**Consultancy Report on Credit Card Fraud Detection: Data Analysis and Model Evaluation**

**Abstract**

As fraudulent activities frequently been reported and can cause huge financial losses, we need to better manage potentially fraudulent transactions. One way is to make full use of the historical transaction data including normal transactions and fraud ones to obtain normal/fraud behavior features based on machine learning techniques. After applying three machine learning methods including logistics regression, XG-boost and random forest, it turns out that random forests can effectively identify fraudulent credit card transactions in a dataset containing both legitimate and fraudulent activity. This report shows how random forests can be applied so successfully of its performance on credit fraud detection.

**Introduction**

In real-world scenarios, credit card datasets have outliers, and fraudulent transactions usually have complex patterns. Random forest algorithm was used for our machine learning model. However, due to limits from given dataset and some practical reasons, certain barriers made it hard to identify exact trends and patterns. Based on some challenges we encountered; some recommendations were made to revise this action in the future.

**Data Cleaning and Data Analysis:**

This dataset from Kaggle contains credit card transactions made by European cardholders in the year 2023. It comprises over 550,000 records, and the data has been anonymized to protect the cardholders' identities.

We kept detecting the missing values，rows with NULL value, duplicates and deleted them. We also converted the ‘Amount’ column into integers，which was originally showing numerical value and might not correctly demonstrate the practical meaning.

We found that the original dataset is the result of Principal Component Analysis (PCA), a statistical method used to simplify the structure of data by reducing the dimensionality of the data. As we find that the data in this column of ‘Amount’ is much more larger than other columns (V1-V28), we standardized the data in this column of ‘Amount’ to the same scale，to avoid the other data with certain features dominating the algorithm, which affects the performance of the model.

The pie chart (Fig1) showcases the dataset with a 50%-50% proportion of fraud and non-fraud transactions, it seems like the data is intentionally balanced. While this equal distribution is helpful in algorithm training, it does not reflect on the real-world prevalence of fraud, which only accounts for 1-5% of transactions (CyberSource, 2024). Because of this reason, the models trained on the dataset may struggle with real-world applications.

**图表, 饼图

描述已自动生成**  **图表, 饼图

描述已自动生成**

***Fig 1: Distribution before cleaning Fig 2: Distribution after cleaning***

After data cleaning, the dataset now contains 240,131 fraudulent transactions, which is 45.9% of the total, and 204,112 non-fraudulent transactions, accounting for 54.1%. Even though the proportion of fraudulent transactions has decreased, it is still biased in terms of real-world representation.

**图表, 条形图

描述已自动生成**

***Fig 3: Fraud and legit transaction count***

**Correlation Heatmap:**

A diagram of a heat map

Description automatically generatedThis correlation heatmap shows the relationship between features V1 to V28 and the target variable ‘Class’, which identifies whether a transaction is fraudulent or not. To focus on the core features for fraud prediction, we have excluded non-predictive columns such as ‘id’ and ‘Amount’ in this heatmap.

***Fig 4: Correlation heatmap***

**Fig 4** shows that high correlation values are primarily clustered among features V1 to V18, which suggests that these features may have similar patterns or describe similar features in the transaction data (Arpit, Xiong and Socher, 2019).

**Transaction Amount Distribution:**

**图表, 箱线图

描述已自动生成**

***Fig 5***

**图表, 箱线图

描述已自动生成*Fig 6***

**图表, 条形图, 直方图

描述已自动生成**

*Fig 7*

The plots in Fig 5-7 show the distribution of transaction amounts for both non-fraudulent and fraudulent transactions. Overall, both types of transactions are distributed evenly across the entire range of amounts (from 0 to around 25,000), without focusing on any specific range. Additionally, the distribution shapes of normal and fraudulent transactions are very similar, with no significant differences.

These observations suggest that the transaction amount alone does not show a clear difference between normal and fraudulent transactions. Therefore, it would be challenging to effectively identify fraud based only on the amount.

**Methodology:**

**Model selection**

We used Logistic Regression, Random Forest, and XG Boost classification algorithms for our Machine-learning models as they are easy to interpret, robust to outliers and good at handling large datasets. We then compared their effectiveness and overall performance in identifying fraudulent transactions in the dataset.

