APPROACH TO CAPSTONE PROJECT

Project Objectives and Background

Credit Card fraudulent transactions result in substantial losses to the credit card companies owing to the fact that they have to compensate their customers in cases when customers inform the company about a fraudulent transaction. So the liability or risk in most cases of fraudulent transactions falls upon the Credit Card companies.

Any effort to mitigate this risk can lead to substantial savings for the companies and is a welcome measure that every Credit Card company can be looking into. Given the huge volume of transactions that happen every second around the globe, it is impossible to manually check and prevent a fraudulent transaction or raise a red flag in the payment system.

Machine learning can contribute to this immensely by a combination of computing speed and intelligent algorithms. **The project aims to develop a robust ML model that can predict fraudulent credit card transactions so that precautionary measures could be taken to stop the transaction and any financial loss occurring therefrom.**

Dataset

This particular case study is one such situation wherein we are presented with dataset obtained from Kaggle. The data set pertains to credit card transactions associated with European credit card holders. These transactions cover two days in September 2013. The data set comprises of 2,84,807 transactions of which 492 transactions (a mere 0.172% of total transactions) were fraudulent. So the data set is highly imbalanced as the positive or minority class is extremely small.

Given the sensitivities of the data involved the dataset has been masked using PCA. Except the "time" and "amount" variables, the rest have been transformed into 28 principal components numbered from V1 until V28.

Steps Leading to Model Building

1) Understanding the Data – this would help us analyse the variables and identify those variables that can contribute to the model building
2) Conduct EDA on the data that includes univariate and bivariate analysis to gain insights for further model building
3) Preparing the Data – taking care of the critical issue of data imbalance through SMOTE or other techniques
4) Test-Train split using 70:30 ratio for train and test respectively
5) Evaluate the model using k-fold cross validation method. It is suggested that in addition to Logistic regression, one or two other algorithms could also be explored.
6) Final Model building – herein the fine-tuning of the model using hyper parameters tuning will be carried out to arrive at the final model.