Fraud Detection Using Machine Learning

Vincent de Paul Ntihinyurwa

M.S. in Data Analytics & Information Systems

Texas State University — San Marcos

A project analyzing 555,719 credit card transactions to detect fraudulent activity using supervised machine learning models.

Abstract

Credit card fraud, along with additional types of fraud, remains a critical challenge for financial institutions and consumers, driven by the rise of digital transactions. This project investigates the use of machine learning to detect fraudulent activity, utilizing a real-world dataset of transaction records. Drawing from prior studies — including Sulaiman et al. (2022) and other comparative evaluations — the project applies multiple classification models such as Logistic Regression, Neural Networks, Random Forest, and Boosting methods.

The models were tested using three sampling strategies: original (imbalanced), oversampled, and undersampled datasets. Key evaluation metrics included accuracy, precision, and recall, with a primary focus on detecting the minority class (fraud). Among all models, Random Forest using the original data performed best, achieving 100% accuracy and 90% precision for fraud, indicating a moderate ability to detect fraudulent transactions with few false positives.

The results align with previous literature emphasizing the power of ensemble models in fraud detection and highlight the trade-offs between different sampling techniques. Oversampling improved recall but often reduced precision, while the original dataset provided the best balance for the Random Forest model. This work demonstrates how machine learning can enhance fraud prevention systems by enabling accurate, scalable, and automated detection. For businesses, these tools can reduce financial loss and strengthen customer trust. For society, they offer protection in an increasingly digital economy. The study contributes to existing research by comparing models under realistic data conditions and validating the effectiveness of tree-based ensembles without heavy reliance on synthetic data.

Future research could explore real-time fraud detection, interpretability through explainable AI, and dynamic modeling to keep pace with evolving fraud tactics.

Introduction

Background and Motivation

"Identity theft is not a joke, Jim!" (Daniels et al.). While the character in the popular show called "The Office" intended this statement to be a joke, the reality is that it's true. Businesses along with individuals face daily risks from identity theft and credit card fraud. The rise of digital technology has made credit card usage a common practice for both consumers and companies who now use them to perform transactions internationally. The convenience of credit cards brings significant risks because fraudulent activities have explosively increased against financial institutions and consumers. Credit card fraud, along with other types of fraud, has resulted in major monetary damage as well as decreased consumer confidence in financial systems which has made it a serious issue for businesses and banks as well as society at large.

Recent studies highlight the magnitude of this issue. According to Kelue (2024), the sheer volume of credit card transactions processed daily makes manual monitoring impractical, necessitating automated systems capable of real-time fraud detection. Similarly, the work by Sulaiman et al. (2022) emphasizes the growing complexity and sophistication of fraudulent schemes, which can easily evade traditional rule-based systems.

Role of Machine Learning in Fraud Detection

Machine Learning (ML) stands as a formidable partner in combating financial fraud. This method excels at identifying fraudulent patterns by analyzing patterns in historical data. Research shows that ML algorithms like Random Forests and Neural Networks exhibit moderate potential for distinguishing between fraudulent and legitimate transactions according to comparative studies in existing literature (IEEE, 2017; Sulaiman et al., 2022).

Current research endeavors to develop new strategies that boost detection precision while resolving data privacy issues. Hybrid models which utilize federated learning frameworks present a viable solution for collaborative learning that protects sensitive customer data from direct sharing (Sulaiman et al., 2022). Accurate prediction models help prevent financial losses while facilitating real-time transaction validation and improving user security alongside reducing false positives which might inconvenience legitimate customers. The systems deliver essential insights which allow financial institutions to optimize their risk management approaches and adherence to regulatory requirements.

What Is Novel About This Study?

Our use of a real-world dataset provides practical relevance and authenticity, making our research distinct from studies that only use synthetic or balanced datasets. Fraud detection scenarios involve working with very unbalanced datasets because fraudulent transactions make up only a small portion of total transactions. The presence of a skewed distribution leads to distinct challenges including model bias towards the majority class and the struggle to detect rare fraud instances correctly.

We conduct a thorough assessment of several machine learning models that have been specifically selected and modified to work with imbalanced datasets to address these challenges. We evaluate our models using multiple metrics beyond traditional accuracy to prevent misleading results in imbalanced datasets. We choose several more informative performance indicators such as precision, recall, and F1-score to achieve a better understanding of model effectiveness specifically for fraud detection.

