CREDIT CARD FRAUD DETECTION REPORT

Introduction

The myriads of plastic cards in use worldwide are a gold mine for criminals, it is estimated that by 2027, there will be about $40 billion hit globally in credit card losses as compared to $27.85 billion in 2018. This significant increase is highly attributed to the rise of electronic transactions and card users in the world. Fraudulent methods are also getting more sophisticated and thus making it harder to spot by traditional fraud detection software and conventional methods.

It is of very high importance that more sophisticated approaches are employed and used in mitigating this massive challenge, high hopes have thus been placed on machine learning models to help credit card companies to recognize credit card transactions which are suspicious and report them to an analyst while letting normal transactions be automatically processed.

Problem Definition and Algorithm

The goal of this project is to build a machine learning model which detects credit card frauds. the tasks involved are the following:

1. Download and preprocess the credit card fraud detection dataset.

2. Perform Exploratory Data Analysis to understand the dataset.

3. Train a classifier that can determine if a card transaction is fraudulent.

Metric

The recall metric has been chosen as the metric of evaluation for the model, the recall metric answers the question: what proportion of actual Positives is correctly classified?

Recall = (TP)/(TP+FN)

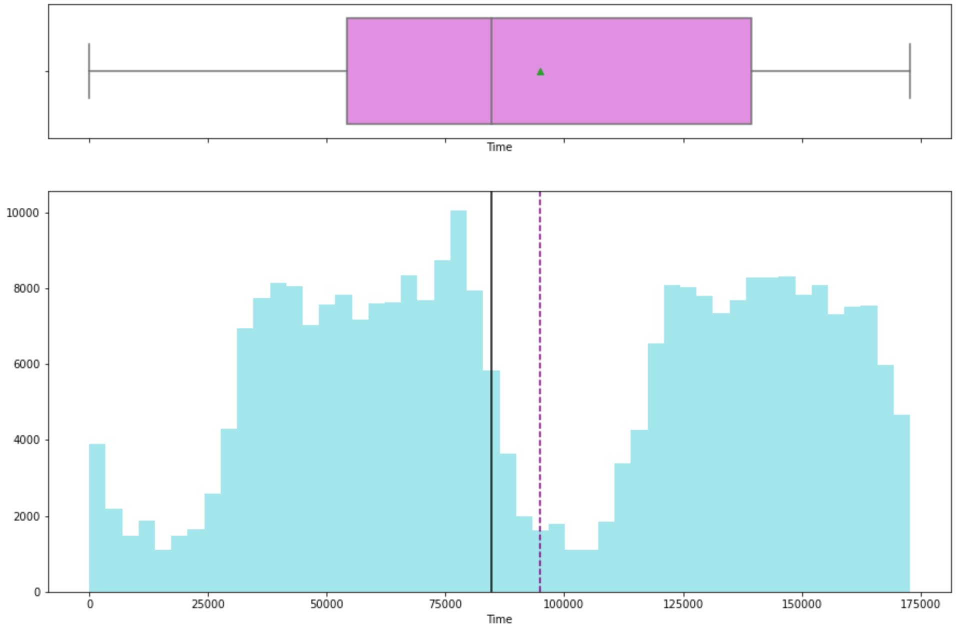
Recall is a valid choice of evaluation metric for the project since we want to capture as many positives as possible.

