VINUNIVERSITY
Data Mining and Big Data Analytics

Project Report

Enhancing Financial Security:
Credit Card Fraud Detection Using XGBoost

Team member: Tran Tue Nhi

1. Project Background

a. Overview:

The increasing prevalence of digital transactions has given rise to sophisticated methods of financial fraud, particularly credit card fraud (Bolton & Hand, 2002). This surge in fraudulent activity not only poses a significant risk to consumer security but also affects the financial integrity of institutions. The central question this project aims to address is: How can credit card fraud be identified quickly and efficiently? Leveraging advanced analytics and machine learning technologies, this project seeks to develop a system capable of detecting fraudulent transactions in real-time, thus minimizing potential losses and enhancing security measures.

b. Motivation:

The motivation for this project stems from the urgent need to combat the growing issue of credit card fraud, which has been exacerbated by the rise in online transactions. Financial losses from fraud are staggering, with billions lost annually worldwide, affecting both consumers and financial institutions. For instance, according to the 2021 Identity Fraud Study by Javelin Strategy & Research, credit card fraud losses increased to $4.2 billion in 2020, up from $3.5 billion in 2019 (Understanding the Threat of Card Transaction Fraud and Its Impact on the Financial Ecosystem, 2023). Additionally, the global cost of fraud to the economy is estimated at $41 billion in 2022 (Payment Fraud Detection and Prevention, 2023). By developing a predictive model that can accurately and swiftly identify fraudulent transactions, this project aims to provide a proactive solution to a pervasive problem. The implementation of such a system not only helps in preventing monetary losses but also secures the trust of customers, which is paramount in the financial services industry. Research by McKinsey emphasizes that integrating fraud detection with cybersecurity measures can significantly improve threat prediction and customer trust (Hasham et al., 2019).

2. Model Background: XGBoost

XGBoost (eXtreme Gradient Boosting) was selected for this project due to several compelling reasons:

- **Performance:** XGBoost is renowned for its high performance on large datasets, which is typical in transaction data scenarios. It excels in delivering superior results in classification problems, particularly when dealing with imbalanced datasets like fraud detection (Chen & Guestrin, 2016). Its implementation of gradient boosting algorithms provides efficient, scalable, and accurate solutions that are essential for real-time fraud detection systems.
- **Interpretability:** One of the significant advantages of XGBoost is its ability to provide insights into feature importance. Unlike more opaque models such as deep neural networks, XGBoost offers an interpretable framework where the contribution of each feature to the final prediction can be understood and analyzed. This interpretability is crucial for iterative model improvement and gaining trust from stakeholders by elucidating the factors driving the predictions.
- **Novelty in Application:** While traditional financial fraud detection models often utilize algorithms like Random Forest, Naive Bayes, and Logistic Regression, there is limited research specifically focusing on the application of XGBoost in this domain. Studies such as those by Kou et al. (2004), Kumar et al. (2019), and Liu et al. (2015) emphasize the widespread use of traditional models in fraud detection but do not extensively cover XGBoost's application. This project explores the relatively under-researched application of XGBoost for fraud detection, presenting an opportunity for novel findings and advancements in the field. By leveraging XGBoost, this project aims to push the boundaries of current fraud detection methodologies and explore new insights into its efficacy and implementation in financial security.

=> XGBoost offers a powerful and interpretable solution for the challenging task of credit card fraud detection. Its robust performance on large, imbalanced datasets, coupled with the ability to gain insights into feature importance, makes it an ideal choice for developing an effective fraud detection system.

3. Dataset Description

a. Data Overview and Sources:

The dataset selected for this project, titled "Credit Card Transactions Fraud Detection Dataset" is obtained from Kaggle and encompasses simulated credit card transactions for the period of January 1, 2019, to December 31, 2020. It documents the financial interactions of 1,000 customers with 800 distinct merchants. The dataset provides a comprehensive mix of legitimate and fraudulent transactions, offering a broad spectrum of data for in-depth analysis. Notably, the data is characterized by a significant imbalance where fraudulent transactions constitute a minor fraction of the total transactions, exemplifying a typical challenge in fraud detection that necessitates sophisticated analytical strategies to accurately identify infrequent fraudulent activities.

