Credit Card Fraud Detection

Name:	**Upaddhye Rugved Milind**
Roll No.:	**21294**
Institute/University Name:	Indian Institute of Science Education & Research Bhopal
Program/Stream:	Electrical Engineering & Computer science
Problem Release date:	17 August 2023
Date of Submission:	17 September 2023

1 Introduction

Problem Statement: Given a set of real bank transactions made by users, the goal is to identify fraudulent transactions which have not been made by the users.
Tasks: Develop novel supervised machine learning algorithms to classify and predict fraudulent transactions. Identify significant features in order to do that. You may use different preprocessing techniques and feature extraction techniques to identify salient features. Show the experimental results of the method proposed (if any) and compare with the state of the arts. Subsequently, analyze the results and report significant findings and scopes of future works. Each row of the text file will contain the class label of an instance of the test data e.g., 0 following the order of the given test data.
About Project: It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Enormous Data is produced every day. The model used must be simple and fast enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible so that the scammers can be traced and caught.
About data: Total number of instances in Training data = 57,116

There are no missing values in the training data.

No. of classes = 2

0 = Legitimate Transactions = 56,974

1 = Fraudulent Transactions = 142

This is a classification problem because we have to classify if the transaction is Fraudulent or non-fraudulent. But Logistic Regression can also be used here, because it can classify the dataset into 0 and 1 labels.

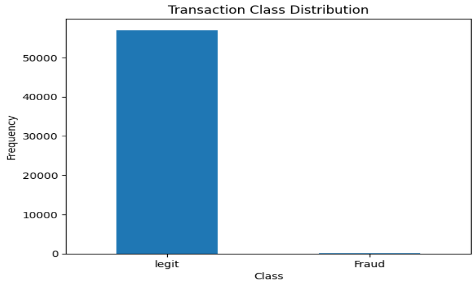

Figure 1: Overview of Data Set

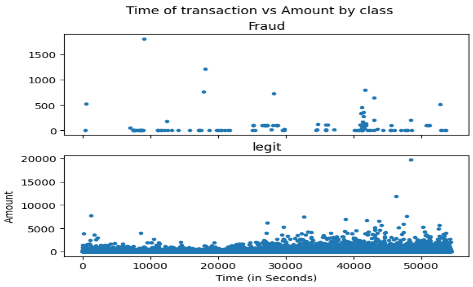

Figure 2: Time of transaction vs Amount by class

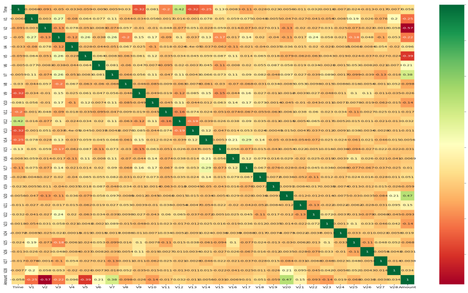

Figure 3: Correlation

2 Methods

As there are no missing values, we don't have to work on this matter. I also discovered the data dependencies like Transaction Class Distribution, Transaction Class Distribution(logarithmic), amount statistics, Comparing the values, Amount per transaction by class, Time of transaction vs Amount by class. Then I used some techniques of feature selection based on Variance, correlation, information gain. But no features were dropped through these techniques. The data is highly Imbalanced, if the model is trained on the same data, it gives results in the bias of legitimate transactions because legitimate transactions are much more than the Fraudulent transactions. I tried using SMOTE and Undersampling, but undersampling is giving better results than smote. I used models namely

Logistic regression

Support vector machine

Adaptive boosting(LR)

with smote and undersampled dataset with hyperparameter tuning. Among these, the combination of undersampled data with logistic regression was fitted best. Its scores were highest. Then I used hypertuning with gridsearch on it and then the scores increased more.

