# Credit card Fraud Detection

The main objective of this project is to to build a model on data set of credit card transaction to detect fraud transactions by carrying out certain processes on the dataset.

In this project necessary libraries like numpy, pandas, scikit learn,matplotlib,seaborn and imblearn are imported.

A csv file is read by using pandas .

The dataset has following features:

1) Account Number

2) CVV

3) Customer Age

4) Gender

5) Marital status

6) Card colour

7) Card Type

8) Domain 9) Amount

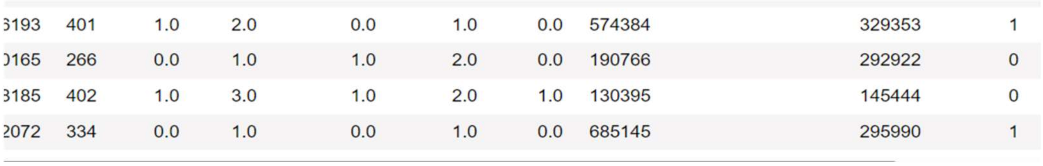
10) Average Income Expendicture

11) Outcome-containing values 0 for valid transaction and 1 for fraud transactions

12) Customer City Address

The dataset is checked for null values and it is found that the feature customer age has many missing values , so the column is dropped from the dataset and as computer can process numeric data more than string data, all values in dataset are converted into integer type using ordinal encoder module from scikit learn also dataset is checked for duplicate values and it is found out that there are no duplicate values.

The final modified dataset looks like this:



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5193 | 401 | 1.0 | 2.0 | 0.0 | 1.0 | 0.0 | 574384 | 329353 | 1 |
| 0165 | 266 | 0.0 | 1.0 | 1.0 | 2.0 | 0.0 | 190766 | 292922 | 0 |
| 3185 | 402 | 1.0 | 3.0 | 1.0 | 2.0 | 1.0 | 130395 | 145444 | 0 |
| 2072 | 334 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 685145 | 295990 | 1 |

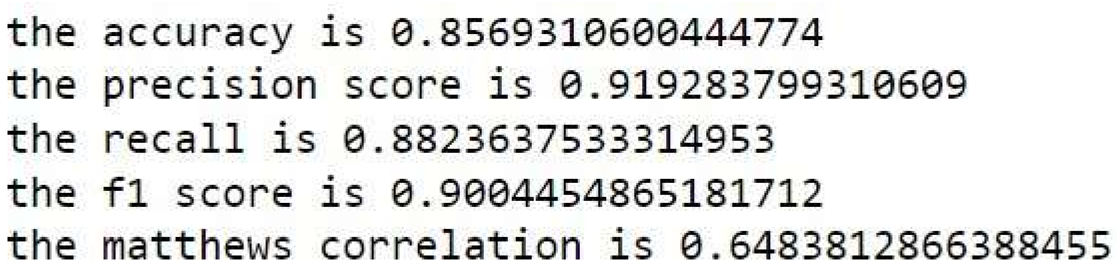
The records of 0 and 1 values from Outcome column are grouped and stored into variable valid and fraud.It is found that the number of fraud and valid cases are 27370 and 9727, Hence the dataset is highly imbalanced.

The dataset is divided into X and Y ,Y containing outcome column and X containing other columns so as to obtain input and output data. Using train_test_split module from scikit learn X and Y datastes are further grouped into xTrain,xTest and yTrain,yTest.

yTrain dataframe is balanced using under_sampling module from imblearn.

The training datasets are trained into models ,decision tree classifier,logistic regression and random forest classifier and performance of model is assessed using metrics like f1 score,accuracy,precission,recall score and matthews correlation coefficient.

It is found that random forest classifier has better performance scores than other model.

