FraudWatch: A Comparative Study on Fraud Detection Techniques

Prashanth Aripirala, Rishik Shekar Salver

1. Introduction

In an era dominated by digital transactions, the rising threat of credit card fraud necessitates advanced detection mechanisms. This project addresses this challenge by employing a systematic approach to preprocess and analyze credit card transaction data. Through meticulous class balance checks, null value handling, and dimensionality reduction techniques, we aim to enhance the efficiency and accuracy of fraud detection models. The exploration involves feature engineering and Principal Component Analysis (PCA), leading to the identification of crucial features and dimensionality reduction for optimal model performance. Additionally, a Mann-Whitney test assesses the impact of missing 'amount' data on fraud classification.

Subsequent sections detail the construction and evaluation of machine learning classifiers, anomaly detection models, and an innovative hybrid approach. This report provides a comprehensive account of our methodology and findings in the realm of credit card transaction fraud detection.

2. Dataset

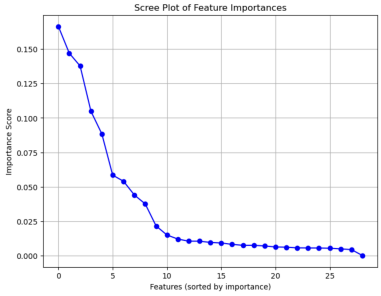
The foundation of our credit card fraud detection project lies in a meticulously curated dataset designed to capture the intricacies of real-world transactions. Sourced from Kaggle, this dataset serves as the bedrock for training and evaluating our fraud detection models. The dataset consists of 30 variables, V1-V28, Amount, and Class Label indicating whether a transaction is fraud or not. The variables V1-V28 are anonymized transaction attributes like time, location, etc. The values in these variables are scaled and transformed.

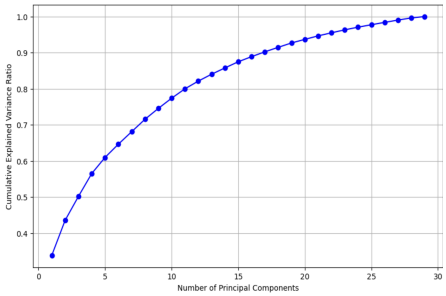
3. Data Preprocessing

This section details the steps taken to refine and optimize the raw data. A comprehensive examination ensued, ensuring class balance, the absence of null values. After performing a quick correlation analysis, it was observed that a lot of variables are correlated. So, dimensionality reduction became an obvious option as our next step. Here two approaches are explored and compared.

a. Feature Selection

Constructing a tree-based model unveils the significance of each feature. The resulting feature importance plot guides us in identifying ten key variables with substantial impact. Beyond this point, additional features contribute minimally to predictive performance.





b. Principal Component Analysis

PCA offers a different approach to dimensionality reduction, converting original variables into uncorrelated principal components. It was found that the first 17 components explain 90% of the total variance.

Datasets from both of the approaches have been extracted and validated against our baseline Logistic Regression model. It was found that the feature Selection dataset gave better results in addition to the advantage that it has lesser dimension compared to the PCA dataset. So we move ahead with the feature selection dataset. ['V2', 'V3', 'V4', 'V7', 'V10', 'V11', 'V12', 'V14', 'V16', 'V17'] are the final set of features

4. Hypothesis Testing

Our exploration into the absence of the 'amount' column prompts a focused hypothesis test using the Mann-Whitney U test. The objective is to assess whether the exclusion of this variable significantly influences the distribution of fraud and non-fraud transactions.

Null Hypothesis (H0):
There is no significant difference in the distribution of fraud and non-fraud transaction amounts

Alternative Hypothesis (H1):
There is a significant difference in the distribution of fraud and non-fraud transaction amounts.

Looking at a p-value of 0.08 for an alpha of 0.05 suggests that there is no significant difference in their distributions.

5. Model Evaluation and Performance

a. Classification

Logistic regression: As our baseline model, logistic regression serves as a fundamental benchmark for evaluating the performance of subsequent classifiers. Trained on the feature-selected dataset, logistic regression provides insights into the linear relationships between features and the likelihood of fraudulent transactions.

