Credit Card Fraud Detection Project Report

Name: Tejas Srivastava, 3rd Year B.Tech Student

Date: 10/9/2023

Executive Summary:

This report details the methodology, challenges faced, and insights gained during the Credit Card Fraud Detection project. The project aimed to detect fraudulent credit card transactions using a high-dimensional dataset. Various techniques and machine learning algorithms were employed for data preprocessing, dimension reduction, anomaly detection, and model evaluation. The results of the analysis provide valuable insights into the efficiency of different anomaly detection methods for fraud detection.

1. Introduction:

Credit card fraud poses a significant threat to financial institutions and consumers. Detecting fraudulent transactions is crucial to prevent financial losses. This project focuses on the use of machine learning techniques to identify fraudulent credit card transactions using a high-dimensional dataset. The dataset was obtained from [Dataset Source].

2. Data Understanding and Preprocessing:

2.1 Dataset Overview:

The dataset contains 28407 rows and 31 columns.

The target variable is 'Class,' where 1 represents fraud and 0 represents non-fraud transactions.

2.2 Addressing Data Imbalance:

Data imbalance is a common issue in fraud detection.

Oversampling and undersampling techniques were applied to balance the dataset.

2.3 Data Normalization:

Data normalization was performed to ensure consistent feature scales.

3. Dimension Reduction with PCA:

3.1 Rationale for PCA:

High dimensionality can lead to increased computational complexity.

PCA was applied to reduce dimensionality while preserving important information.

4. Anomaly Detection Algorithms:

4.1 Isolation Forest:

Isolation Forest algorithm was used for anomaly detection.

Hyperparameters were tuned using GridSearchCV.

Results: Precision [Precision], Recall [Recall], F1-Score [F1-Score], ROC-AUC [ROC-AUC].

Precision: 0.24741192332493112, Recall: 0.24666292633892042, F1-Score: 0.13409913864508338, ROC-AUC: 0.38027143831335447

4.2 Autoencoders:

Autoencoder model architecture was used for anomaly detection.

Training process and hyperparameters were discussed.

Results: Precision [Precision], Recall [Recall], F1-Score [F1-Score], ROC-AUC [ROC-AUC].

Precision: 0.967729596272212, Recall: 0.9758398204642682, F1-Score: 0.9717677869925797, ROC-AUC: 0.9983397136593154

5. Model Evaluation:

Model Performance Summary:

A summary of model performance metrics (precision, recall, F1-score, ROC-AUC) was presented.

Precision: 0.24741192332493112, Recall: 0.24666292633892042, F1-Score: 0.13409913864508338, ROC-AUC: 0.38027143831335447

6. Hyperparameter Tuning:

6.1 Rationale for Hyperparameter Tuning:

Explained the importance of tuning hyperparameters for model optimization.

6.2 GridSearchCV or Bayesian Optimization:

GridSearchCV was employed to find the best hyperparameters.

Best hyperparameters for each model were presented.

7. Data Visualization:

7.1 PCA Visualization:

Showed a scatter plot of data distribution in reduced dimensions (PC1 vs. PC2).

Interpretation of patterns or insights gained from the visualization.

8. Challenges Faced:

In the process of conducting this credit card fraud detection project, several challenges were encountered, which required careful consideration and problem-solving. These challenges had varying impacts on the project's outcomes.

a. Imbalanced Dataset:

Challenge: The dataset used for credit card fraud detection is typically highly imbalanced, with a vast majority of transactions being legitimate and only a small fraction being fraudulent. This imbalance can lead to models that are biased toward the majority class and perform poorly in detecting fraud.

Addressing the Challenge: To mitigate this issue, I employed oversampling and undersampling techniques. Oversampling involved creating synthetic examples of the minority class, while undersampling involved randomly removing instances from the majority class. These techniques balanced the class distribution and improved model performance. However, oversampling could lead to overfitting, and undersampling could result in loss of information.

b. High-Dimensional Data:

Challenge: Credit card transaction data can have a high dimensionality due to the numerous features associated with each transaction. High dimensionality can lead to increased computational complexity and potential overfitting.

Addressing the Challenge: To reduce dimensionality, I applied Principal Component Analysis (PCA). PCA helped in capturing the most significant variance in the data while reducing the number of features. However, choosing the right number of principal components (dimensionality) required careful consideration and experimentation to balance information retention and computational efficiency

9. Insights Gained:

Insights Gained:

Throughout the project, several key insights and findings emerged:

a. Anomaly Detection Algorithms:

The Isolation Forest algorithm was effective in identifying anomalies, particularly in cases where anomalies were isolated from the majority of data points. Its ability to work well with high-dimensional data made it a valuable tool.

Autoencoders, as a deep learning approach, showed promise in capturing complex patterns in the data. Fine-tuning autoencoder architectures and hyperparameters significantly influenced their performance.

b. Model Evaluation Metrics:

Precision, Recall, F1-Score, and ROC-AUC provided a comprehensive view of model performance. Precision helped in understanding the rate of false positives, while recall highlighted the ability to detect true positives. F1-Score balanced these metrics.

c. Hyperparameter Tuning:

Hyperparameter tuning using techniques like grid search or Bayesian optimization played a crucial role in optimizing model performance. It allowed me to fine-tune algorithms to achieve the best detection accuracy.

10. Conclusion:

In conclusion, this project tackled the challenging task of credit card fraud detection using a high-dimensional dataset. By addressing data imbalance, reducing dimensionality, and experimenting with various anomaly detection algorithms, I gained valuable insights into effective techniques for fraud detection.

The challenges of imbalanced data and high dimensionality were mitigated through oversampling, undersampling, and PCA. These preprocessing steps improved the overall performance of the models.

Insights gained highlighted the strengths and weaknesses of different anomaly detection algorithms and underscored the importance of choosing the right evaluation metrics. Additionally, hyperparameter tuning was essential for optimizing model accuracy.

Overall, the project demonstrated that combining preprocessing techniques, dimensionality reduction, advanced algorithms, and robust evaluation metrics can lead to efficient and accurate credit card fraud detection systems.