b. Types of Data: The dataset is comprised of a variety of data types, which include:

- **Numerical Data:** This includes fields such as transaction amount (amt), latitude (lat), longitude (long), city population (city_pop), transaction coordinates (merch_lat, merch_long), and fraud status (is_fraud).
- **Categorical Data:** Includes zip code (zip), credit card number (cc_num), merchant names, transaction categories, first and last names of the cardholder, gender, street, city, state, occupation (job), and transaction identifiers (trans_num).
- **Time Series Data:** Features like transaction date and time (trans_date_trans_time) and date of birth (dob) of the cardholder.

c. Compatibility with XGBoost:

To utilize the dataset with the XGBoost model, several preprocessing steps are required to prepare the data adequately:

- **Data Cleaning:** Ensuring the dataset is free from errors or irrelevant information.
- **Normalization:** Adjusting the scale of the numerical features to bring consistency.
- **Encoding:** Transforming categorical variables into a machine-readable format.
- **Feature Engineering:** Developing new features or modifying existing features to enhance model performance and predictive capability.

4. Project Pipeline

You can access to the Github of the project including the source code, dataset and project presentation here: http://github.com/ellynnhitran/credit-card-fraud-detection-xgboost.git

4.1 Import Libraries: The project begins by importing essential libraries for data manipulation, visualization, and machine learning.

4.2 Data Collection:

- **Collect Data:** The dataset, sourced from Kaggle, includes simulated credit card transactions from January 1, 2019, to December 31, 2020. It consists of records from 1,000 customers and 800 merchants, providing a rich diversity of data points for analysis.
- **Clean Data:** Checking and handling missing values ensures data integrity.
- **Feature Engineering:** New feature was created to capture patterns that might indicate fraud. hour_of_day were created by converting 'unix_time' to 'transaction_time' and extract 'hour_of_day' to understand temporal patterns.

4.3 Exploratory Data Analysis (EDA):

EDA is conducted to uncover patterns and inform feature engineering.

- **Spending Category vs. Fraud:** EDA showed significant variation in fraud occurrences across different spending categories. This suggests that 'category' might be a strong predictor of fraud.

