LEAD SCORE CASE STUDY

GROUP MEMBERS :

AKSHAY NAGPAL

NIMYA GEORGE

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘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 know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

METHODOLOGY

- ❑ DATA CLEANING
- ❑ EDA
- ❑ DUMMY VARIABLE CREATION
- ❑ FEATURE SCALLING
- ❑ MODEL BUILDING
- ❑ FEATURE SELECTION
- ❑ VALIDATION
- ❑ EVALUATION OF METRICES
- ❑ CONCLUSION

CLEANING THE DATA

- The dataset contains 9241 rows and 37 columns.
- Columns with more than 40% of missing values were dropped.
- Unique categorical columns having columns have 99% as 'No' in the data. So these columns were dropped and are:

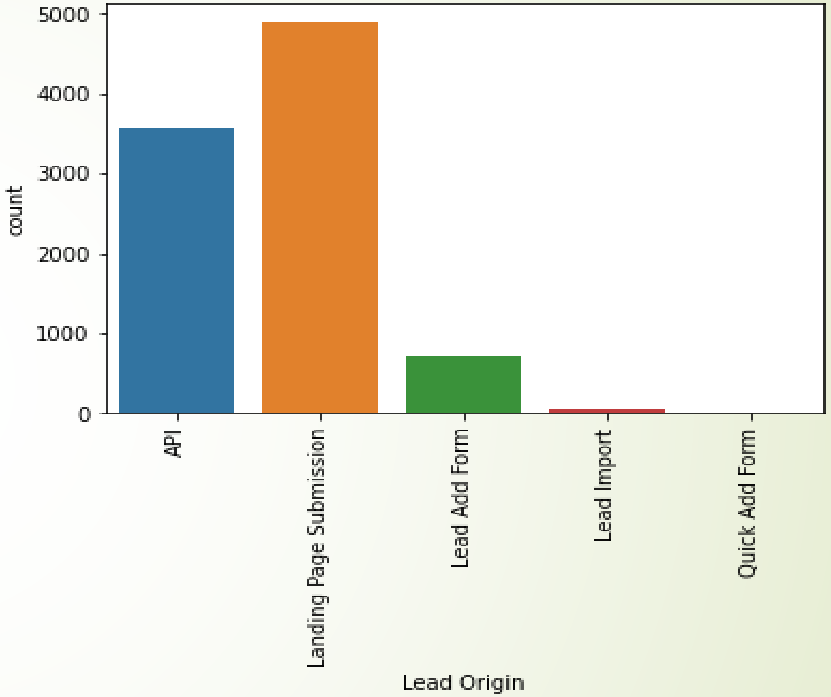
  * Do Not call
  * Search
  * Magazine
  * Newspaper Article
  * X Education Forums
  * Newspaper
  * Digital Advertisement
  * Through Recommendations
  * Receive More Updates About Our Courses
  * Update me on Supply Chain Content
  * Get updates on DM Content
  * I agree to pay the amount through cheque.

- 'What matters most to you in choosing a course' and 'Country' skewed towards only one category. So these columns are also dropped.
- Replace the following columns with mode values:

  * Lead Source
  * TotalVisits
  * Page Views Per Visit
  * Last Activity
  * Specialization
  * What is your current occupation
  * Tags
  * City

