Fraud Detection in a Financial Payment System

Dushani Perera[1], Manisha Rajaratne[1], Damitha Sandaruwan[1] and Nihal Kodikara[1]

[1] University of Colombo School of Computing, Colombo, SRI LANKA
{2015cs102.stu, 2015cs107.stu, dsr, ndk}@ucsc.cmb.ac.lk

Abstract. Many financial payment systems have to face fraudulent activities due to the fast-paced development of the technology. Fraud detection is essential for the proper management of fraud control. It automates the manual checking processes and helps the detection be done conveniently. It is important to research and find ways and means of proper methodologies which will help serve the purpose of fraud detection effectively. Machine Learning Approach becomes more popular and accurate compared to a rule-based approach in this scenario. This paper presents such a performance comparison among a few methods which were tested with a dataset.

Keywords: Fraud Detection. Financial Systems. Machine Learning

1 Introduction

Fraud means maltreatment of a framework of a profit making organisation without fundamentally prompting direct lawful outcomes. In an aggressive domain, fraud detection can turn into a business basic issue in the event that it is pervasive and if the anticipation methodologies are not safeguard. Fraud detection is a part of the overall fraud control mechanism. It mechanizes and decreases the manual pieces of a screening/checking process [2].

With the rising ascent of innovation today, the reliance on web-based business has developed exponentially. As the credit card gives accommodation to the clients yet frauds caused because of these exercises causes bother. The credit card data is private, the bank and the other budgetary undertakings wouldn't like to reveal the data about their clients. The temporary misfortune emerges due to bank loans the cash to clients who in the long run don't have the ability to pay back [3].

2 Data Set Description

The chosen dataset for the task is ‘Synthetic data from a financial payment system’. This dataset is generated using BankSim which is an agent-based simulator of bank payments based on a transactional data provided by a bank in Spain. This synthetically generated dataset consists of payments from various customers made in

different time periods and with different amounts. The source of this dataset is Kaggle [1]. In this dataset, the target variable consists of two labels; 1 for fraudulent transactions and 0 for normal transactions. Thus, the methodology used is Supervised learning. In this dataset the target variable is Fraud. By exploring the dataset we can find out which variables have an effect on the target variable and the correlation between those variables.

Category vs Transaction Class; the attribute category has 15 categories. As visible in Fig. 1, most of the fraudulent payments are done in the health category. None of the fraud payments is done in transportation, food or contents categories. So we can say that there is a relation between the categories and fraudulent transactions. As shown in Fig. 2, most of the purchases done for transportation when it comes to normal transactions. There is a significant difference between the purchases done in fraudulent transactions and normal transactions.

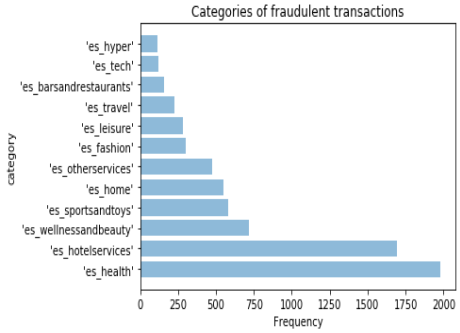

Fig. 1. Categories of fraudulent transactions.

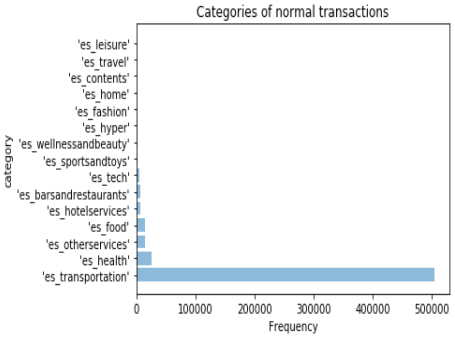

Fig. 2. Categories of normal transactions.

Amount vs Transaction Class; experimented on the amount attribute, to check whether it has a correlation between transaction class. As shown in Fig. 3, the minimum transacted amount when fraud is set is $0.03. The maximum transacted amount when fraud is set is $8329.96. When it comes to the non fraud transactions minimum is $0 and the maximum is $2144.86. Thus, the fraud amounts are higher compared to normal transaction amounts.