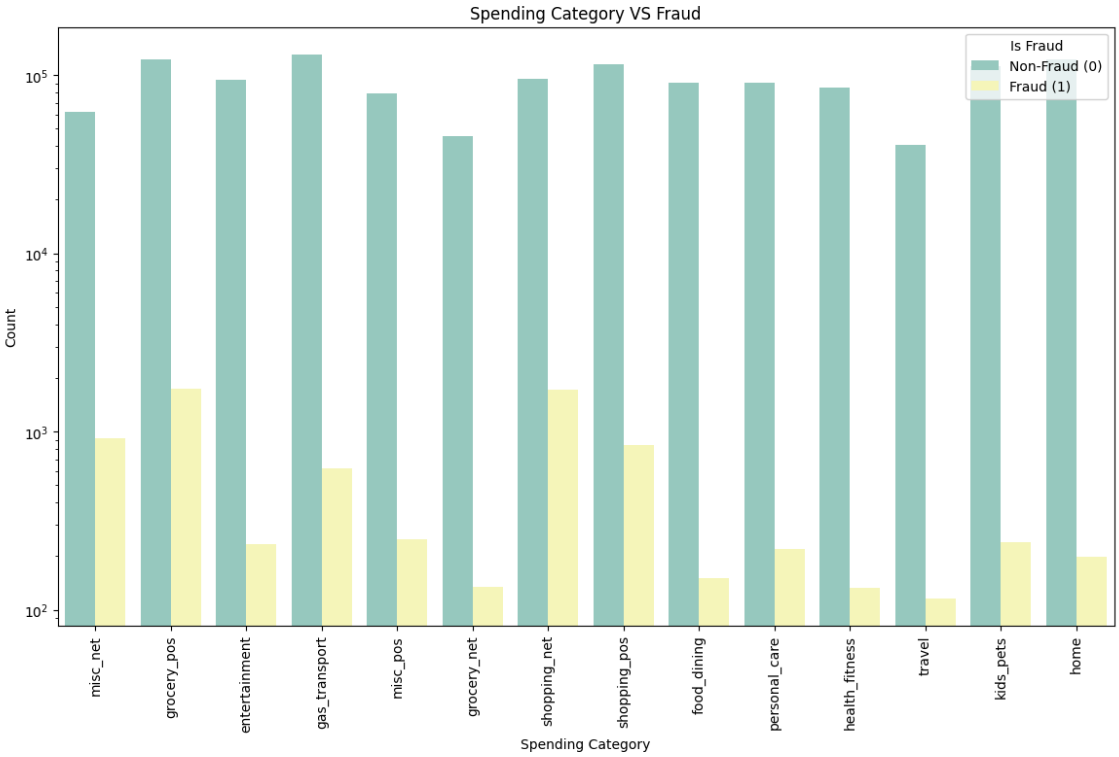


Figure 1: Spending categories susceptible to fraud

- **Gender vs. Fraud:** Gender did not show a significant difference in fraud rates, indicating that this is not a crucial factor in deciding fraud possibilities.

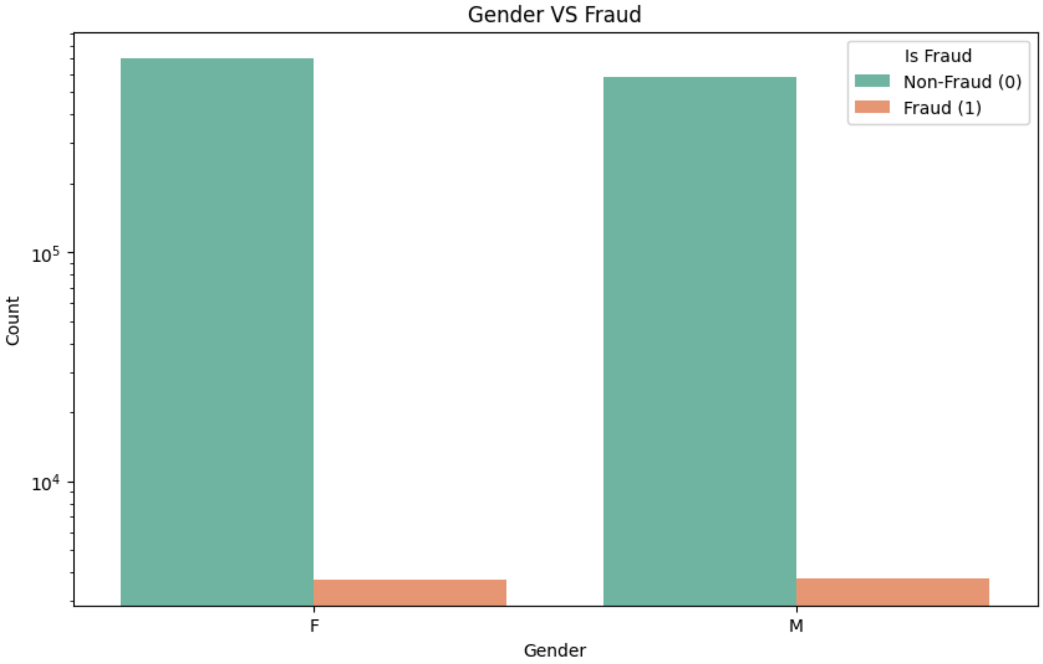


Figure 2: Is there gender bias in fraudulent transactions?

- **States vs. Fraud:** The distribution across states did vary, indicating some geographical patterns in fraud occurrences.

