**Credit Card Fraud Detection**

### **Problem statement:**

Credit Card Fraud is one of the biggest issues faced by the government and the amount of money involved in this is generally enormous. As world is getting more towards digitalization, the risk of online fraud is also increasing. The websites with online payment mode contribute to rise in online frauds. Also, due to this pandemic situation(COVID-19), everyone prefers to do cashless transaction which increases the chances of people getting trapped into such frauds.

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. Among all of the online frauds, one such fraud is credit card fraud which is an ever-growing menace in the financial industry. Detecting fraudulent transaction is of great importance for any credit card company.

We are going to approach this real-life problem using Data Science.

### **Proposal:**

The development of a model that provide best results in identifying credit card fraudulent transactions.

This helps both, the credit card company and the customers from getting charged unnecessarily.

### **Data set**

The dataset is obtained from Kaggle. <https://www.kaggle.com/mlg-ulb/creditcardfraud>

The datasets contain 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. Due to confidentiality issues, 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'.

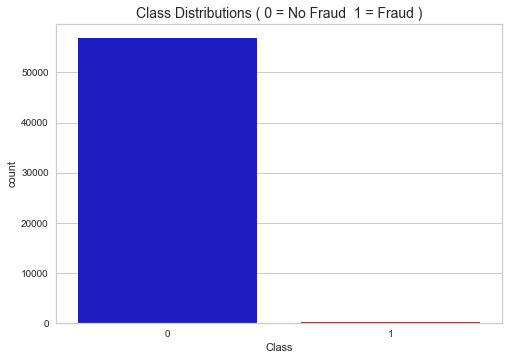
There are 284807 number of transactions(rows) and 31 features in this dataset.

Time: It contains the seconds elapsed between each transaction and the first transaction in the dataset.

Amount: It is the transaction Amount.

Class: It is the response variable and it takes value 1 in case of fraud and 0 otherwise

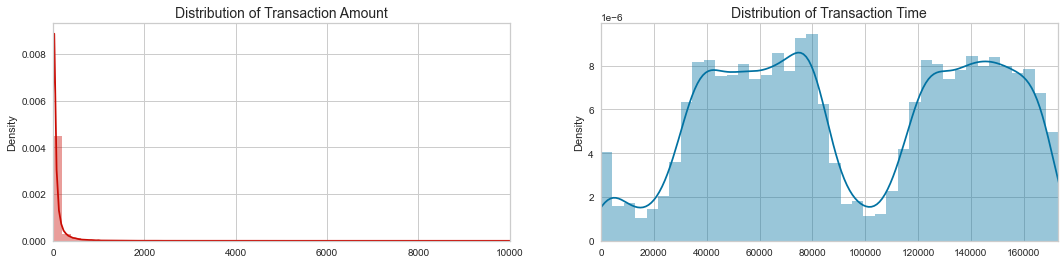
1. Class disctribution



As you can see from the graph, the data is very skewed. I will have to deal with the data skewness later. Most of the transactions are non-fraud. If we use this dataframe as the base for our predictive models and analysis, we might get a lot of errors and our algorithms will probably overfit since it will "assume" that most transactions are not fraud. But we don't want our model to assume, we want our model to detect patterns that give signs of fraud!