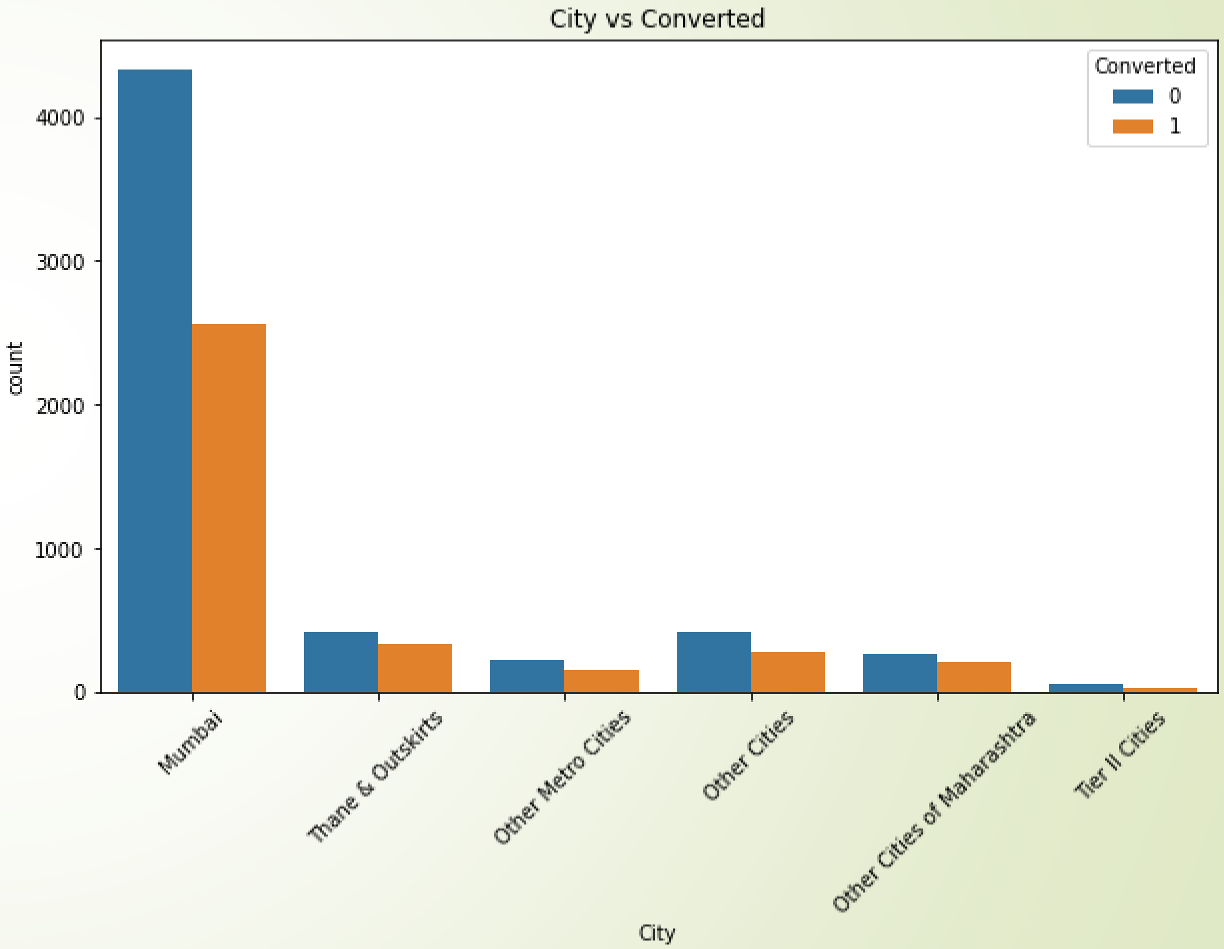
EDA

- ❑ Landing Page Submission have highest count followed by API.
- ❑ Quick Add Form had the least count.