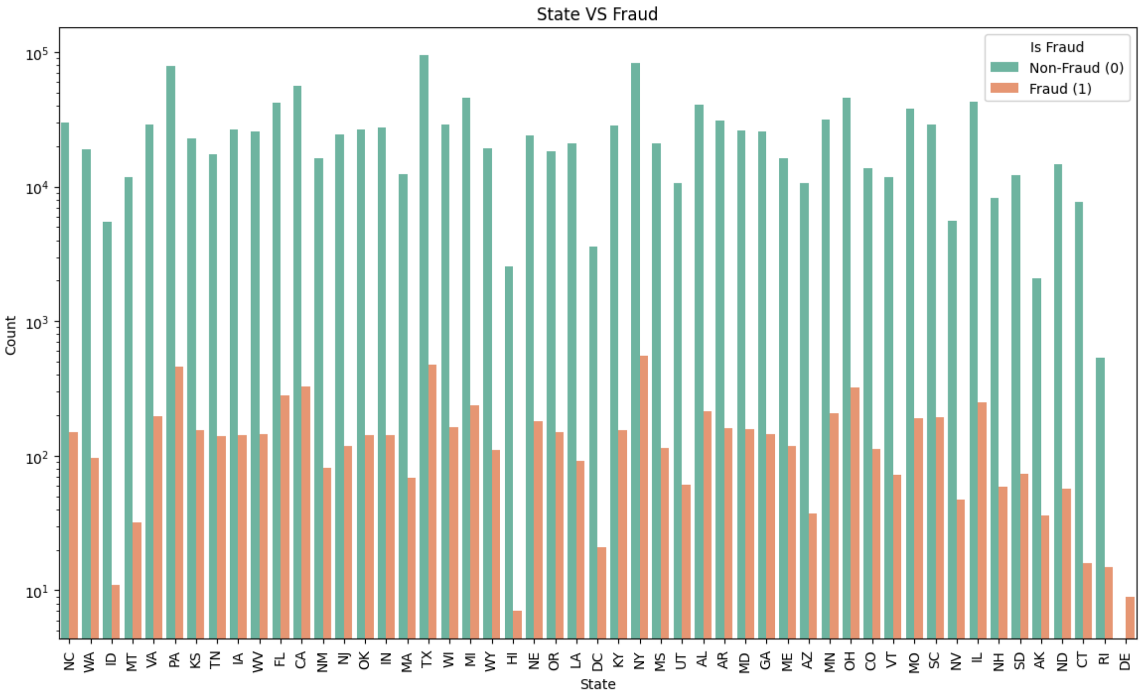


Figure 3: States with its fraud rates for geographically targeted interventions

- **Hour of Day vs. Fraud:** There was a noticeable trend of increased fraudulent transactions during late-night hours from 10PM to 3AM.

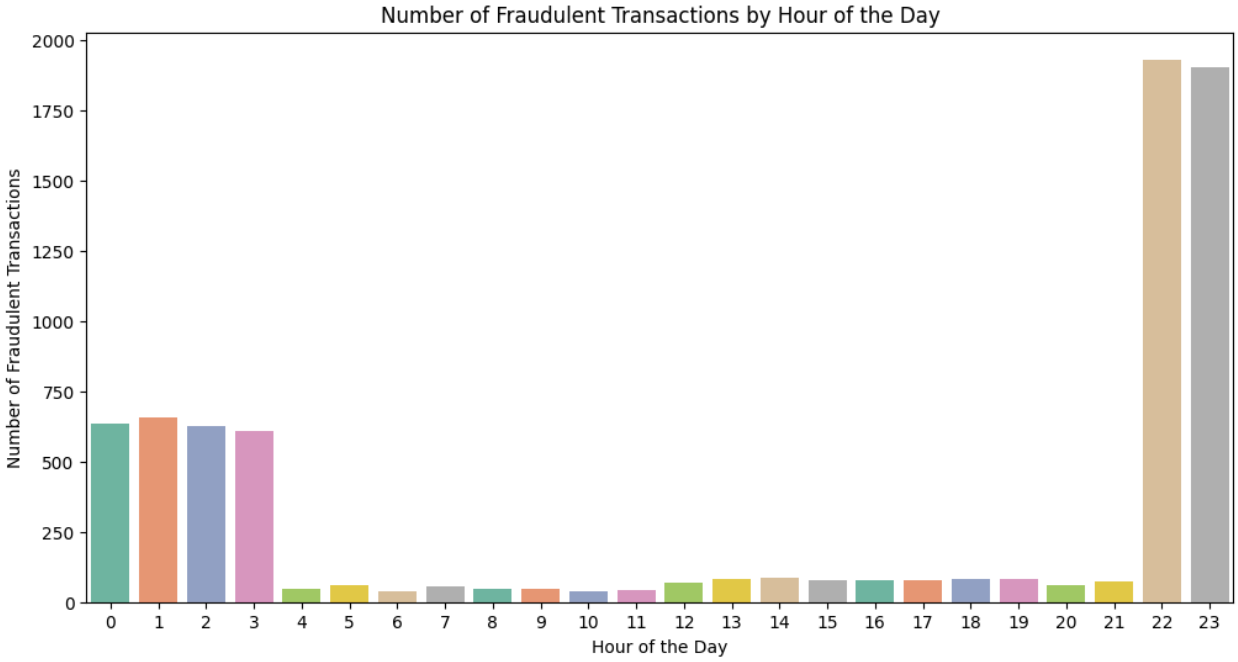
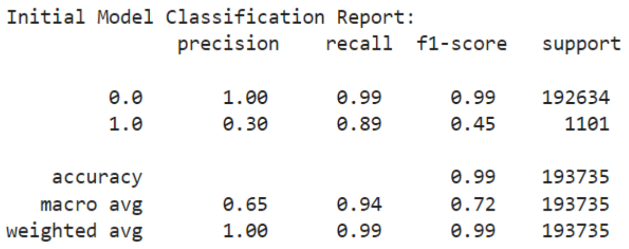


