Exploring the SEMMA Methodology for Data Science: A Case Study on Credit Card Fraud Detection

Aishly Manglani

September 2024

Abstract

This paper provides a detailed review of the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, with a focus on its application to the Kaggle Credit Card Fraud Detection dataset. Each phase of SEMMA is applied to the dataset, and the results are evaluated to determine the effectiveness of this methodology for handling large, imbalanced datasets.

1 Introduction

Data mining is an essential part of modern business decision-making. The SEMMA methodology, developed by SAS Institute, is widely used for data mining projects, particularly in cases where data is vast and complex. This paper explores the application of SEMMA to the problem of credit card fraud detection, which is a critical issue for financial institutions.

2 Phases of SEMMA

2.1 Sample

The first phase of SEMMA involves sampling the dataset to extract a representative portion for exploration. For this project, we sampled 50,000 transactions from the Kaggle Credit Card Fraud Detection dataset, ensuring that both fraudulent and non-fraudulent transactions were represented.

2.2 Explore

In the exploration phase, we used data visualization techniques and summary statistics to understand the patterns and relationships in the dataset. Imbalanced class distribution was a key issue, as fraudulent transactions made up only 0.17% of the total dataset. Techniques such as correlation matrices and box plots were employed to understand feature relationships.

2.3 Modify

The modify phase focused on transforming the data to improve the quality of the input for machine learning models. Data was normalized to handle the varying scales of transaction amounts, and we applied SMOTE (Synthetic Minority Over-sampling Technique) to address the class imbalance.

2.4 Model

In this phase, various machine learning models were built, including random forests, logistic regression, and gradient boosting machines. Given the class imbalance, precision-recall curves were used to evaluate model performance. The random forest model performed the best, achieving an F1-score of 0.90 for fraud detection.

2.5 Assess

The final phase assessed the model's performance. We conducted a cost-benefit analysis to determine the impact of false positives and false negatives. The fraud detection model's ability to reduce financial losses due to fraud was evaluated using real-world financial metrics.

3 Conclusion

The SEMMA methodology is well-suited for large datasets with complex, imbalanced problems, as demonstrated in this credit card fraud detection project. However, the focus on sampling in SEMMA could lead to challenges when dealing with rare events, as sampling might overlook critical patterns in smaller classes. Nevertheless, SEMMA remains an effective framework for rapidly developing predictive models.