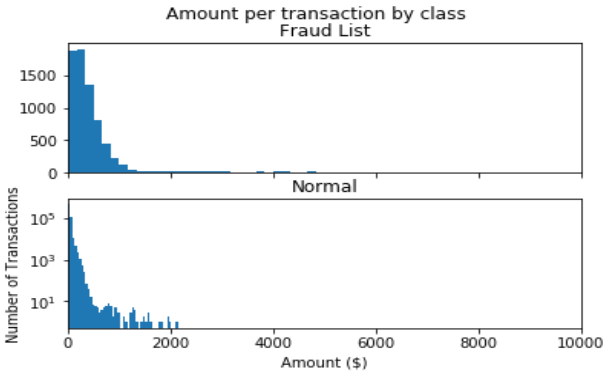

Fig. 3. Amount per transaction by class

Gender vs Transaction Class; only three gender types are involved in fraudulent transactions: ["'M'", "'F'", "'E'"] From them Males are more likely to involve in fraudulent transactions than other genders (Fig. 4 and Fig. 5).

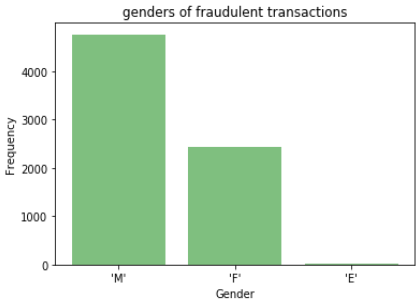

Fig. 4. Genders in fraud transactions

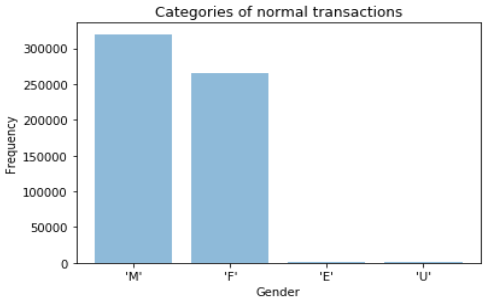

Fig. 5. Genders in non-fraud transactions

Age vs Transaction Class; as shown in Fig. 6 and Fig. 7, category 3 of age seemed more involved in fraudulent transactions. Category 3 means the people who are in 35-45 ages.

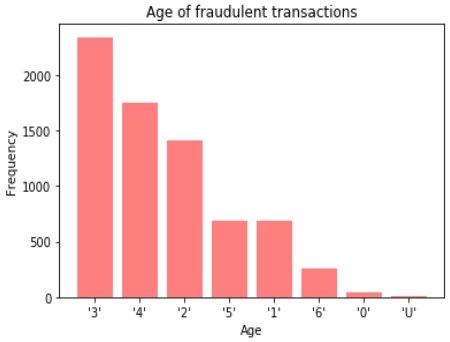

Fig. 6. Age in fraudulent transactions

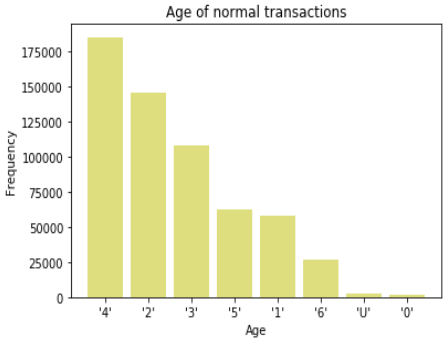

Fig. 7. Age in non-fraudulent transactions

Gender vs Amount; since the category, amount, age and gender have a correlation with the transaction class, then whether these variables have a correlation with each other was checked. As shown in Fig. 8 and Fig. 9, there are a considerable number of outliers in gender 'M' and 'F' categories and age. Fraud transactions are done with a large amount in gender Male and Female categories.