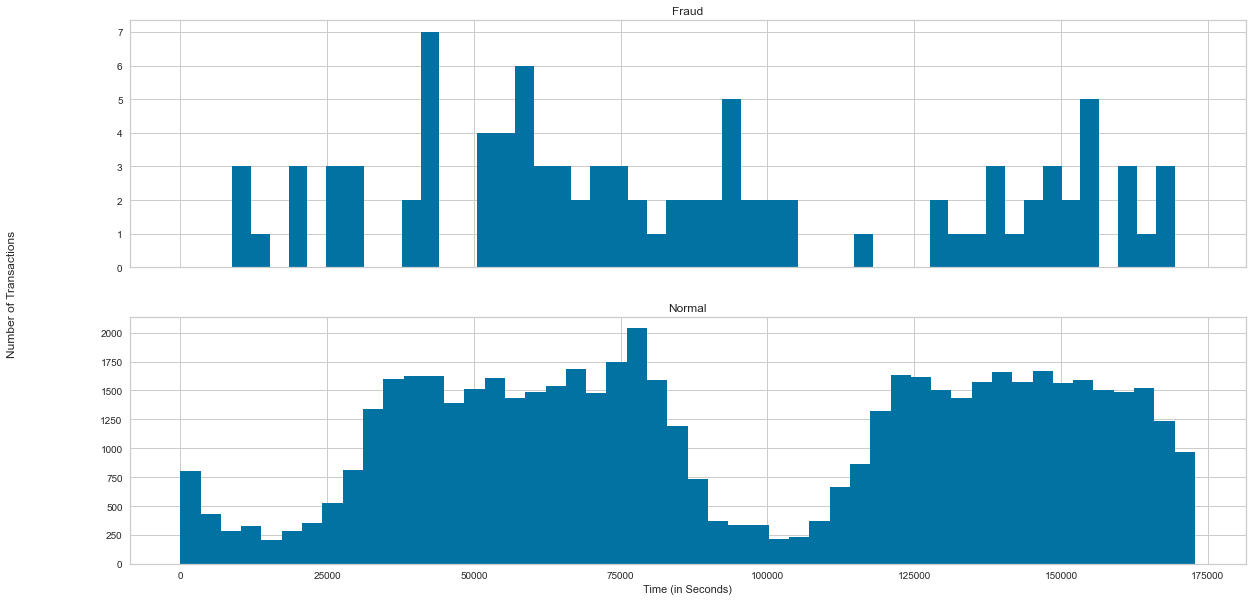
No Frauds 99.83 % of the dataset

Frauds 0.17 % of the dataset



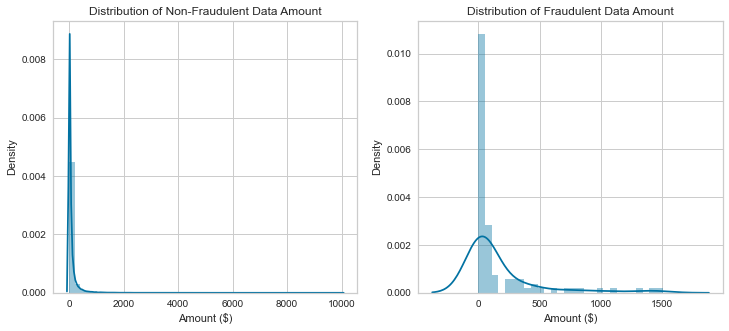
From the above graph, we have noticed that the distribution of time is bimodal in nature which in turn also indicates that there is a sudden fall in the volume of transactions after 28 hours of the first transaction been made. As the timing of the transactions are not provided, we can assume that the reduction in volume occurred during night.

1. Fraud VS Non-Fraud Time Distribution



The graph indicates that Fraud transaction time are evenly distributed all over the place while non-fraud transaction time is bimodal in nature.

1. Amount Distribution



The above graph gives us a quick understanding on how the amounts of fraudulent and Non-fraudulent transactions are distributed.

1. Distribution of anomalous features.

These helps visualizing all the other features from the dataset on graphs.

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

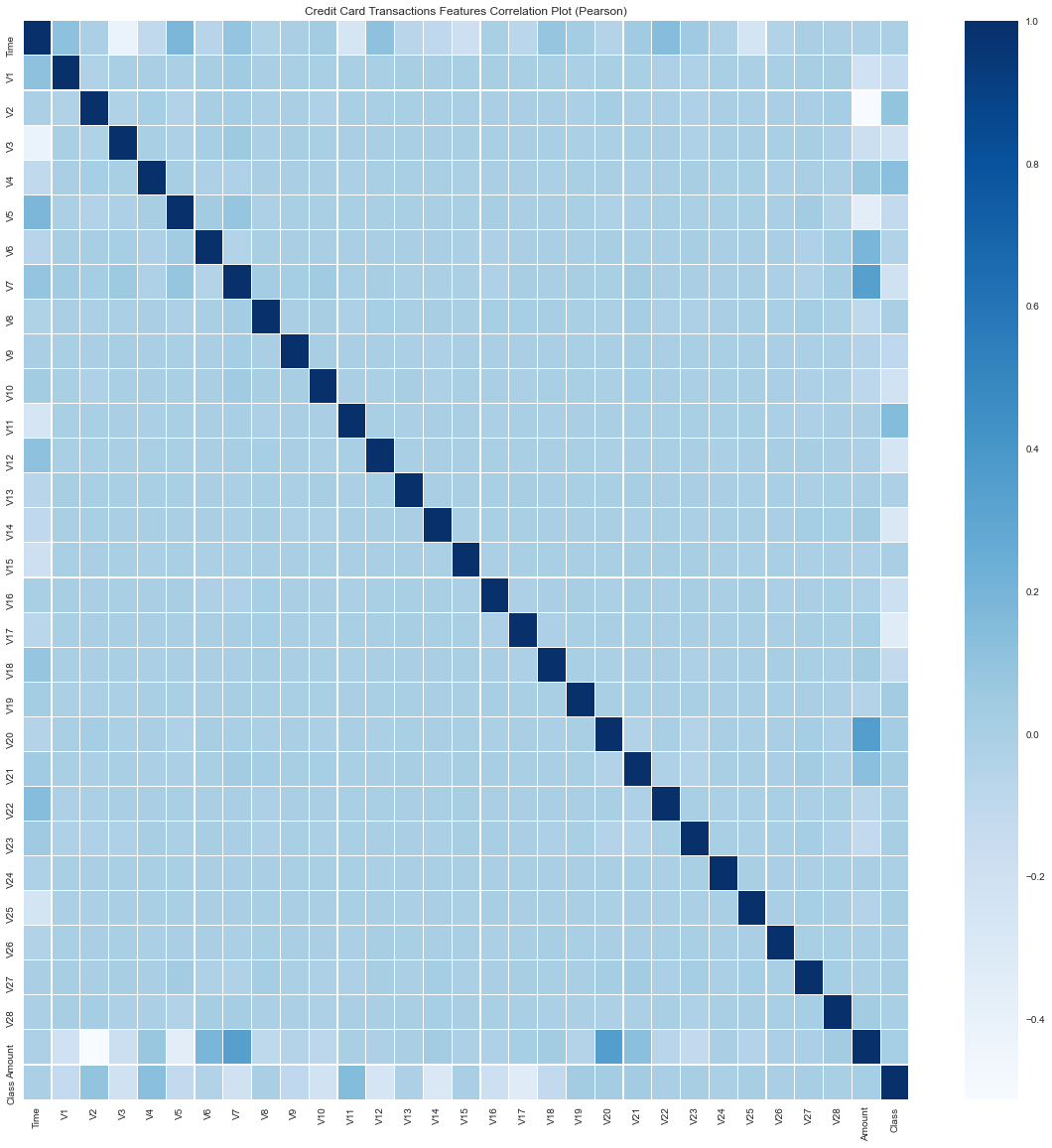
Chart, radar chart

Description automatically generated

Chart, line chart

Description automatically generated

6. Correlation between features



In the HeatMap we can clearly see that most of the features do not correlate to other features but there are some features that either has a positive or a negative correlation with each other.

For example, we can see some correlation with “V20” and “Amount”. This gives us a deeper understanding of the Data available to us.

1. Pearson Ranking

