Data Science Project Report

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Predicting Fraud in Financial Payment Services

Anjali, Riya
Department of Physics
Indian Institute of Technology Hyderabad

Roll No: PH22MSCST11021, PH22MSCST11007

Project Advisor: Prof. Shantanu Desai
Department of Physics
Indian Institute of Technology Hyderabad

January 21, 2025

Abstract

As we know that financial fraud is great threat in this digital age, that impact all individuals, institutions etc. This thesis helps to tackle this challenge by using a dataset from Kaggle which is a paysim simulation . With the help of this data we want to predict the pattern of fraudulent transactions. In this data set we have 11 features we will analyse them using EDA and to improve the model performance we will do data cleaning and data visulization.

Contents

1 Introduction

In digital age we know that financial fraud is a great threat for individuals as well as institutions. To combat this challenge we need datasets but this type of datasets are rare to find. The one we are using here is one of only four datasets present on kaggle. This data gives us an opportunity to find an effective method for the financial fraud detection. But there are some challenges, the data is highly imbalanced. So, the main goal here is to resolve the issues by detailed data exploration and to develop a practical solution for identifying fraud, and make digital finance more secured and safe. [1]

2 Dataset

The dataset that we are using here is a PaySim dataset i.e. a simulated dataset for mobile money transaction. But the dataset that we are using here is scaled down to 1/4 of the orignal dataset in research papers.
There are total 11 features in the dataset that are explained below:

1. **step:** It represent a unit of simulated time and here 1 unit represent 1 hour in real-world.
2. **type:** It tells about the type of transaction. There are 5 type here:
   - **CASH-IN:** It tells that money is added to a mobile money account.
   - **CASH-OUT:** It tells that money is withdrawled from mobile money account.
   - **DEBIT :** Money debited from a mobile account(like for purchases).
   - **PAYMENT:** Sending money to the another mobile account.
   - **TRANSFER:** Transfer of money between mobile accounts.
3. **amount:** It shows the amount of transaction in local currency .
4. **nameOrig:** Customer ID of the customer who initiated the transaction.
5. **oldbalanceOrig:** It shows the balance amount of the customer who started transaction before the transfer occurred.
6. **newbalanceOrig:** It shows the balance amount of the customer who started transaction after the transfer occurred.
7. **nameDest:** Recipient ID of the transaction.
8. **oldbalanceDest:** It shows the balance amount of the recipient before the transfer occurred.
9. **newbalanceDest:** t shows the balance amount of the recipient after the transfer occurred.
10. **isFraud:** It tells about the fraudulent transactions. Where 1 indiactes fraudulent transaction and 0 indicates non-fraudulent transaction.
11. **isFlaggedFraud:** It flags the illegal attempts of transferring more then 200,000 in single transactions.

3 EDA (Exploratory Data Analysis)

EDA is an important step in data science. EDA helps us to understand the data in better way , to find the relationship between features, it tells that do we require data cleaning or not for further analysis. It can helps us to find a better model and can suggest the need of feature engineering to increase the model performance.

3.1 Which type of transactions are fraudulent?

We find out that out of 5 type of transactions fraud occurs only in two transaction types. TRANSFER ,CASH-OUT and we also got to know that the number of fraudulent TRANSFER (4097) is nearly equal to the number of fraudulent CASH-OUT (4116), suggests that there is a two-step pattern , first transferring the money to another account then cashing it out.

3.2 What feature and factors sets the isFlaggedFraud feature?

As we mentioned in above section the isFlaggedFraud get set when the transfer exceeds 200,000 , but we found out that is not always true. We encountered many example when its not set despite of condition being met. So we tried to analyse correlation between isFlaggedFraud and any other feature in dataset.

- First we tried to find relation of isFlaggedFraud with oldbalanceDest and newbalanceDest. We noticed that where ever the old and new balance is identical the isFlaggedFraud is set. This may be due to the transaction is halted before completion but this is not always the case.
- We found that it do not depend upon oldbalanceOrg as the range of value where isFlaggedFraud is set overlaps with the range of value where it is not set. That indicates it is not enough to determine that the transaction is fraudulent.
- There is not specific time pattern for flagged transaction or transaction frequency of senders and receivers is not related to flagged transaction as well.