CITY Vs CONVERTED

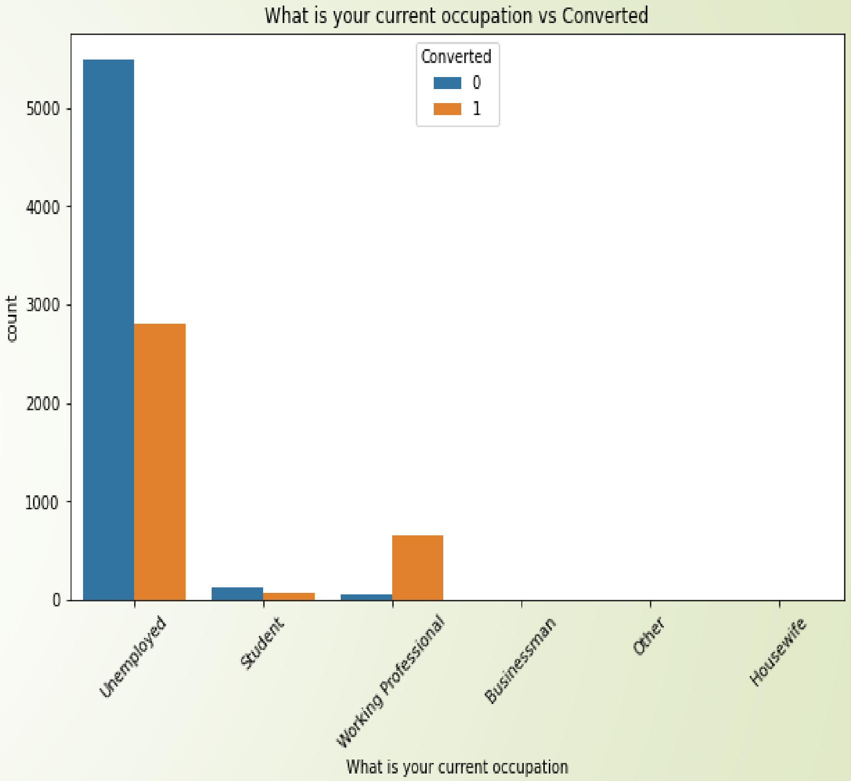
- ❑ Mumbai have the highest conversion rate.
- ❑ The tier 2 cities hold the least.



Analysing the OCCUPATION Column

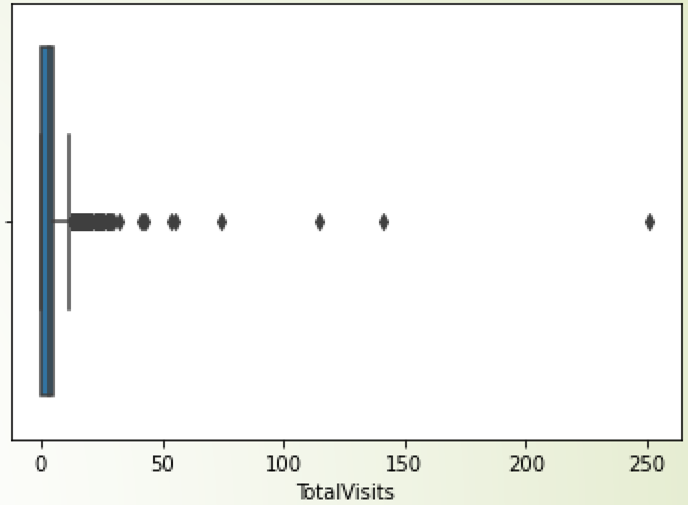
- Unemployed contributed to most of the data.

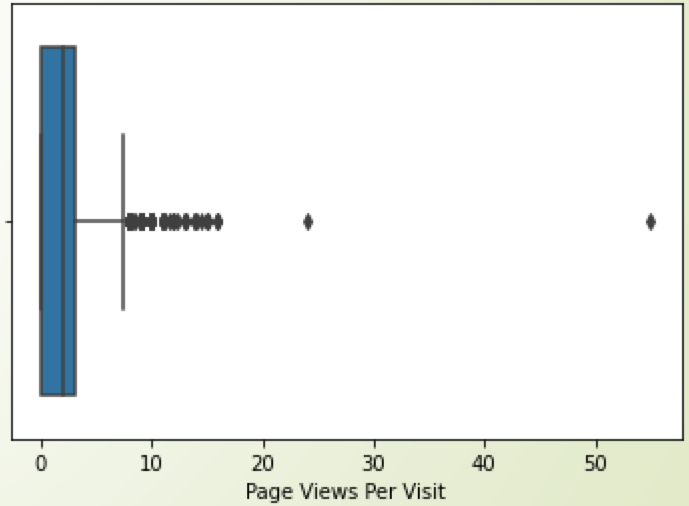
- High conversion rate is seen among students and working professionals.