Figure 4: Temporal patterns in fraud transactions

4.4 Model Training:

- **Incorporating EDA Insights into Model Training:**
  + **Feature Engineering Based on Time of Day:** Create a new feature that flags high-risk hours which could potentially increase model sensitivity to these periods.
  + **Prioritize Spending Category in Feature Engineering:** Given the strong predictive power of the 'category', ensure it is prominently featured in the model through targeted feature engineering like interaction terms with other features.
  + **Feature Engineering Geographical Patterns:** Geographical feature is adequately encoded and utilized in the model.
- **Train-Test Split**: Divide the data into 80% training and 20% testing sets.
- **Handling Imbalance:** Utilize SMOTE to address class imbalance.

4.5 Model Evaluation: Initial evaluation of the model's performance is conducted using classification reports and confusion matrices.



```
Initial Model Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99    192634
         1.0       0.30      0.89      0.45      1101

    accuracy                           0.99    193735
   macro avg       0.65      0.94      0.72    193735
weighted avg       1.00      0.99      0.99    193735
```

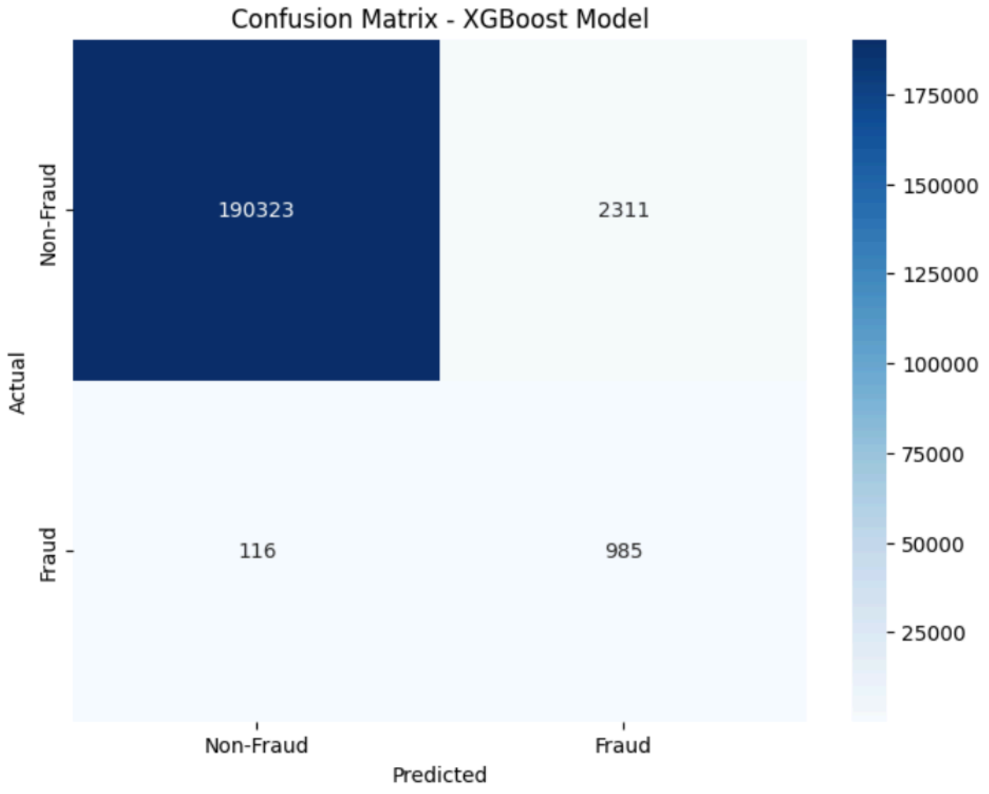


Figure 5: Initial Model Evaluation

The XGBoost model's performance demonstrates high accuracy in identifying non-fraudulent transactions, achieving near-perfect precision and recall for the non-fraud class. However, it shows a significant challenge in detecting fraudulent transactions, with a low precision of 0.30 for the fraud class, indicating many false positives. The model's high recall of 0.89 for fraudulent transactions suggests it captures most fraud cases, but the overall effectiveness, reflected in the F1-score of 0.45 for fraud, indicates room for improvement. This imbalance highlights the need for further tuning and possibly additional feature engineering to enhance the model's capability in accurately identifying fraudulent activities.

4.6 Model Optimization and Tuning: Hyperparameter tuning is performed using RandomizedSearchCV to enhance model performance.

4.7 Tuned Model Evaluation: The performance of the tuned model is re-evaluated to assess improvements.

5. Challenges and Recommendations

- Challenges:

- Class Imbalance: The dataset is highly imbalanced, with fraudulent transactions being significantly less frequent than legitimate ones. This can lead to a model that is biased towards predicting non-fraudulent transactions.