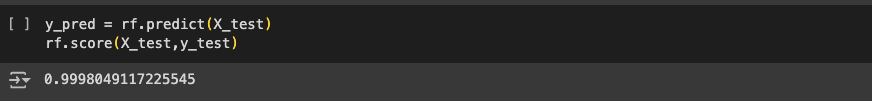
In real-world scenarios, credit card datasets contain outliers, and fraudulent transactions usually have complex patterns. Unlike logistic regression, random forest classifier is robust to outliers and nonlinear relationships. Random forest algorithm uses feature bagging and ensemble learning techniques to make complex decisions based on multiple features and their relationship to each other. Additionally, the random forest classifier is computationally less intensive than the XG Boost algorithm on large datasets while giving approximately equal output therefore we have used random forest algorithm for our machine learning model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic regression | Random Forest | XG boost |
| True positive (TP) | 61114 | 72040 | 72046 |
| True negatives (TN) | 71984 | 61212 | 61197 |
| False positive (FP) | 56 | 6 | 9 |
| False negative (FN) | 119 | 15 | 21 |

***Table 1 Comparison of error rate across different algorithm***

**Results and Analysis:**

The random forest algorithm performs exceptionally well in identifying and segregating legitimate and fraudulent transactions. The diagrams below (Fig 9) show the accuracy of the ML (machine learning) and (Fig 10) shows the model precision, recall and F1- score



***Fig 9: Accuracy score***

表格

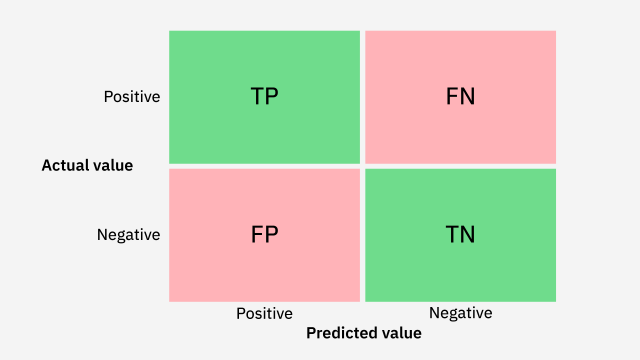
描述已自动生成

***Fig 10: Classification report (precision, recall, f1- Score)***

A blue squares with white text

Description automatically generated

***Fig 11: Confusion matrix***



***Fig 12: Detailed Confusion matrix***

**True positive (TP)** - correctly identifies 72,035 legitimate transactions.

**True negative (TN)** - correctly identifies 61,212 fraudulent transactions.

**False negative (FN)** - Identifies 15 legitimate transactions as fraud.

**False positive (FP)** - Identifies 6 fraudulent transactions as legitimate.

**Fig 11,12** shows the confusion matrix of the model that explains the error rate of the model. It explains how many transactions it identifies as legitimate and fraudulent and how many errors it makes.

Since the data is anonymized using PCA, we lack contextual understanding, making it challenging to interpret the features directly. However, we can still identify feature importance, revealing which columns have the greatest impact on the model's predictions. In our case, the feature V14 shows significant influence on model’s predictions as shown in **Fig 13.**

日程表

低可信度描述已自动生成

**Fig 13: *Feature Importance***

**Fig 14**, Shows you the ROC Curve and the AUC score, which explains the overall performance of the model. Our machine learning’s AUC score is 1. which means the model can predict approximately 100% of fraudulent transactions.

图表, 折线图

描述已自动生成

***Fig 14:*** *ROC curve and AUC score*

A screenshot of a computer

Description automatically generated**Dashboard:**

***Fig 8: Dashboard***

The dashboard ***fig 8*** was designed to enable the filtering of data to compare fraudulent (Class 1), non-fraudulent (Class 0) and both transactions, simplifying the data to reveal key insights for fraud detection.

**Challenges** **encountered:**

1. Firstly, in fraud detection, model adaptability and real-time performance are crucial. Fraud patterns may change over time. In the future, it is recommended to increase the adaptability of the model by regularly updating the data and retraining the model to respond to new fraud behaviors in a timely manner.
2. Secondly, anonymization of data made it hard to interpret the data. Important information such as transaction type, date, time, gender, age and geographic location etc. were hidden which made it hard to identify trends and patterns and affect the model’s performance.
3. Thirdly, after using PCA and reducing the dimensionality of data, certain features with high correlation was found. However, it’s hard to combine or remove the feature since the nature of the data are unknown. This limits our expansion possibilities in feature engineering. If the data had not been anonymized by PCA, we could have performed predictive analyses on the original features to uncover more feature patterns related to fraudulent transactions.