Boxplot and Outlier Treatment

The columns (page views per visit and total visit) contains the outlier and should be limited within the Range.





DUMMY VARIABLE CREATION

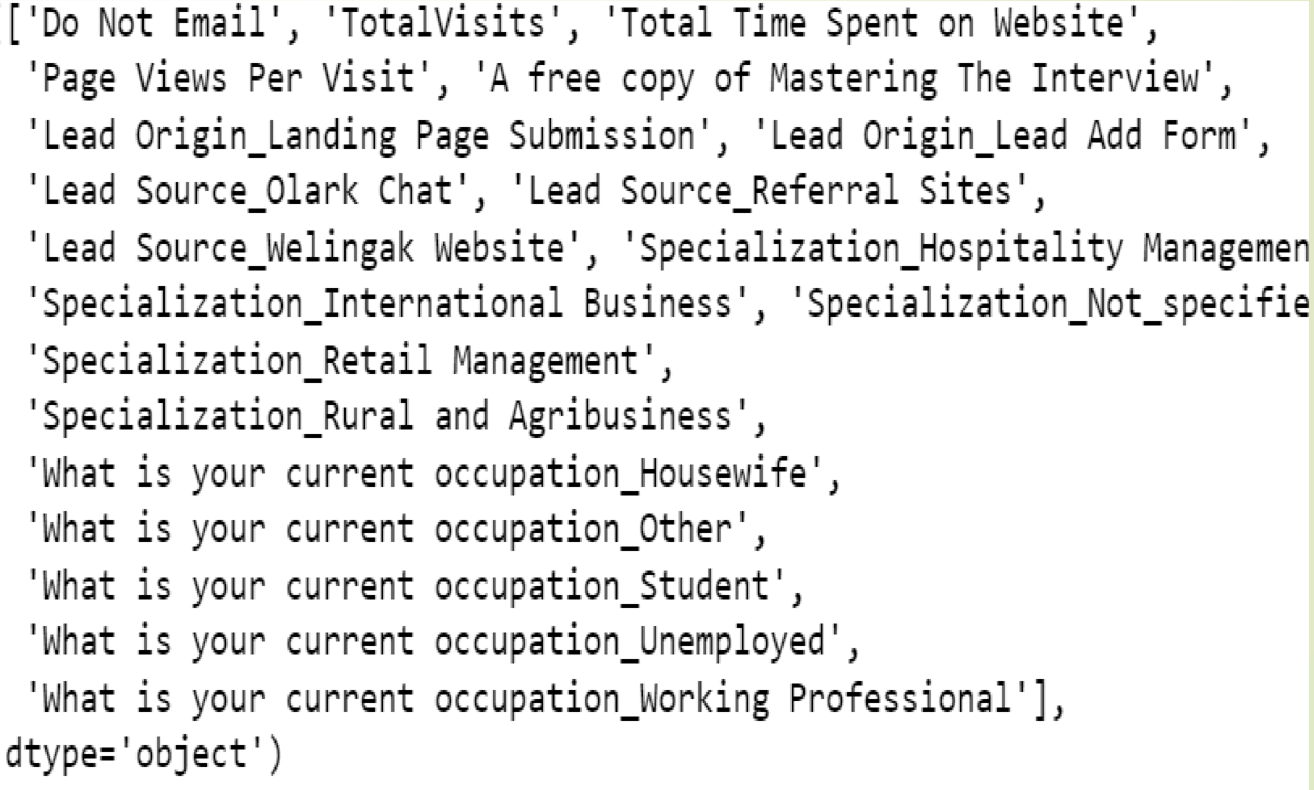
- After dropping the irrelevant categorical columns there were 5 categorical variables in the data.
- Created the dummy variables for each of the columns and dropped the first columns for each category.
- Total of 52 dummy variables were created.
- After deleting the repeated columns there is 46 columns were present in the final data.