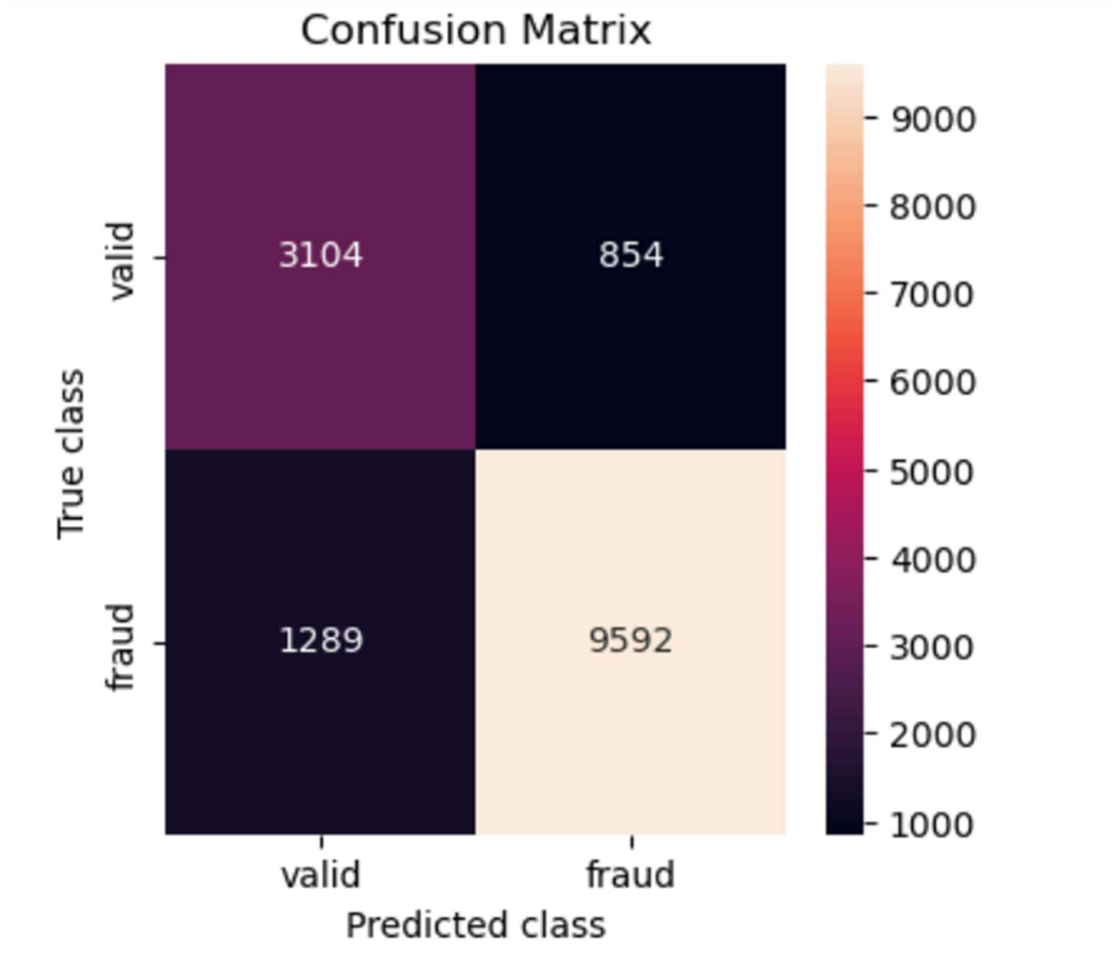


```
the accuracy is 0.8569310600444774
the precision score is 0.919283799310609
the recall is 0.8823637533314953
the f1 score is 0.9004454865181712
the matthews correlation is 0.6483812866388455
```

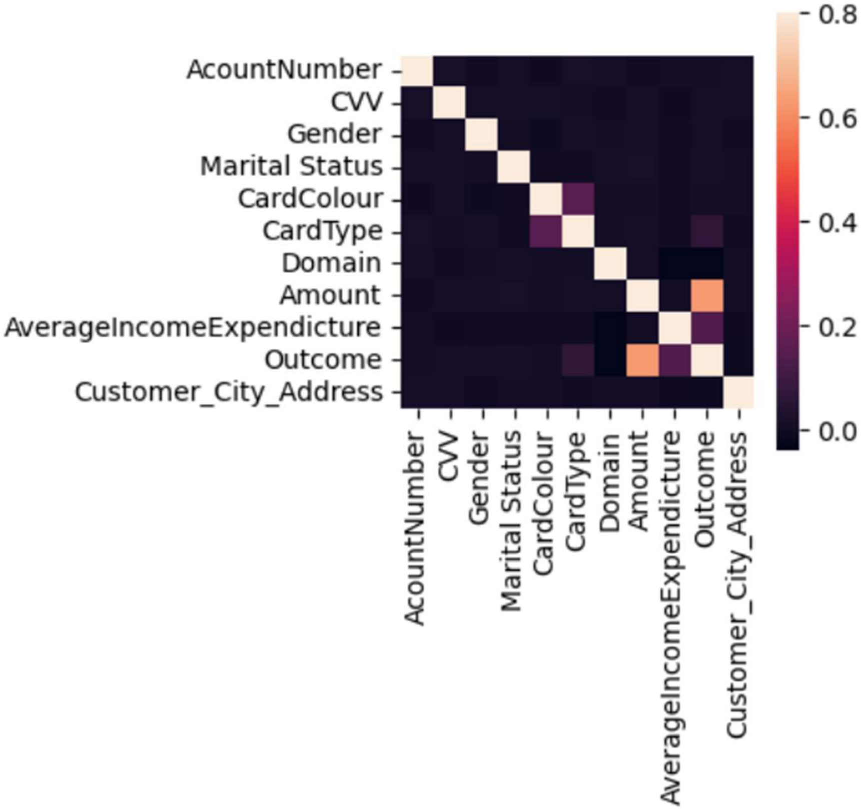
The dataset is further cross validated using kfold cross validation the model is trained 5 times by dividing training and testing datasets differently each time in same given ratios, in this case the given ratio is 60% and 40 %.

The new accuracy is found to be 85.93 %

Further by confusion matrix the number of correct predictions regarding fraud and valid transactions are found out.It is found that a pretty good amount of fraud transactions can be detected.



And by correlation matrix it is found that Amount and outcome are closely related ,other features aren't related to outcomes.


AcountNumber
CVV
Gender
Marital Status
CardColour
CardType
Domain
Amount
AverageIncomeExpendicture
Outcome
Customer_City_Address
0.8
0.6
0.4
0.2
0.0