Credit Card Fraud Detection

Context

Credit card fraud model is designed to predict fraudulent transactions in the data and prevent them in future for bank's users

Data

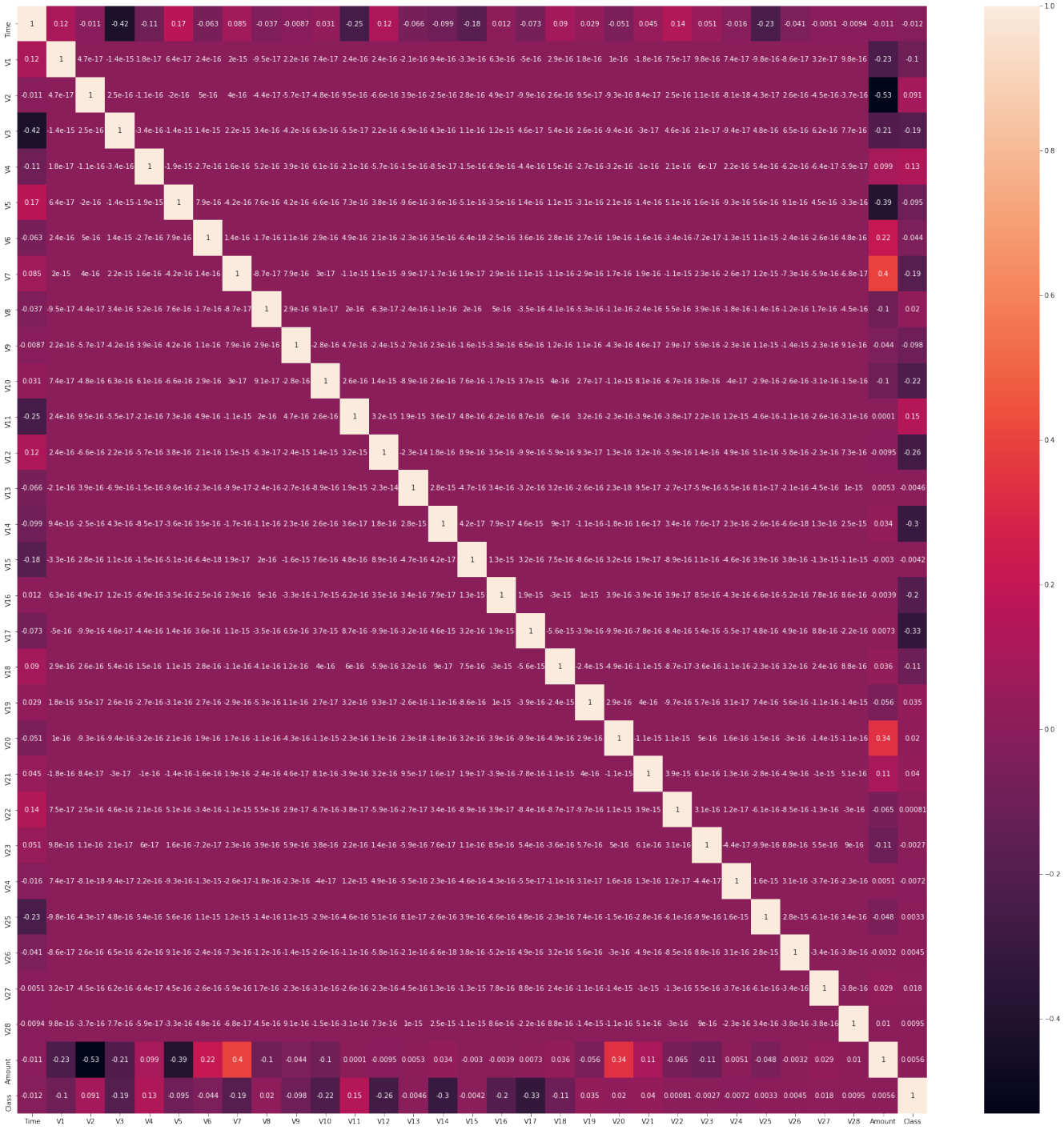
The dataset obtained from kaggle contains transactions made by credit cards in September 2013 by European cardholders.

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-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Data wrangling:

There are no missing values in the data. There are 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Correlation heat map



Preprocessing:

In preprocessing class and amount columns are removed and remaining data is standardized using StandardScaler. The data is split in to training and test datasets and is ready for evaluation

Modeling:

Logistic Regression, Gradient Boosting algorithm and Random Forest Classifier models are used to fit the training data
Model evaluation metrics are as below

Logistic Regression:

```
accuracy: 0.9990695551420246
f1 score: 0.7225130890052357
Precision: 0.8414634146341463
Recall: 0.6330275229357798
```

Gradient Boosting

```
Learning rate:  0.75
Accuracy score (training): 0.999
Accuracy score (validation): 0.999
F1 score: 0.7096774193548386
Precision: 0.7129629629629629
Recall: 0.7064220183486238
```

Random Forest:

```
Accuracy=1.000
f1 score 0.999583442406364
Precision: 0.9479166666666666
Recall: 0.8348623853211009
```

Accuracy in this case can be higher as there are a very few fraudulent transactions compared to total number of transactions. Precision and recall metrics have to be considered to evaluate the model.

Precision is true positives out of total positives

Recall is true positives out of (true positives + false negative)

Recall is a significant measure here as false negative are the transactions that were classified as not fraudulent but is actually fraudulent.

Random forest classifier has better metrics overall and is identified as the best model for this project.