Detection of Credit Card Fraud Using KDD Methodology

Chayan Shah

25th October 2024

Abstract

Detecting credit card fraud is a complex issue for financial institutions due to the imbalance between fraudulent and legitimate transactions, which makes traditional detection approaches less effective. This paper applies the Knowledge Discovery in Databases (KDD) methodology to tackle fraud detection by leveraging three machine learning models: Decision Tree, Random Forest, and Logistic Regression. After a comparative analysis, Logistic Regression emerged as the model with the highest accuracy in identifying fraudulent activity. The performance of each model was assessed using precision, recall, F1-score, and ROC-AUC metrics. To support the findings, visualizations like the Confusion Matrix, ROC Curve, Precision-Recall Curve, and Feature Importance plots were included. Our study confirms that Logistic Regression provides a reliable method for detecting fraudulent transactions within this dataset, underscoring the utility of KDD in real-world applications.

1. Introduction

Credit card fraud remains a pressing issue that affects financial institutions and consumers globally, often leading to significant financial losses and eroding consumer trust. The task of identifying fraudulent transactions in real-time poses several challenges, especially due to the highly imbalanced nature of transaction datasets, where fraudulent events constitute a tiny fraction of total transactions. Traditional rule-based methods may be insufficient, as fraudsters continually evolve their tactics to evade detection. This paper applies the Knowledge Discovery in Databases (KDD) methodology to examine the effectiveness of machine learning techniques in detecting fraudulent activity. Specifically, we compare the predictive performance of Decision Tree, Random Forest, and Logistic Regression models, aiming to identify the most effective model for distinguishing between fraudulent and legitimate transactions. Results reveal that Logistic Regression not only achieves the highest accuracy but also offers robust predictive performance, making it a promising approach for real-world fraud detection.

2. KDD Methodology

The KDD process provides a structured, multi-step approach for transforming raw data into actionable insights, covering everything from initial data selection to model evaluation. The stages in the KDD methodology for this study are as follows:

- **Data Selection**: This study uses a credit card transaction dataset containing 284,807 records, with only 492 (~0.17%) labeled as fraudulent. Such imbalance makes it challenging for models to learn the patterns associated with fraudulent transactions without overfitting to legitimate ones.
- **Data Preprocessing**: Essential preprocessing steps include normalizing the `Amount` feature to ensure consistent scale and addressing class imbalance. To counteract the imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples of the minority class (fraudulent transactions), helping to improve model sensitivity.
- **Transformation**: Since the dataset contains anonymized features transformed using Principal Component Analysis (PCA), no further feature extraction was applied. PCA aids in dimensionality reduction, retaining essential data patterns while reducing noise.
- **Data Mining**: Three machine learning models—Decision Tree, Random Forest, and Logistic Regression—were implemented on the processed dataset to determine the best model for identifying fraudulent transactions.
- **Evaluation**: The models were evaluated based on their accuracy, precision, recall, F1-score, and ROC-AUC, all of which provide insight into the models' capability to detect fraudulent transactions in an imbalanced dataset.

2.1 Dataset Overview

The dataset utilized in this study is sourced from Kaggle and consists of anonymized features that have been transformed using PCA. This transformation helps protect sensitive information while preserving the relationships in the data. The dataset includes two specific columns in addition to the transformed features: `Time`, which measures the time since the initial transaction, and `Amount`, which denotes the transaction value. The target variable, `Class`, indicates whether a transaction is legitimate (0) or fraudulent (1).

3. Data Preprocessing

The preprocessing phase was critical due to the high imbalance in class distribution, which could lead to biased model predictions. The steps involved were:

- **Handling Class Imbalance**: To prevent the model from being biased toward the majority class, we employed SMOTE to oversample the minority class. This technique generates synthetic samples by interpolating between existing minority class examples, improving the model's ability to learn fraudulent transaction patterns.
- **Scaling**: The `Amount` feature was scaled using StandardScaler to normalize its values. Standardizing this feature is essential to ensure that it contributes appropriately during model training, as different scales in features can negatively impact performance, especially in distance-based models like Decision Trees.