Test train split and Feature scaling

- The given data is split in the ratio of 70:30 for training and testing respectively.
- Scaled the data using minmax scaler.
- We have got a conversion rate of 38.5

FEATURE SELECTION

- Some of the most relevant features are selected using the RFE algorithm.
- Top 20 variables are retained and all others are dropped
- The significant variables are shown here:

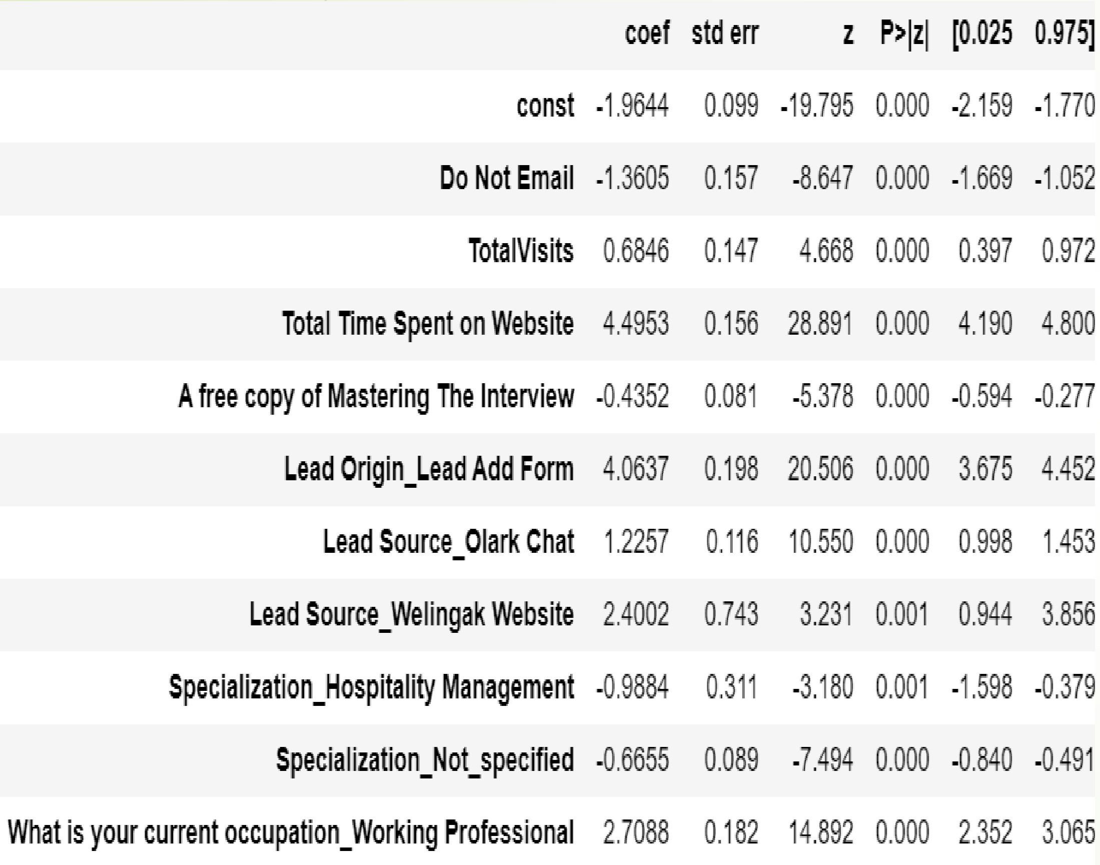


```
['Do Not Email', 'TotalVisits', 'Total Time Spent on Website',
 'Page Views Per Visit', 'A free copy of Mastering The Interview',
 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
 'Lead Source_Olark Chat', 'Lead Source_Referral Sites',
 'Lead Source_Welingak Website', 'Specialization_Hospitality Managemen
 'Specialization_International Business', 'Specialization_Not_specifie
 'Specialization_Retail Management',
 'Specialization_Rural and Agribusiness',
 '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'],
dtype='object')
```