Based on this we can tell isFlaggedFraud is unreliable and we can discard it from our dataset.

3.3 Are expected accounts accordingly labelled?

According to the expected behaviour of dataset the merchant should be identified with a name prefixed ”M”. As in CASH-IN customer receives money from the merchant with a name prefixed ”M” and in CASH-OUT customers pay a merchant with a named prefixed ”M”. But through analysis we observed that there are no transaction where the recipient is a merchant (prefixed by ”M”). This shows that there may be an error in labelling merchants in the dataset.

3.4 Are there accounts labels common to fraudulent transfers and CASH-OUTs?

According to the data description the fraud might follow a two step pattern, by first making a transfer to a fraudulent account that cash it out. In this two step process the destination in a transfer and originator in a CASH-OUT would be the same fake account. But there are no such common accounts between TRANSFER and CASH-OUTs. Thus , the fraudulent transaction are not related to nameOrig and nameDest features and in above section as well we got to know that both of these features d not label merchant accounts in expected way, so we can drop both of these features.

4 Data Cleaning

Now we will clean up the data to prepare it for fraud detection. From the EDA section we got to know that only two type of transaction are involved in fraud i.e. TRANSFER and CASH-OUT .Thus we will filter out the data and only keep these two type of transaction for further analysis.

4.1 Imputation of missing values

From the analysis we discovered an interesting trend that alot of transaction have zero balance before and after transfer even if the transaction amount is non-zero and the main thing to notice is that this is much more common in fraudulent transactions (50%) compare to the genuine transaction(0.06%). Since, this zero balance in destination account after a transfer of non-zero amount can be a sign of fraud. Now imputing these values with our standard statistical approach may hinder the fraud and them appear as genuine transaction. SO, we used a new approach, we replaced these 0 values with -1, as it can interpreted by ML models for fraud detection. In the same way data also have a lot of transactions with zero balance in the originating account before and after transaction. But here its is very less common in fraudulent transaction (0.3%) compared to genuine ones(47%). So for imputing these values instead of standard approach replace them with 'null'. So, that ML models can understand the fraud pattern better.

5 Feature Engineering

Feature engineering is the technique to create new features from the existing data to improve model performance. In the previous we discovered that zero balance account can differentiate between fraudulent and genuine transaction, in this section we will go a step further and we will create two new features to capture the errors in originating and destination account for each transaction. There two new features will improve the performance of ML models used for fraud detection.

6 Data Visualization

The best way to know that ML model can predict well using the data, is to directly visualize the difference in the fraudulent and non-fraudulent transactions. To visualize it properly we used several plots.

