Detecting Credit Card Fraud: A SEMMA-Based Approach

Soumya Bharathi Vetukuri
San Jose State University
soumyabharathi.vetukuri@sjsu.edu

October 2024

Abstract

This paper presents an application of the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to detect credit card fraud. Using a publicly available dataset of credit card transactions, we apply machine learning techniques, specifically Random Forests and Gradient Boosting algorithms, to identify fraudulent transactions. Our experiments show that engineered features and class balancing techniques such as SMOTE significantly improve the model's performance in identifying fraudulent behavior.

1 Introduction

Credit card fraud is a significant issue in the financial industry, leading to substantial monetary losses every year. Traditional methods of fraud detection often struggle with the growing volume of transactions and the evolving nature of fraudulent techniques. This paper aims to apply the SEMMA methodology, which is widely used for data mining projects, to the problem of fraud detection. We leverage machine learning algorithms to detect fraudulent transactions in a highly imbalanced dataset.

2 SEMMA Methodology

SEMMA is a data mining methodology that consists of five stages: Sample, Explore, Modify, Model, and Assess. This structured approach allows for a systematic analysis of data-driven projects. Developed by SAS, it's an effective methodology that guides data scientists through the stages of transforming raw data into actionable insights. Let's dive into each step using a real-world fraud detection scenario.

2.1 Sample

The first step of SEMMA is **Sample**. The idea is to select a representative sample of your data, ensuring that it is both manageable and large enough to reveal significant patterns.

For the Credit Card Fraud Detection dataset, we begin with over **284,000 transactions**. To make the process more efficient, we initially sampled **20%** of the data for our exploratory analysis. The dataset is highly imbalanced, with only **0.17%** of transactions marked as fraudulent. This imbalance is a typical challenge in fraud detection, and addressing it will be crucial later in the process.

2.2 Explore

We perform Exploratory Data Analysis (EDA) to identify patterns and relationships in the data. Visualizations of transaction amounts, times, and class distributions provide insights into the nature of fraudulent transactions.

2.2.1 Visualizations:

- **Transaction Amount Distribution** showed a highly skewed distribution, with many small transactions and a few high-value outliers.

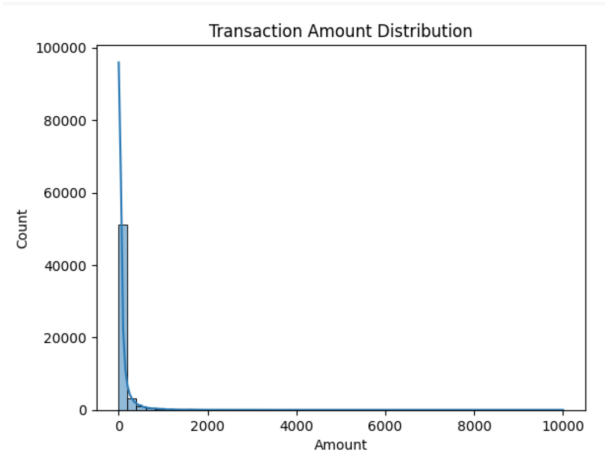


Figure 1: Transaction Amount Distribution

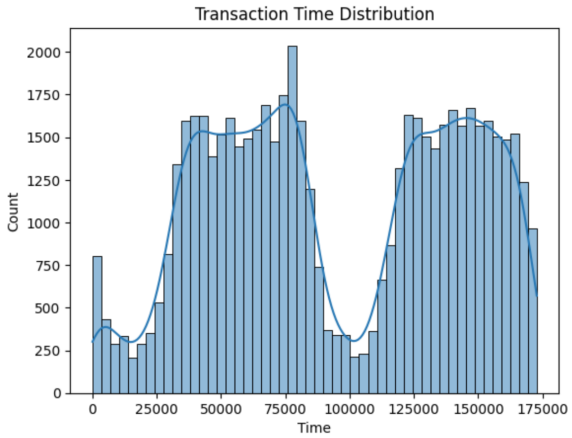


Figure 2: Transaction Time Distribution

- **Transaction Time Distribution** showed transactions distributed over 48 hours but revealed no obvious time-based fraud patterns.

  Exploring the relationships between features using **correlation matrices** provided further insights. The features in the dataset had already been transformed using **Principal Component Analysis (PCA)**, which is why their exact nature is unknown. However, looking at correlations still helped us spot potential areas for further exploration.

