Project -2

Credit Card Fraud Detection in Financial Transaction

Name : ***kratika soni***

Batch – DST20823

Batch Date: 01/09/2023

Submitted to : Kevin shah

Problem Statement :

The challenge is to recognize fraudulent credit card transactions by the credit card companies so that customers are not charged for items that they did not purchase or being into a helplessness situation.

Content:

The dataset contains transactions made by credit cards in September 2013 by European cardholders.

This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.

Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.

The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning.

Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

Source: https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_3_GettingStarted/SimulatedDataset.html.

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (University Libre de Bruxelles) on big data mining and fraud detection.

A comma separated value file containing the dataset in tabular format –creditcard.csv.

Solution:

Main challenges involved in credit card fraud detection are:

a)Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.

b)Imbalanced Data i.e most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones

c)Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.

d)Adaptive techniques used against the model by the scammers.

How to tackle these challenges?

a)The model used must be simple and fast enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible.

b)To tackle Imbalance data I taken a data from dataset which contains a good number of fraud and Non Fraud transactions, so data can be used further.

c)We can make the model simple and interpretable so that when the scammer adapts to it with just some tweaks we can have a new model up.

d)We make few models and categorize best model for fraud detection whose accuracy level is higher and faster too.

Technology Stack:

a) Anaconda - Jupyter notebook
b) Excel for dataset

Libraries: numpy, pandas, matplotlib, scikitplot, sklearn.

Knowledge: Python and its libraries, Machine Learning concepts and various Model building.

Dealing with dataset:

The data contains 492 fraud transactions.

And, 284K valid transactions .

This makes the data highly imbalance because the fraud transactions are very less as compare to valid ones for detection

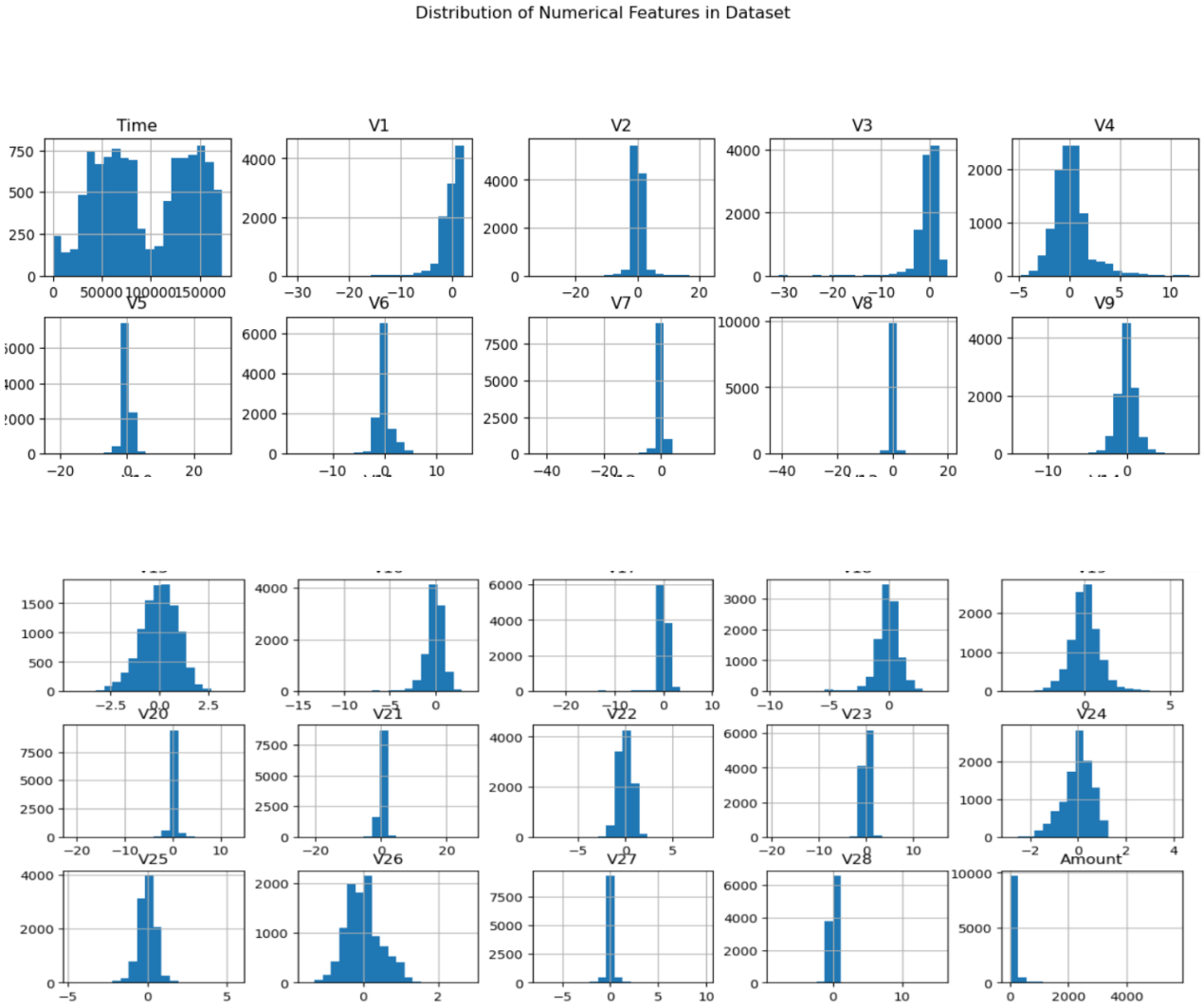
So, selected a sample data of fraud transactions size 469 and Valid transactions 10000

