MS4610 Project Team-28

Yash Mathur CE19B106

Hritik Bohra CE19B054

Rushikesh Gulve CE19B051

Rishav Dhakad CE19B079

Ashish Pawar CE19B003

Parikshit Gaikwad CE19B071

AMEX CREDIT DEFAULT DETECTION :

We were given with a data which constituted information about 83000 people. For every person several information were provided in 47 columns. The information included Past Credit Crad payments made by the person, his/her application id etc.

Missing Data Evaluation :

The data provided to us wasn’t clean and it needed to be cleaned first. There were a lot of missing values in various columns. The columns which had very less missing values as compared to the length of data set were removed since they won’t affect the model learning much. On the other hand columns which had comparitively larger number of missing values were taken care of by replacing them with mean of the data.

Question Arises Why not replace them with mode or median ??

Ans : We tried doing that as well but on training the model the accuracy wasn’t affected much there was just a plus minus 0.4% gap. So we decided to stick with mean value only.

Some columns had missing values more than 50% of the data size. The columns included 11,23,30, 31, 40, 41, 45. Since these columns were not giving sufficient information about the data so we decided to remove them.

Model Training :

First we normalised the data so that values in all the columns were in similar range. After that we tried running models like Logistic Regression, SVM Classifier, Random Forest Classifier, Nueral Network, Decision Tree Classifier. We got maximum accuracy using Random Forest Classifier around 76%. So we decided to make our final predictions on the data set using it.

Testing :

We took the testing data set and applied all the operations which were made on the training data set. Finally we made predictions using the finalised model. We received an accuracy of 50.53% in total.