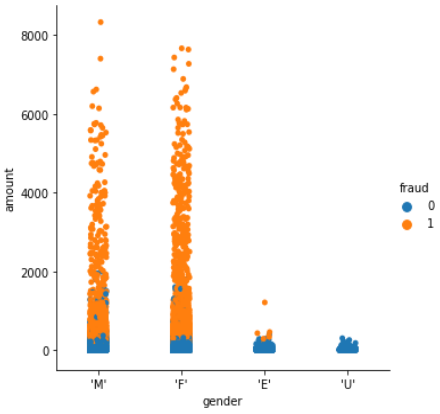

Fig. 8. Gender vs amount

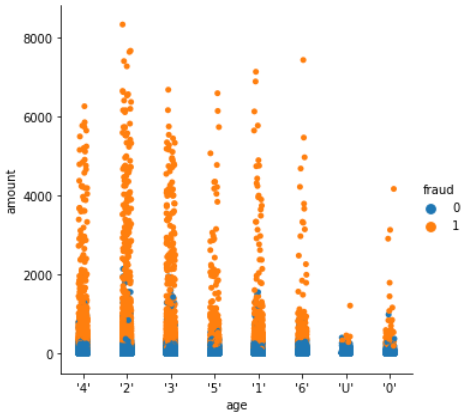

Fig. 9. Age vs amount

3 Planned Preprocess

First and foremost the dataset is checked for missing values. Then checked for unique entries in the input variables. If there is only one unique entry, those attributes are dropped. Then checked for categorical variables. These variables are transformed into numerical values. Oversampling the dataset with SMOTE (Synthetic Minority Oversampling Technique).

Fraud datasets are highly imbalanced datasets. Because of this it is difficult to get a better result using the original dataset. By using SMOTE, make copies of minority class and balance the dataset before splitting the dataset.

4 Candidate Machine Learning Models

There are several machine learning models that could be used:

a. Autoencoders - this is a data compression algorithm which takes the input and going through a compressed representation and gives the reconstructed output. It is a neural network that helps to identify anomalous data points in the dataset. Four layers were used for the neural network. First two for the encoder and last two for the decoder. The activation function used is 'Tanh' and as for the metric, Mean Squared Error (MSE) was used.

Model Evaluation; By using cross-validation for autoencoders, the model couldn't perform well. The loss it gave shown in the Fig. 10 and Fig. 11.

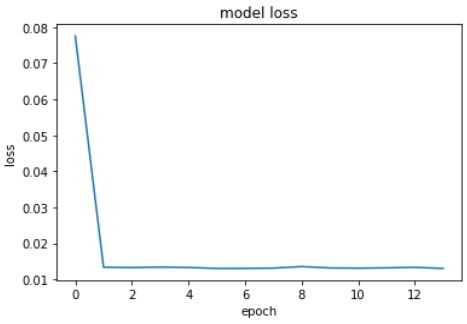

Fig. 10. The loss of the model

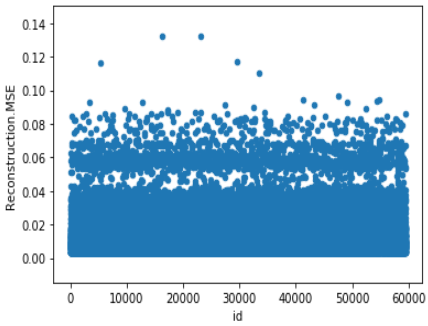

Fig. 11. Reconstruction error

Evaluation was done without using cross validation as well. First the data was split the dataset into two sets, 80% for the training set and 20% for the testing set. From the training set, only the normal transactions were sent to the model. The reason is because, by letting the model train for the nonfraud transaction, it will be able to learn the pattern of such transactions. The model was evaluated using the testing set. The testing set has both normal and fraudulent transactions in it. From this training method, The model will learn to identify the pattern of the input data. If an anomalous test point does not match the learned pattern, the autoencoder will likely have a high error rate in reconstructing this data, indicating anomalous data. So that we can identify the anomalies of the data. As shown in Fig. 12 and Fig. 13, reconstruction error is high in fraudulent data. As shown in Fig. 14, the ROC curve gave a 70% accuracy.

