**Fraud Credit Card Transactions**

**Term Project Question:**

Statistics pulled from s <https://www.self.inc/info/credit-card-fraud-statistics> shows that the losses from fraud involving cards used for payment worldwide reached $27.85 billion in 2018. These projected to rise to $35.67 billion in five years and 40.63 billion in 10 years. To address these rising credit card fraudulent transactions, credit card companies and merchants put many measures in place to prevent credit card frauds. According to statistics from cardrates.com, here were more than 108.6 million credit card transactions occurred in US every day and there were $39.6 billion combined purchase transaction occurred in the US in 2019. Based on millions of credit card transaction, it is very hard to analyze fraudulent transaction manually. However, by using the Machine Learning, we expect to get sufficiently information by using the EDA and applying the appropriate models, which will help to identify and block any suspicious fraudulent transaction.

**Outcome of your EDA:**

I have used given dataset and wanted to use machine learning to develop a model, which can predict a fraudulent transactions. I wanted to train a random forest model to establish some milestones and then loop back to EDA, I wanted to look at the important predictive variables and use other models. I have downloaded the credit card dataset, which had transactions for September 2013 by European cardholders. The data had transactions of two days and my analysis found 492 fraudulent transactions out of total 284,807 transactions. I had following variables

Transaction Country

Transaction Method

Transaction currency

Transaction amount

Transaction Time

Merchant City

Merchant Type

Merchant Code

According to my data analysis, the given data was unbalanced and there were only 0.17% of fraud account transactions based on above variables.

**What do you feel was missed during the analysis?**

The available credit card dataset was unbalanced, a null classifier which always predict class = 0 would obtain over 99% accuracy on this task. Hence, simple measure of mean accuracy are not be used due to insensitivity to false negatives.

I wanted to figure out the appropriate measures for this task

I wanted to know how I could transform the given data into over or under sampling

**Were there any variable you felt could have helped in the analysis?**

Additional variables of Merchant type, Merchant City, Transaction country, online or store transaction variables could help more to understand the transaction. These additional features could help to apply different algorithms for fraudulent transactions.

**What Challenges did you face, what did you not fully understand?**

I understood that detecting the fraudulent transaction is a combination of complex algorithms and this requires a careful analysis of data, more time of planning to create a model and apply the machine learning algorithms. I wanted to use the random forest algorithm and the available dataset had few important variables missing.