Investigating Fraudulent Credit Card Activity

Theo Delettre (21262281), Albertas Ovodas (21268870), Paul McBrien (17715295), Rohan Bhatkande (20210678)
School of Computing, Dublin City University, Ireland
theo.delettre2@mail.dcu.ie, albertas.ovodas2@mail.dcu.ie, Paul.McBrien2@mail.dcu.ie, rohan.bhatkande2@mail.dcu.ie

***Abstract*—Credit cards are now being used more then cash all over the world and it is becoming easier to fall victim to fraudulent activity. in turn not only do the users suffer financially but also big banking companies. The detection of fraud is a huge problem and being worked diligently for permanent solutions. A synthetic dataset allows for exploratory analysis without GDPR interaction. Looking for the optimal solution to maximize fraud detection the methods incorporated are logistic regression, random forest and XGBoost models. The integration of principle component analysis within the models further accelerates the model run time however, does not necessarily increase f1-score. The most effective model was found to be logistic regression with a f1-score of 92.8 percent without PCA, in this case PCA is not worth implementing since the run efficiency does not justify the loss of performance. If however this model was implemented to a real time data feed PCA would almost be a necessity. Ultimately by exploring the different classification models a suitable system was created to tackle the ever growing problem of credit card fraud.**

***Index Terms*—Fraud Detection, GDPR, Logistic Regression, Random Forest, XGBoost, Principle Component analysis**

Introduction

In a more paperless world, credit cards are the predominant way of purchasing and interacting with financial products. Customers are more likely to be declined the use of cash than credit cards. The swiftness and ease of use have created an enormous amount of transactions showing 2.14 billion transactions equivalent to 7.58 trillion euros in the year 2020 in Ireland alone, with an increase of 52 per cent since 2019 [1]. Even at a tiny percentage, fraudulent activity will lose people and banks a hefty amount of currency. With the likes of stolen or lost credit card fraud and the upcoming cryptocurrency card scams, fraud is at an all-time high. Manual intervention would never succeed due to the sheer volume of transactions; hence a sophisticated and powerful algorithm is needed to detect fraudulent activity. Investigating the different methods to optimise accuracy and f1-score is the direction purpose of the paper. The problem, however, is not that simple since there are many forms of fraud, some of which include stolen/loss, laundering and abuse (no intention of paying back and claiming bankruptcy). However, within our dataset, the central basis of fraud was detected when funds were being transferred out or taken out.

Related Work

The primary focus for banks and other large money organizations is to minimize fraud and maximize automation. This creates a common problem for banks and is the main target for risk control as it loses banks a huge amount of currency every year. Detecting using transaction data and speeds is the simplest way of looking for these outliers [2].

One strain tree models are not sufficient; hence, an ensemble of classification trees, also known as random forests, is used to detect fraud. Primarily influenced by Amit and Geman in 1997 on a geometric selection of random methods and the random split selection approach by Dietterich in 2000, Breiman developed the random forest model [3]. The model works exceptionally well as it generates a massive amount of uncorrelated trees and, in turn, tries to outclass the previous generation of trees to generate the best results.

To further build upon the tree classification models xgboost is a sufficient and and common method used to scale and further enhance upon the tree system. Tianqi Chen developed the model in 2014. It uses gradient boosting within the tree models to generate state of the art results in most cases. [4]

Logistic regression can be used for fraud detection since there are two primary variables either it is a fraud or it is not. Developed by Joseph Berkson in 1944 and since has been one of if not the most popular model for classification problems [5]. This model uses statistics and, based on probability, calculates if the transaction is fraud or not based on the regression curve.

Principle component analysis was invented in 1901 by Karl Pearson but later developed and implemented by Harold Hotelling in the 1930s. It uses matrices transformation to reduce the dimensions of datasets.[6] By creating new variables that maximize variance, it can summarize datasets with a large number of features and is particularly useful when trying to maximize the accuracy or f score of models; hence it will, in turn, be used for most models in this paper.

As further mentioned in the 'dataset and exploratory' analysis taken from a synthetically generated state, the dataset chosen overcomes the GDPR and principles, making it easier to work on the data. However, working with sensitive information can lead to unintentional personal leaks and cause a severe backlash; furthermore, if the dataset was not synthetic, the data protection acts of 1988 and 2003 take effect in particular in section 2(1) part 7, where an appropriate security and data protection team must be present when dealing with unlawful forms of processing. In turn, this creates an extra workforce to tackle this problem, e.g. (the presence of a data protection officer). Hence by eliminating working with real-life data, the work is simplified and can be engaged sooner and with less stress. [7]

Dataset and Exploratory Analysis

Credit card transaction data is intrinsically confidential. No credit card provider is willing to release customers' information online publicly. Therefore, the only kind of real-world data available has been pre-processed to anonymize it. This is the case, for instance, with [8], which applies PCA to transform nearly all its features. The resulting dataset completely loses its interpretability. Additionally, it ruins the point of the exercise by skipping a lot of the feature engineering and pre-processing steps. Therefore, we decided to forgo a real-world example to use more significant synthesized data.