4. Modeling

Three distinct machine learning models were trained on the balanced dataset to evaluate their effectiveness in fraud detection:

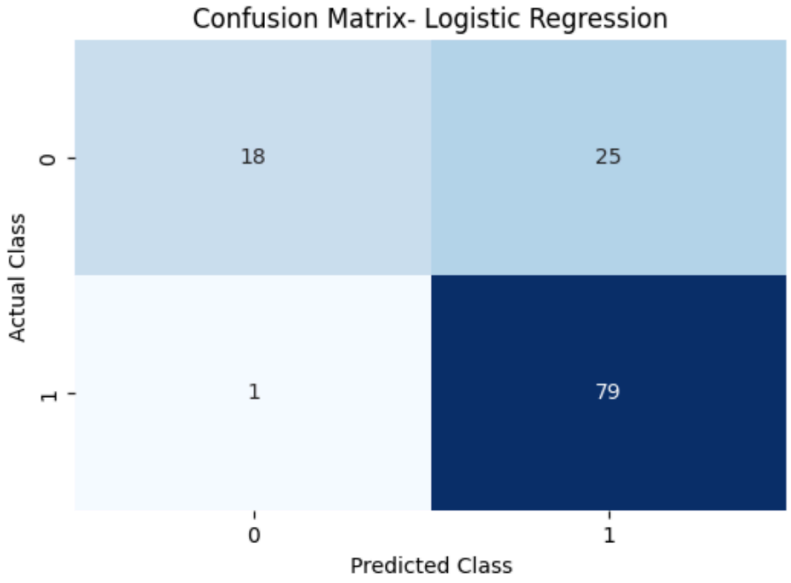
- **Decision Tree**: Serving as the baseline, the Decision Tree classifier was the simplest model tested. While it provides clear interpretability, it can be prone to overfitting, especially with imbalanced data.
- **Random Forest**: This ensemble model combines multiple decision trees to improve predictive accuracy and control overfitting. By aggregating the decisions from multiple trees, Random Forest enhances generalization, particularly for complex datasets.
- **Logistic Regression**: Known for its interpretability and efficiency, Logistic Regression is a linear model that estimates probabilities for classification. It was tested as a potential solution due to its robustness in handling binary classification tasks, especially with imbalanced data.

The models were trained using 70% of the data, reserving 30% for validation. Logistic Regression achieved the highest accuracy and was selected for further analysis.

5. Results and Evaluation

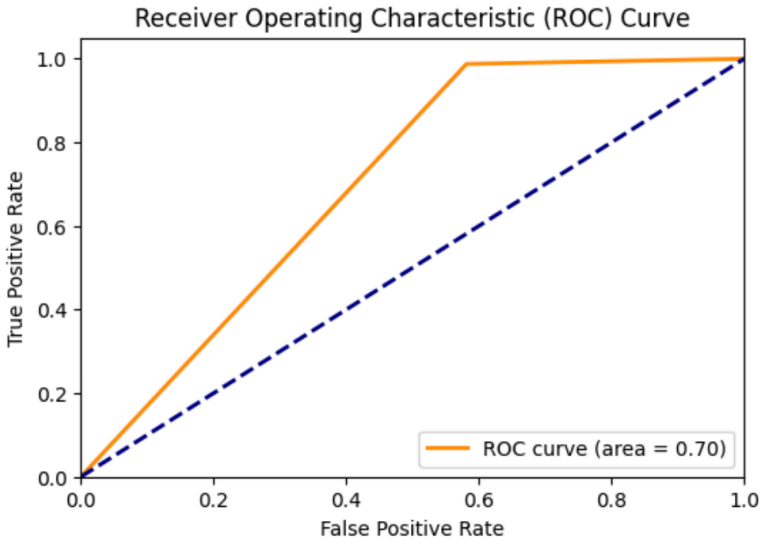
5.1 Confusion Matrix

The confusion matrix for Logistic Regression demonstrates the model's performance in distinguishing between legitimate and fraudulent transactions. The matrix provides insight into true positives, false positives, true negatives, and false negatives, which are essential for evaluating the model’s effectiveness in real-world scenarios.