Dataset Description and Preprocesses

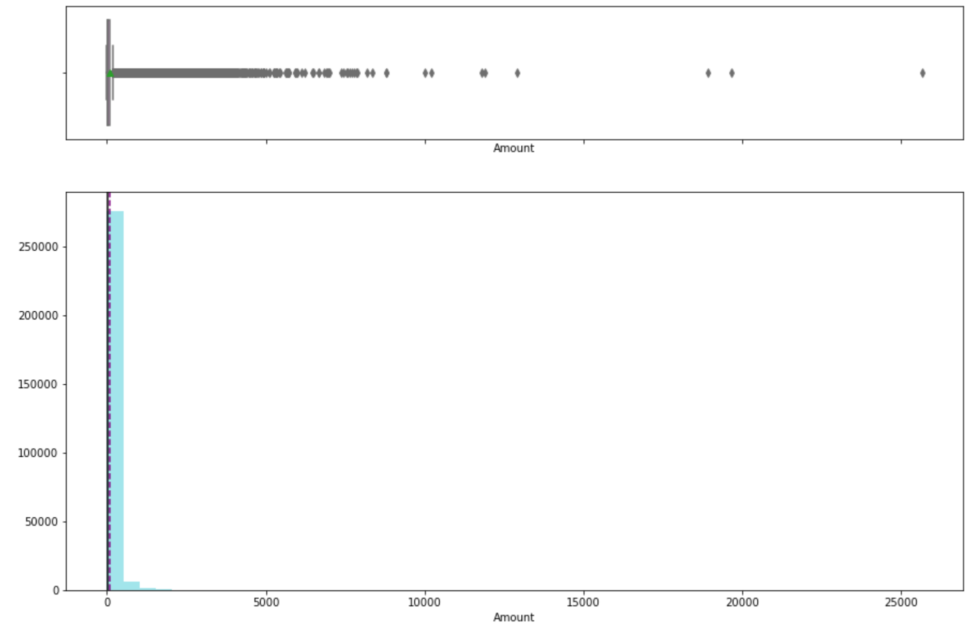
The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. 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.

Exploratory Data Analysis

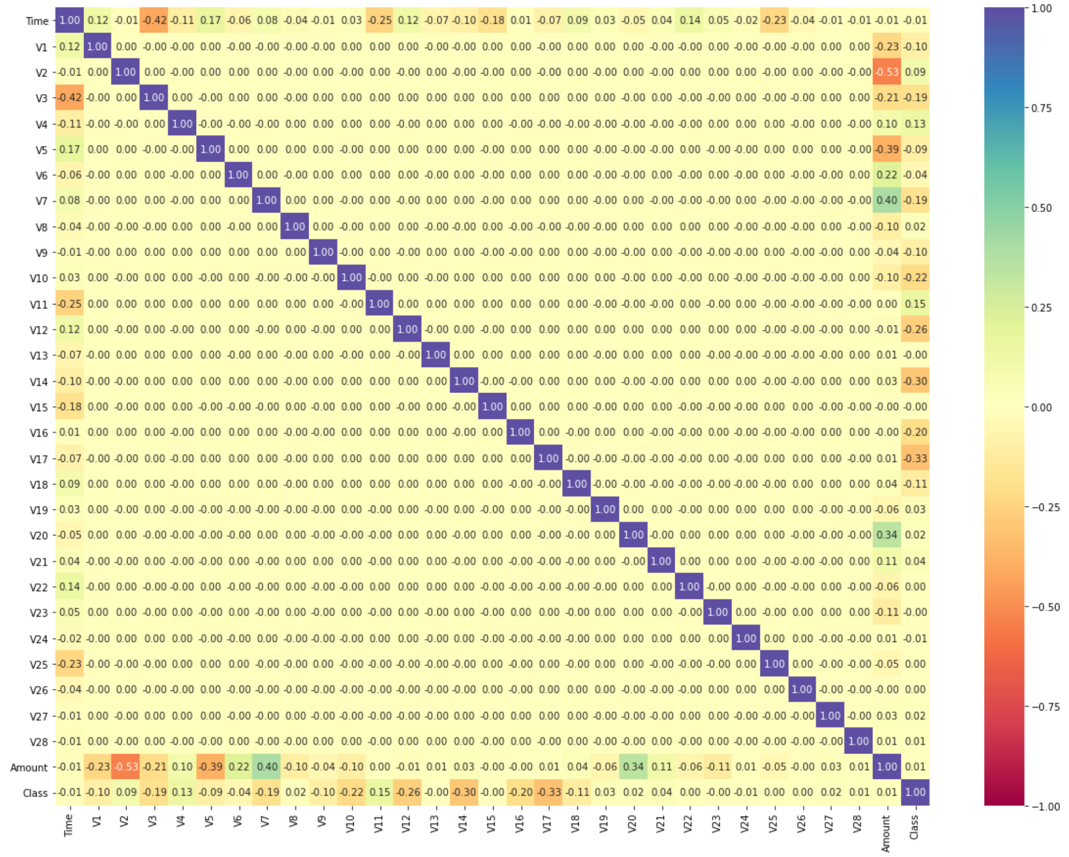
- **Observations on Time**: The box-plot and histogram charts below gives information on the univariate analysis of the time column, it shows that the time column is symmetrical at around 100,000 and at around 12,500.