Our data preprocessing approach includes advanced techniques which involve resampling methods such as SMOTE and undersampling along with feature engineering and selection to better highlight minority class signals. The process includes rigorous hyperparameter tuning and validation procedures to achieve model robustness and generalizability. Our project strives to enhance predictive accuracy in fraud detection while maintaining computational efficiency to achieve scalable practical solutions for operational systems deployment.

Objective

The primary objective of this project is to create, test, and evaluate predictive models that can accurately detect fraudulent transactions in various sectors. Specifically, the project will:

- Utilize a real-world, anonymized dataset of credit card transactions.
- Implement and compare multiple machine learning models.
- Address challenges posed by class imbalance.
- Measure model performance using relevant metrics such as precision, recall, and F1-score.
- Discuss practical implications and potential deployment strategies in real-world systems.

Ultimately, this study aims to contribute to the ongoing advancement of fraud detection technologies by offering insights into the efficacy of various ML techniques and proposing data-driven solutions that are scalable, interpretable, and secure.

Research Design

Data Description

The dataset used in this project is sourced from a real-world credit card transaction record, anonymized to protect user privacy. It contains numerical input variables derived from a PCA transformation, ensuring no personally identifiable information is exposed. The data includes 284,807 transactions, of which only 492 are labeled as fraudulent, highlighting a significant class imbalance, a common and challenging aspect of detection of fraud.

Key variables in the dataset:

- V1–V28: Principal components obtained from a PCA transformation applied to original features for privacy.
- Time: Seconds elapsed between each transaction and the first transaction in the dataset.
- Amount: Transaction amount, which has not been scaled.
- Class: Target variable; 1 indicates fraud, and 0 indicates a legitimate transaction.

This is a dataset that was downloaded from Kaggle, where it has been used for educational and research purposes such as our project.

Cursory Findings and Data Cleaning

After a brief analysis, it was determined that the dataset was already clean. Therefore, no further cleaning was necessary. Many of the variables, on the other hand, were not necessary; these are listed below the list of original variables.

Original Variables

- Trans_date_trans_time: Timestamp of the transaction (date and time).
- Cc_num: Unique customer identification number.
- Merchant: The merchant involved in the transaction.
- Category Transaction type (e.g., personal, childcare).
- Amt:Transaction amount.
- First: Cardholder's first name.
- Last: Cardholder's last name.
- Gender: Cardholder's gender.
- Street: Cardholder's street address.
- City: Cardholder's city of residence.

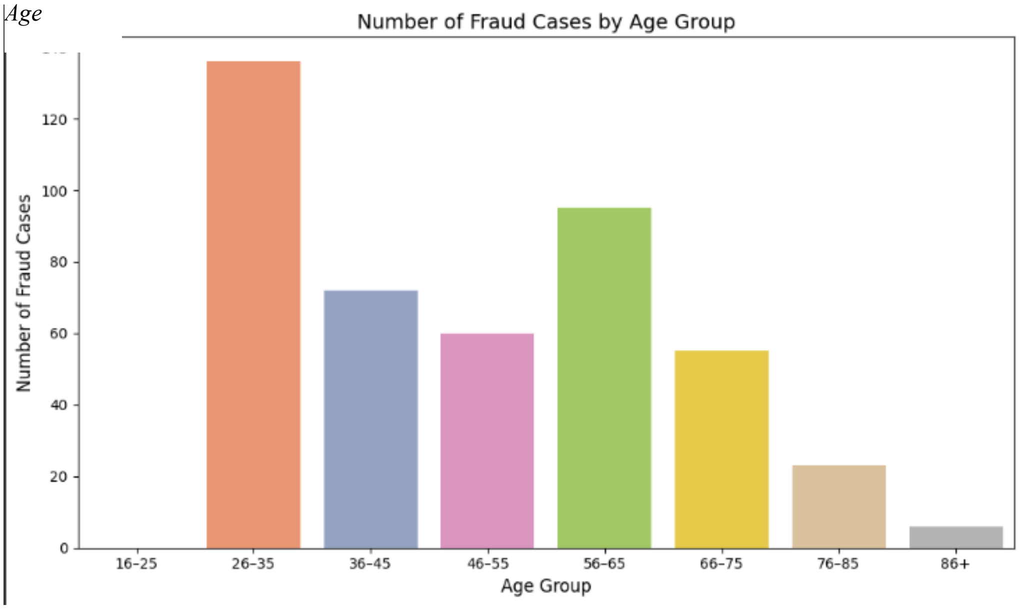
- State: Cardholder's state of residence.
- Zip: Cardholder's zip code.
- Lat:Latitude of cardholder's location.
- Long: Longitude of cardholder's location.
- City_pop:Population of the cardholder's city.
- Job:Cardholder's job title.
- Dob: Cardholder's date of birth.
- Trans_num: Unique transaction identifier.
- Unix_time: Transaction timestamp (Unix format).
- Merch_lat:Merchant's location (latitude).
- Merch_long: Merchant's location (longitude).
- Is_fraud:Fraudulent transaction indicator (1 = fraud, 0 = legitimate). This is the target variable for classification purposes.