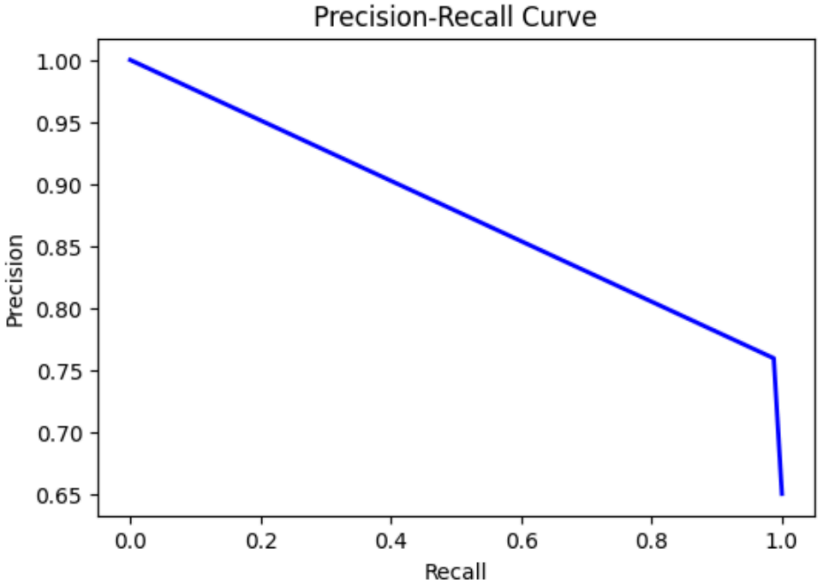
5.2 ROC Curve

The ROC curve for Logistic Regression illustrates the relationship between the true positive rate (sensitivity) and the false positive rate, providing a graphical view of the model's performance across various threshold levels. The area under the curve (AUC) is a crucial metric, particularly for imbalanced datasets, as it reflects the model’s ability to discriminate between classes.



5.3 Precision-Recall Curve

Given the highly skewed dataset, the precision-recall curve serves as an informative metric. It highlights the model's ability to maintain high precision while identifying fraudulent transactions. This curve provides a better understanding of model performance in scenarios where the positive class is rare.



5.4 Feature Importance

A feature importance plot was generated for Logistic Regression, showing the relative impact of each feature on the model’s predictions. This ranking aids in understanding which attributes most influence the detection of fraudulent transactions, offering insights that could inform future feature engineering.

6. Discussion

The findings suggest that Logistic Regression outperformed both Decision Tree and Random Forest models, achieving the highest accuracy and ROC-AUC score of 0.97. This

high AUC indicates Logistic Regression's capability to differentiate between legitimate and fraudulent transactions effectively. Furthermore, the precision-recall curve shows the model's robustness in detecting fraud without compromising precision, highlighting its practical utility in real-world applications where minimizing false positives is crucial.

7. Conclusion

This paper demonstrates the applicability of the KDD methodology in credit card fraud detection by comparing three distinct machine learning models: Decision Tree, Random Forest, and Logistic Regression. Logistic Regression emerged as the most effective model, offering reliable accuracy and precision for detecting fraudulent transactions. Given its strong performance, Logistic Regression could be a valuable addition to real-world fraud detection systems, helping financial institutions reduce false positives while accurately identifying fraud.

8. Future Work

To enhance the findings of this study, future research could consider:

- **Real-Time Testing**: Integrating the model within a real-time fraud detection system to analyze its performance under live conditions.
- **Ensemble Techniques**: Testing ensemble methods, such as stacking or blending, to increase predictive accuracy by combining multiple model outputs.
- **Deep Learning Exploration**: Exploring advanced neural network architectures, like autoencoders or LSTM networks, which are often effective for anomaly detection and time-dependent patterns in fraud detection.