- **Observations on Amount:** The box-plot and histogram charts below gives information on the univariate analysis of the amount column, it shows that the amount column has lots of outliers which are potentially fraudulent activities.

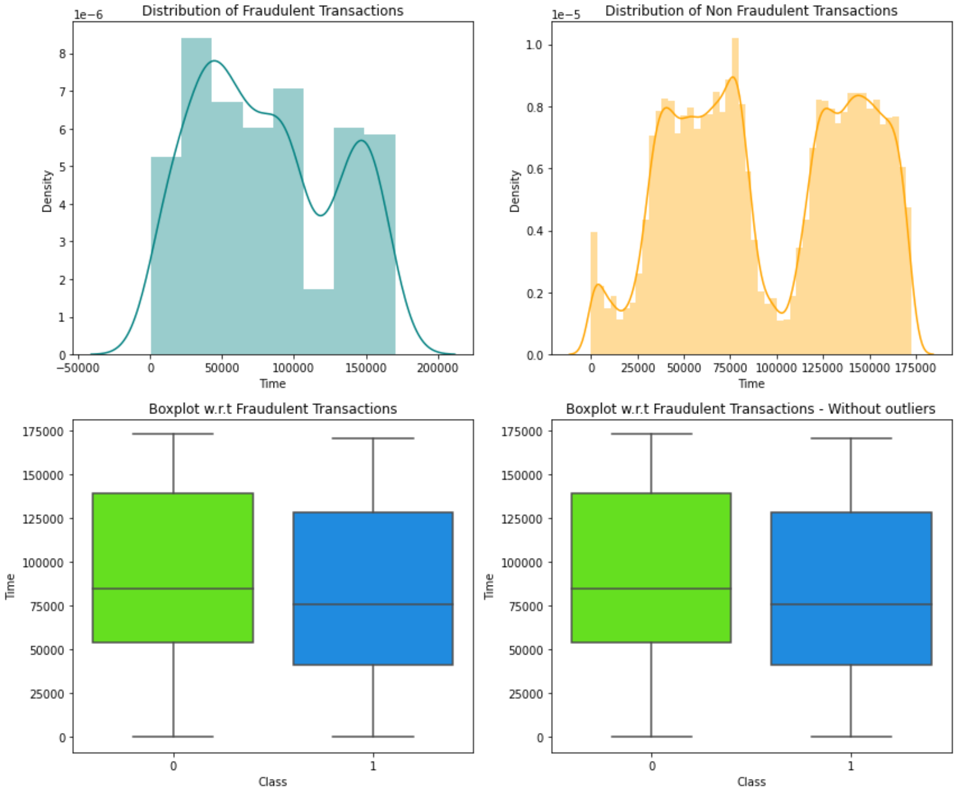


- **Correlation**: The features are not very correlated with one another.

