Enhancing Credit Card Fraud Detection Using Knowledge Discovery in Databases: A Comprehensive Study

Ronak Urvish Malkan
Student at San Jose State University

October 24, 2024

Abstract

This research paper explores the application of the Knowledge Discovery in Databases (KDD) methodology to credit card fraud detection. The study provides a comprehensive analysis encompassing data selection, preprocessing, transformation, mining, evaluation, and presentation applied to a Kaggle dataset. We aim to enhance the model's predictive accuracy and efficiency through sophisticated feature engineering, model optimization, and handling of imbalanced datasets. Emphasis is placed on the practical implications and potential improvements in fraud detection systems.

1 Introduction

Credit card fraud is a significant challenge faced by the financial industry, costing billions of dollars each year. As fraudsters develop more sophisticated techniques, the need for efficient fraud detection models has never been greater. This paper discusses the implementation of Knowledge Discovery in Databases (KDD) methodology for building a predictive fraud detection model. We aim to highlight how the KDD process helps in tackling the complex nature of credit card fraud detection and improving prediction accuracy.

The dataset used in this study is publicly available on Kaggle, containing credit card transactions over a period, with a class label indicating whether a transaction is fraudulent or not. By applying the KDD methodology, we systematically approach the data mining task, from initial exploration to model evaluation.

2 KDD Methodology

The KDD process provides a structured approach to extracting useful patterns from large datasets. It includes six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

2.1 Business Understanding

The goal of this study is to build a model that can accurately predict fraudulent transactions. Fraud detection is a highly imbalanced problem, with fraudulent transactions representing a small proportion of the total dataset. The key business objective is to minimize false negatives (missed frauds) while maintaining a low false positive rate to avoid inconveniencing legitimate customers.

2.2 Data Understanding

The dataset used in this study contains anonymized transaction features, labeled V1 to V28, derived from a PCA transformation, along with features like ’Amount’ and ’Class’, where ’Class’ indicates whether a transaction is fraudulent. Initial exploration of the data revealed a significant class imbalance, with fraudulent transactions accounting for less than 1% of the total transactions.

We performed a thorough exploration to assess feature distributions, correlations, and outliers, helping us understand the nature of the data before proceeding with preprocessing.

2.3 Data Preparation

Data preprocessing steps included handling missing values, scaling numerical features, and removing irrelevant features such as ’Time’. We also standardized the ’Amount’ feature to reduce scale differences between features.

2.4 Feature Engineering

We dropped the ’Time’ feature initially due to its minimal contribution to fraud detection, but later reconsidered this decision as part of our iterative modeling approach. We also created a normalized version of ’Amount’ to help improve model performance by standardizing its values.

3 Modeling and Optimization

3.1 Initial Model

We implemented a RandomForestClassifier due to its ability to handle imbalanced datasets. The model was trained using an 80-20 train-test split, with performance evaluated on various metrics, including accuracy, precision, recall, and F1-score.

```
Initial Model:
precision    recall  f1-score   support
       0       1.00      1.00      1.00     56864
       1       0.87      0.75      0.80        98
```

The model showed high accuracy but struggled with recall for the fraudulent class, highlighting the importance of further optimization.

3.2 Model Optimization

We enhanced the model by tuning hyperparameters and addressing the class imbalance through techniques such as adjusting class weights and applying SMOTE (Synthetic Minority Oversampling Technique). The optimized model demonstrated a notable improvement in recall for the fraud class:

```
Optimized Model:
precision    recall  f1-score   support
       0       1.00      1.00      1.00     56864
       1       0.94      0.85      0.89        98
```

The improvement in recall for fraud cases suggests a more balanced trade-off between capturing fraudulent transactions and avoiding false positives.

4 Evaluation

4.1 Performance Metrics

The model's performance was evaluated using several key metrics:

- **Accuracy**: Proportion of correct predictions over total predictions.
- **Precision**: The proportion of correctly predicted fraud cases out of all predicted fraud cases.
- **Recall**: The proportion of correctly predicted fraud cases out of all actual fraud cases.
- **F1-score**: The harmonic mean of precision and recall.

We also examined the confusion matrix to understand the trade-offs between true positives, false positives, true negatives, and false negatives.

4.2 Confusion Matrix

The confusion matrix for the optimized model is as follows:

```
Confusion Matrix:
[[56863, 1],
 [   15, 83]]
```

The results show a substantial improvement in detecting fraudulent transactions while maintaining a low false positive rate.

5 Discussion

Our model's high precision and recall suggest it is well-suited for real-world deployment in credit card fraud detection systems. The KDD methodology played a critical role in systematically approaching the problem and iteratively improving model performance.

One of the challenges we encountered was the extreme class imbalance, which we addressed through SMOTE and class weight adjustments. Future work could explore more advanced techniques such as deep learning models or ensemble methods to further improve fraud detection accuracy.

6 Conclusion

This study demonstrates the effective use of the KDD methodology in building a predictive model for credit card fraud detection. The structured approach of KDD allowed us to systematically analyze, preprocess, and model the data, ultimately resulting in a robust fraud detection system. Our findings emphasize the importance of balancing recall and precision in fraud detection systems, ensuring that fraudulent transactions are identified while minimizing the impact on legitimate customers.

7 Future Work

Future research will explore advanced anomaly detection techniques, such as autoencoders or graph-based approaches, to further enhance the model's ability to detect rare fraudulent activities. Additionally, deploying the model in a real-time environment would provide valuable insights into its practical performance and the potential for continuous learning in evolving fraud detection scenarios.