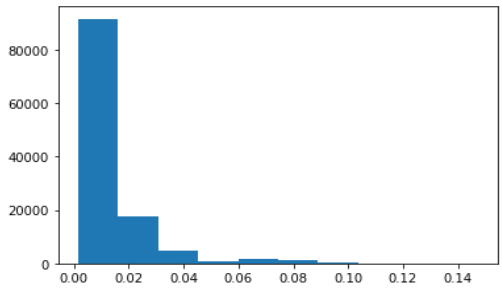

Fig. 12. Reconstruction error for normal transactions

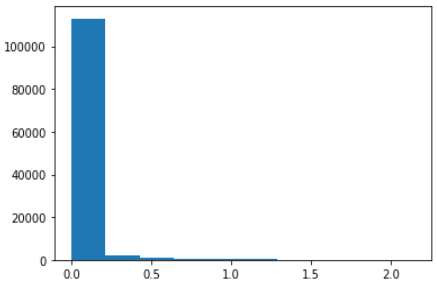

Fig. 13. Reconstruction Error; fraud transactions

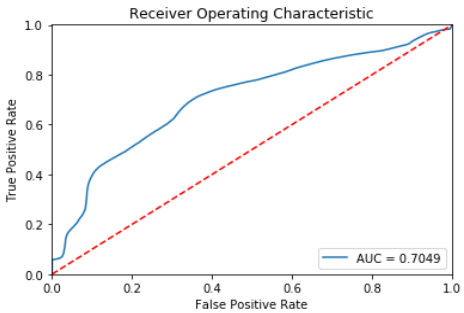

Fig. 14. ROC curve for Autoencoders

b. K-Nearest Neighbors - a supervised learning technique which assumes that similar things exist in close proximity. In other words, similar things are near to each other. It is a very good classifier.

Model Evaluation; 85% of accuracy for the model was obtained by using K-Fold Cross-validation. Generated the classification report and the confusion matrix to get a more detailed report of the model training. The model without cross validation gave a 99% accuracy for the model as shown in Fig. 15.

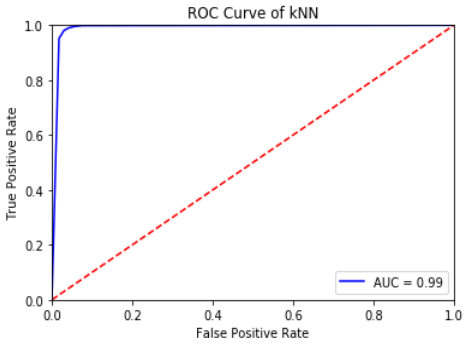

Fig. 15. ROC curve for KNN model

c. XGBoost - a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

Model Evaluation; 91% of accuracy for the model was obtained by using K-Fold Cross-validation. Without using cross-validation, the model gave a 97% of accuracy as shown in Fig. 16.

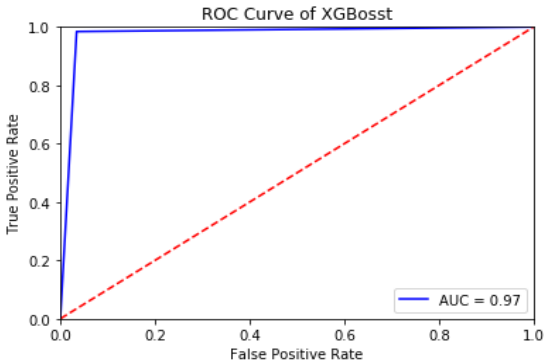

Fig. 16. ROC Curve for XGBoost

5 Conclusion

From the model evaluation results, it shows that the best performing model is XGBoost. A precise result wasn't received from autoencoders with the cross-validation technique. But without using cross-validation it gave a 70% accuracy. As for the KNN, it gave a 85% accuracy with cross-validation and 99% without using cross-validation. As for the XGBoost, it gave a 91% accuracy with cross-validation and 97% without using cross-validation. Even though KNN performed well without using cross-validation, XGBoost performed well with both of the techniques. This can be confirmed by looking at the confusion matrix (Table 1) of both techniques.

Table 1. Value Comparison

	Autoencoders	K-Nearest Neighbors
Using Cross-validation		0.8562
Without Cross-validation	0.7049	0.99

References

1. https://www.kaggle.com/ntnu-testimon/banksim1.
2. Jeong, Seong Hoon & Kim, Hana & Shin, Youngsang & Lee, Taejin & Kim, Huy Kang, A Survey of Fraud Detection Research based on Transaction Analysis and Data Mining Technique, Journal of the Korea Institute of Information Security and Cryptology, 25. 1525-1540. 10.13089/JKIISC.2015.25.6.1525, (2015).
3. Patidar, Raghavendra & Sharma, L., Credit card fraud detection using neural network, International Journal of Soft Computing and Engineering (IJSCE). 1. 32-38, (2011).