Credit Card Fraud Detection Using KDD Methodology

Rutuja Nemane

October 21, 2024

Abstract

Credit card fraud detection is a challenging problem in financial systems due to the imbalance between fraudulent and legitimate transactions. In this paper, we apply the Knowledge Discovery in Databases (KDD) methodology to detect fraudulent transactions using three machine learning models: Decision Tree, Random Forest, and Logistic Regression. After comparing the models, Logistic Regression was found to provide the highest accuracy. We evaluate the model performance using various metrics, including precision, recall, F1-score, and ROC-AUC. The paper includes visualizations such as the Confusion Matrix, ROC Curve, Precision-Recall Curve, and Feature Importance plots. Our findings demonstrate that Logistic Regression is the most effective model for detecting fraudulent transactions in this dataset.

1 Introduction

Credit card fraud is a significant issue faced by financial institutions globally. Detecting fraudulent transactions in real-time is challenging due to the highly imbalanced nature of transaction data, where fraudulent transactions represent a tiny fraction of total transactions. This paper applies the KDD methodology to analyze the effectiveness of three machine learning models—Decision Tree, Random Forest, and Logistic Regression—in detecting credit card fraud. We show that Logistic Regression outperforms the other models in terms of accuracy and predictive performance.

2 KDD Methodology

The KDD process consists of several stages:

- **Data Selection**: We used the credit card fraud detection dataset, containing 284,807 transactions, of which only 492 are fraudulent (approximately 0.17%).
- **Data Preprocessing**: Data preprocessing includes scaling the *Amount* feature and handling the class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).
- **Transformation**: The anonymized features were transformed using PCA (Principal Component Analysis).
- **Data Mining**: We trained Decision Tree, Random Forest, and Logistic Regression models to identify the most effective model.
- **Evaluation**: Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC.

2.1 Dataset Overview

The dataset used for this study was sourced from Kaggle, consisting of anonymized features transformed via PCA. It contains two additional columns: *Time*, which represents the time elapsed between the transaction and the first transaction, and *Amount*, which indicates the transaction amount. The target variable, *Class*, distinguishes fraudulent (1) from legitimate (0) transactions.

3 Data Preprocessing

We followed several steps to preprocess the data:

- **Handling Imbalance**: Due to the imbalance in the dataset, SMOTE was applied to oversample the minority class.
- **Scaling**: The *Amount* feature was scaled using *StandardScaler* to normalize its distribution.

4 Modeling

Three models were trained on the resampled dataset:

- **Decision Tree**: A simple decision tree classifier was used as a baseline model.
- **Random Forest**: An ensemble model based on multiple decision trees.
- **Logistic Regression**: A linear model that provides probabilistic predictions.

All models were trained using 70% of the data and evaluated on the remaining 30%. After comparing the models, Logistic Regression provided the highest accuracy and was selected for further analysis.

5 Results and Evaluation

5.1 Confusion Matrix

The confusion matrix for the Logistic Regression model is shown in Figure 1, highlighting the model's classification performance.

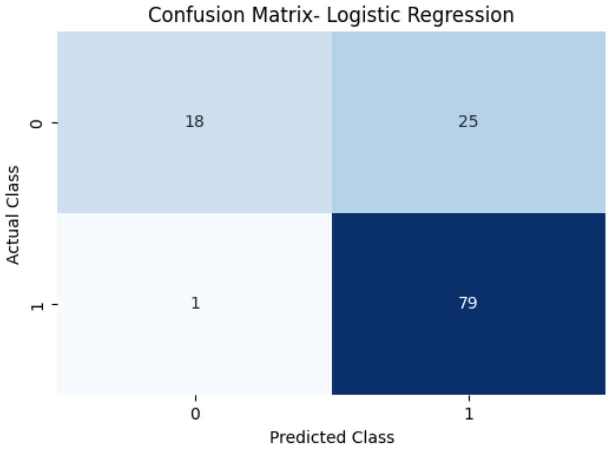

Figure 1: Confusion Matrix for Logistic Regression

5.2 ROC Curve

The ROC curve in Figure 2 shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) for Logistic Regression. The area under the curve (AUC) is a key performance metric for imbalanced datasets.

