# Fraud Detection in Mobile Payment System

Ashish Jain
*M.Tech Computer Science and Engineering*
*IIIT Delhi*
ashish18052@iiitd.ac.in

Shubham Gupta
*M.Tech Computer Science and Engineering*
*IIIT Delhi*
shubham18055@iiitd.ac.in

## I. Problem Statement

Classification of synthetic datasets generated by the PaySim mobile money simulator into two classes: fraud and not fraud.

## II. Literature Review

This dataset is obtained from kaggle, and as per our knowledge, only one published work is there on this dataset. [1] use deep learning framework for detecting fraud transactions in PaySim data. They use restricted Boltzmann machines (RBM) and shows that it performs better than the technique stacked autoencoder (SAE) in terms of accuracy and ROC. We refer to the kernels of Kaggle for other literature. Because the dataset is highly skewed, [3] uses extreme gradient-boosted (XGBoost) algorithm for the fraud classification. XGBoost allows giving more weight to the positive class compared to the negative category. Since the data is highly skewed, he uses the metric AUPRC rather than AUROC as the latter is more sensitive towards the parameter settings of different algorithms [5]. The literature on credit card fraud detection presented by [6] shows a comparative study using Bayesian and neural networks. The work shows that Bayesian performs better than neural networks. The main problem in any fraud detection dataset is skewness of the data towards not fraud class. The skewed data leads to overfitting of the model. [4] use Synthetic Minority OverSampling Technique (SMOTE) which oversamples the minority class. Their results showed that the proposed approach could improve the accuracy of the minority class. [7]use a hybrid sampling of the dataset for credit card fraud detection. They achieve two sets of distribution (10:90 and 34:64) for analysis.

## III. Database details

Money transactions details contain various confidential information. Since there is a scarcity of available public datasets on money transactions primarily in the domain of mobile payment system. Therefore we are using synthetic dataset generated by a mobile money simulator PaySim [9]. The dataset is obtained from Kaggle [2]. Some sample of original datasets is shown in Fig.2 The dataset contains more than 6 million transactions each having 11 attributes. The dataset is highly skewed as shown in Fig.1 The details of the attributes are mentioned in Table I on Page 1.

TABLE I
Dataset Features Description

| Features | Description |
|---|---|
| Step | It maps a unit of time in the real world.In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation) |
| Type | CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER |
| Amount | amount of the transaction in local currency |
| NameOrig | customer who started the transaction |
| OldBalanceOrg | initial balance before the transaction |
| NewBalanceOrg | new balance after the transaction |
| NameDest | customer who is the recipient of the transaction |
| OldBalanceDest | initial balance recipient before the transaction. |
| NewBalanceDest | new balance recipient after the transaction. |
| IsFraud | Fraud, Not Fraud (1, 0) |
| IsFlaggedFraud | Marked if transaction attempt to transfer more than 200,000 in a single transaction |

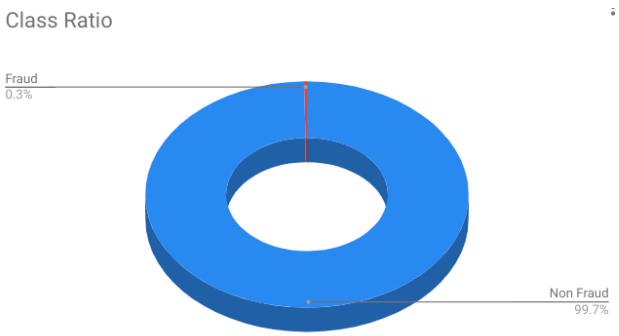


Fig. 1. Class wise Data Distribution

## IV. Task Completed

### A. Data Preprocessing and Feature Engineering

The original data from Kaggle contains more than six million samples. We preprocessed the data and reduces the number of samples to 2.7 million (Fig.3) according to the observations we made. We removed some irrelevant features without the loss of useful information. We perform dimensionality reduction by observing the data and not by any standard algorithms. The observations which lead to data preprocessing and obtaining relevant features are mentioned below:

- The attribute isFlaggedFraud is not useful:
  - It is set only 16 times in 6 million transactions.

| | step | type | amount | nameOrig | oldbalanceOrg | newbalanceOrig | nameDest | oldbalanceDest | newbalanceDest | isFraud | isFlaggedFraud |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | PAYMENT | 9839.64 | C1231006815 | 170136.0 | 160296.36 | M1979787155 | 0.0 | 0.0 | 0 | 0 |
| 1 | 1 | PAYMENT | 1864.28 | C1666544295 | 21249.0 | 19384.72 | M2044282225 | 0.0 | 0.0 | 0 | 0 |
| 2 | 1 | TRANSFER | 181.00 | C1305486145 | 181.0 | 0.00 | C553264065 | 0.0 | 0.0 | 1 | 0 |
| 3 | 1 | CASH_OUT | 181.00 | C840083671 | 181.0 | 0.00 | C38997010 | 21182.0 | 0.0 | 1 | 0 |
| 4 | 1 | PAYMENT | 11668.14 | C2048537720 | 41554.0 | 29885.86 | M1230701703 | 0.0 | 0.0 | 0 | 0 |

