Approach

The process is started with importing the required libraries and loading the data into a data frame for analysis. The Data was then analysed for Null checks and NaN value checks. In some of the columns there were more than 25% of data missing. Therefore, those columns were removed from the data frame. Some of the columns where having less than 5% of missing data, therefore different techniques were used to fill the missing data with Mode. Some of the Column's data from transform into 1 and 0 for better processing. Once the Data cleaning steps are completed EDA is performed to find correlations between the columns. For the categorical columns dummies were created and appended to the data frame. The columns which were used for dummy creation is dropped to avoid redundancy.

Scaling

The Data is scaled using StandardScaler and Feature selection is done. After this Model building process is started

Model building

Using the statsmodel library the model is built and the training data is used. RFE was used to initially filter top 20 features. The logistic regression model was used, result is obtained and analysed. To finetune the model the columns with P>|z| > 0.05 is dropped. After dropping the column, the model is built again. This process is followed until all the columns/ feature has P>|z| <= 0.05 VIF is calculated for the remaining features and once all results are ready. Model is evaluated. In the model built there were 8 iterations until the required result are visible.

Evaluation

The results from the last model is use for evaluation by trying the predict the results and the accuracy, specificity , sensitivity , recall and precision is calculated using the metrics in sklearn library.

ROC curve was plotted to identify the optimal cutoff point which came out to be 0.3.

The model was applied on the test data and it was able to predict the results and the accuracy, specificity, sensitivity, recall, and precision using the metrics in sklearn library.