- Feature Selection: Identifying which features are most relevant for fraud detection is challenging and requires extensive EDA and domain knowledge.
- Real-Time Detection: Implementing a system that can detect fraud in real-time without significant delays is technically demanding.

- **Recommendation**:
  - Advanced Sampling Techniques: Utilize advanced oversampling and undersampling techniques beyond SMOTE to address class imbalance more effectively.
  - Feature Engineering: Continuously develop and test new features that can capture the subtleties of fraudulent behavior.
  - Incremental Learning: Implement models that can update themselves with new data, improving over time without the need for retraining from scratch.

References

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. Statistical Science, 17(3), 235–249. https://www.jstor.org/stable/3182781

Understanding the Threat of Card Transaction Fraud and its Impact on the Financial Ecosystem | Waylay Blog. (2023). https://www.waylay.io/articles/understanding-the-threat-of-card-transaction-fraud-and-its-impact-on-the-financial-ecosystem

Payment fraud detection and prevention: A how-to guide | Stripe. (2023). https://stripe.com/resources/more/payment-fraud-detection-and-prevention

Hasham, S., Joshi, S., & Mikkelsen, D. (2019, October 1). Financial cybercrime and fraud | McKinsey. Www.mckinsey.com. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/financial-crime-and-fraud-in-the-age-of-cybersecurity

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). DOI:10.1145/2939672.2939785.

Miller, T. (2022). Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature.

Kou, Y., Lu, C.T., Sirwongwattana, S., Huang, Y.P. (2004). Survey of fraud detection techniques. In IEEE International Conference on Networking, Sensing and Control (Vol. 2, pp. 749–754). DOI:10.1109/ICNSC.2004.1297040.

Kumar, M.S., Soundarya, V., Kavitha, S., Keerthika, E.S., Aswini, E. (2019). Credit card fraud detection using random forest algorithm. In 3rd International Conference on Computing and Communications Technologies (pp. 149–153). DOI:10.1109/ICCCT2.2019.8824930.

Liu, C., Chan, Y., S. Hasnain, A. Kazmi, and H. Fu (2015). Financial Fraud Detection Model: Based on Random Forest. International Journal of Economics and Finance, 7(7), 178. DOI:10.5539/ijef.v7n7p178.

Credit Card Transactions Fraud Detection Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/kartik2112/fraud-detection?datasetId=817870&sortBy=voteCount