LEAD SCORE ASSIGNMENT

SUBMITTED BY:-

ROHAN KULKARNI

ADIL LAKHANI

CONTENT

- Problem Statement
- Process Flow
- Meta Data & Data cleaning
- EDA
- Model building
- Model Evaluation
- Conclusion

PROBLEM STATMENT

- Assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted

PROCESS FLOW

- Importing DataSet
- Meta Data
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Data Modelling
- Model Evaluation

META DATA & DATA CLEANING

Shape of Data Set

9240 Rows and 37 columns

31 Data types are Object

6 Data types are Int/Float

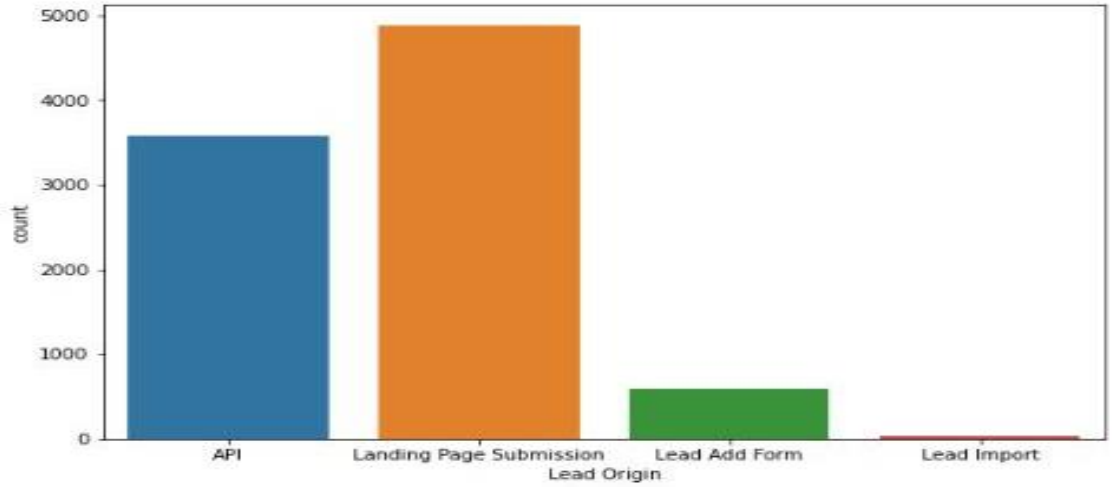
Data Cleaning

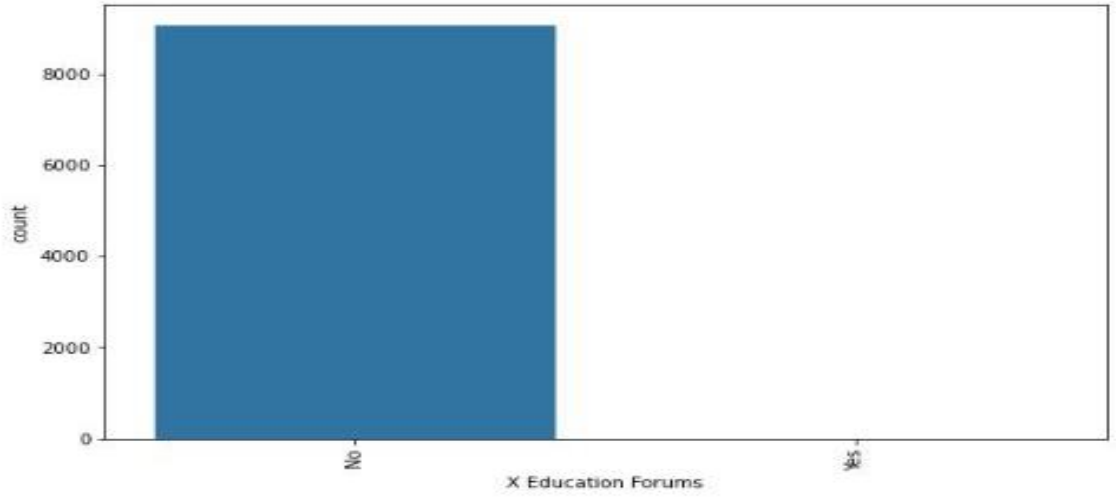
Removing columns with missing value >40%

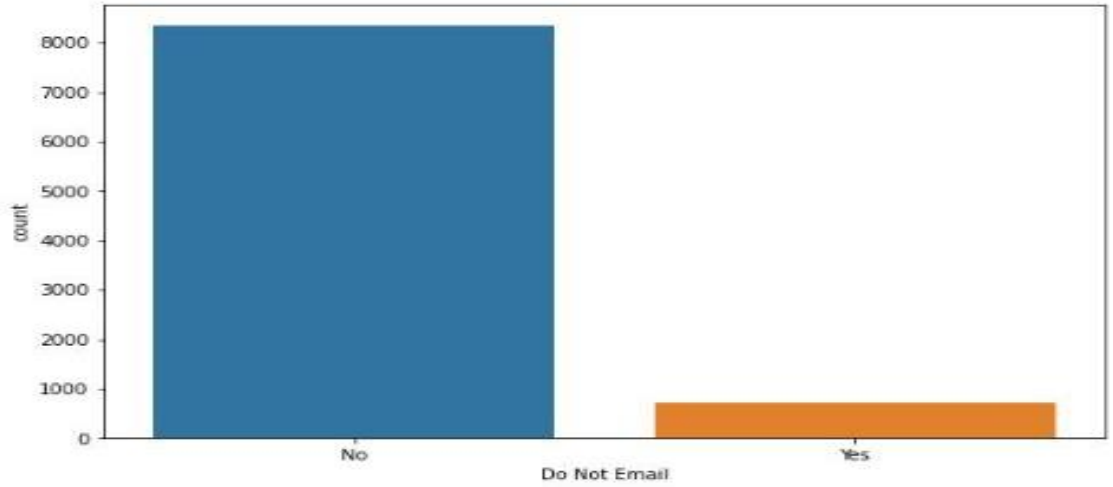
Removing rows with missing value <2%

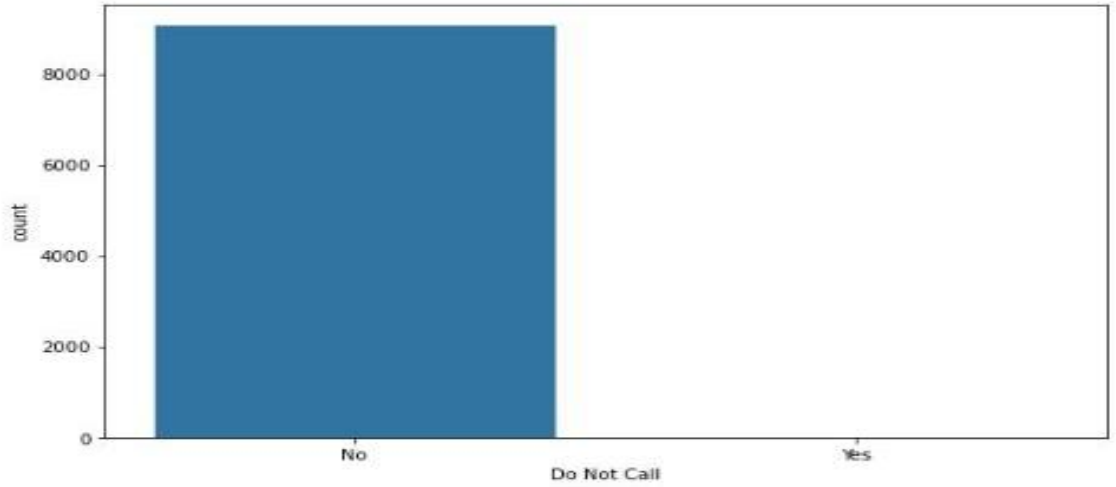
Imputing most common/mean/median/ “Not Available” In place of missing values

EXPLORATORY DATA ANALYSIS

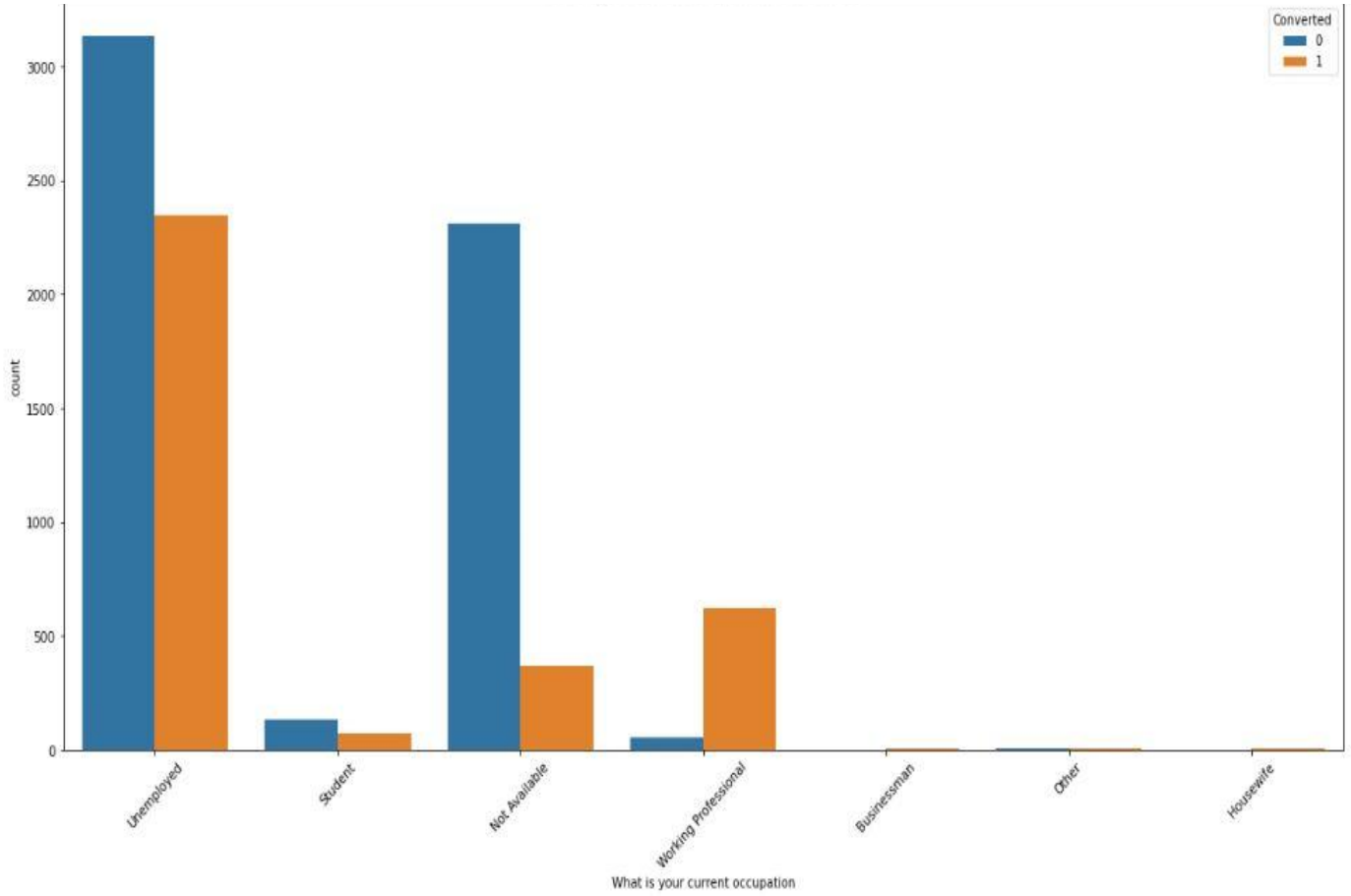


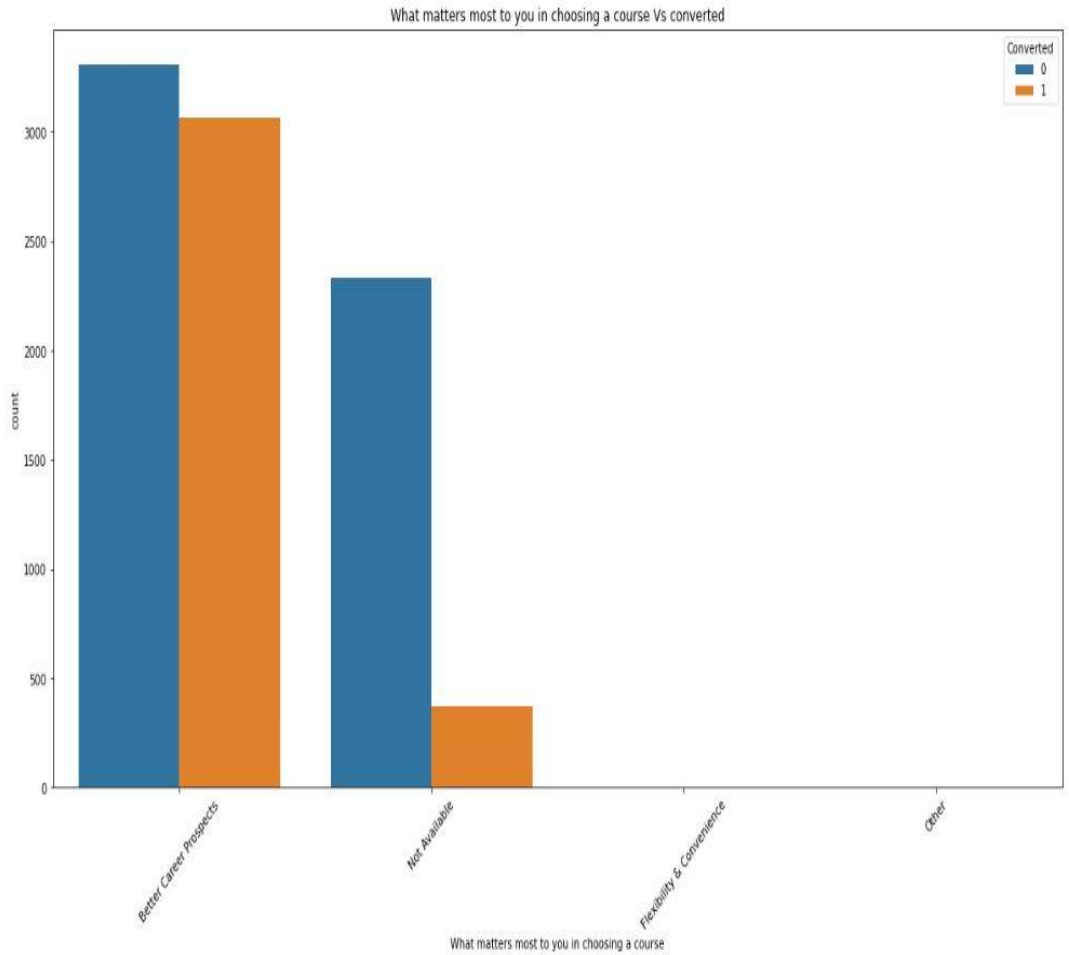




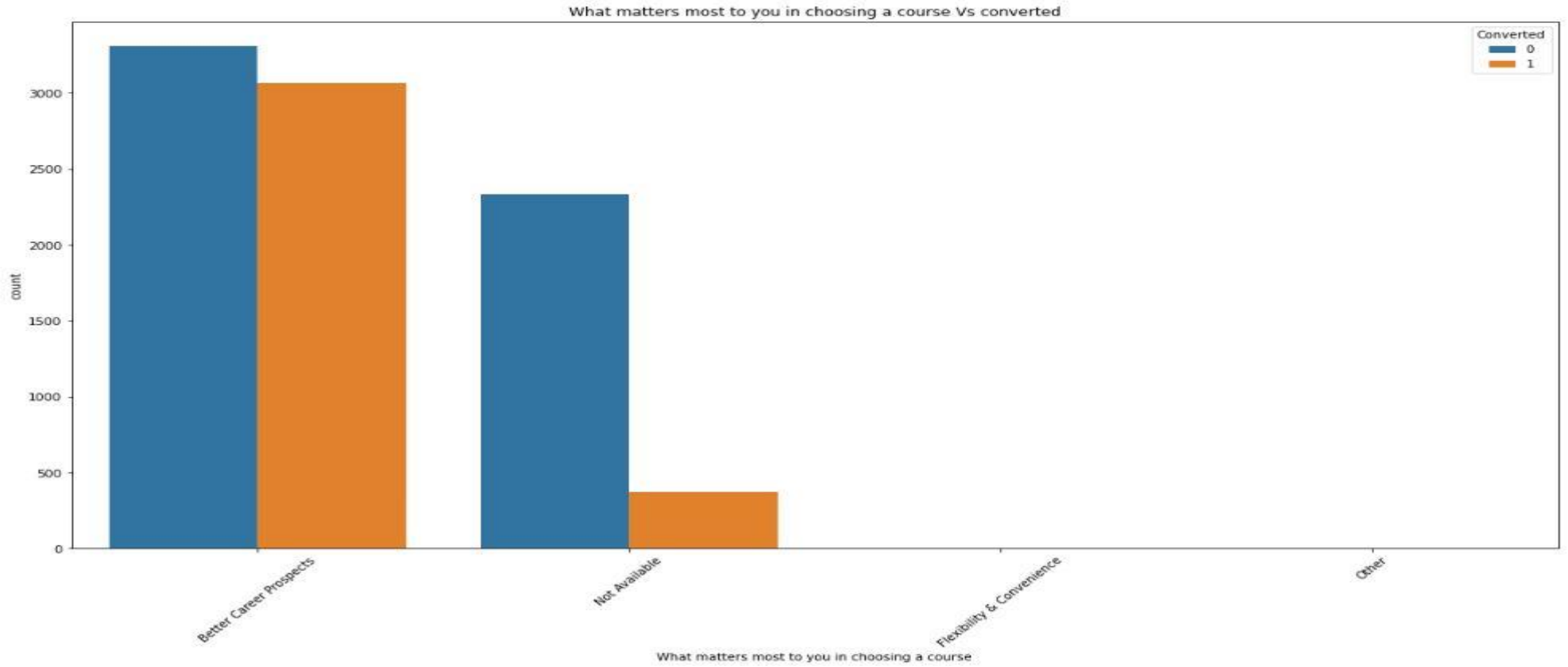


Landing page submission shows maximum count, which eventually shows most leads are generated from website.
-Most leads are welcoming email conversations and dont wish to have telephonic disturbances which are "converted"
-Conversion rate is is 38%
-Most Leads are From India and other countries count is insignificant





- Lead source is maximum through direct traffic and google
- Specialization field is not selected, as there might be confusion on what field is to be chosen.
- while asking purpose about wanting to join institute was better career opportunities.
- Newspaper, search,recommendation, digital advertisment can be eliminated as it doesnt help in finding hot leads



DATA PREPARATION

- Created Dummy variables for Object data type
- There are 83 columns and 9074 rows
- Used Recursive Feature Elimination(RFE) for feature selection
- Total 15 features are selected through RFE.

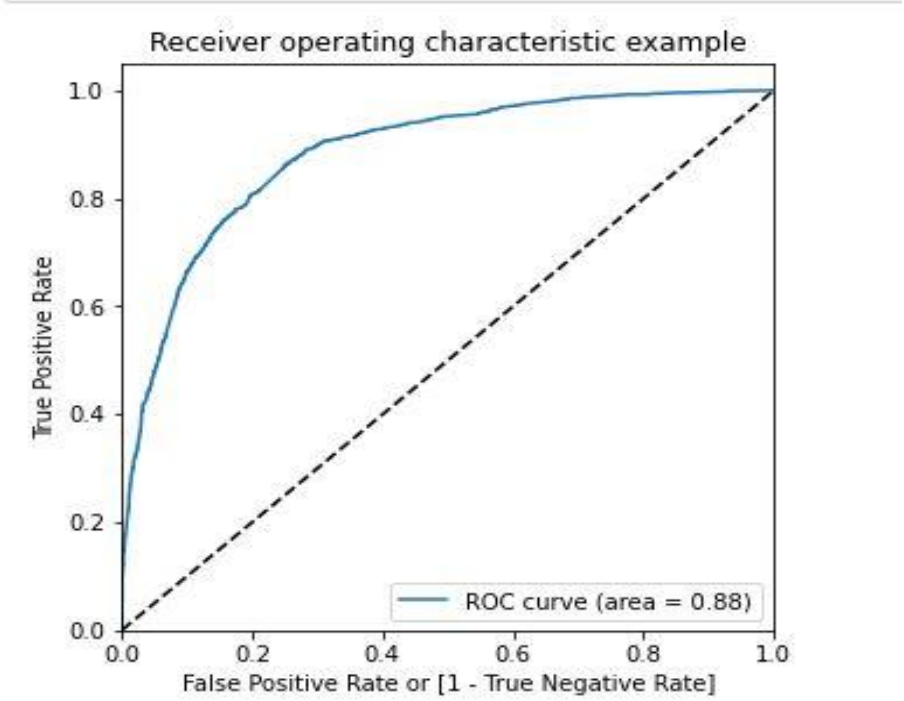
```
: # RFE Supported Columns
cols = X_train.columns[rfe.support_]
cols
```

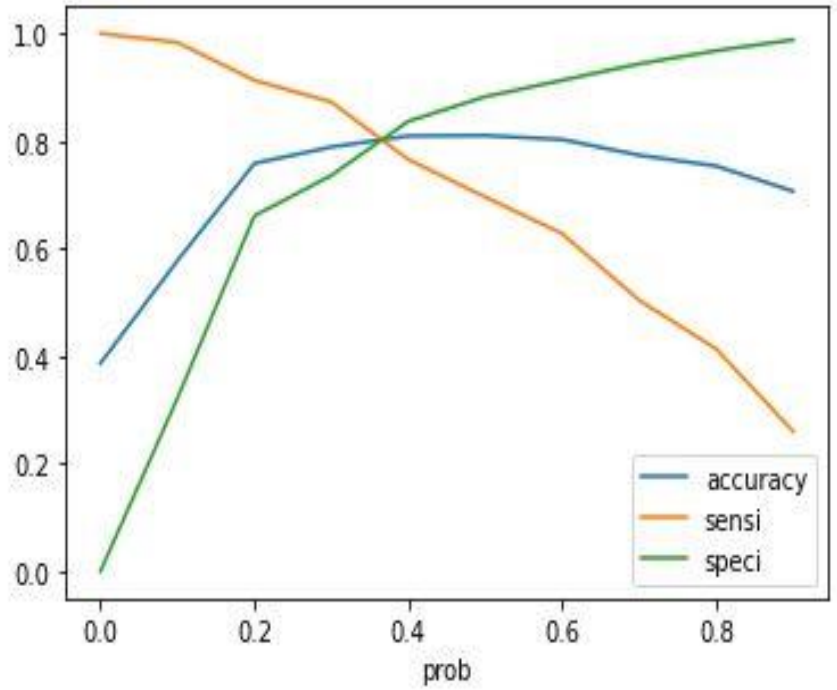
```
: Index(['TotalVisits', 'Total Time Spent on Website',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Welingak Website', 'Do Not Email_Yes',
       'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent',
       'What is your current occupation_Housewife',
       'What is your current occupation_Other',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Unreachable'],
      dtype='object')
```

MODEL BUILDING

- Dropping the columns for which dummies are created
- Scaling the features using StandardScaler
- Splitting the data set into Train-Test Set.
- Ratio for split is 70:30.
- Building Models with Statsmodels
- Eliminate features with P-value >0.05 and VIF >5

ROC CURVE





Optimal Cutoff can be observed as 0.38

MODEL EVALUATION

Train

Accuracy : 0.81

Sensitivity : 0.78

Specificity : 0.82

Test

Accuracy: 0.813

Sensitivity : 0.78

Specificity : 0.82

	Prospect ID	Converted	Conversion_Prob	final_predicted	Lead_Score
2	2085	1	0.982741	1	98
3	4048	1	0.878240	1	88
15	3917	1	0.873302	1	87
17	8088	1	0.994747	1	99
18	3192	1	0.919472	1	92

CONCLUSION

- Overall Accuracy =81%
- Important Variables

TotalVisits	5.542672
Total Time Spent on Website	4.604821
Lead Origin_Lead Add Form	3.750105
What is your current occupation_Working Professional	3.679731
Lead Source_Welingak Website	2.582057
What is your current occupation_Other	2.156716
Last Notable Activity_Unreachable	1.815310
Lead Source_Olark Chat	1.580159
Last Activity_SMS Sent	1.267234
What is your current occupation_Student	1.245642
What is your current occupation_Unemployed	1.163151
Last Activity_Olark Chat Conversation	-1.397360
Do Not Email_Yes	-1.436043
const	-3.453287

dtype: float64