The Paytm Financial simulator generates synthetic data that is similar to real-world data by using agent-based simulation [9]. The dataset we are using in this paper was generated by PaySim using logs from a mobile money service implemented in an African country as input for the simulation.

The dataset contains information on 6.4 million transactions. The features include the type of transaction, the ID of origin and destination accounts, their corresponding balances before and after the transaction, the transaction amount, the isFraud label for fraudulent transactions, and isFlaggedFraud, which is a simple heuristic for flagging illegal attempts. We discard the ID features as they will not be used in our models. In addition, after investigation, isFlaggedFraud only has a value of 1 for transfers of amount>200k. Since this is something our model should detect, we also discard this feature.

TABLE 1
Number of records per type. Only includes types which include fraudulent transactions.

Transaction Type	Count Fraud	Count Not Fraud
CASH_OUT	4116	2233384
TRANSFER	4097	528812
Total	8213	2762196

The first thing to notice is that fraud is only present in transactions of types cash withdrawal and cash transfer, as seen in table 1. This means that we can remove all transactions of other types as they are irrelevant to our models. We also dummy encode the type feature to be an isTransfer feature.

The resulting dataset consists of 8.2k fraudulent transactions compared to 2.8M non-fraudulent ones. Therefore, we need to consider the imbalance of the dataset in the following steps.

Pre Processing

Multicollinearity

When a predictor variable in the dataset can be predicted linearly using other predictor variables, there is multicollinearity in the dataset. While this work aims to predict credit card fraud, the process of understanding what influence the predictor variables have on the outcome variables is susceptible to multicollinearity [10]. By plotting the correlation of each of the predictor variables in a 2D grid, the collinearity between all variables can be visualised. Relatively large collinearity values between distinct variables indicate that those variables may have colinearity between them. In this case, it may be helpful to remove these variables.

While these correlation values are good indications of multicollinearity, a more reliable measure is found by using a quantity known as the variance inflation factor (VIF), defined in Equation 1.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{1}$$

where R_i^2 is the coefficient of determination for a linear regression of the other variables (and excluding the variable in question). Typically, if the VIF is larger than ten then it is strongly advisable to drop variables from the

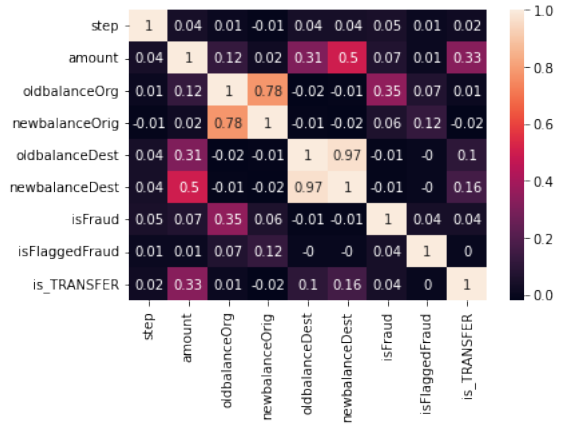

Fig. 1. The correlation values for 'oldbalanceOrig' and 'newbalanceOrig' are relatively high, and the 'oldbalanceDest' and 'newbalanceDest' are very high.

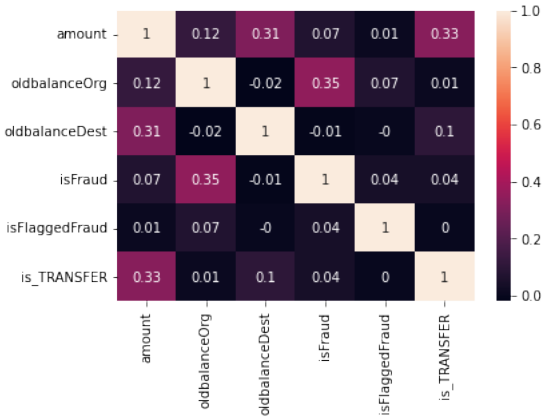

Fig. 2. The variables for the new balances have been dropped.

model. If they are between 5 and 10, it is still important to check for multicollinearity. The VIF for each of the dataset variables is calculated and displayed in Table 2.