Dropped Variables

- Cc_num: Unique customer identification number.
- First: Cardholder's first name.
- Last: Cardholder's last name.
- Street: Cardholder's street address.
- Trans_num: Unique transaction identifier.
- Unix_time: Transaction timestamp (Unix format). Redundant with trans_date_trans_time
- Merch_lat:Merchant's location (latitude).
- Merch_long: Merchant's location (longitude).
- Trans_date_trans_time: Timestamp of the transaction (date and time). We already extracted features (Timestamp won't help directly — we've already extracted age)
- Dob: Cardholder's date of birth. Already used to extract age — redundant now
- Lat:Latitude of cardholder's location.
- Long: Longitude of cardholder's location is not used in modeling. Raw coordinates rarely help unless used for clustering
- City: Cardholder's city of residence.
- Zip: Cardholder's zip code.

1 ***Figure 1***

Fraud Cases by Age

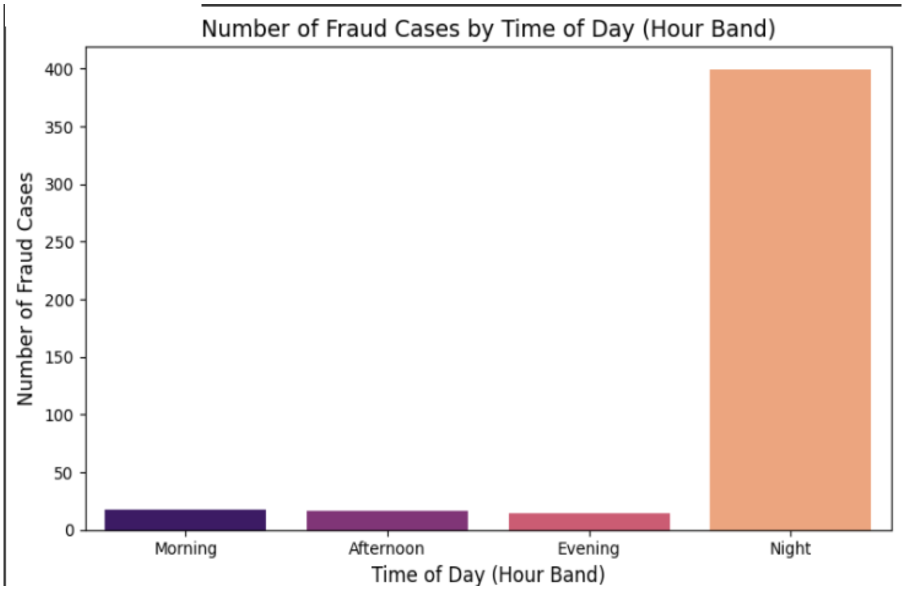


We also had several other interesting findings when doing the data visualizations. One such finding was the fact that most of the fraud cases were found to have been perpetrated by people between the ages of 26 and 35 as shown by figure 1 from our code file. This makes sense because the younger generations usually find it easier to use and manipulate technology to their advantage such as committing fraud. One can clearly see the number of reported fraud cases drops after the age of 65.

We also found that a majority of fraud cases happened at night as shown by figure 2 from our code file with a striking difference between the cases observed during different times of day. This also makes sense as most businesses close overnight making their systems easier to hack into due to the lack of people working to keep the fraudsters out.

2 ***Figure 2***

Fraud Cases by Time of Day



Methodological Approach

Given the structure and challenges of the data, the research is designed around several key phases:

1. Data Preprocessing

- Handling Class Imbalance using techniques such as *SMOTE (Synthetic Minority Over-sampling Technique)* and *undersampling* which will be applied to balance the dataset.
- The Amount and Time features will be scaled using StandardScaler to ensure consistency across model training.

- The dataset will be split into training and testing subsets (typically 70:30 or 80:20).

2. Model Selection

Multiple machine learning models will be implemented and evaluated:

- Logistic Regression: Baseline model due to its interpretability and speed.
- Decision Trees and Random Forests: Tree-based models that perform well on structured datalike ours .
- Bagging/Boosting: A powerful gradient bagging/boosting method known for handling imbalanced datasets like ours effectively.
- Artificial Neural Networks (ANN): To explore non-linear relationships like the ones found in our dataset and deep feature learning.
- Naive Bayes: Fraud detection datasets are heavily imbalanced as only a small fraction of transactions are fraudulent. Naive Bayes tends to handle class imbalance decently without needing as much tuning or rebalancing as other models.