Fig. 2. Original Data Sample

| | step | type | amount | nameOrig | oldbalanceOrg | newbalanceOrig | nameDest | oldbalanceDest | newbalanceDest |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 181 | 1 | 181 | 0 | 2 | 0 | 0 |
| 1 | 1 | 1 | 215310 | 3 | 705 | 0 | 4 | 22425 | 0 |
| 2 | 1 | 1 | 311686 | 5 | 10835 | 0 | 6 | 6267 | 2.71917e+06 |
| 3 | 1 | 1 | 62610.8 | 7 | 79114 | 16503.2 | 8 | 517 | 8383.29 |

Fig. 3. Processed Data Sample

  - Whenever isFlaggedFraud is set the attribute isFraud is always set.
  - It is always set in payment Type Transfer. Moreover, the features oldbalanceDest and newbalanceDest are always 0.0 when isFlaggedFraud is set.

  So we decided to discard isFlaggedFraud attribute without loss of any useful information.
- We have five types of payment in the original dataset CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER. Only two payment types CASH-OUT and TRANSFER out of these five have fraud transactions. The rest three payment types have always isFraud set to 0. So we discarded the samples which have payment type other than CASH-OUT and TRANSFER. Fig.4 tells the ratio of fraud transactions with their payment types.

### B. Classification

Applied GaussianNB of sklearn on our preprocessed data. Obtained training accuracy. Split the preprocessed data into 80:20 and obtained the accuracy. Evaluation metrics used:

- Accuracy
- Confusion Matrix

## V. Task need to be completed

- Handling skewness of data by applying different undersampling or oversampling techniques for avoiding overfitting.

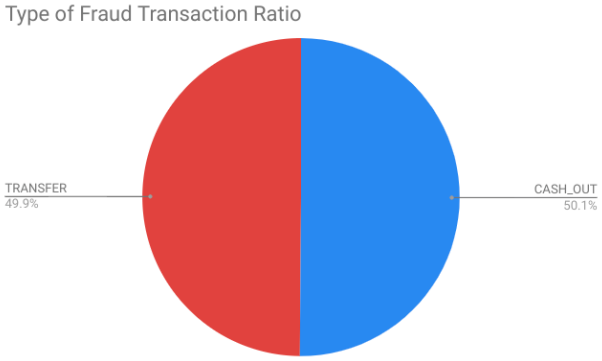


Fig. 4. Types of Transactions that are Fraud

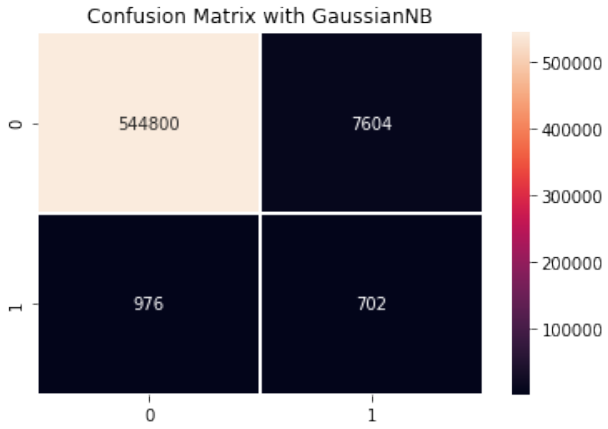


Fig. 5. Confusion Matrix Using Gaussian Naive Bayes on Test Data(split 80:20)

- Comparing the performance of different classifiers after handling the unbalanced problem.
- There are transactions which have destination original balance and new balance are zero when the amount is non zero. It is actually missing values which are fixed as zero in the dataset. Similarly for the source account. We will try to fix these missing values by using different techniques that dealt with missing data.

## VI. Results

Training Accuracy = 98.47. Its confusion matrix is reported in separate file. Testing Accuracy on a split (80:20) = 98.45. Its confusion matrix is shown in Fig. 5

### A. Analysis

Our model is overfitted due to highly unbalance between classes. The observations related to data preprocessing and feature engineering are reported above.

## References


[1] Mubalaike, A. M., & Adali, E. (2018, September). Deep Learning Approach for Intelligent Financial Fraud Detection System. In 2018 3rd International Conference on Computer Science and Engineering (UBMK) (pp. 598-603). IEEE.

[2] https://www.kaggle.com/ntnu-testimon/paysim1

[3] https://www.kaggle.com/arjunjoshua/predicting-fraud-in-financial-payment-services

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[5] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.

[6] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

[7] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-9). IEEE.

[8] http://cs229.stanford.edu/proj2018/report/261.pdf

[9] E. A. Lopez-Rojas , A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016