TABLE 2
THE VIF VALUES FOR THE ORIGINAL DATASET. THE VALUES EXCEEDING 10 INDICATE STRONG MULTICOLLINEARITY.

Variable	VIF Value
step	1.303365
amount	5.321682
oldbalanceOrg	2.783066
newbalanceOrig	2.631296
oldbalanceDest	65.281776
newbalanceDest	80.740306
is_TRANSFER	1.338878

Removing multicollinearity from the dataset will ensure orthogonal input features for any models used for predictions. This also prevents sensitivities that might occur for small changes in input data. In order to ensure robustness in a real-world scenario, it is crucial to mitigate this. It will also allow the most critical variables that affect fraud to be identified correctly. For example, the 'oldbalanceDest' and 'newbalanceDest' values well exceed the value of ten. The correlation between these variables can be intuited because the amount variable summed with the old balances provides the destination balances. Therefore, the transaction information is fully preserved by excluding the new balances. In Figure 2, the new balances show a reduction in the correlation values between predictor variables, and this is clear from the VIF values in Table 3.

TABLE 3
THE VIF VALUES AFTER REMOVING THE NEW BALANCES ARE WELL BELOW THE THRESHOLD OF FIVE FOR CONSIDERING DROPPING FURTHER VARIABLES.

Variable	VIF Value
amount	1.397963
oldbalanceOrg	1.032239
oldbalanceDest	1.195571
is_TRANSFER	1.230799

Undersampling

Since fraudulent transactions make up a tiny fraction of overall transactions, a randomly chosen training split will have approximately the same proportion of classes as the original dataset. In this case, the classes are represented by the binary variable 'fraud' or 'not fraud'. It is found that over 99.8% of the transactions in the dataset are 'not fraud'. By trivially predicting that all transactions will not be fraudulent, it is possible to achieve an accuracy of 99.8%. However, using the F1 score as the evaluation metric, such a trivial model would achieve a score of zero. Any models applied to the unbalanced dataset are likely to learn the tendency of each output to be non-fraudulent rather than learning how the feature values can contribute to a fraudulent transaction [11].

For this reason, class balancing can be used to make the proportion of classes more even in the dataset. To achieve this, the dataset is balanced using undersampling. This is a method where the class with the most number of samples (in this case, not fraud) is randomly sampled until a sample set is obtained with the same number of samples as the class in the dataset with the least number of samples. For this dataset, the total number of samples is 6362620. Of those, only 8213 are fraudulent. So by

only choosing 8213 non-fraudulent transactions, the best case accuracy for the trivial model will be 50%.

Principal Component Analysis

Principal component analysis (PCA) is a form of factor analysis that can be used to manufacture variables that are not directly observable from the input data. In summary, it is a method that allows new features to be generated using already existing features [12]. It is advantageous because it can reduce the dimensionality of the dataset while preserving important information in the features. It also ensures that newly generated features are orthogonal. This is achieved using a combination of linear algebra methods [13]. Using PCA will reduce the amount of information available in the dataset, but reduce the amount of features. It is only appropriate for numerical data, and so the categorical variable 'is_TRANSFER' is removed before applying PCA. For this work, it may also have the added effect of obscuring sensitive data about bank customers. When applying these models in the real world, using PCA can create data that does not contain information about customers and their financial data. This can be used, at least in contribution, to handle data to respect the privacy of banking customers.

METHODS

Logistic Regression

Logistic regression allows us to perform a binary classification based on continuous data inputs. The model classifies input data according to its value on a logistic function. This logistic function is created by trying to reduce error rates in predicted values through maximum-likelihood estimation [14].

Random Forest

Random Forest Algorithm is a supervised learning model. It is usually trained on labelled data (independent variables) to classify the unlabelled data (dependent variable). Random Forest is used in both regressions as well as classification. This algorithm uses the final decision of each node to come to its conclusion, i.e. it looks through the results of multiple decision trees and takes an average from it. Also, in the classification, it uses gini index, which uses class and probability to determine the gini of each node, which in turn shows which branches will occur. It is also possible to use entropy that uses probability to decide how the nodes should branch[15].

XGBoost

XGBoost is a state of art machine learning algorithm which is a standard form of gradient boosting; a gradient boosted decision tree. It learns from the residual error rather than updating the weights of data points. One advantage of XGBoost is that it handles missing data internally. In addition, XGBoost is more optimized and enhanced when compared to Gradient Boosting Machine[16].

RESULTS

A comparative analysis was performed using Logistic Regression, Random Forest and XGBoost Algorithm on normal data and data on which Principle Component Analysis was performed. The results of the same is represented in Table 4.