6.1 Separating out fraud from non-fraud transaction

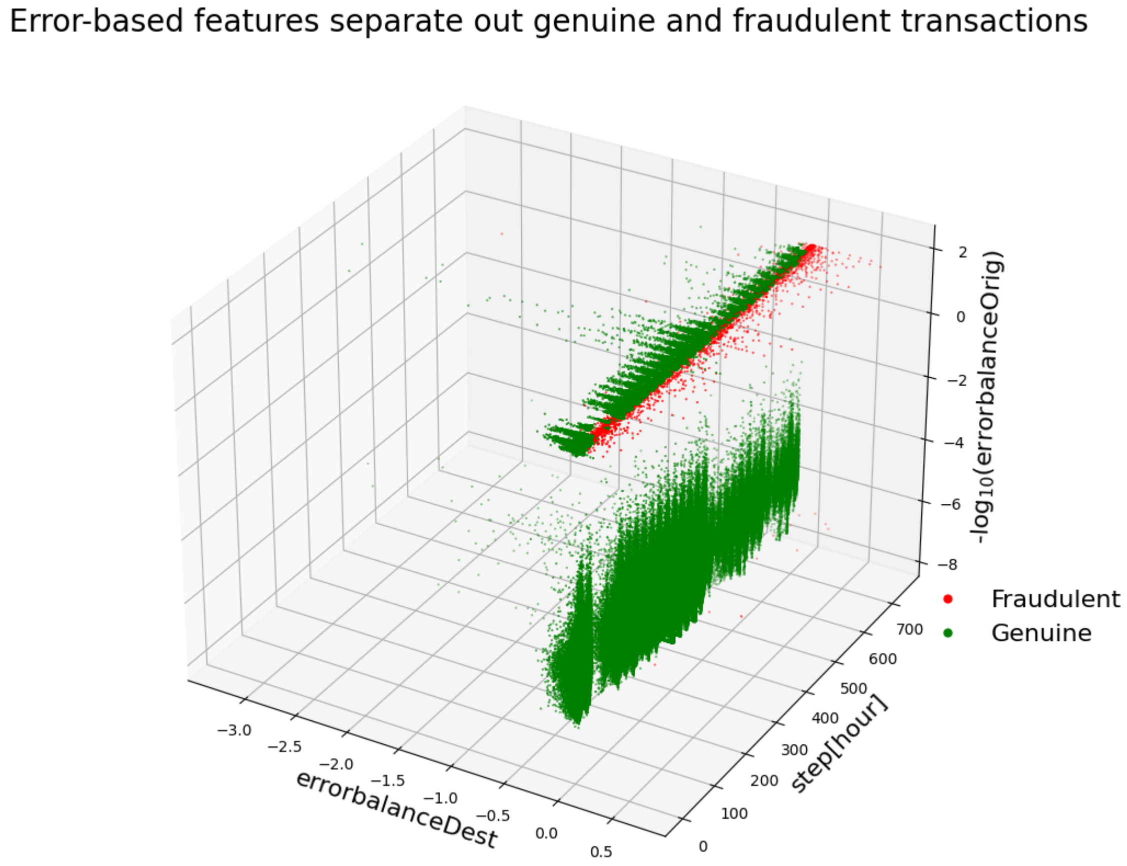

Figure 1: 3D data Visualization

We used a 3D plot to distinguish between fraud and non-fraud transactions by using both of the engineered error based features.

6.2 Fingerprints of genuine and fraudulent transaction

In this section we used heatmaps to explore the unique characteristics of fraudulent and non-fraudulent transactions. We can get to know about the correlations between the features from the below heatmaps.

