Credit card Fraud Detection

Name:	**Bhavana M R**
Registration No./Roll No.:	20335
Institute/University Name:	IISER Bhopal
Program/Stream:	Economic Science
Problem Release date:	August 17, 2023
Date of Submission:	November 17, 2023

1 Introduction

With technological developments, credit cards have become an integral part of our day-to-day life. A credit card comes with an option to buy anything without the constraints of paying the full amount at the moment. With this popularity of credit cards comes the other side, which is fraudulent credit card transactions. They involve stealing one's credit card information and making unauthorised transactions. The high frequency of such cases invokes the urgency to find a suitable mechanism to detect such fraudulent transactions. The objective of this project is to find the best model that can detect fraudulent transactions from a given dataset. The dataset contains features V1, V2, ... V28, which are PCA transformed; the features that are not transformed are time and amount. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction. There are 57116 entries in the training set, including 30 columns. The test set consists of 14280 entries. There are no null values in the dataset. The 'Class' is the response variable, which takes value 1 in case of fraud and 0 otherwise. The data is highly unbalanced, as there are only 142 fraudulent transactions out of the 57116 instances in the training data.(Figure 1). Since this is a classification problem, logistic regression is primarily performed. Other machine learning models such as SVM, Decision tree algorithm and Random Forest Classifier are also performed. The relevant features are selected through chi square method and the parameters are tuned by grid search cv. This is followed by a model evaluation in order to find out which is the best-performing model. The results show that the Random Forest algorithm is the best-performing model with an accuracy score of 99.89 per cent out of the rest three models.

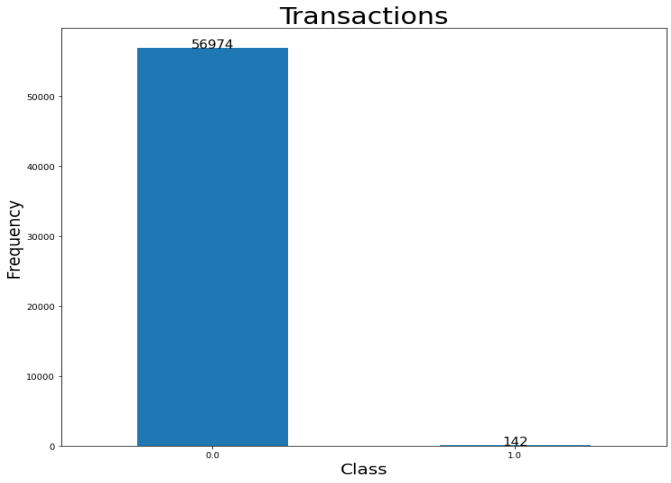

Figure 1: Overview of Data Set

2 Methods

I begin by analyzing and interpreting the data. The distribution of normal and fraudulent transactions is checked. A bargraph is plotted to represent the overview of the dataset. There is a high imbalance in the dataset, as there are only 147 fraudulent transactions. The percentage distribution of normal classes is 99.75 and the percentage distribution of fraud classes is 0.24. There are no missing values or null values in the dataset. I then use the individual statistics of each classes to understand the trends in the dataset. A correlation heatmap is also plotted to see the correlation between various features and to identify the important features.

The data is scaled using a MinMaxScaler. Then the data is split into training and testing data following 20 per cent criteria for the test and 80 per cent for the training set. I then define various classifiers such as logistic regression, SVM, Decision tree algorithm and Random forest classifier with proper grids and hyperparameters.[1] I used the Chi-square method for feature selection.[2] The hyperparameter tuning is done by using grid search cv.[3] The scoring used is the f1 score. The classification models are evaluated, and the classification results are printed in the end[4].

This is the link to my GitHub repository: `https://github.com/bhavanamr123/Credit-card-fraud-detecti`

3 Experimental Setup

I have used different evaluation criteria to evaluate the various classifiers, such as accuracy, precision, recall and F measure. The accuracy of a classifier is defined as the number of samples from the data set correctly predicted to the desired classes divided by the number of total samples in the data set.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

The effectiveness of a classifier in class assignment can be measured by the standard precision, recall and f-measure. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall is the ratio of correctly predicted positive observations to all observations in actual class

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The classification report is produced for each classifier. The confusion matrix is also printed along with each classifier. The hyperparameters are tuned using grid search cv. The size of cv=5 and the scoring method used is f1 macro.

4 Results and Discussion

The results of the various models which were evaluated are given in Table 1. The best-performing model is the Random Forest model with an accuracy of 99.89 per cent among the rest of the classifiers. The macro averaged precision score of random forest model is 0.92, recall score is 0.86 and the f1 score is 0.88. These scores indicate that random forest model is a pretty good model for this classification problem. The confusion matrix of all the models is also given in the Table 2.

Table 1: Performance Of Different Classifiers

Classifier	Accuracy	Precision	Recall	F-measure
Logistic Regression	0.9852	0.57	0.97	0.62
Support Vector Machine	0.9911	0.60	0.94	0.66
Decision Tree	0.9983	0.82	0.86	0.84
Random Forest	0.9989	0.92	0.86	0.88

Table 2: Confusion Matrices of Different Classifiers

Actual Class	Predicted Class	
	Normal	Fraud
Normal	11228	168
Fraud	1	27

Logistic Regression

Actual Class	Predicted Class	
	Normal	Fraud
Normal	11298	98
Fraud	3	25

SVM

Actual Class	Predicted Class	
	Normal	Fraud
Normal	11385	11
Fraud	8	20

Decision Tree

Actual Class	Predicted Class	
	Normal	Fraud
Normal	11392	4
Fraud	8	20

Random Forest

5 Conclusion

The results show that the Random Forest model is the best classifier for this dataset. The best parameters that we got from hyperparameter tuning are as follows: criterion: gini, maximum depth of the tree: 30, number of estimators:100. The limitation of the model that I have implemented is that it is not proficient in handling the imbalance in the dataset. An also the code assumes stationarity in the dataset, which might not hold in a real-world scenario. Future studies have great potential in this area as credit card fraud cases are reported very frequently these days. The future scope is to extend the model for real-time monitoring, as fraud patterns can change rapidly.

References

[1] Vaishnavi Nath Dornadula and S Geetha. Credit card fraud detection using machine learning algorithms. *Procedia Computer Science*, 165:631–641, 2019. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Andrea Dal Pozzolo, Olivier Caelen, Yann-Aël Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Syst. Appl.*, 41(10):4915–4928, 2014.

[4] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.