K-Nearest Neighbours: Expanding our repertoire, KNN harnesses the power of proximity-based classification. This model evaluates transactions based on the characteristics of their neighboring data points, providing a non-linear perspective on fraud detection.

Random Forest Classifier: Random Forest emerges as a powerful ensemble learning technique, leveraging multiple decision trees to enhance accuracy and robustness. Its capacity to handle complex relationships within the data positions it as a key player in our ensemble of classifiers.

The performance of each classifier is rigorously assessed using a suite of metrics, including accuracy, f1 score, and precision. These metrics provide a comprehensive understanding of each model's ability to accurately classify both fraudulent and non-fraudulent transactions.

Model/Metric	Accuracy	Precision	F1-Score
Logistic Regression	0.95	0.97	0.95
KNN	0.99	0.99	0.99
Random Forest Classifier	0.99	0.99	0.99

b. Anomaly Detection

Anomaly detection, crucial in data analysis, identifies deviations from the norm, particularly vital in credit card transactions for early fraud detection. These systems adapt to evolving tactics, offering dynamic defense without frequent retraining. The two models of anomaly detection we implemented are:

Isolation Forest:
The Isolation Forest method, an unsupervised anomaly detection algorithm, proves highly efficient in isolating anomalies within high-dimensional datasets like credit card transactions. During training, the algorithm strategically selects features and swiftly splits data points, identifying anomalies with fewer splits than normal instances. This approach ensures a rapid and effective detection of fraudulent transactions deviating from the majority.

One-Class SVM:
The One-Class SVM, a specialized support vector machine, offers a unique approach by learning the representation of normal instances during training. Capturing the characteristics of normal transactions, it classifies new instances based on their similarity to this learned representation. Anomalies, deviating from this learned norm, are efficiently identified. Particularly valuable for datasets like credit card transactions, the

One-Class SVM excels in flagging transactions with unusual patterns, indicating potential fraud.

Both anomaly detection models exclusively trained on non-fraud data distinguish norms, categorizing out-of-norm points as fraud. Predicting fraud for each test dataset point, we applied key metrics for performance assessment. The table below illustrates anomaly model performance.

Model/Metric	Accuracy	Precision	F1-Score
Isolation Forest	0.913	0.919	0.912
One-Class SVM	0.919	0.946	0.916

c. Hybrid Architecture

In our pursuit of enhanced fraud detection, we developed a hybrid model that integrates a Random Forest Classifier with the anomaly detection capabilities of One-Class SVM. The process involved training the Random Forest Classifier on the dataset, obtaining predicted probabilities for both train and test data, and feeding these probabilities into the One-Class SVM anomaly detection model. By utilizing anomaly scores derived from this approach, potential fraudulent transactions were identified.

The hybrid model is a promising strategy for credit card fraud detection, merging a traditional classification model's strengths with anomaly detection's adaptability. It offers resilience to new fraud tactics. Here is the hybrid model's performance.

Accuracy	Precision	F1-Score
0.54	0.57	0.54

Contrary to expectations, the hybrid model showed weaker performance than standalone models, prompting a need for deeper optimization exploration in future iterations. Ongoing refinement is crucial to fully leverage the hybrid model's potential advantages.

6. Conclusion and Future Scope

Tree-based feature selection significantly contributed to robust standalone model performance. Principal Component Analysis (PCA) provided limited improvements, with the feature selection dataset outperforming the PCA-transformed dataset. The hybrid architecture, integrating Random Forest and One-Class SVM, faced challenges, displaying weaker performance than individual models. This highlights the need for nuanced refinement in the hybrid model's design and application.

To advance our project, prioritizing additional feature engineering techniques is crucial. Exploring diverse methods to enrich captured information and enhance discriminative capabilities will be a key focus. Investigating alternative ways to combine the strengths of Random Forest and One-Class SVM, beyond the explored hybrid architecture, presents a promising avenue for improving model synergy.

7. References

https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023/data
https://blog.paperspace.com/anomaly-detection-isolation-forest/
https://www.datatechnotes.com/2020/04/anomaly-detection-with-one-class-svm.html
https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425

8. Appendix

Link to Code: https://drive.google.com/file/d/1rA8mGuArZm6k5y-5O1Ik1d3UebkH0yOr/view?usp=drive_link