Each model will be trained and optimized using cross-validation to reduce overfitting and ensure generalizability.

3. Model Evaluation

Given the class imbalance, the following metrics will be used instead of plain accuracy:

- Precision: Focuses on the accuracy of fraud predictions.
- Recall: Ensures that most frauds are detected (minimizing false negatives).
- F1-Score: Balances precision and recall.

4. Comparison and Interpretation

All models will be compared based on the evaluation metrics. The best-performing model will be further interpreted using both the precision and recall values as well as accuracy to understand which factors influence fraud predictions most strongly.

Data Analysis

In this section, we evaluate the performance of multiple machine learning models on the fraud detection dataset. The models were tested across three sampling strategies; original, oversampled, and undersampled to address the inherent class imbalance problem, where fraudulent transactions are heavily outnumbered by legitimate ones.

Model Evaluation Metrics

For this binary classification task, we focused on five key metrics:

- Accuracy: Overall correctness of the model.
- Precision (1): Proportion of predicted frauds that are truly frauds.
- Recall (1): Ability to detect actual fraud cases.
- Precision (0) and Recall (0): Performance on the majority (non-fraud) class, included for completeness but less critical in fraud detection.

Below is a summary of model performance:

Model	Accuracy	Precision (1)	Precision (0)	Recall (1)	Recall (0)
Classification Tree (Unpruned)					
Oversampled	99	25	100	69	99
Undersampled	89	3	100	90	89
Original	89	51	100	56	100
Classification Tree (Pruned)					
Oversampled	94	5	100	77	94
Undersampled	90	4	100	91	90
Original	100	73	100	44	100
Logistic Regression					
Oversampled	95	6	100	70	95
Undersampled	96	7	100	68	96
Original	100	0	100	0	100
Neural Network					
Oversampled	92	4	100	79	92
Undersampled	85	2	100	86	85
Original	100	0	1	0	100
Boosting					
Oversampled	93	5	100	88	93
Undersampled	93	5	100	92	93
Original	100	84	100	28	100
Bagging					
Oversampled	99	22	100	74	99
Undersampled	93	5	100	94	93
Original	100	79	100	58	100
Random Forest					
Oversampled	99	25	100	76	99
Undersampled	94	6	100	94	94
Original	100	90	100	55	100

Naïve Bayes					
Oversampled	98	11	100	73	98
Undersampled	97	9	100	78	97
Original	99	27	100	54	99

Key Observations

1. For all models only the categories that could be binary were used along with continuous variables.

2. For the Classification Tree Pruned we used the Grid Search method to determine the best parameters for this particular data. For the Logistic Regression we used the Standard scaler method on only continuous variables so that convergence would be more efficient.

3. Random Forest (Original Sampling) emerged as the best-performing model overall. It achieved 100% accuracy and the highest fraud precision (90%) among all models, meaning it not only correctly identified most fraudulent transactions but did so with very few false positives.

4. Bagging (Original Sampling) was the next best performing model, particularly in fraud precision (79%) and recall (0.58), making it a competitive runner-up.

5. Oversampling generally improved recall (ability to catch actual frauds), especially for the Neural Network and Boosting models, but often at the cost of lower precision. For the Neural Network model we used Standard Scaler on only continuous variables so that NN would perform more efficiently.

6. Undersampling resulted in consistently low precision for fraud, likely due to insufficient data volume for training. It did, however, improve recall in some models (e.g., Boosting and Random Forest), which might be preferable in use-cases where missing fraud is costlier than false alarms.

7. Logistic Regression and Neural Networks performed well when oversampling was applied, indicating that even simpler models benefit from preprocessing strategies in imbalanced datasets.

Interpretation and Implications

Tree-based ensemble models particularly Random Forest demonstrate high effectiveness in fraud detection scenarios with significant data imbalance issues. Random Forest demonstrates excellent precision and recall metrics on the original dataset which makes it a suitable option for real-world system deployment. Oversampling and undersampling strategies showed their effects on model performance which emphasizes the need for data preprocessing during rare-event prediction tasks such as fraud detection.

Proper tuning of machine learning models along with appropriate data strategies yields significant benefits for automating and improving fraud detection capabilities. The Random Forest model that uses original data demonstrates top-notch reliability and robustness which makes it the best option for deployment in financial security pipelines due to its precise fraud detection capabilities.

