying fraudulent patterns.

The next step was to apply several machine-learning models to classify transactions. The models included logistic regression, which is an easily interpreted model suitable for binary classification, and random forest, which performs well on imbalanced datasets and captures complex relationships between features. The last two additional models used were k-nearest neighbors, a distance-based algorithm that classifies transactions based on the majority class, and neural networks that use deep learning models capable of understanding non-linear relationships in data. Each model was trained on the preprocessed dataset, and performance metrics were evaluated using precision, recall, F1-score, and AUC-ROC. Lastly feature importance was analyzed to determine the most impactful variables.

1. Outline of Findings

The evaluation metrics determined that random forest and neural networks were the most effective at differentiating the transaction classifications. The analysis displayed that the transaction amount and account balance change were the most influential features in detecting fraudulent transactions. Balance changes, as in the originating and destination accounts, have the highest feature importance in the Random forest model. Transaction type contributed significantly to fraud detection, especially “Cash\_Out”. The models performed well for non-fraudulent transactions, though there were issues with detecting fraudulent transactions due to low recall. Suggesting that improvements are needed in detecting fraudulent transactions.

1. Limitations

There were several limitations while completing this project. The class imbalance was the biggest issue. Due to the non-fraudulent transactions outnumbering the fraudulent one, the models had difficulty distinguishing the fraudulent transactions. The imbalance led to high accuracy scores but poor recall for detecting fraudulent transactions. SMOTE addressed some of the imbalance, but it could still be an issue because synthetic data is used. Logistic Regression struggles to capture the complex relationships between features, making it less effective than more sophisticated models like random forests or neural networks.

1. Proposed Actions

Feature engineering can be explored further to create more meaningful features that may improve the model’s predictive capabilities. By looking into adding transaction frequency or time-related features, we can predict patterns of behavior over time. Model selection and tuning can also be a good next step in furthering research with more sophisticated techniques like ensemble methods like gradient boosting ( Barnard, 2024). These models can capture complex relationships within the data that simpler models like logistic regression might miss (Barnard, 2024). Integrating anomaly detection techniques like isolation forest can help identify fraudulent transactions deviating from typical patterns (Barnard, 2024). Model calibration could also help improve the accuracy of predicting models, creating more reliable fraud risk assessments. By focusing on these areas, the model can better adapt to the evolving nature of fraud and enhance its effectiveness in real-world applications.

1. Benefits of the Study

Enhancing recall from 46% (observed in certain models) to approximately 70-80% will make the system more efficient at detecting fraudulent transactions, thus minimizing the risk of financial losses. Moreover, utilizing time-series data alongside ensemble techniques can enhance the model's robustness and accuracy, transforming it into a more trustworthy tool for real-time fraud detection in financial systems. By pinpointing the critical characteristics that contribute to fraudulent transactions, financial organizations can more effectively distribute resources to oversee transactions with high-risk signs. Despite the decrease in fraudulent transactions, this focused strategy allows for a more effective utilization of technology and resources.

The research further emphasizes the constraints and potential enhancements, enabling adopting actionable strategies, like altering decision thresholds or incorporating sophisticated methods, to identify fraud more efficiently in financial systems.

**Presentation of Findings**

B.  Attached: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=47eb5e3c-7dae-4241-8b41-b27300644c83>

C. Citations

Alonso Lopez-Rojas, E., Axelsson, S., & Elmir, A. (2016, September). *PAYSIM: A financial mobile money simulator for fraud detection*. Research Gate. https://www.researchgate.net/publication/313138956\_PAYSIM\_A\_FINANCIAL\_MOBILE\_MONEY\_SIMULATOR\_FOR\_FRAUD\_DETECTION

Barnard, J. (2024, December 17). *What is anomaly detection?*. IBM. https://www.ibm.com/think/topics/anomaly-detection

Nanda, A. (2024, October 12). *What is fraud analytics? techniques, workflows, and Tools*. DataCamp. https://www.datacamp.com/blog/fraud-analytics