**Strategic recommendations:**

**1. Dynamic data update and retraining**

The patterns and characteristics of fraudulent behavior will change over time, and static models will gradually lose their effectiveness in the face of new types of fraud. Therefore, it is recommended to establish a regular data update and model retraining process to ensure that the model can continue to identify the latest fraudulent behavior. This dynamic update helps to improve the continued accuracy of the model, but also ensures that the model maintains a high degree of sensitivity when dealing with new types of fraudulent behavior.

**2. Anomaly detection methods**

Considering the actual fraud problem, new types of fraud will appear. In this situation, there are some anomalies that do not fit the model, so the model may not be able to identify. It is recommended that the anomaly detection method be added in addition to the main model. This method helps to capture anomalous transactions outside the distribution and enhance the detection coverage of the system.

**3. Feature analysis**

Since the 28 features of the current dataset have been anonymized by PCA processing, the specific meaning of the original features is lost. This limits our expansion possibilities in feature engineering. If the data is not anonymized by PCA, we can perform exploratory and predictive analysis on the original features to identify deeper associations, patterns and trends, further improving the efficiency of the model in fraud detection.

**Conclusion and Reflections:**

1. **Conclusion**

This project significantly improves fraud detection accuracy by identifying highly correlated transaction features and focusing the model on the most valuable features. The Random Forest model excelled in identifying fraudulent transactions and controlling false positives, helping to reduce the negative impact of the customer experience while effectively controlling financial risk, saving the company potential risk management costs and supporting business security objectives.

**2) Reflections**

For visual analytics, current visual analytics have helped identify some important features, but in the future, more interactive visualization tools can be applied so that business teams can explore the distribution and relationships of data in different dimensions to further deepen their understanding of fraud detection

**References:**

1. Wang Yundong, A. Z. G. A., 2023. *Wasit Journal of Computer and Mathematics Science.* [Online]   
   Available at: <https://wjcm.uowasit.edu.iq/index.php/wjcm/article/view/184/109> [Accessed 25 October 2024].
2. AlsharifHasan Mohamad Aburbeian, H. I. A., 2023. *researchgate.* [Online]   
   Available at: <https://www.researchgate.net/publication/369199151_Credit_Card_Fraud_Detection_Using_Enhanced_Random_Forest_Classifier_for_Imbalanced_Data> [Accessed 25 October 2024].
3. M.Suresh Kumar, V. ,. S. ,. E. ,. E., 2019. *IEEE.* [Online]   
   Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=8824930&tag=1> [Accessed 25 October 2024].
4. Arpit, D. Xiong, C. and Socher, R. 2019. Predicting with High Correlation Features. *arXiv: Machine Learning*.

**Appendix:**

**Dataset -** [**https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023/data**](https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023/data)

**Data analysis and Cleaning:** [**https://colab.research.google.com/drive/1TQL8qJGaOfdOY5gaea0Mo0BQMoHa\_zeW**](https://colab.research.google.com/drive/1TQL8qJGaOfdOY5gaea0Mo0BQMoHa_zeW)

**Machine learning Models：**

Random forest ML model: <https://colab.research.google.com/drive/1YwSFj9kq-QmMFZfAglj9hZx32xyUN_Cf?usp=sharing>

Logistic regression ML model: <https://colab.research.google.com/drive/1Q9e2j49Qux3e5PGs0oAD731Sy14QLk6b?usp=sharing>

XG Boost ML model: <https://colab.research.google.com/drive/1Qiv0cNk1s_yCO99qEbGaoDGY6w4pF0Cq?usp=sharing>

**Data visualization:** [**https://colab.research.google.com/drive/1er7bS7J5S6vlTQu0qYyEwKlQ9Vgc8o\_Y**](https://colab.research.google.com/drive/1er7bS7J5S6vlTQu0qYyEwKlQ9Vgc8o_Y)