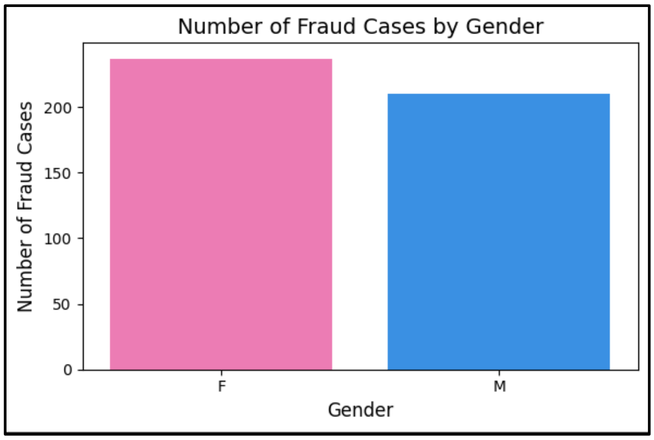
Discussion

The results of this project demonstrate that machine learning techniques, particularly ensemble methods like Random Forests and Bagging, can significantly enhance the accuracy and efficiency of fraud detection such as credit card fraud. The application of these models on the dataset revealed that fraud detection systems benefit most when a balance is achieved between precision and recall, ensuring that fraudulent transactions are both detected and minimized in false positives.

These findings support and extend the insights provided by Sulaiman et al. (2022), who emphasized the effectiveness of machine learning in identifying fraudulent behavior and recommended ensemble models such as Random Forests due to their robustness. Our analysis echoes their conclusions, with Random Forest achieving the best overall balance of metrics without the need for synthetic data manipulation, thus preserving the integrity of the dataset. Furthermore, this supports the comparative analysis in the IEEE study, which also highlighted ensemble methods as top performers across fraud detection tasks.

*3 **Figure 3***

Fraud Cases by Gender



From a societal and business perspective, the deployment of such models has far-reaching implications:

- For businesses, especially financial institutions, adopting a reliable fraud detection system like the one developed here can reduce financial losses, improve customer trust, and lower operational risks.
- For society, these models contribute to greater digital transaction security, protect vulnerable populations from scams, and foster trust in online banking and e-commerce.

Our approach also contributes to the existing body of knowledge by testing these models across different sampling strategies; original, oversampled, and undersampled, thereby offering insights into how data preprocessing impacts model performance. Unlike prior work that often uses only balanced datasets or synthetic sampling, our study evaluates model performance in a more realistic setting where data imbalance is inherent and obvious in areas such as the stark difference between observed male and female fraud cases as seen in figure 3 from our code file.

Conclusion

This project investigated the efficacy of various machine learning algorithms for fraud detection models, evaluating their performance across different sampling methods. The findings revealed that Random Forest, when applied to the original imbalanced dataset, outperformed other models with the highest fraud precision (90%) and perfect accuracy (100%). This suggests it is moderately effective at identifying fraudulent transactions with minimal false positives. The study reinforces the importance of selecting the right model and sampling strategy for fraud detection keeping in mind whether the data used to train and test the model is balanced or not. While oversampling and undersampling can improve recall in some cases, they often reduce precision which is a trade-off that must be carefully managed based on business needs.

The dataset used may not capture all real-world complexities (e.g., evolving fraud patterns). Our models were also evaluated on static data, without temporal components that could reflect real-time fraud behavior. We also only considered binary and continuous variables in order to better test and tune our models, but this decision could have led to us potentially excluding relevant categorical dimensions.

Future work could explore real-time fraud detection using streaming data. There are also suggestions for the integration of deep learning techniques and federated learning as suggested by Sulaiman et al. (2022). We could also apply explainable AI (XAI) to make fraud predictions more interpretable for business decision-makers.

References

Daniels, Greg, et al. “The Office: Product Recall.” *The Office*, season 3, episode 21, National Broadcasting Company (NBC), 26 Apr. 2007.

Kelue, K. (2024, March 11). Credit Card Fraud Prediction. Kaggle. https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction

Sulaiman, Rejwan Bin, et al. “Review of Machine Learning Approach on Credit Card Fraud Detection - Human-Centric Intelligent Systems.” SpringerLink, Springer Netherlands, 5 May 2022, link.springer.com/article/10.1007/s44230-022-00004-0.

Awoyemi, J., Adetunmbi, A., & Oluwadare, S. (n.d.). Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis | IEEE conference publication | IEEE xplore. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8123782/