For the results that included under-sampling, a total of ten under-sampled data subsets were found and the F1 score was calculated for each subset. These scores were then averaged to produce the values in Table 4.

As expected, the use of undersampling is a necessity. Without it, the models are biased towards classifying values as not fraud, leading to a high accuracy, but terrible F1 score.

The use of PCA improves the speed of the model however the trade off in our case is not worth it. The run time of our model is relatively fast and transactions do not suffer from high-dimensionality. For our usage the 6 percent loss in F1 score cannot be justified. However with a live data feed of millions of transactions PCA would most likely be necessary for the model to function efficiently.

With the differences in Training and Testing classification scores being negligible, we can confidently say our models aren't overfitted in reference to the training data.

In the case of credit card fraud, we are generally aiming for a high recall score. False Positives, preventing a legal transaction that we identify as fraud, are a problem but one that can be resolved by contacting the client. On the other hand False Negatives, allowing fraudulent transactions to go through, is a much bigger issue that is costlier to resolve. Our logistic regression model, with a recall score of 0.9612, manages to identify 96% of all fraudulent transactions.

If the bank decides the personnel cost of resolving False Positives is higher than the simple monetary cost of reimbursing a fraudulent transactions, we would be aiming instead for a high precision score. The same

TABLE 4
RESPECTIVE F1 SCORE FOR TRAIN & TEST IN EACH MODEL. FOR RF AND XGBOOST, THE RESULTS FOR THE UNDERSAMPLED

Model	Training F1 Score	Testing F1 Score
Logistic Regression (no undersampling | no PCA)	0.0006	0.0004
Logistic Regression (undersampling | no PCA)	0.9267	0.9253
Logistic Regression (undersampling | PCA)	0.8680	0.8620
Random Forest (no PCA)	0.8742	0.8725
Random Forest (PCA)	0.8235	0.8226
XGBoost (no PCA)	0.8821	0.8813
XGBoost (PCA)	0.8432	0.8420

logistic regression model holds a precision score of 0.8919.

Combining both of these scores into a single metric, our best logistic regression model has an F1 score of 0.9281.

CONCLUSION

Ours is one of many ways to approach the challenge of fraud detection. Other approaches may focus on analysing transaction times to understand time patterns present in fraud. Suspicious IDs could also be identified by analysing their transaction patterns before they even commit fraud.

For our approach which considers transactions independent of time or account IDs, we were able to evaluate the performance of multiple models and pre-processing methods. We found undersampled Logistic Regression with no PCA gave us the best results for identifying fraud.

Future work for this paper could involve creating a model that can intake any dataset including real time and calculate its accuracy and f1-score in comparison to our synthetic dataset. Furthermore if there was a more varied selection of data, new feature engineering opportunities could develop, since all datasets will have different features and corresponding approaches.

REFERENCES

[1] Central Bank Statistics. payment statistics ireland. https://www.centralbank.ie/statistics/data-and-analysis/payments-services-statistics, journal=Centralbank.ie, year=2022.

[2] Wen-Fang Yu and Na Wang. Research on credit card fraud detection model based on distance sum. In *2009 International Joint Conference on Artificial Intelligence*, pages 353–356, 2009.

[3] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[4] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016.

[5] Jan Salomon Cramer. The origins of logistic regression. 2002.

[6] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[7] Intersoft Consulting. Art. 9 gdpr – processing of special categories of personal data - general data protection regulation (gdpr), 2022.

[8] Machine Learning Group ULB. Credit card fraud detection, Mar 2018.

[9] Edgar Lopez-Rojas, Ahmad Elmir, and Stefan Axelsson. Paysim: A financial mobile money simulator for fraud detection. In *28th European Modeling and Simulation Symposium, EMSS, Larnaca*, pages 249–255. Dime University of Genoa, 2016.

[10] Jamal I. Daoud. Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, 949:012009, 2017.

[11] Julián Luengo, Alberto Fernández, Salvador García, and Francisco Herrera. Addressing data complexity for imbalanced data sets: Analysis of smote-based oversampling and evolutionary undersampling. *Soft Computing*, 15(10):1909–1936, 2010.

[12] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[13] Benjamin W. Wah and Tao Wang. *Journal of Global Optimization*, 14(1):1–25, 1999.

[14] Y. Sahin and E. Duman. Detecting credit card fraud by ann and logistic regression. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 315–319, 2011.

[15] A. S. More and Dipti P. Rana. Review of random forest classification techniques to resolve data imbalance. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, pages 72–78, 2017.

[16] Asha Latha. xgboost as regressor and classifier. https://blog.ineuron.ai/XGBoost-as-Regressor-and-Classifier-4IVqLERTn8, 2021. Accessed: 2021-03-07.