In the sampled data –

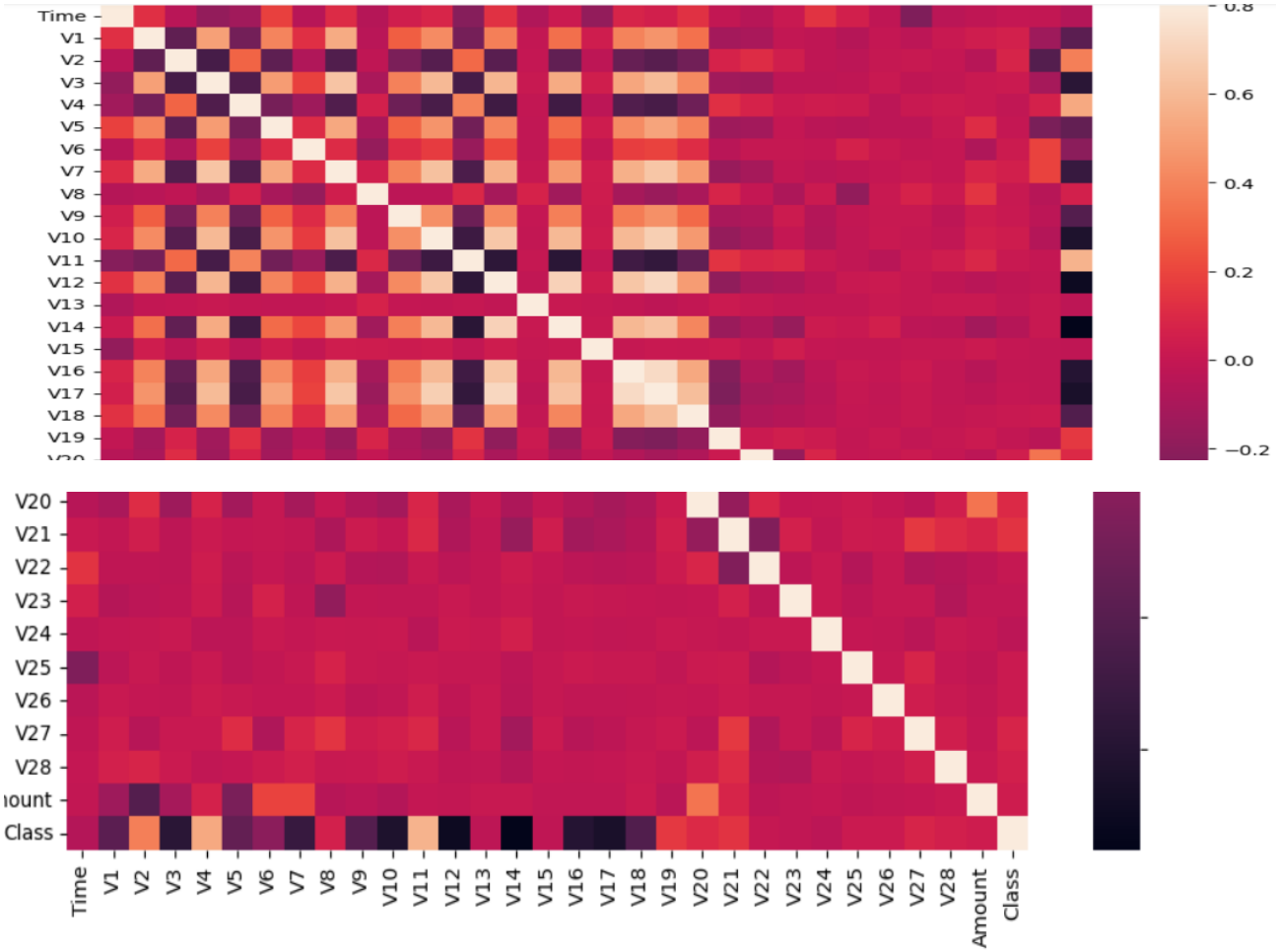
1)No null values

2)And, 19 duplicated transactions which I deleted.

Visualization:



This is a distribution of the numerical features from time to amount column of the dataset.



In this heatmap V20 and V24 features are correlating whereas V18 and V20 are negatively correlating with the amount ..Also some features are correlating and some are not either positively or negatively.

Splitting the data: XTrain, YTrain, XTest, YTest

Model Building and their evaluation or validation:

1)Decision tree model:

Result:

Accuracy: 0.9919971204382554

Precision: 0.9634615497224193

Recall: 0.7804924242424243

F1 Score: 0.8606806245424

2)Gradient Boosting Model

Result:

Accuracy: 0.9905154922762975

Precision: 0.8921101092291304

Recall: 0.8053030303030303

F1 Score: 0.844004335379414

3) *Logistic Regression Model*

Result:

Accuracy: 0.9908115447795263

Precision: 0.9236246716040932

Recall: 0.7805871212121213

F1 Score: 0.8425476969348331

4) *Random Forest Model*

Result:

Accuracy: 0.9930836867754763

Precision: 0.980536431310625

Recall: 0.7990530303030303

F1 Score: 0.8787176402188015

5)Naive Bayes Model

Result:

Accuracy: 0.9884400032775499

Precision: 0.8697564041446133

Recall: 0.7499053030303031

F1 Score: 0.8044362424321294

6) KNN Model

Result:

Accuracy: 0.9658170073627819

Precision: 0.392968142968143

Recall: 0.11448863636363635

F1 Score: 0.1750226547491974

Results:

From all the models Random forest model is faster as well as give high accuracy & precision.

So it is considered to be the best model for coping the challenge of credit card fraud detection.