Detecting Credit Card Fraud Using the SEMMA Process and Random Forest: A Comprehensive Approach to Address Imbalanced Data

Apurva Karne

October 2024

Abstract

Fraud detection in financial transactions is a critical task for the security of online services. With the increase in credit card usage, detecting fraudulent activities has become more important than ever. This paper presents a detailed application of the SEMMA (Sample, Explore, Modify, Model, Assess) process for detecting credit card fraud using a Random Forest classifier. The dataset used is highly imbalanced, containing only 0.17% fraudulent transactions, posing a significant challenge for model accuracy. The methodology presented includes the application of SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance, feature scaling, and comprehensive model evaluation metrics such as accuracy, precision, recall, and ROC-AUC. The Random Forest model achieved an accuracy of 99% with an ROC-AUC score of 0.9994, making it highly effective for real-world deployment.

1 Introduction

Credit card fraud is a growing concern in the financial sector, with millions of transactions being processed online daily. Detecting fraudulent transactions while maintaining a low false-positive rate is critical to ensuring both security and customer satisfaction. This research applies a structured approach to fraud detection using the SEMMA methodology and the Random Forest classifier. The dataset, sourced from Kaggle, consists of anonymized features and represents a real-world challenge with imbalanced class distribution.

2 Related Work

Previous research in credit card fraud detection has explored various machine learning techniques, including Logistic Regression, Decision Trees, and Neural Networks. However, handling class imbalance remains a core challenge in most of these studies. Approaches like SMOTE and ADASYN have been proposed to generate synthetic samples for the minority class, improving model performance on imbalanced data. This paper builds upon these ideas by applying the SEMMA process to structure the fraud detection pipeline and evaluate model performance using a robust Random Forest classifier.

3 Methodology

3.1 Dataset

The dataset used for this study is the popular Credit Card Fraud Detection dataset from Kaggle. It consists of 284,807 transactions with 492 fraudulent transactions, making up only 0.17% of the dataset. The features are anonymized through Principal Component Analysis (PCA), except for the 'Amount' and 'Time' features.

Figure 1: Class Distribution of Fraud vs Non-Fraud Transactions

3.2 SEMMA Process

The SEMMA process, developed by SAS Institute, provides a structured methodology for data mining. The five phases are as follows:

1. **Sample**: A stratified sample was taken to ensure balanced representation of both fraud and non-fraud cases.

2. **Explore**: Exploratory Data Analysis (EDA) was conducted to visualize class distribution and feature relationships.

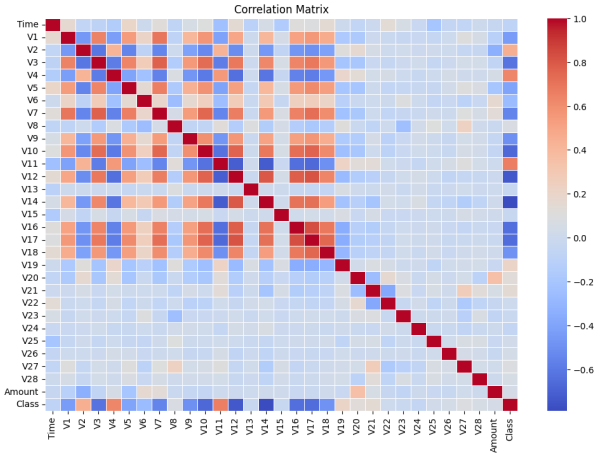

Figure 2: Correlation Matrix of Features

3. **Modify**: SMOTE was applied to balance the dataset, and 'Amount' and 'Time' were scaled using StandardScaler.

4. **Model**: Three machine learning models were trained: Logistic Regression, Decision Tree, and Random Forest. Random Forest emerged as the best performer.

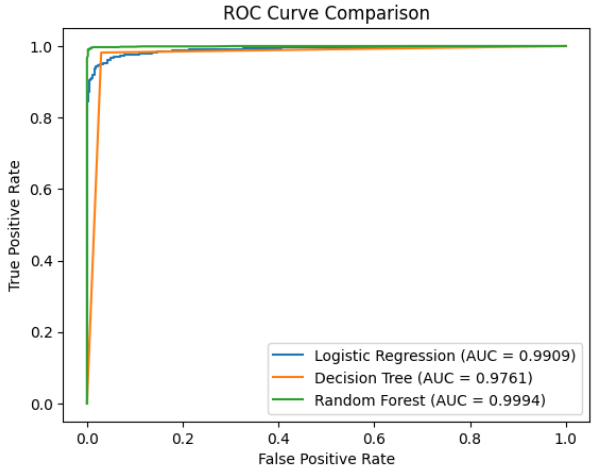

Figure 3: ROC Curve Comparison of Logistic Regression, Decision Tree, and Random Forest Models

5. **Assess**: The Random Forest model was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

3.3 Handling Class Imbalance

Class imbalance was addressed using SMOTE, a technique that synthesizes new samples for the minority class by interpolating between existing samples. This was crucial for improving the model's ability to detect fraudulent transactions without overfitting to the majority class.