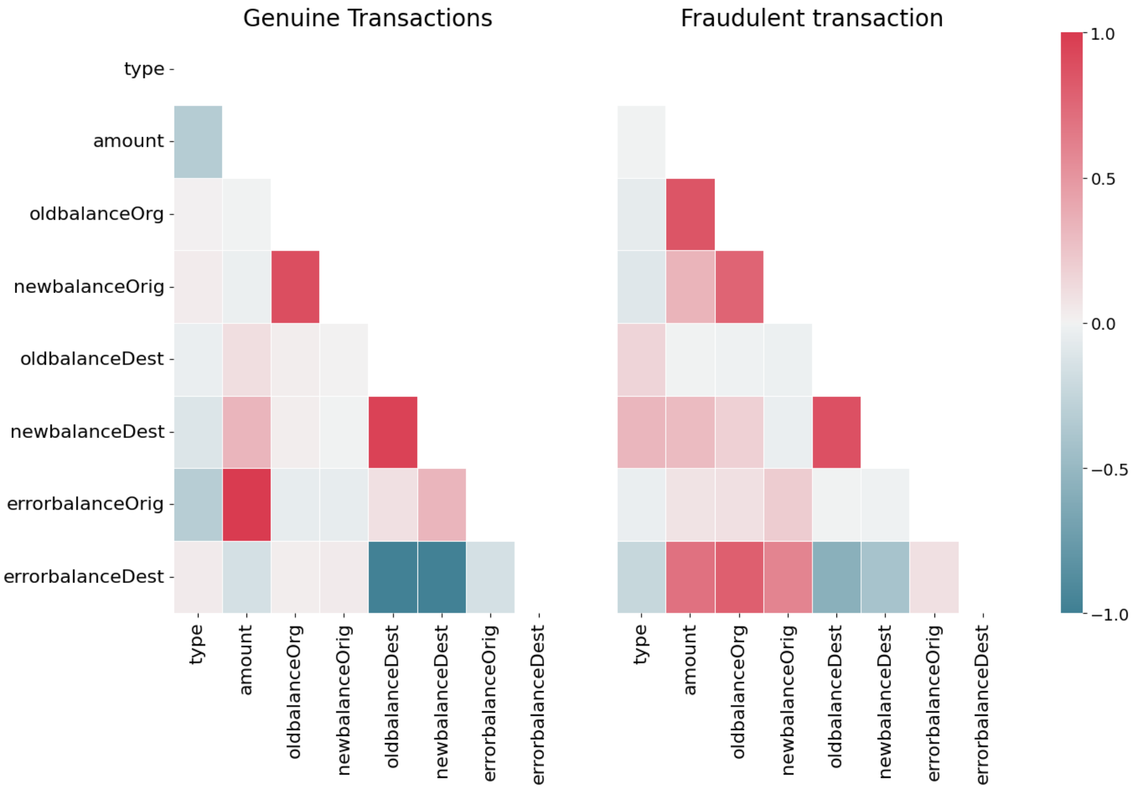

Figure 2: feature importance

7 Machine Learning to detect fraud in Skewed data

7.1 XGBoost

XGBoost or Extreme Gradient Boosting algorithm is an ensemble learning method that helps to build a predictive model with high performance and accuracy by combining the predictions of multiple individual models iteratively. Initially, a model is built using training data, then the next model is built in such a manner that it tries to correct the errors of the previous model. This process continues until, either the complete training data has been predicted correctly or the limit to the maximum number of models is reached. Key features of XGBoost are enhanced speed and performance of the model, efficient handling of missing data, built-in support for parallel processing, and fine-tuning that allows model flexibility according to the problem statement.

XGBoost performs better than the other models because it can deal with the missing values and works well for unbalanced data hence, it works well with real data that is unbalanced and has missing entries.