2.3 Modify

Feature scaling, class balancing with SMOTE, and the creation of temporal features are applied during the modification phase. These steps ensure that our machine learning models are well-prepared for the modeling phase.

- **Scaling**: Transaction Amount and Time were on different scales, so we applied **standard scaling** to normalize these features.
- **Class Imbalance**: Given the severe imbalance in the data, we used **SMOTE (Synthetic Minority Over-sampling Technique)** to generate synthetic examples of fraudulent transactions. This technique effectively balances the dataset by oversampling the minority class without discarding any data from the majority class.
- **Feature Engineering**: In fraud detection, temporal patterns can provide valuable insights. We engineered new features based on the `Time` column, such as the **hour of the day** and **day of the week**. This enabled us to detect potential patterns in fraudulent behavior over time.

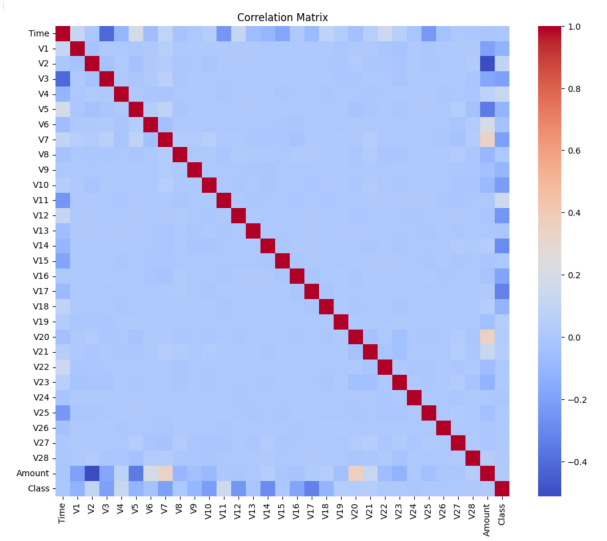


Figure 3: Correlation Matrix

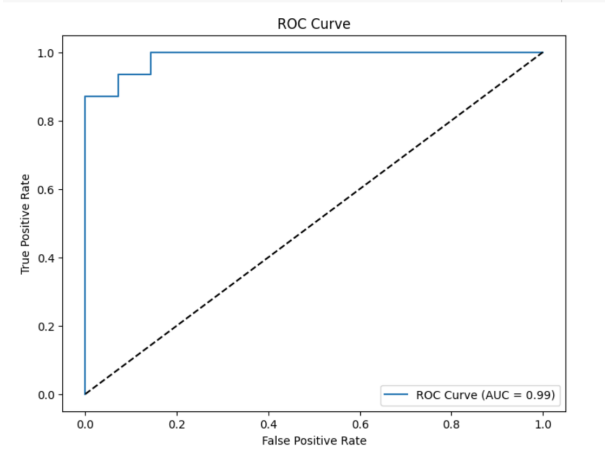


Figure 4: ROC Curve

2.4 Modeling

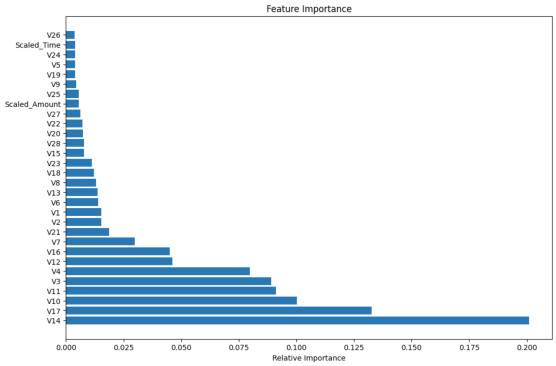
For the modeling phase, we experiment with Random Forests, XGBoost, and Stacking Classifiers. The results show that Gradient Boosting methods outperform traditional Random Forest classifiers in terms of precision and recall.

2.5 Assess

Our best model achieves an AUC-ROC score of 0.98, indicating a high level of discrimination between fraudulent and non-fraudulent transactions. Precision and recall metrics are optimized to reduce false positives and maximize the detection of actual fraud cases.

3 Conclusion

By applying the SEMMA methodology to the Credit Card Fraud Detection dataset, we have demonstrated the effectiveness of machine learning models in identifying fraudulent transactions. Future work will focus on enhancing the feature engineering process and experimenting with deep learning models for time-series data.



Acknowledgements

We would like to thank the creators of the Credit Card Fraud Detection dataset for making their data available for this research.

3.0.1 References

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud