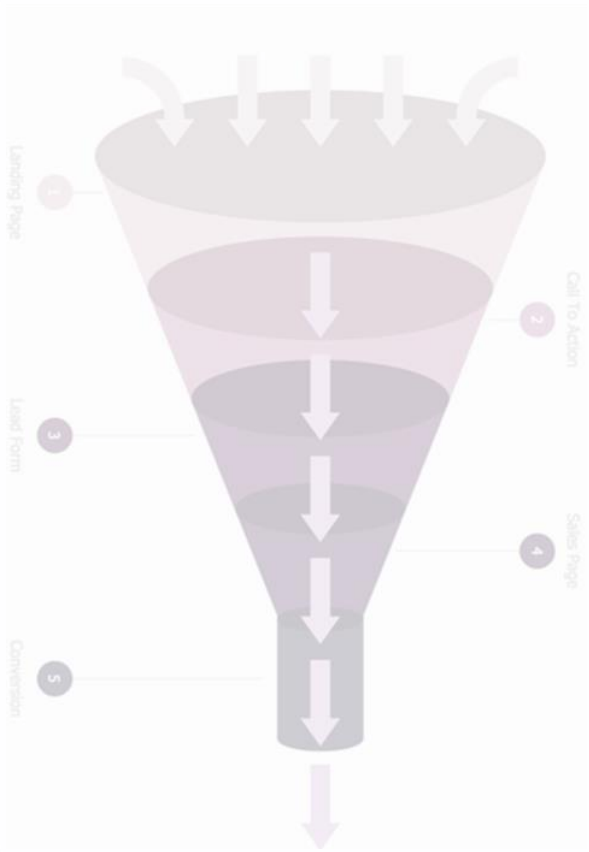
LEAD SCORING CASE STUDY



DS C58 JULY 23 BATCH : NAYAN , NAVEEN , NINAD

PROBLEM STATEMENT

- X Education sales Online Courses to Industry Professionals
- Though they are gathering lot of leads, their lead conversion rate is very poor at 30%.
  for example lets say, If they are acquiring 100 leads in a day, only about 30 of them are actually interested in their course
- To make this process more efficient, the company wishes to identify the most potential leads -'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE

- X Education wants to improve their Leads conversions ratio i.e. Lead Acquisition to Hot Leads, from 30% to targeted 80% .
- This will help in minimizing their marketing spends and maximize their bottom-line

PLAN OF ACTION

DATA SANITY & CLEANING

Checking Data using Head/Tail. Shape , Info, describe etc.

Dropping Columns with 30-40% null values

Imputing necessary columns data

EXPLORATORY DATA ANALYSIS

Univariate Analysis

Bi-Variate Analysis

MODEL BUILDING

Feature Scaling

Dummy Variables

Feature Selection

Creating X & Y Data frames containing Features and Target respectively

MODEL EVALUATION

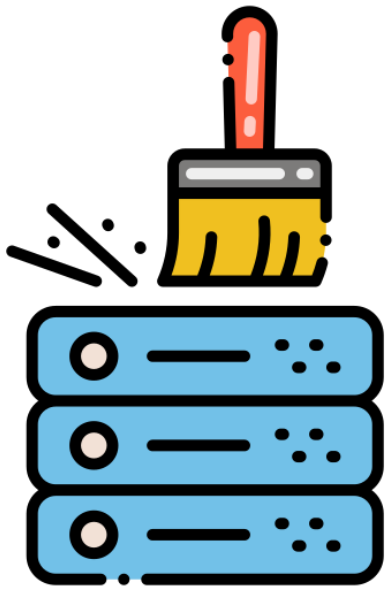
Using Logistic regression to iterate & select best features

Model evaluation using P-values , VIF , Accuracy Score , F1 , Recall , ROC , AUC , Sensitivity and Specificity

MODEL PREDICTION & SUMMARY

Prediction of Lead Score in Train , test , Recheck Model accuracy in test

Final Important variable , accuracy summary



DATA SANITY AND CLEANING

- Original data has Total Number of Rows =37, Total Number of Columns =9240.
- We dropped columns with more than 35% missing values : 'Asymmetrique Profile Index', 'Asymmetrique Activity Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'Lead Profile', 'Tags' , 'Lead Quality', 'How did you hear about X Education', 'City', 'Lead Number' etc.
- We removed Unique value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply" as they were not drawing any inference
- Other columns like "Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped as they were not adding value to analysis.
- "Prospect ID" and "Lead Number" columns were also dropped as the data in it was not making any sense.
- Later with remaining columns by checking value counts of some of the object type variables, we found some of the features which did not had enough variance, were dropped: "Do Not Call", "Search", "Newspaper," Article", "X Education Forums", "Digital Advertisement" etc.
- We also imputed values for following columns which were important from Analysis point of view : " Specialization" , "What matters most to you in choosing course" etc.

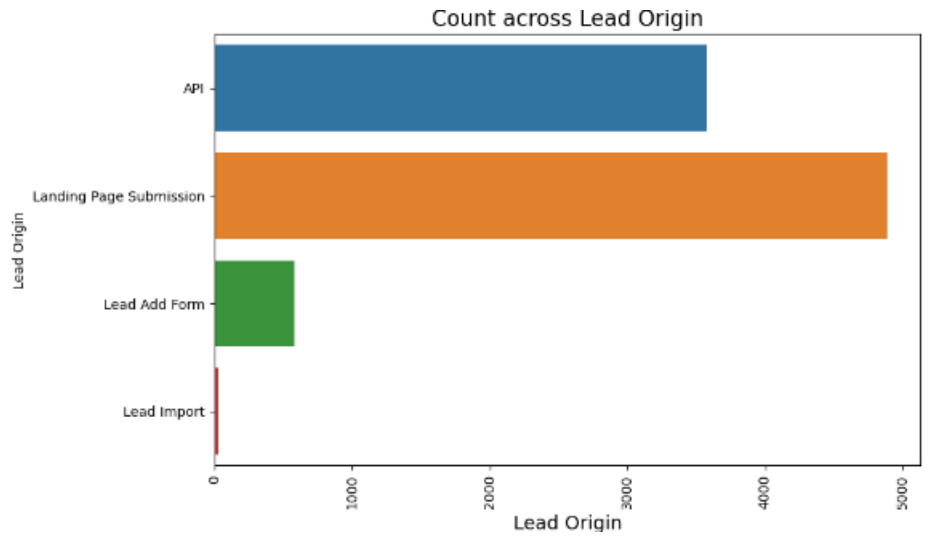
EXPLORATORY DATA ANALYSIS

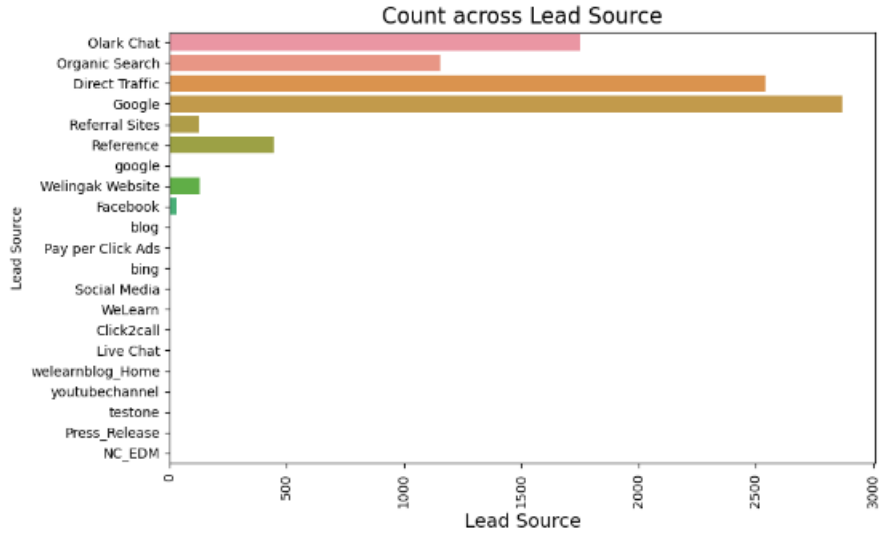
EDA is split between Univariate Analysis and Bivariate Analysis

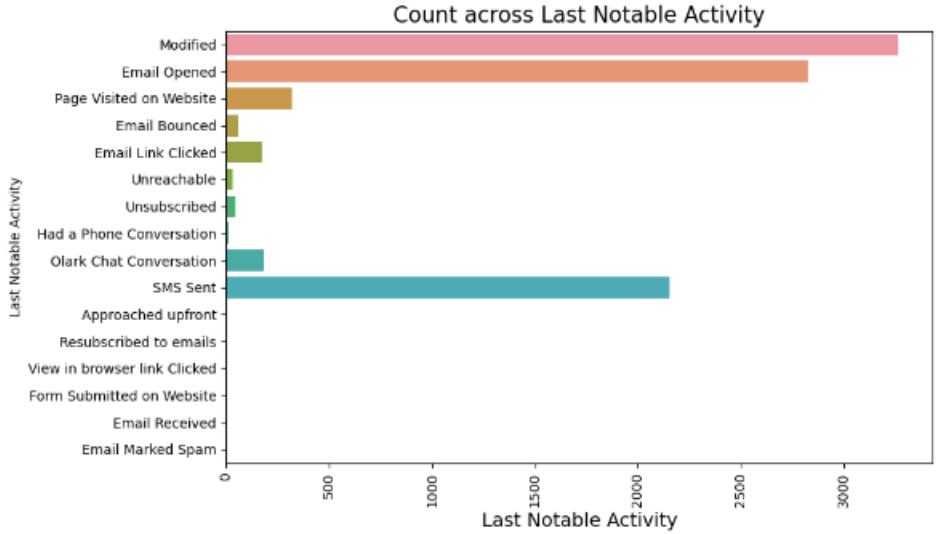
1. In Univariate data analysis we checked value count and distribution of data for variables etc.

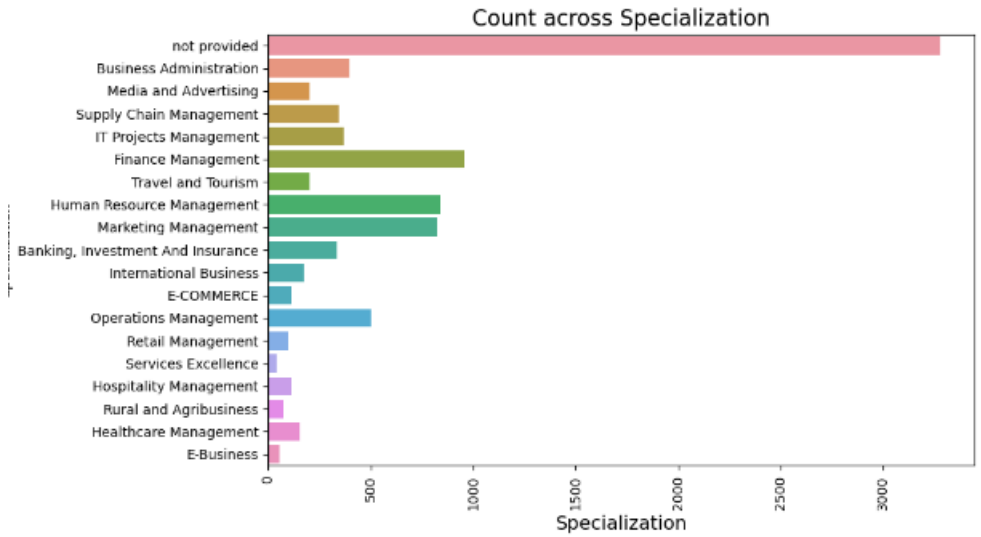
2. In Bivariate data analysis we checked different categorical variables against “Converted” variable & correlation between the variables etc.

UNIVARIATE ANALYSIS

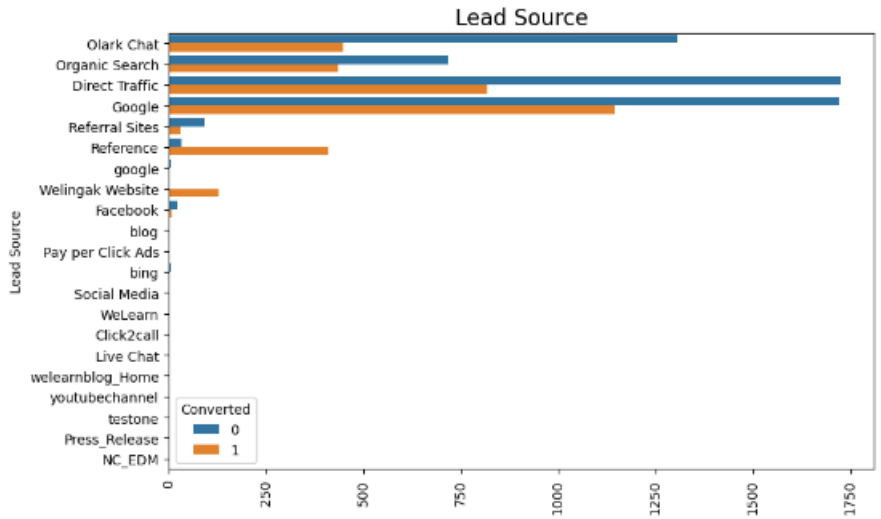


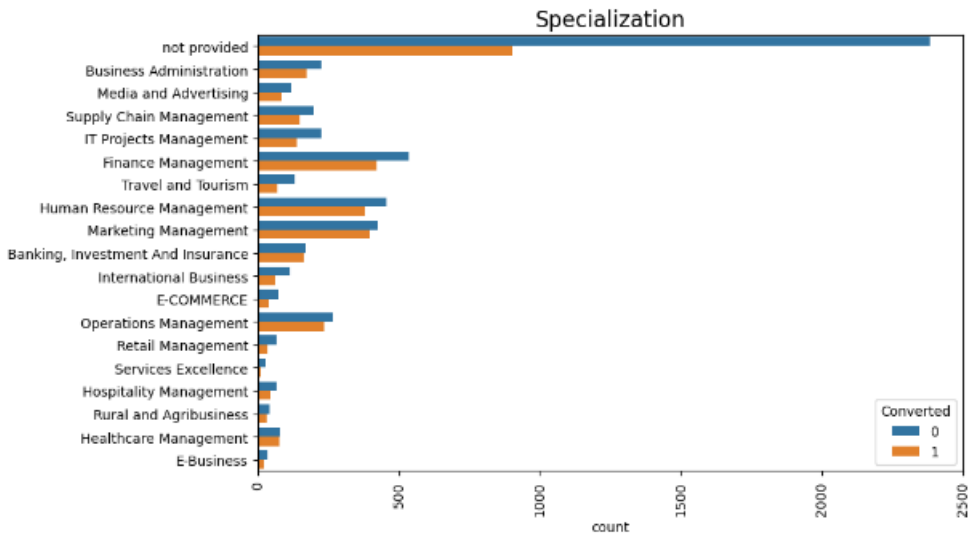


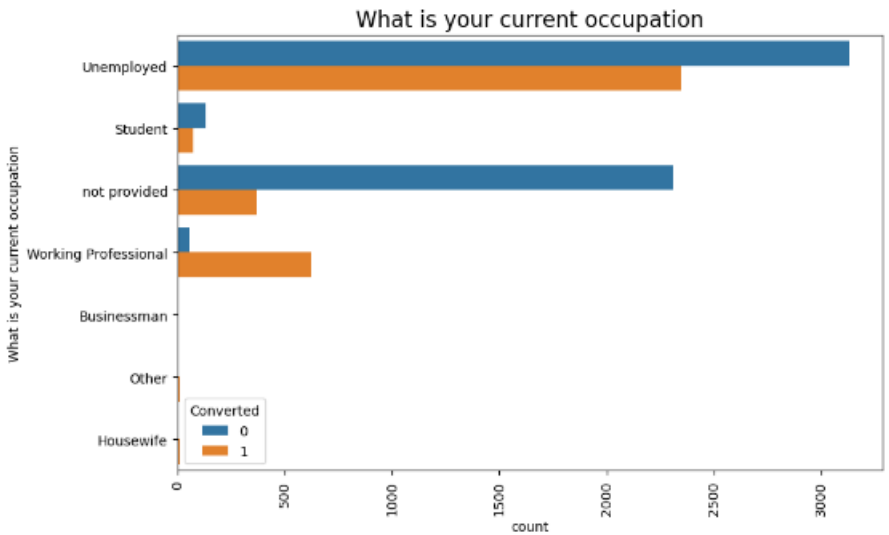


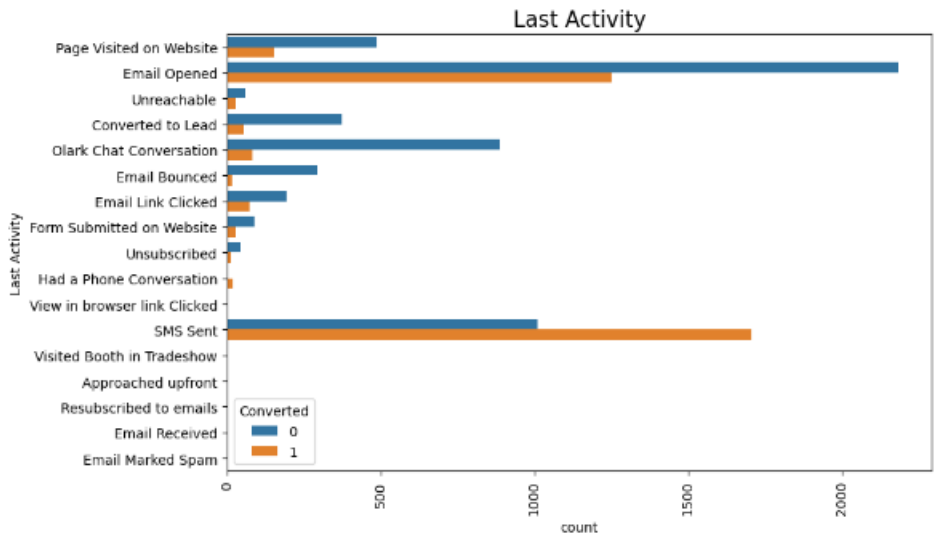


BIVARIATE ANALYSIS

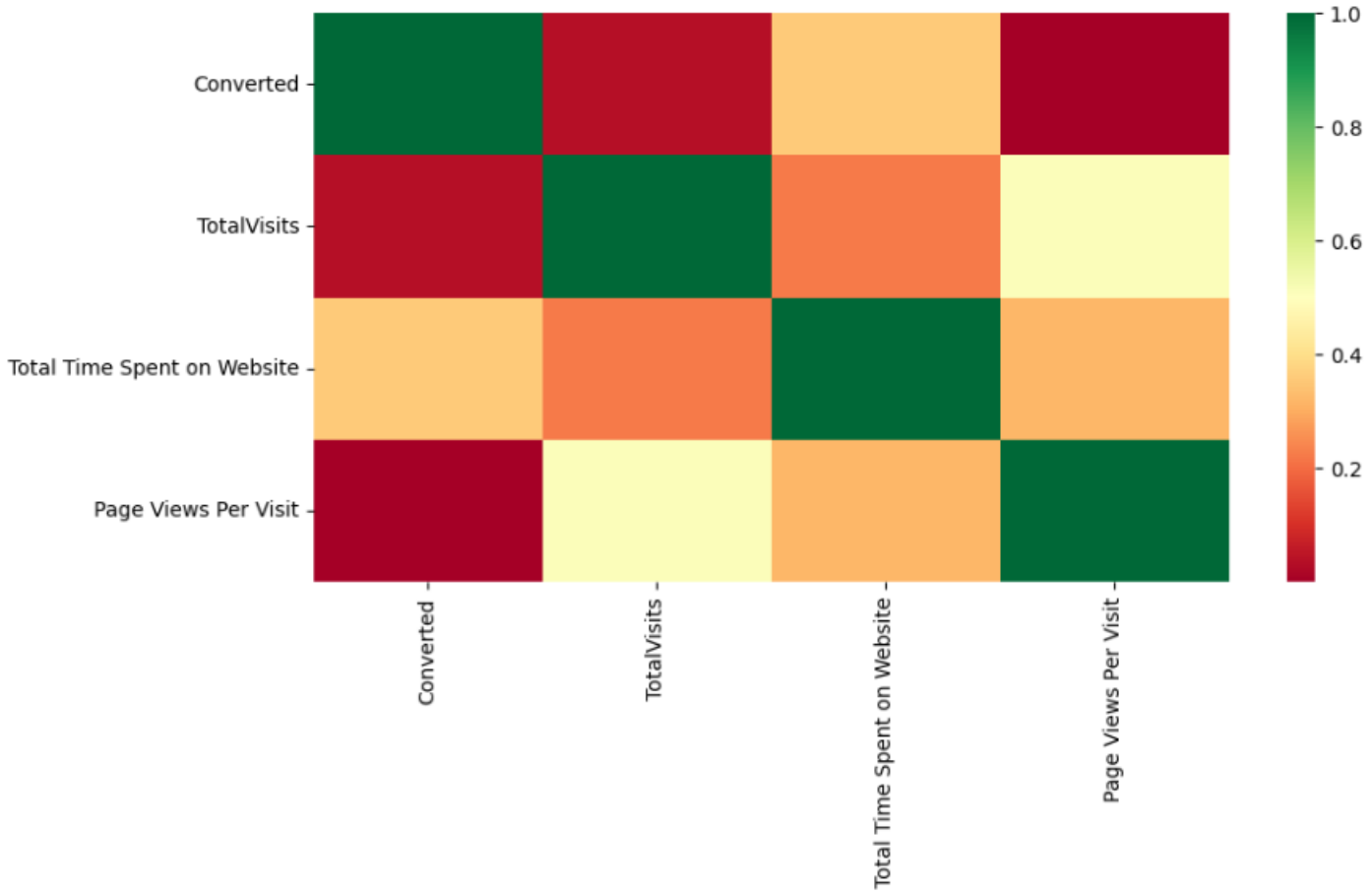


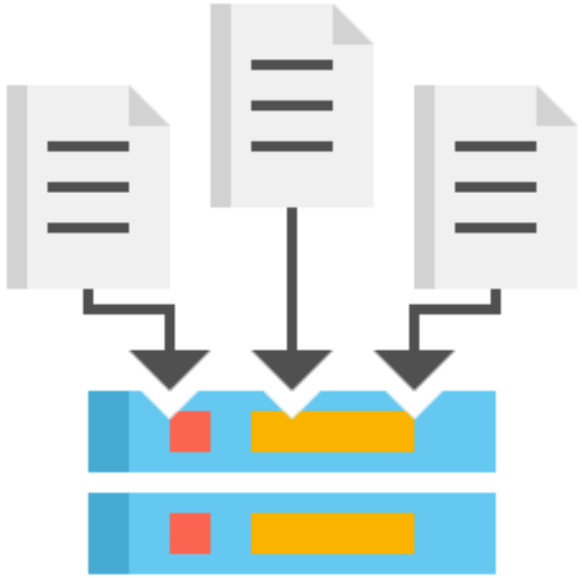






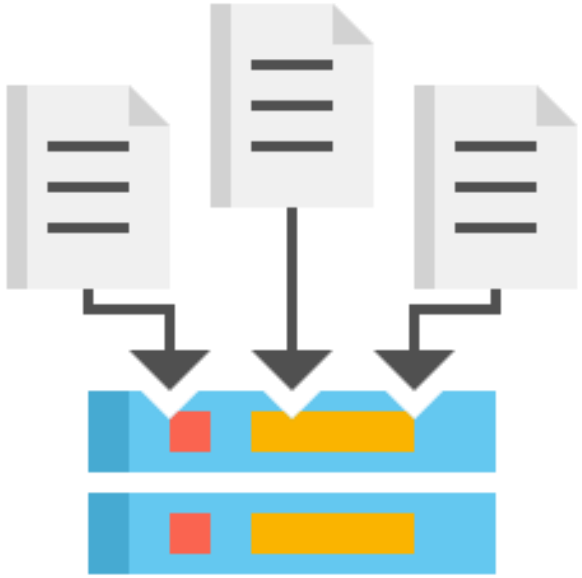
BIVARIATE ANALYSIS





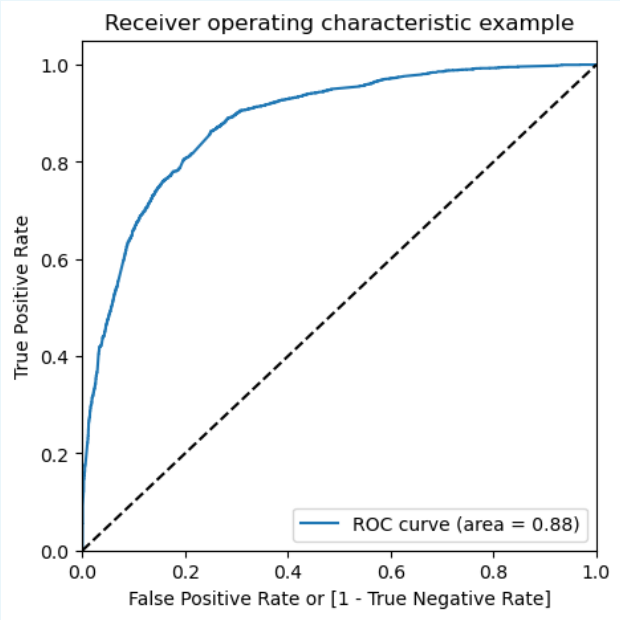
MODEL BUILDING

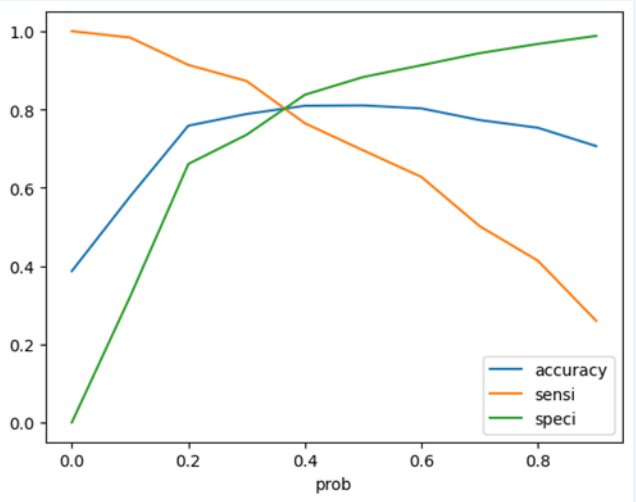
- Splitting the Data into Training and Testing Sets