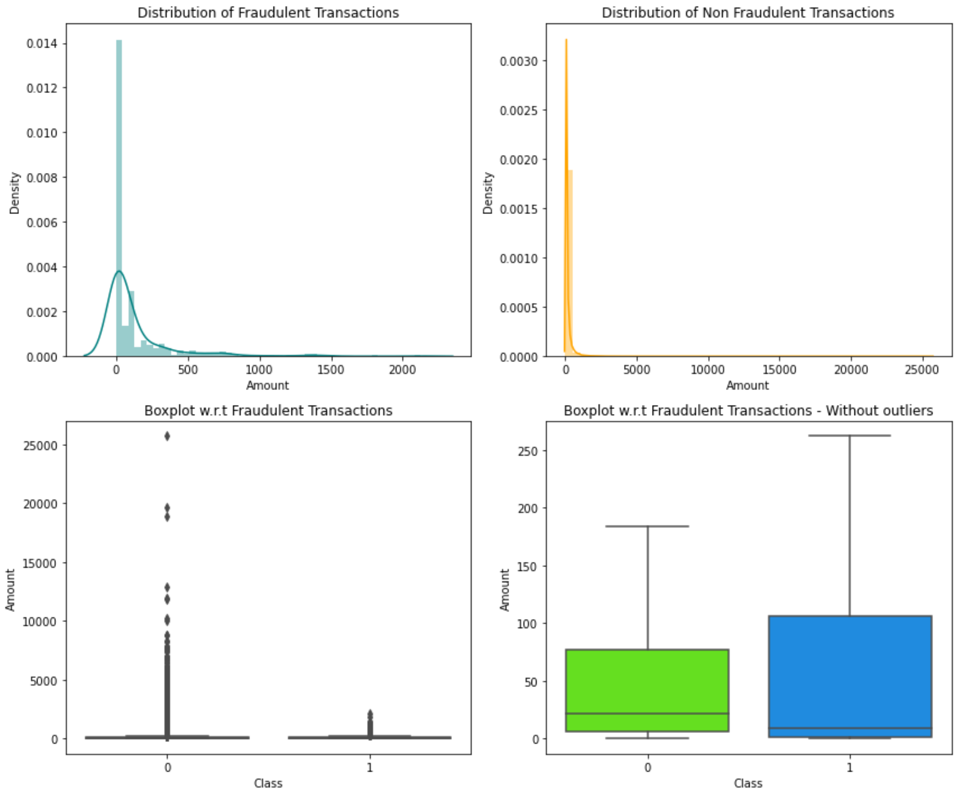


	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
Time	1.00	0.12	-0.01	-0.42	-0.11	0.17	-0.06	0.08	-0.04	-0.01	0.03	-0.25	0.12	-0.07	-0.10	-0.18	0.01	-0.07	0.09	0.03	-0.05	0.04	0.14	0.05	-0.02	-0.23	-0.04	-0.01	-0.01	-0.01	-0.01
V1	0.12	1.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.23	-0.10
V2	-0.01	0.00	1.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.53	0.09
V3	-0.42	-0.00	0.00	1.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	-0.21	-0.19
V4	-0.11	-0.00	-0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.10	0.13
V5	0.17	0.00	0.00	-0.00	-0.00	1.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.39	-0.09
V6	-0.06	-0.00	0.00	0.00	-0.00	0.00	1.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	0.22	-0.04
V7	0.08	-0.00	0.00	0.00	-0.00	0.00	0.00	1.00	0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.40	-0.19
V8	-0.04	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	1.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.10	0.02
V9	-0.01	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.04	-0.10
V10	0.03	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	1.00	-0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.10	-0.22
V11	-0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.15
V12	0.12	0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	1.00	-0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.01	-0.26
V13	-0.07	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	1.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.01	-0.00
V14	-0.10	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	1.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.03	-0.30
V15	-0.18	0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	1.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00
V16	0.01	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	1.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.20
V17	-0.07	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.01	-0.33
V18	0.09	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	1.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.04	-0.11
V19	0.03	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	1.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.06	0.03
V20	-0.05	0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	1.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.34	0.02
V21	0.04	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	1.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.11	0.04
V22	0.14	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	1.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.06	0.00
V23	0.05	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	1.00	0.00	-0.00	0.00	0.00	0.00	-0.11	-0.00
V24	-0.02	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	-0.00	-0.00	0.01	-0.01
V25	-0.23	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	1.00	0.00	-0.00	-0.00	-0.05	0.00
V26	-0.04	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00
V27	-0.01	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	1.00	-0.00	0.03	0.02
V28	-0.01	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	1.00	0.01	0.01
Amount	-0.01	-0.23	-0.53	-0.21	0.10	-0.39	0.22	0.40	-0.10	-0.04	-0.10	0.00	-0.01	0.01	0.03	-0.00	-0.00	0.01	0.04	-0.06	0.34	0.11	-0.06	-0.11	0.01	-0.05	-0.00	0.03	0.01	1.00	0.01
Class	-0.01	-0.10	0.09	-0.19	0.13	-0.09	-0.04	-0.19	0.02	-0.10	-0.22	0.15	-0.26	-0.00	-0.30	-0.00	-0.20	-0.33	-0.11	0.03	0.02	0.04	0.00	-0.00	-0.01	0.00	0.00	0.02	0.01	0.01	1.00

- **Distribution of time column**:



- **Distribution of amount column:**



Algorithms

The classifier was built using Logistic Regression and Decision Trees.