3 Experimental Setup

In this project I experimented with many techniques and models, and then finally I am submitting the most accurate method that I have found so far. During hyper-parameter tuning these are the best found parameters "Tuned Logistic Regression Parameters: 'C': 0.23357214690901212, 'max$_iter'$: 1000,$'penalty' :' l2',' solver' :' newton - cg'$"

Best score is 0.9577081000373274

4 Results and Discussion

From the tables above, we can say the logistic regression is the best fitted model for this data set. The given data was obtained from the PCS of original data, so it is difficult to correlate the classes with the given dataset. Moreover, the results can be summarised with some popular metrics such as

Table 1: Performance Of Different Classifiers Using All Features

Classifier	Precision	Recall	F-measure
Adaptive Boosting	0.99	0.94	0.96
Logistic Regression	0.93	1.00	0.96
Support Vector Machine	0.65	0.59	0.55

Table 2: Confusion Matrices of Different Classifiers

Actual	Predicted Class	
Class	0	1
0	140	8
1	2	134

Adaptive Boosting

Actual	Predicted Class	
Class	0	1
0	43	3
1	0	40

Logistic Regression

Actual	Predicted Class	
Class	0	1
0	28	21
1	15	22

SVM

Accuracy: The overall correctness of the model, is calculated as the ratio of correctly predicted transactions (both fraud and non-fraud) to the total number of transactions. Here, We have got accuracy of 0.97.

Precision: Precision is the ratio of true positive predictions to the total number of positive predictions. In the context of credit card fraud detection, precision represents the ability of the model to correctly identify fraudulent transactions without labeling too many legitimate transactions as fraudulent.

Precision= True Positives+False Positives True Positives Here, we have got precision of 0.93.

Recall (Sensitivity or True Positive Rate): Recall is the ratio of true positive predictions to the total number of actual positive instances. It measures the ability of the model to capture all instances of fraud.

Recall= True Positives+False Negatives True Positives

Here, we have got Recall of 0.93.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.

F1 Score= Precision+Recall 2×Precision×Recall

Here, WE have got the F1 score of 0.97

Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC): The ROC curve illustrates the trade-off between sensitivity (recall) and specificity (true negative rate). AUC-ROC measures the area under the ROC curve and is used to assess the model's ability to distinguish between fraud and non-fraud transactions.

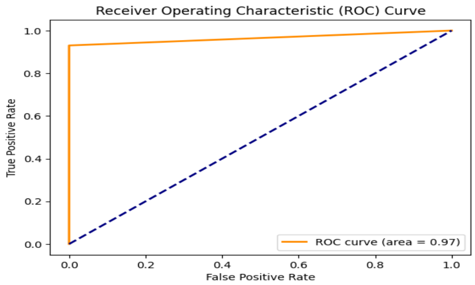

Figure 4: Receiver Operating Characteristic (ROC) Curve

This type of curve is very similar to the curve obtained by perfect classifier. So, we can conclude that our model is best fit fot this case.

Confusion Matrix: A confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. Confusion matrix for all models is mentioned above. When evaluating the credit card fraud detection model, it's important to consider a balance between precision and recall based on the specific requirements of the application. For example, in fraud detection, false negatives (missed fraud) are more critical than false positives (legitimate transactions mistakenly flagged as fraud).

5 Conclusion

The project revealed that the machine learning model achieved commendable performance in detecting credit card fraud, with a balanced trade-off between precision and recall. Key features contributing to the model's decision-making process were identified, shedding light on the characteristics of fraudulent transactions.

Future endeavors could focus on improving the model's robustness by addressing specific challenges such as data imbalance or exploring advanced ensemble methods. Additionally, incorporating real-time transaction data and leveraging anomaly detection techniques may enhance the model's responsiveness to emerging fraud patterns. Collaboration with domain experts and continuous model monitoring are essential for adapting to evolving fraud tactics and maintaining the model's effectiveness over time. GITHUB LINK : https://github.com/rugved-upaddhye/Credit-Card-Fraud-Detection-ML-Model

References

[1] Emmanuel Ileberi, Yanxia Sun Zenghui Wang : A machine learning based credit card fraud detection using algorithm