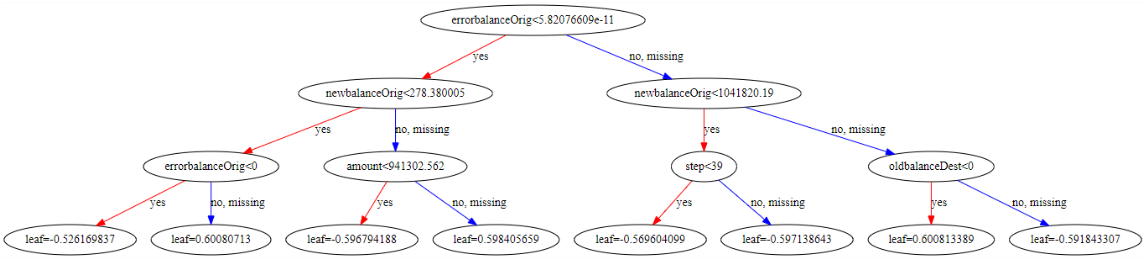

Figure 3: Model Visualization

7.2 Performance Evaluation

7.2.1 Correlation Matrix

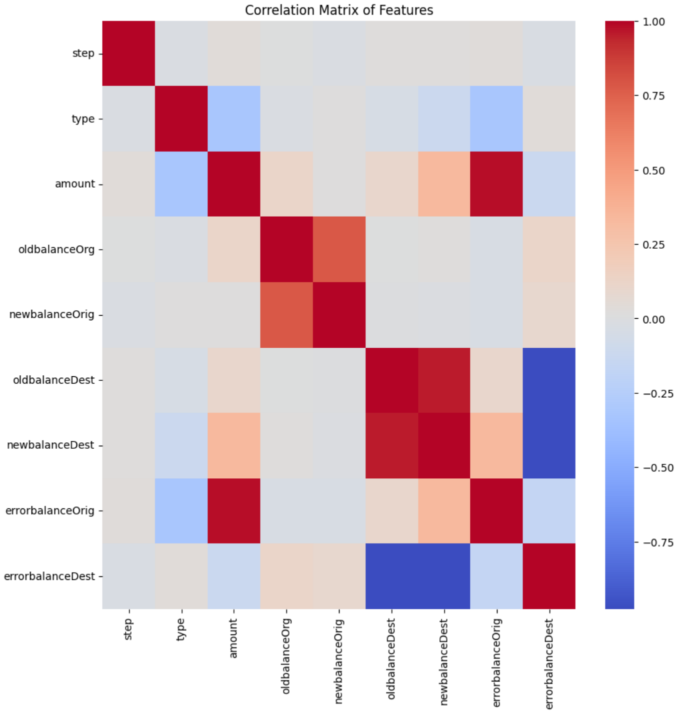

Figure 4: Correlation Matrix

Here in the correlation matrix, the correlation coefficient varies from -0.75 to 1, where positive and negative numbers indicate positive and negative correlation respectively. When a positive correlation is there that implies that the variables are moving in the same direction i.e. two variables with positive correlation increase and decrease simultaneously. If a negative correlation is there that implies the variables are moving in the opposite direction. A number closer to zero signifies that there is no correlation between the variables.

7.2.2 AUPRC

We split the data into 80:20, where 80% data is utilized for the model training. Since the data is highly skewed with skewness = 0.00296, we'll use AUPRC i.e. area under the precision-recall curve rather than the AUROC i.e. area under the receiver operating characteristic because AUPRC is highly sensitive to the differences between algorithms and their parameter settings in comparison to AUROC. After the model training the value for AUPRC was calculated as 0.991.

7.2.3 Feature Importance

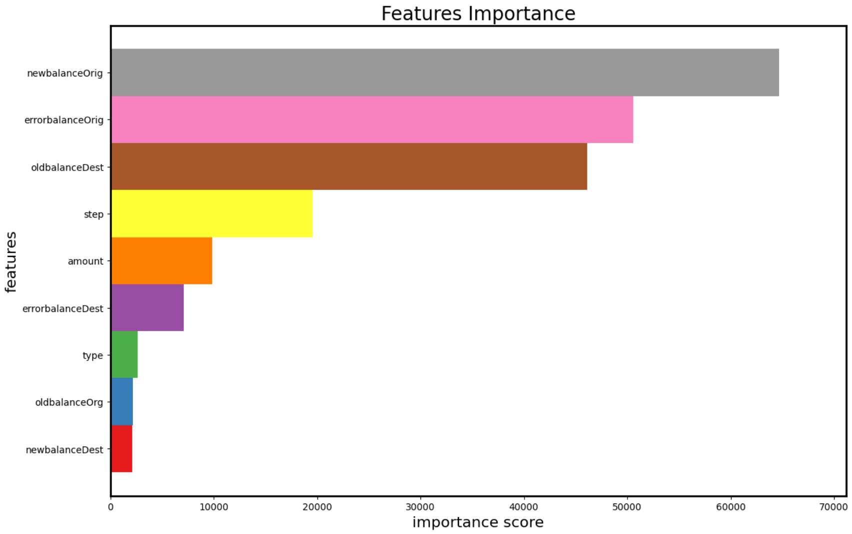

Figure 5: Correlation Matrix

As we can observe here that the new feature "errorBalanceOrig" that we have created is there in the top three most relevant features for the model.

8 Source Code

Source Code for the project can be found on the following link: https://github.com/anjali-anjalii/Predicting-Fraud-in-Financial-Payment-Services

9 Conclusion

In conclusion , this thesis contribute in the detection of fraud in this digital era using Data analysis. We took a paysim stimulated data from Kaggle and analysed it before model training. For the training purpose we used XGBoost model and we used AUPRC as the data was highly skewed and after training we got AUPRC as 0.991.

$$Keff =$$

References

[1] Predicting fraud in financial payment services. `https://www.kaggle.com/code/boscochanam2/predicting-fraud-in-financial-payment-services`. Accessed: 2024-04-30.