Knowledge Discovery in Databases (KDD) for Credit Card Fraud Detection

Ankit Ojha

October 22, 2024

Abstract

This research paper presents a comprehensive approach to detecting credit card fraud using the Knowledge Discovery in Databases (KDD) process. The study utilizes the Credit Card Fraud Detection dataset from Kaggle and implements various machine learning techniques to predict fraudulent transactions. Each step of the KDD process—Data Selection, Data Preprocessing, Data Transformation, Data Mining, and Interpretation/Evaluation—is described in detail, and the results are analyzed to assess model performance and effectiveness.

1 Introduction

Credit card fraud is a significant issue impacting financial institutions and consumers globally. As digital transactions become more prevalent, the need for robust fraud detection systems has increased. This paper applies the KDD methodology to develop a predictive model capable of identifying fraudulent transactions efficiently. The primary goal is to minimize false negatives, ensuring that fraudulent activities are detected while maintaining an acceptable rate of false positives.

2 Methodology

The KDD process consists of five key steps: Data Selection, Data Preprocessing, Data Transformation, Data Mining, and Interpretation/Evaluation. Each step is crucial for extracting meaningful patterns and insights from the data.

2.1 Data Selection

The dataset used in this study is the Credit Card Fraud Detection dataset from Kaggle, which contains 284,807 transactions, of which 492 are fraud cases. The dataset is highly imbalanced, making it a suitable candidate for exploring anomaly detection and classification challenges.

2.2 Data Preprocessing

Data quality is essential for building effective models. The following preprocessing steps were implemented:

- **Handling Missing Values**: The dataset was checked for missing values, and none were found.
- **Removing Duplicates**: Duplicate records were identified and removed to ensure data integrity.
- **Data Type Verification**: All features were verified for correct data types.

2.3 Data Transformation

Feature scaling was performed to standardize the numerical features using the StandardScaler from scikit-learn. This step ensures that all features are on the same scale, which is crucial for distance-based algorithms and models sensitive to feature magnitudes.

2.4 Data Mining

Two machine learning models, Logistic Regression and Random Forest, were selected for initial experimentation. The dataset was split into training (70%) and testing (30%) sets to evaluate model performance.

- **Logistic Regression**: A simple yet effective model for binary classification problems.
- **Random Forest**: An ensemble method known for its robustness and ability to handle class imbalances.

The models were trained on the scaled data, and hyperparameter tuning was performed to optimize performance.

3 Results

The performance of both models was evaluated using accuracy, precision, recall, and F1-score. Given the imbalanced nature of the dataset, recall and F1-score were prioritized as key metrics to assess the models' effectiveness in detecting fraudulent transactions.

3.1 Logistic Regression

The Logistic Regression model achieved the following performance metrics:

- **Accuracy**: 99.92%
- **Precision**: 88.10%
- **Recall**: 55.22%
- **F1 Score**: 67.89%

While the model achieved high accuracy, the recall was relatively low, indicating that it missed a significant number of fraudulent transactions. This trade-off highlights the model's tendency to prioritize precision over recall.

3.2 Random Forest

The Random Forest model performed better in terms of both precision and recall:

- **Accuracy**: 99.95%
- **Precision**: 95.10%
- **Recall**: 72.39%
- **F1 Score**: 82.20%

The Random Forest model provided a more balanced performance, with a significant improvement in recall compared to Logistic Regression. This makes it a more suitable choice for fraud detection, where catching as many fraudulent cases as possible is crucial.

3.3 Comparison

Overall, the Random Forest model outperformed Logistic Regression, especially in terms of recall and F1-score, making it more reliable for identifying fraudulent transactions. The trade-off between precision and recall remains a critical consideration in the deployment of fraud detection systems.

4 Discussion

The high recall scores achieved by both models suggest that the models are effective at identifying fraudulent transactions. However, the relatively lower precision indicates a higher rate of false positives, which could lead to inconvenience for legitimate users. The imbalanced nature of the dataset posed a challenge, which was partially addressed using resampling techniques like SMOTE. Future work could explore additional methods, such as cost-sensitive learning and ensemble techniques, to further improve model performance.

5 Conclusion

This research demonstrates the effectiveness of the KDD process in developing a fraud detection model. By following a structured approach, meaningful insights and predictive models were developed. The study highlights the importance of handling class imbalances and the need for continuous model evaluation and improvement. Future research could incorporate more advanced algorithms and real-time data processing techniques.

6 Acknowledgment

We would like to express our gratitude to Kaggle for providing the Credit Card Fraud Detection dataset, which was instrumental in conducting this research. The availability of high-quality, open-source datasets on platforms like Kaggle greatly supports data-driven research and advancements in the field of data science and machine learning.

7 References

References

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

[2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[3] 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.