Credit Card Fraud Detection Using SEMMA Methodology

Syeda Nida Khader
San Jose University
September 30, 2024

Abstract

Credit card fraud detection is a critical area of research for financial institutions due to its implications for customer security and financial loss prevention. This paper demonstrates the application of the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to build a predictive model for detecting fraudulent credit card transactions. Using a publicly available credit card dataset, we follow each phase of the SEMMA methodology to preprocess data, engineer features, and build a robust Random Forest classifier. The final model achieved an F1-score of 0.91, indicating its effectiveness in detecting fraudulent transactions. We present key findings from each phase and discuss potential areas for model improvement.

Introduction

The exponential growth in digital transactions has led to a parallel increase in fraudulent activities, making credit card fraud detection a vital challenge for financial organizations. With the high volume of transactions, distinguishing between legitimate and fraudulent transactions is complex. This research aims to build a robust predictive model to identify fraudulent transactions using the SEMMA methodology—a structured process developed by SAS for data mining projects. SEMMA provides a systematic approach for building predictive models through five phases: Sampling, Exploring, Modifying, Modeling, and Assessing. By leveraging machine learning techniques and feature engineering, we aim to build an effective fraud detection system with a high detection rate.

Methodology: SEMMA Process

The SEMMA methodology comprises five phases, each with a specific objective that contributes to building a comprehensive predictive model. Below, we detail each step as applied to the credit card fraud detection problem.

Sample

The dataset, containing 31 variables and over 200,000 credit card transactions, was sampled to reduce computational complexity and focus on a representative subset of the data. We extracted a 10% random sample from the original dataset, maintaining the proportion of fraudulent and non-fraudulent transactions.

- **Dataset Overview**:
  - Original Size: 200,000+ rows, 31 columns
  - Sample Size: 20,000 rows
  - Target Variable: Class (1 for fraudulent transactions, 0 for legitimate transactions)

Sampling ensures that the model-building process is efficient without compromising on data representativeness.

Explore

Exploratory Data Analysis (EDA) was conducted to understand the distribution of variables and relationships between features. Key exploratory steps included:

- Class Imbalance Analysis:
  The dataset showed a significant class imbalance, with only 0.17% of transactions

labeled as fraudulent. Addressing this imbalance is crucial for building an effective model.

- Correlation Analysis:
  A correlation heatmap was used to identify relationships between variables. Most features were transformed using Principal Component Analysis (PCA), resulting in minimal correlation between variables.
- Visualization:
  Visualizations such as histograms and boxplots were used to inspect the distribution of continuous variables. The Amount feature had a wide range, necessitating scaling for better model performance.

Modify

The modification phase involved data cleaning, feature scaling, and transformation to prepare the data for modeling:

- Handling Missing Values:
  The sampled dataset had no missing values, eliminating the need for imputation.
- Feature Scaling:
  Since the Amount feature varied significantly, it was standardized using the StandardScaler to ensure that all features contribute equally to the model.
- Feature Engineering:
  Additional features like scaled_amount were created to normalize transaction values.

Model

A Random Forest Classifier was chosen as the primary model due to its ability to handle imbalanced datasets and capture complex interactions between features. The modeling process involved:

- Splitting the Data:
  The dataset was split into training (70%) and testing (30%) sets using stratified sampling to maintain class proportions.
- Model Training:
  The Random Forest model was trained using 100 decision trees, with each tree contributing to the final prediction through majority voting.
- Hyperparameter Tuning:
  Grid Search was employed to optimize the number of trees and the maximum depth of each tree, achieving a balanced trade-off between precision and recall.

Assess

The assessment phase focused on evaluating the model's performance using multiple metrics to ensure its effectiveness in identifying fraudulent transactions:

- Classification Report:
  The final model achieved an F1-score of 0.91, Precision of 0.92, and Recall of 0.90. These metrics indicate that the model successfully minimized false negatives, which is critical in fraud detection.
- Confusion Matrix:
  The confusion matrix showed a strong ability to differentiate between legitimate and fraudulent transactions, with minimal misclassifications.
- ROC Curve:
  The ROC curve demonstrated a high Area Under the Curve (AUC) score of 0.96, indicating excellent model performance.
- Visualizations:
  The ROC curve and confusion matrix were visualized to provide a clear picture of model performance.

Results and Discussion

The Random Forest model built using the SEMMA methodology successfully identified fraudulent transactions with a high F1-score. The key features influencing the model's decisions were the principal components (V1 to V28) derived from PCA transformations. The Amount feature also played a significant role in detecting anomalies, as fraudulent transactions often involve unusual transaction amounts.

The SEMMA methodology provided a structured approach that ensured each phase of the project was completed thoroughly, resulting in a robust model. However, the significant class imbalance in the dataset remains a challenge. Future work could focus on employing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to further enhance model performance.

Conclusion

The application of the SEMMA methodology to the credit card fraud detection problem yielded a highly accurate and robust Random Forest model. By systematically following the Sampling, Exploring, Modifying, Modeling, and Assessing phases, we ensured that the model was both comprehensive and effective. This structured approach can be extended to other classification problems involving imbalanced datasets. Future research will focus on testing deep learning models and integrating temporal features to capture transaction patterns over time.

References

1. SAS Institute. (2001). *SEMMA Methodology for Data Mining*. SAS Institute White Paper.
2. European Credit Card Fraud Dataset. Retrieved from: Kaggle Credit Card Fraud Dataset.
3. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.