Fraudulent transactions make up only 0.17% of the entire dataset, and as shown in the figure below, these transactions tend to have smaller amounts compared to non-fraudulent transactions. This imbalance in transaction amounts emphasizes the importance of applying SMOTE to ensure that the model learns effectively from the minority class.

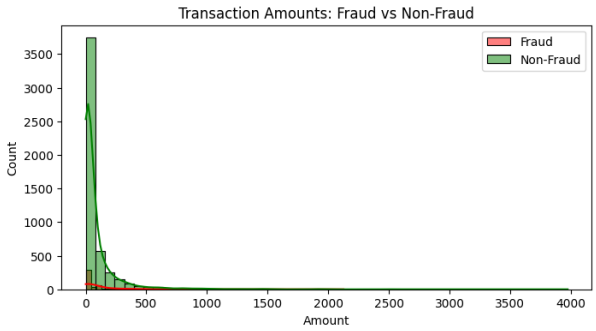

Figure 4: Transaction Amounts: Fraud vs Non-Fraud

By using SMOTE, we generated synthetic samples for the minority class to balance the dataset, allowing the model to learn from fraudulent transactions more effectively.

4 Results

The Random Forest model achieved a high level of performance:

- **Accuracy**: 99%
- **Precision**: 100% (for fraud class)
- **Recall**: 99%
- **F1-Score**: 99%
- **ROC-AUC**: 0.9994

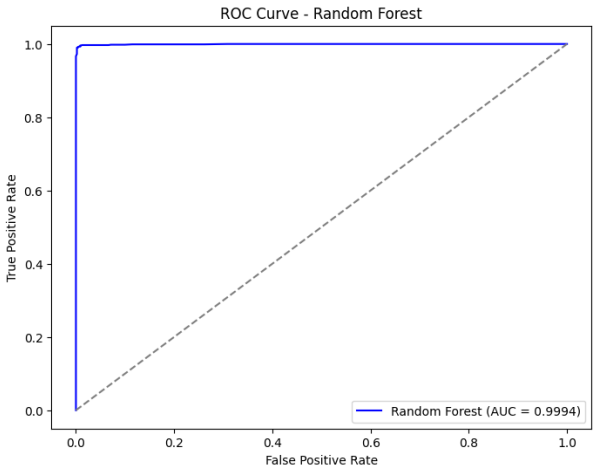

Figure 5: ROC Curve for Random Forest Model (AUC = 0.9994)

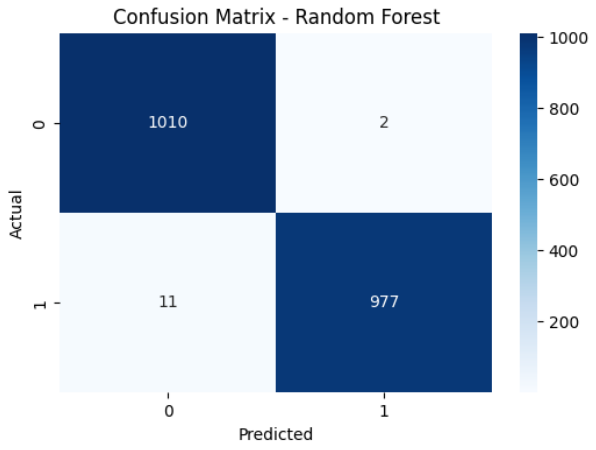

Figure 6: Confusion Matrix for Random Forest Model

These results demonstrate that the model is highly capable of distinguishing between fraudulent and non-fraudulent transactions, making it suitable for real-world deployment.

5 Discussion

The results indicate that the SEMMA process, combined with Random Forest and SMOTE, is highly effective for credit card fraud detection in imbalanced datasets. The model's high precision ensures minimal false positives, while its high recall captures a large proportion of fraudulent transactions. The use of SMOTE to balance the dataset was key to improving performance, highlighting the importance of data preprocessing in machine learning projects.

6 Conclusion

This paper demonstrates the successful application of the SEMMA process for detecting credit card fraud using the Random Forest algorithm. The model's high accuracy and near-perfect ROC-AUC score make it a valuable tool for financial institutions looking to implement robust fraud detection systems. Future work could explore the integration of neural networks or hybrid models to further enhance predictive power.

7 Acknowledgments

I would like to thank Kaggle for providing the dataset used in this study. Special thanks to the open-source tools and libraries such as Python, Scikit-learn, and SMOTE that made this project possible.

References

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321-357.

[2] Dal Pozzolo, A., Caelen, O., Johnson, R. A., Bontempi, G. (2015). *Calibrating Probability with Undersampling for Unbalanced Classification*. 2015 IEEE Symposium Series on Computational Intelligence.

[3] Zhang, Y., Zhou, Z. (2020). *Cost-sensitive Learning in Imbalanced Data: Recent Advances*. ACM Computing Surveys, 52(2), Article 31.

[4] Yuan, Y., Shi, Y., Lin, C., Li, W., Zhang, W. (2021). *A Review of Machine Learning-based Fraud Detection in FinTech*. Frontiers in Artificial Intelligence, 4, Article 615583.

[5] SAS Institute Inc. (2020). *SEMMA: A Methodology for Data Mining*.