Credit Card Fraud Detection

Dexter Schincke

Bellevue University

DSC: 680

Amirfarrokh Iranitalab

1 December 2024

Credit Card Fraud Detection

Topic:

This project focuses on developing a machine learning model to accurately identify fraudulent credit card transactions, enabling financial institutions to minimize financial losses and protect consumers.

Business Problem:

Credit card fraud is pervasive in the financial sector, resulting in billions of dollars in losses annually. As digital transactions grow, so does the sophistication of fraudulent schemes. Financial institutions need robust, efficient, and scalable tools to detect real-time fraudulent transactions. This project addresses this critical challenge by building a machine-learning model to classify transactions as fraudulent or legitimate based on transaction data.

- Research Questions:
  - What machine learning techniques are most effective for detecting fraudulent transactions in highly imbalanced datasets?
  - How can class imbalance in the dataset be effectively managed to improve fraud detection accuracy?
  - Which features in the transaction data contribute most significantly to identifying fraudulent behavior?
  - How does model performance vary across different machine learning algorithms, and which provides the best trade-off between precision and recall in fraud detection?

Dataset:

This project will use a Kaggle Credit Card Fraud Detection dataset specifically designed for fraud detection tasks.

- Features:
  - id: A unique identifier for each transaction.

    - V1-V28: These are anonymized features that represent different transaction attributes (like time, location, or other patterns).
    - Amount: The transaction amount in monetary units.
    - Class: This is the target variable, where 1 means the transaction is fraudulent and 0 is legitimate.
- Key Details:
    - Size: The dataset has about 570,000 transactions.
    - Date: It uses data from 2023
    - Fraud Cases: I don’t know the exact number of fraudulent cases yet, but I’ll figure that out during my exploratory data analysis.
    - Privacy: Since the features are anonymized, there’s no risk of exposing sensitive information, which makes it easier to focus on building the model.

This dataset is great because it mirrors the real-world challenge of fraud detection, especially with its imbalance between fraudulent and legitimate transactions. It’ll push me to explore techniques for handling these imbalances while building an effective model.

Methods:

The first step will involve preprocessing the data to ensure it's ready for modeling. One of the main challenges is class imbalance, where fraudulent transactions are much less frequent than legitimate ones. To address this, I will use techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or under-sampling to balance the classes. Additionally, I will normalize and scale numerical features, such as transaction amounts, to ensure that all features contribute equally to the model and improve overall performance.

After preprocessing, I will perform exploratory data analysis (EDA) to better understand the dataset. This will include visualizing the distribution of fraudulent and legitimate transactions to identify trends or outliers. I will also explore the anonymized features (V1-V28) to look for any significant patterns or correlations that could help differentiate between the two transaction classes. This step is crucial for gaining insights that may guide the selection of features for the models.

For the modeling phase, I will begin with Logistic Regression as a baseline model due to its simplicity and interpretability. I will explore more advanced machine learning algorithms, including Decision Trees, Random Forests, and Gradient Boosting Models such as XGBoost, which have performed well on imbalanced datasets. I will evaluate model performance using precision, recall, F1-score, and ROC-AUC, with a strong emphasis on minimizing false negatives to reduce the risk of overlooking fraudulent transactions.

Ethical Considerations:

Data privacy is crucial, even though the dataset is anonymized. In real-world applications, strict privacy measures are necessary to protect user information. Additionally, the model must be fair and unbiased, ensuring that no demographic group is disproportionately penalized. Transparency is also essential, and the model should be interpretable to allow financial stakeholders to understand how decisions are made. This is important for building trust and ensuring the model is used responsibly.

Challenges/Issues:

One major challenge is the class imbalance, with fraudulent transactions being much less frequent than legitimate ones, potentially leading to biased models. To address this, I'll use techniques like resampling or ensemble methods. Another issue is overfitting, where the model might memorize the training data, so I'll apply cross-validation and regularization to improve generalization. Scalability is also important, as the model must quickly handle large volumes of data for real-time fraud detection, ensuring efficiency and timely processing.

References

Projects/Papers:

Abueltouh, A. (2024). Credit Card Detection. Kaggle. https://www.kaggle.com/code/abdallahabuelftouh/credit-card-detection

Fatima, S. (2024). Credit card fraud detection: Achieving 99% accuracy. Kaggle. https://www.kaggle.com/code/samanfatima7/credit-card-fraud-detection-achieving-99-acc

GeeksforGeeks. (2024, September 6). ML credit card fraud detection. GeeksforGeeks. https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning-based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data, 9, 24. https://doi.org/10.1186/s40537-022-00573-8

Datasets:

Elgiriye Withana, N. (2023). Credit card fraud detection dataset 2023. Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023/data