  we used 70:30 ratio.
- We used RFE (Recursive Feature Elimination) method for Feature Selection
- We started model building by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- We built the prediction data set finalizing on 12 variables and checked their accuracy on test module which came out to be 81.5 %



MODEL BUILDING & EVALUATION

- Area under ROC curve =88%
- Model Evaluation done using Confusion Matrix , results are as follows :
- TP : True Positives , TN : True Negatives ,
- FP : False Positive , FN : False Negative.
- Sensitivity : TP / (TP+FN) :80.45%
- Specificity : TN / (TN+FP) : 80.25 %

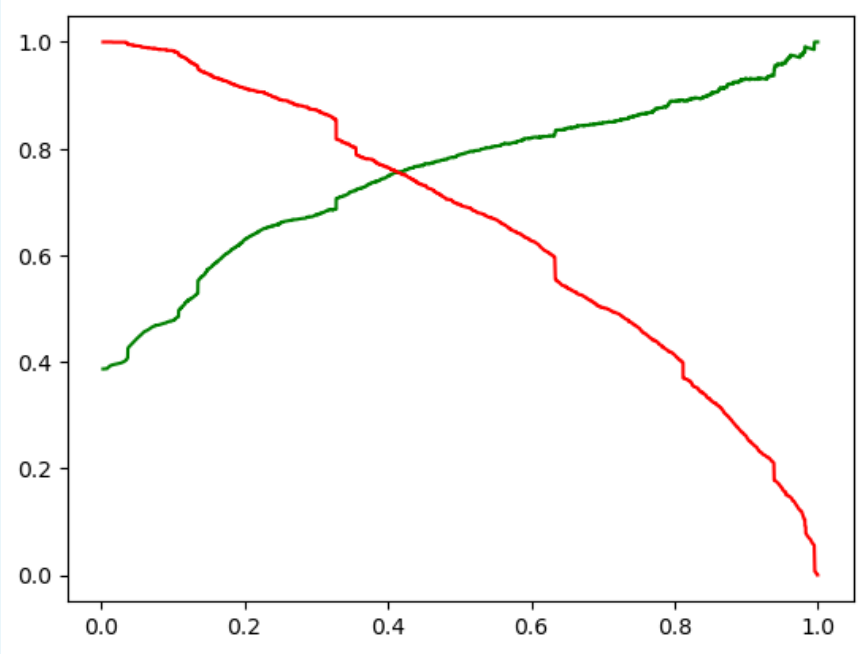




MODEL PREDICTION ON TEST SET

Using trained module result we evaluated on Test Data and found following results

Precision Recall Tradeoff
At 41% cut off
Precision : TP / (TP+FP) : 73.30%
Recall : TP / (TP+FN) : 76.30%



SUMMARY : AREAS OF FOCUS

- Improve Total number of visits to website & The total time spend on the Website as they will lead to conversion ultimately
- It has been observed that maximus conversions happening through SMS activity conducted and via Chat bot conversions -“Olark Chat” on their website , so keep it upto date with all FAQ and reduce TAT for team addressing queries real-time.
- Google , Direct traffic , Organic search , Website, SMS, Olark Chat etc. leads sources are doing well and must be capitalized
- Nurture leads of Working Professionals as they are most like to convert

THANK YOU