Classifier Training and Results

✓ Logistic Regression

The Logistic Regression classifier was first trained on the preprocessed training data, the following coefficients (coef) were obtained for each variable:

```
               coef     std err        z         P>|z|     [0.025      0.975]
------------------------------------------------------------------------------
const      -8.1197     0.286    -28.420     0.000     -8.680      -7.560
Time      -7.872e-06   2.75e-06     -2.862      0.004   -1.33e-05   -2.48e-06
V1          0.0929     0.051      1.837     0.066     -0.006       0.192
V2         -0.0250     0.066     -0.377     0.706     -0.155       0.105
V3         -0.0183     0.065     -0.279     0.780     -0.146       0.110
V4          0.6626     0.085      7.810     0.000      0.496       0.829
V5          0.1335     0.080      1.667     0.096     -0.024       0.291
V6         -0.1124     0.092     -1.221     0.222     -0.293       0.068
V7         -0.1056     0.077     -1.368     0.171     -0.257       0.046
V8         -0.1442     0.039     -3.742     0.000     -0.220      -0.069
V9         -0.5118     0.132     -3.879     0.000     -0.770      -0.253
V10         -0.8191     0.107     -7.661     0.000     -1.029      -0.610
V11         -0.1302     0.100     -1.307     0.191     -0.325       0.065
V12          0.2746     0.118      2.333     0.020      0.044       0.505
V13         -0.4406     0.105     -4.189     0.000     -0.647      -0.234
V14         -0.6528     0.081     -8.062     0.000     -0.812      -0.494
V15         -0.0603     0.103     -0.583     0.560     -0.263       0.142
V16         -0.1957     0.144     -1.358     0.175     -0.478       0.087
V17         -0.0724     0.082     -0.882     0.378     -0.233       0.088
V18          0.0078     0.149      0.052     0.958     -0.285       0.301
V19          0.1119     0.113      0.991     0.322     -0.109       0.333
V20         -0.4982     0.095     -5.223     0.000     -0.685      -0.311
V21          0.4589     0.075      6.106     0.000      0.312       0.606
V22          0.7242     0.164      4.416     0.000      0.403       1.046
V23         -0.1048     0.071     -1.485     0.138     -0.243       0.034
V24         -0.0259     0.179     -0.145     0.885     -0.376       0.325
```

V25	-0.1586	0.156	-1.019	0.308	-0.464	0.146
V26	0.1407	0.222	0.634	0.526	-0.294	0.576
V27	-0.9115	0.133	-6.847	0.000	-1.172	-0.651
V28	-0.3228	0.100	-3.214	0.001	-0.520	-0.126
Amount	0.0010	0.000	2.202	0.028	0.000	0.002

This shows that features: Time, v2, v3, v6, v7, v8, v9, v10, v11, v13, v14, v15, v16, v17, v20, v23, v24, v25, v27, v28 have a negative impact on the classifier i.e. the higher these variables/features, the more likely the transactions are not fraudulent and vice-versa while the other features: v1, v4, v5, v12, v18, v19, v22, v21, v26 have a positive impact on the classifier meaning that the higher these variables/features the more likely the transactions are fraudulent.