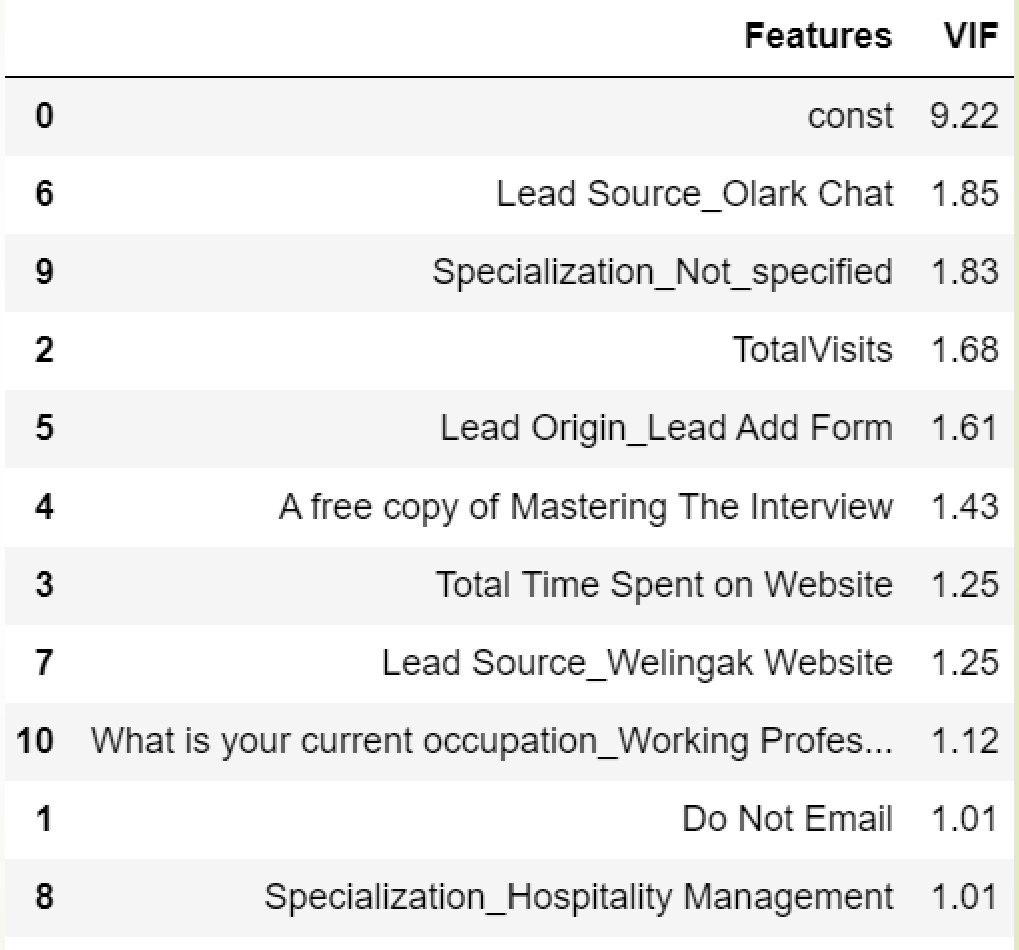
MODEL BUILDING

- An initial model is created using the all variables which shows some poor statistics.
- Evaluated the model parameters.
- Rebuild the model again.
- Check for VIF and p_values.
- Continue the step until a stable model is obtained.

The Final Model

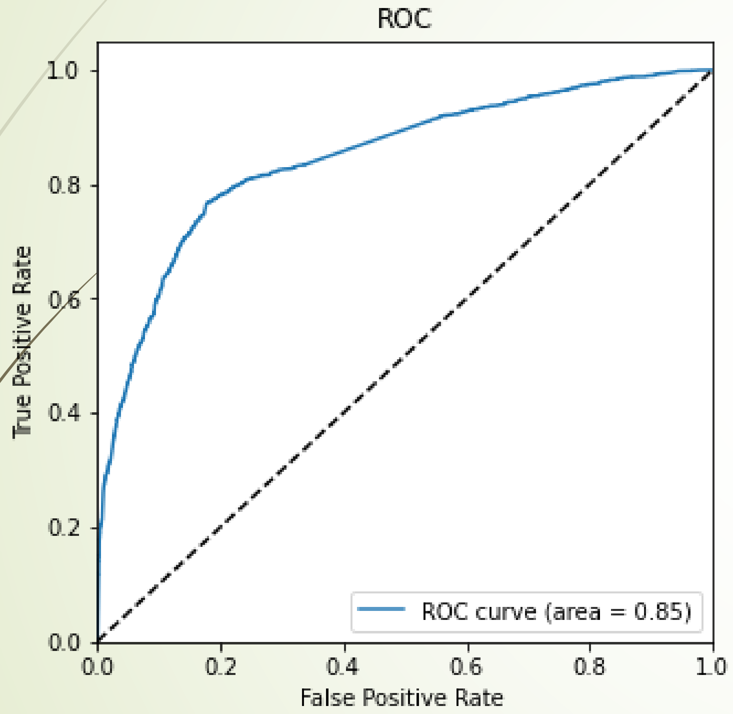


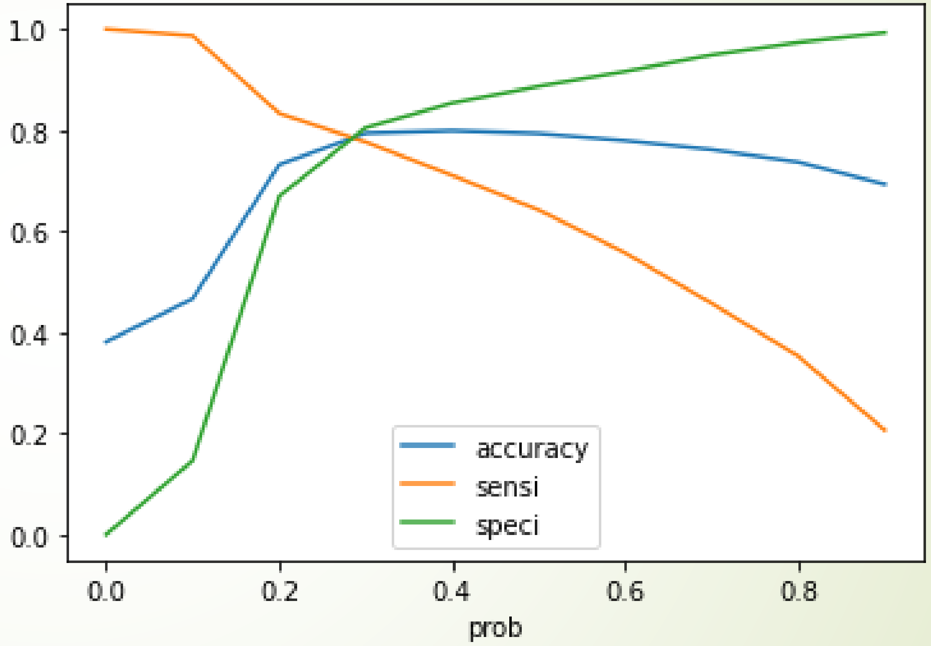
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9644	0.099	-19.795	0.000	-2.159	-1.770
Do Not Email	-1.3605	0.157	-8.647	0.000	-1.669	-1.052
TotalVisits	0.6846	0.147	4.668	0.000	0.397	0.972
Total Time Spent on Website	4.4953	0.156	28.891	0.000	4.190	4.800
A free copy of Mastering The Interview	-0.4352	0.081	-5.378	0.000	-0.594	-0.277
Lead Origin_Lead Add Form	4.0637	0.198	20.506	0.000	3.675	4.452
Lead Source_Olark Chat	1.2257	0.116	10.550	0.000	0.998	1.453
Lead Source_Welingak Website	2.4002	0.743	3.231	0.001	0.944	3.856
Specialization_Hospitality Management	-0.9884	0.311	-3.180	0.001	-1.598	-0.379
Specialization_Not_specified	-0.6655	0.089	-7.494	0.000	-0.840	-0.491
What is your current occupation_Working Professional	2.7088	0.182	14.892	0.000	2.352	3.065



	Features	VIF
0	const	9.22
6	Lead Source_Olark Chat	1.85
9	Specialization_Not_specified	1.83
2	TotalVisits	1.68
5	Lead Origin_Lead Add Form	1.61
4	A free copy of Mastering The Interview	1.43
3	Total Time Spent on Website	1.25
7	Lead Source_Welingak Website	1.25
10	What is your current occupation_Working Profes...	1.12
1	Do Not Email	1.01
8	Specialization_Hospitality Management	1.01

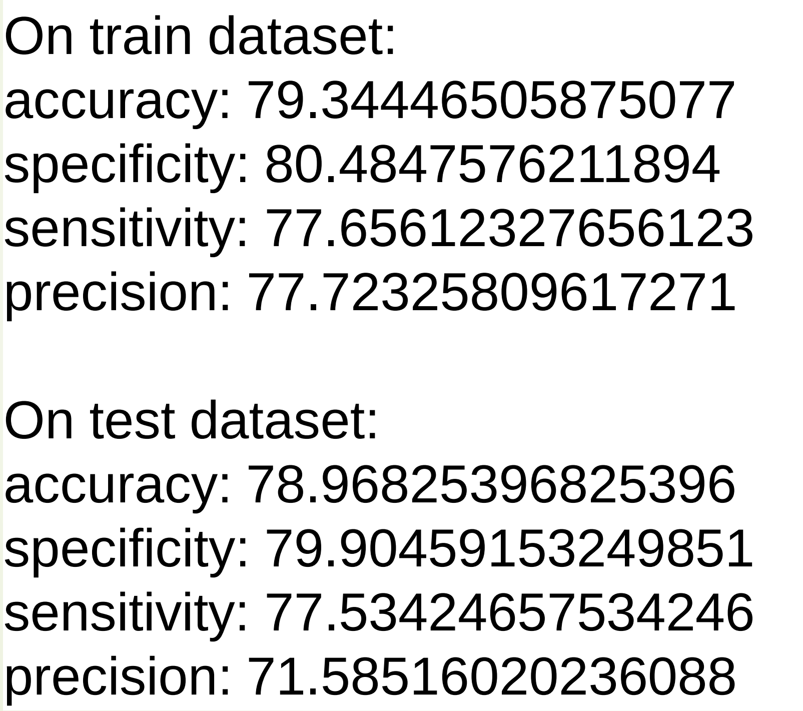
ROC CURVE AND OPTIMAL VALUE





- From the curve optimal value is taken as 0.3

EVALUATION OF THE MATRICES ON TEST AND TRAIN DATASET



```
On train dataset:
accuracy: 79.34446505875077
specificity: 80.4847576211894
sensitivity: 77.65612327656123
precision: 77.72325809617271

On test dataset:
accuracy: 78.96825396825396
specificity: 79.90459153249851
sensitivity: 77.53424657534246
precision: 71.5851602023608
```

CONCLUSIONS

It is found that the variables which contributes to conversion of the leads are:

- ➢ Total time spent on website.
- ➢ Lead Origin.
  - * Lead Add Form
- ➢ What is your current occupation.
  - * Working Professional