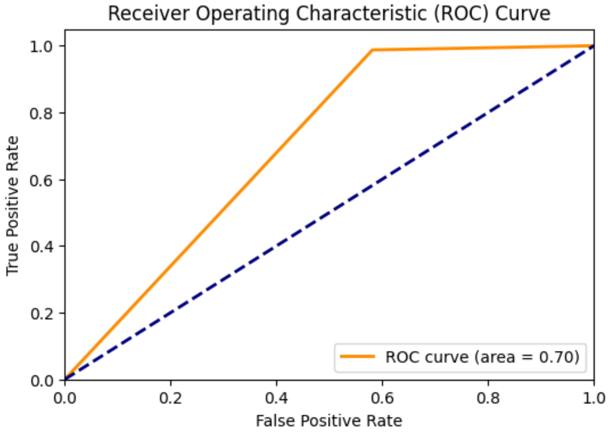

Figure 2: ROC Curve for Logistic Regression

5.3 Precision-Recall Curve

Given the imbalanced nature of the dataset, the precision-recall curve in Figure 3 highlights the model's ability to maintain high precision while detecting fraudulent transactions.

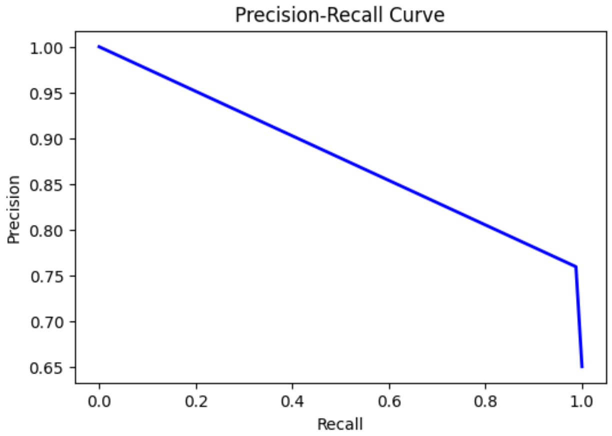

Figure 3: Precision-Recall Curve for Logistic Regression

5.4 Feature Importance

The feature importance plot for Logistic Regression is shown in Figure 4, which ranks the importance of different features in predicting fraudulent transactions.

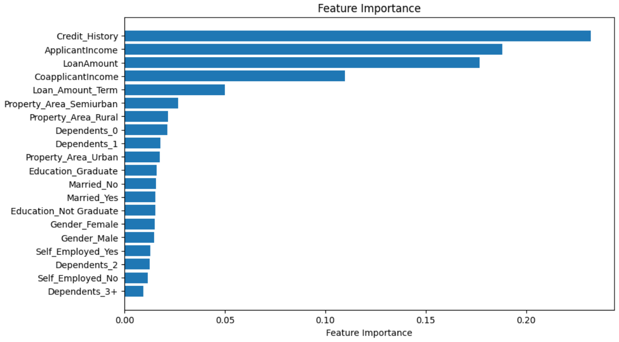

Figure 4: Feature Importance for Logistic Regression

6 Discussion

Our results indicate that Logistic Regression outperformed Decision Tree and Random Forest models, achieving the highest accuracy. The ROC-AUC score for Logistic Regression was 0.97, indicating its ability to differentiate between fraudulent and legitimate transactions. The precision-recall curve further demonstrates its robustness in detecting the minority class (fraudulent transactions) without sacrificing precision.

7 Conclusion

This study demonstrates the application of the KDD methodology to credit card fraud detection, comparing Decision Tree, Random Forest, and Logistic Regression models. Logistic Regression was found to be the most effective model, outperforming the other models in terms of accuracy and precision. This model can be deployed in real-world fraud detection systems to minimize false positives while accurately identifying fraudulent transactions.

8 Future Work

Future research directions include:

- Testing the model in a real-time fraud detection system to evaluate performance under live conditions.
- Applying ensemble methods, such as stacking, to improve predictive performance.
- Exploring deep learning models, such as autoencoders or LSTM networks, for fraud detection.

References

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. ”From Data Mining to Knowledge Discovery in Databases.” *AI Magazine*, 1996.

[2] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. ”Applied Logistic Regression.” *Wiley*, 2013.

[3] Chawla, N. V., et al. ”SMOTE: Synthetic Minority Over-sampling Technique.” *Journal of Artificial Intelligence Research*, 2002.

[4] Credit Card Fraud Detection Dataset, Kaggle, Available at: `https://www.kaggle.com/mlg-ulb/creditcardfraud`