This Logistic Regression model gives the following results:

Accuracy on training set: 0.9992676711943982

Accuracy on test set: 0.9991573329588147

Recall on training set: 0.6666666666666666

Recall on test set: 0.5777777777777777

Precision on training set: 0.8981132075471698

Precision on test set: 0.8387096774193549

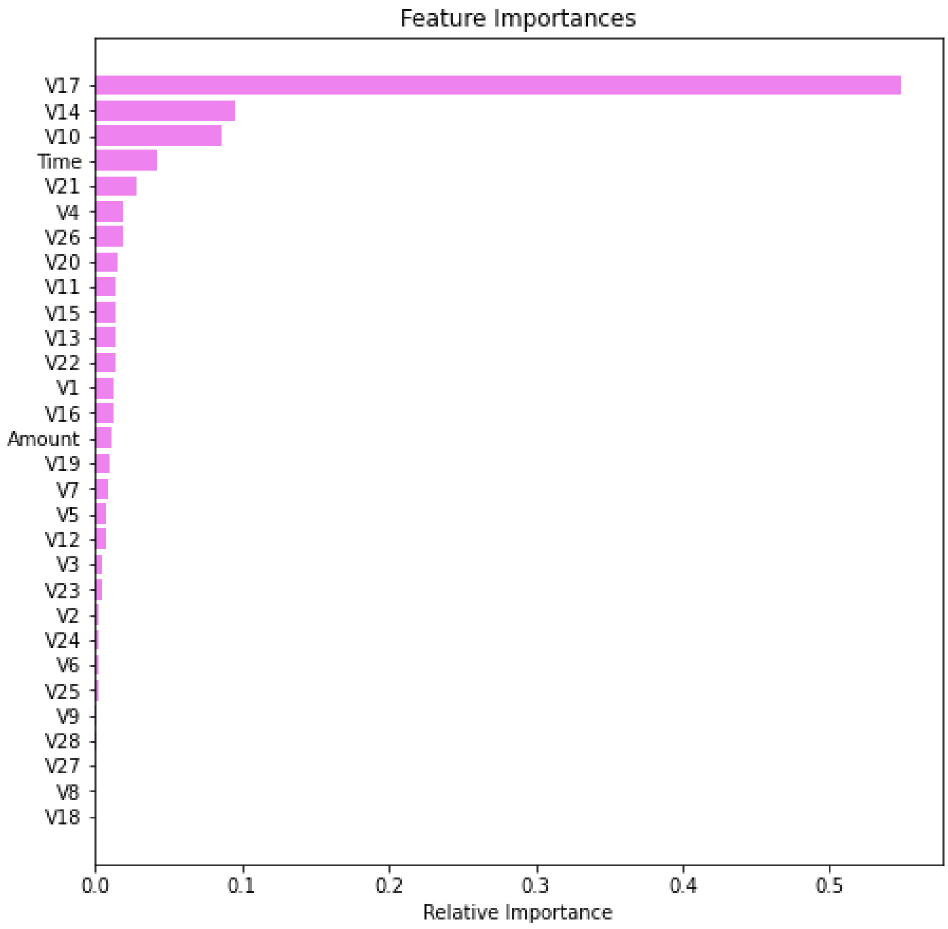
ROC-AUC Score on training set: 0.8332654965235728

ROC-AUC Score on test set: 0.788800972163611

✓ **Decision Tree**

A decision tree classifier was also trained on the preprocessed training data, setting the default 'gini' criteria to split.

The feature importance of the decision tree classifier is as follows:



Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable, this shows us that V17, V14, V10, Time and V21 contributes most significantly to the model while V9, V18, V8, V28, V27 contribute least and have little to no impact on the model.

The decision Tree model gives the following Results:

Accuracy on training set: 0.999197447884272

Accuracy on test set: 0.9991807403766253

Recall on training set: 0.7703081232492998

Recall on test set: 0.7037037037037037

Precision on training set: 0.7790368271954674

Precision on test set: 0.76

ROC-AUC Score on training set: 0.8849580886186752

ROC-AUC Score on test set: 0.8516760184012963.

Improvement

The created classifier models can be improved to give the best possible result by subjecting the dataset to further pre-processing, this amidst other things include: More Feature engineering, Feature selection, hyperparameter tuning for better optimization and finally use of Ensembling techniques.

Conclusion

A machine learning classifier have been built to detect fraudulent card transactions; the overall model score is as shown: