Lead Score Case Study

Abhilasha Garg

Priyanka Vashisht

Lead Score Case Study

Problem Statement

Data Cleaning and Analysis

Data Transformation

Model Building

Hot Leads

Model Evaluation

Model Prediction

Accuracy Score

Lead Score

Score Prediction

Conclusion

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

- Build a logistic regression model to 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
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Approach to case study

- Read and Understand the Data
- Data Cleaning & Preparation
- Exploratory Data Analysis (EDA)
- Prepare the data for modelling
- Model Building
- Model Evaluations
- Making predictions on test set
- Optimalization and Accuracy Score

Data Cleaning

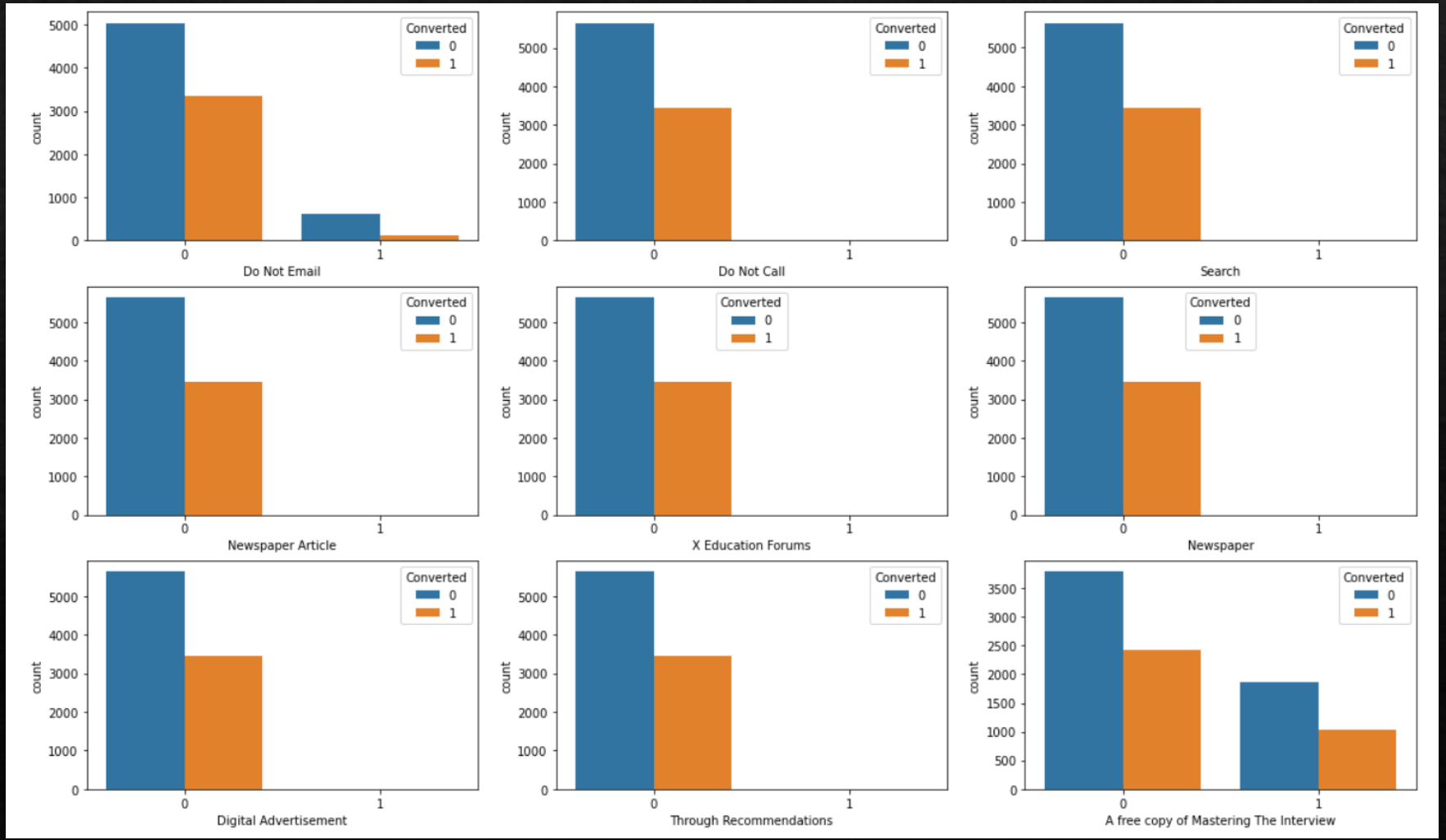
- Dropping the columns having only one unique value.
- Replacing ‘Select’ with not provided.
- Dropping columns having more than 35% missing values.
- Dropping columns that would not be in use for analysis.
- Checking & Removing Null values

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9103 entries, 0 to 9239
Data columns (total 18 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin                             9103 non-null   object
 1   Lead Source                             9103 non-null   object
 2   Do Not Email                            9103 non-null   object
 3   Do Not Call                             9103 non-null   object
 4   Converted                               9103 non-null   int64
 5   Total Time Spent on Website             9103 non-null   int64
 6   Last Activity                           9103 non-null   object
 7   Country                                 9103 non-null   object
 8   Specialization                          9103 non-null   object
 9   What is your current occupation         9103 non-null   object
 10  Search                                  9103 non-null   object
 11  Newspaper Article                       9103 non-null   object
 12  X Education Forums                      9103 non-null   object
 13  Newspaper                               9103 non-null   object
 14  Digital Advertisement                   9103 non-null   object
 15  Through Recommendations                 9103 non-null   object
 16  A free copy of Mastering The Interview  9103 non-null   object
 17  Last Notable Activity                   9103 non-null   object
dtypes: int64(2), object(16)
memory usage: 1.3+ MB
```

```
Lead Origin                               0.0
Lead Source                               0.0
Do Not Email                              0.0
Do Not Call                               0.0
Converted                                 0.0
Total Time Spent on Website               0.0
Last Activity                             0.0
Country                                   0.0
Specialization                            0.0
What is your current occupation           0.0
Search                                    0.0
Newspaper Article                         0.0
X Education Forums                        0.0
Newspaper                                 0.0
Digital Advertisement                     0.0
Through Recommendations                   0.0
A free copy of Mastering The Interview    0.0
Last Notable Activity                     0.0
dtype: float64
```

EDA – Exploratory Data Analysis

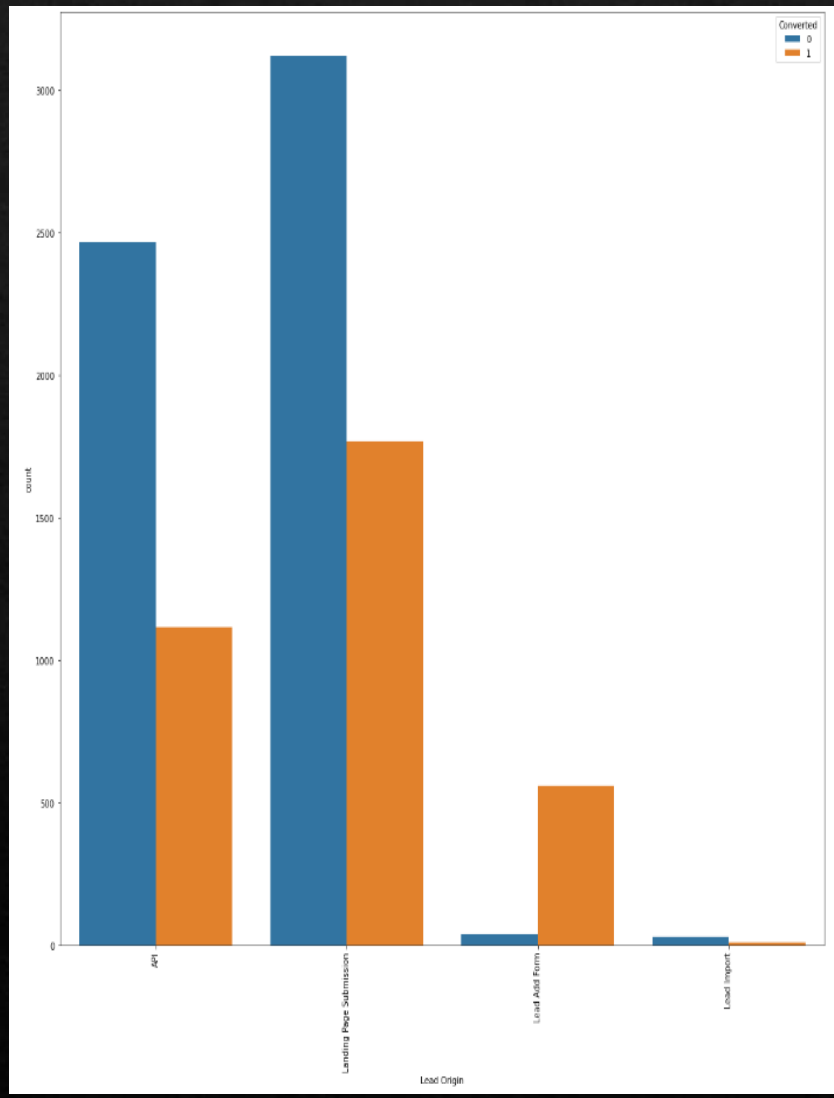
- Bivariate Analysis

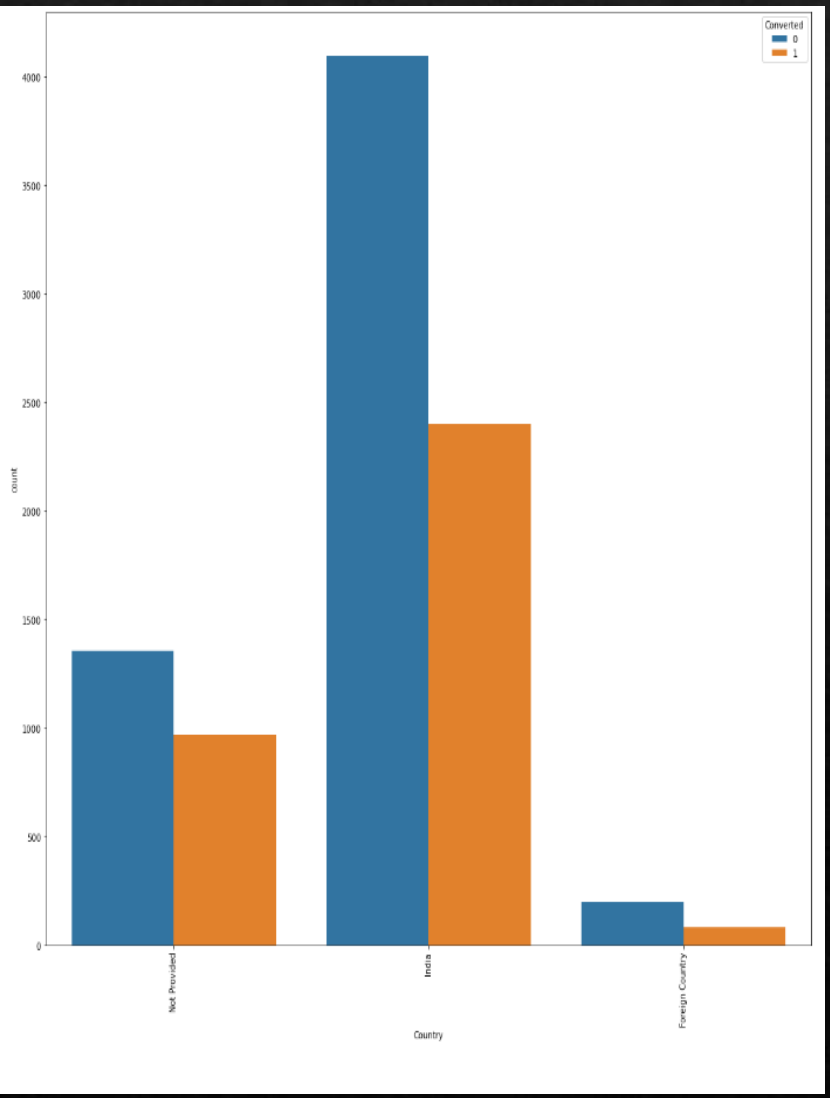


EDA – Exploratory Data Analysis

Inference/ results

- API and Landing Page Submission have higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form have very few leads.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

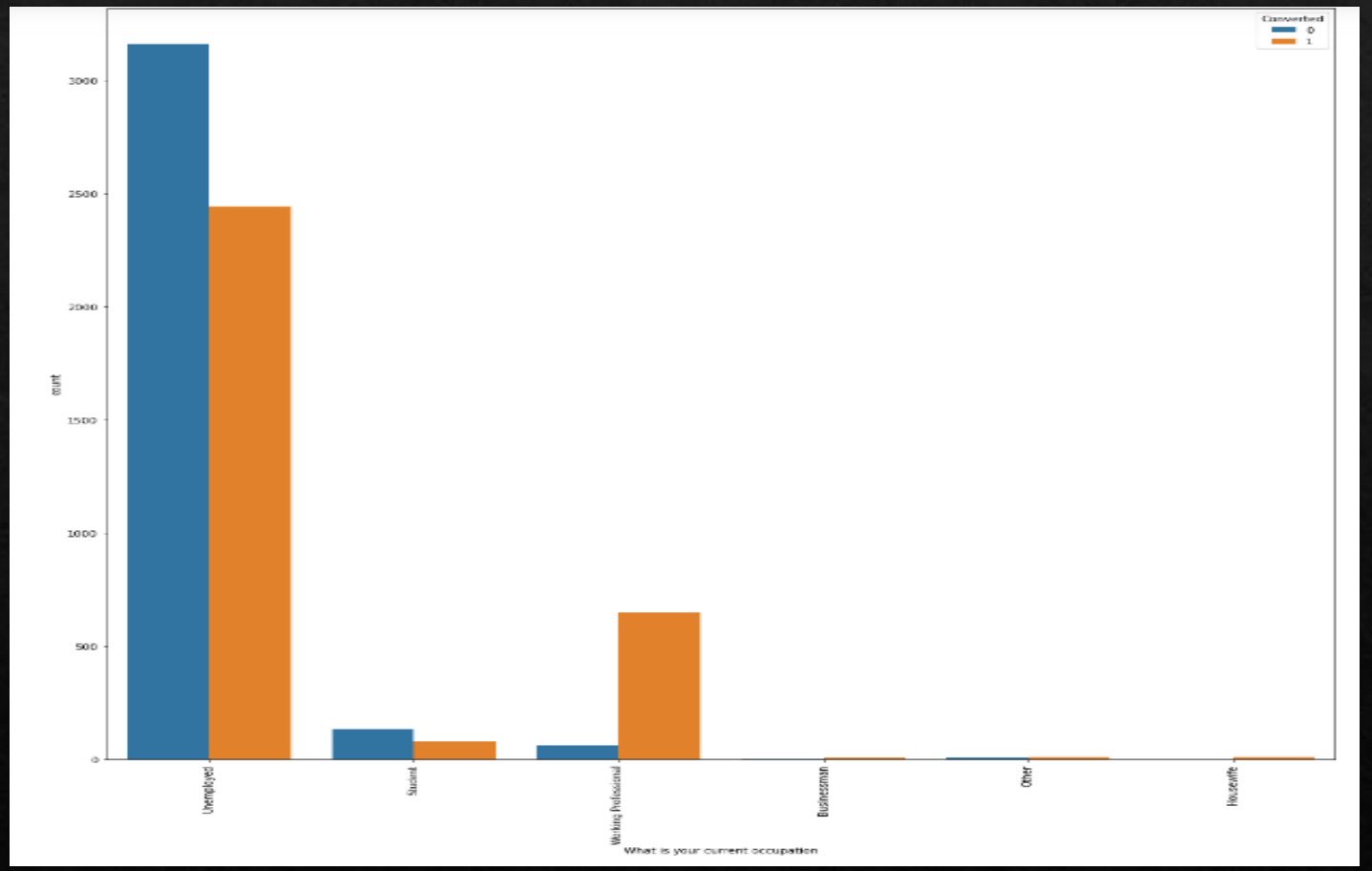




EDA – Exploratory Data Analysis

Inference/ results

- Unemployed people are more inclined towards converting into leads



EDA – Exploratory Data Analysis

- Multivariate Analysis:

Model Building – Logistic regression model

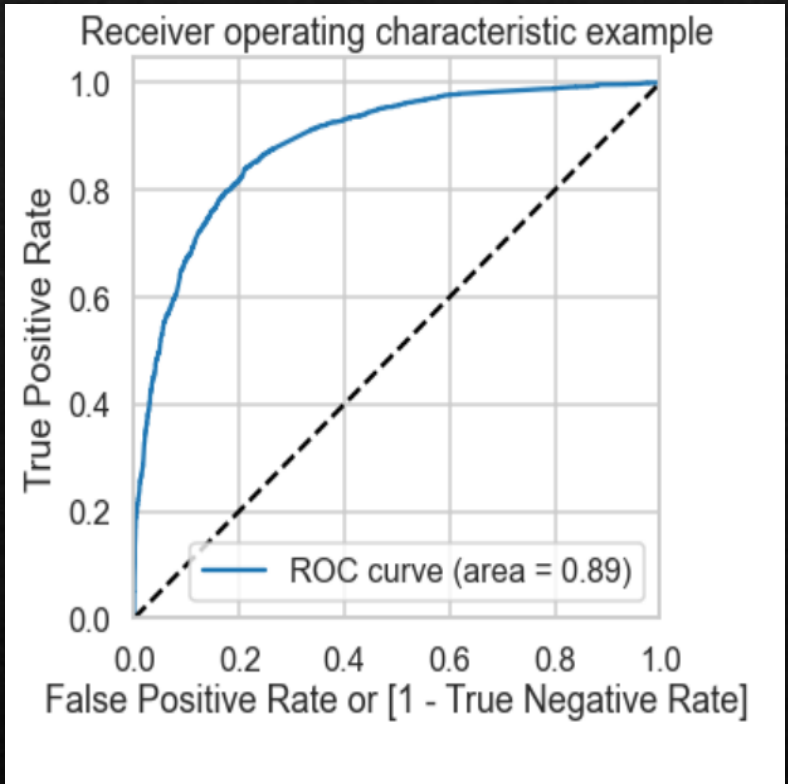
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6361
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2618.5
Date:	Sun, 12 Jun 2022	**Deviance:**	5236.9
Time:	09:38:31	**Pearson chi2:**	6.79e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

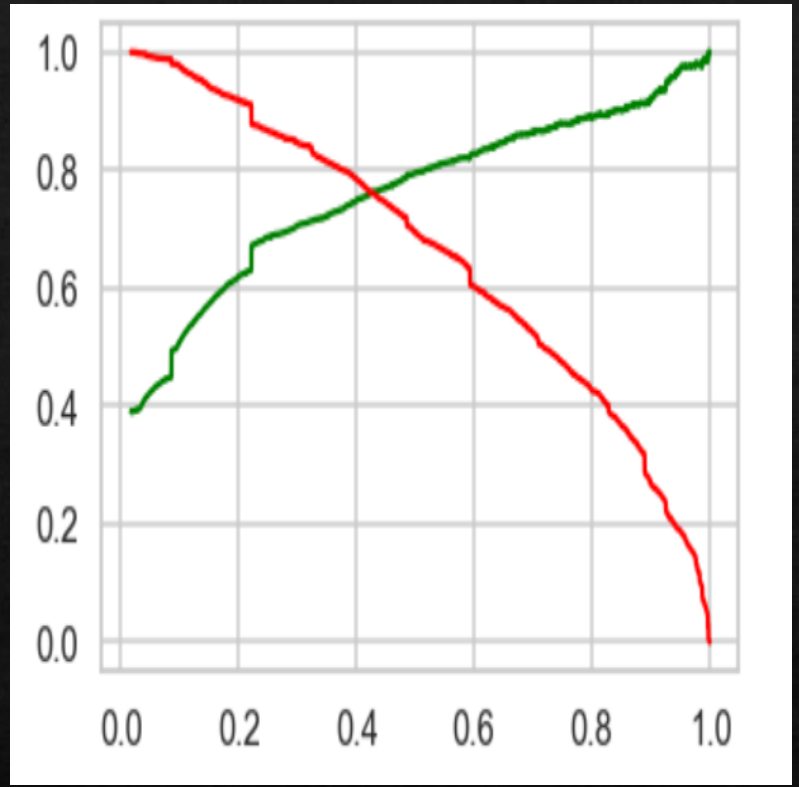
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1470	0.126	-9.111	0.000	-1.394	-0.900
Do Not Email	-1.2450	0.165	-7.545	0.000	-1.568	-0.922
Total Time Spent on Website	4.6072	0.167	27.534	0.000	4.279	4.935
Lead Origin_Landing Page Submission	-1.1171	0.122	-9.134	0.000	-1.357	-0.877
Lead Origin_Lead Add Form	2.1394	0.229	9.345	0.000	1.691	2.588
Country_Not Provided	1.0872	0.119	9.123	0.000	0.854	1.321
Specialization_Not Provided	-1.1977	0.124	-9.662	0.000	-1.441	-0.955
What is your current occupation_Not Provided	-1.1096	0.087	-12.713	0.000	-1.281	-0.938
What is your current occupation_Working Professional	2.2393	0.190	11.785	0.000	1.867	2.612
Last Notable Activity_SMS Sent	1.6314	0.080	20.443	0.000	1.475	1.788
Lead Source_Welingak Website	3.1317	1.033	3.031	0.002	1.107	5.157

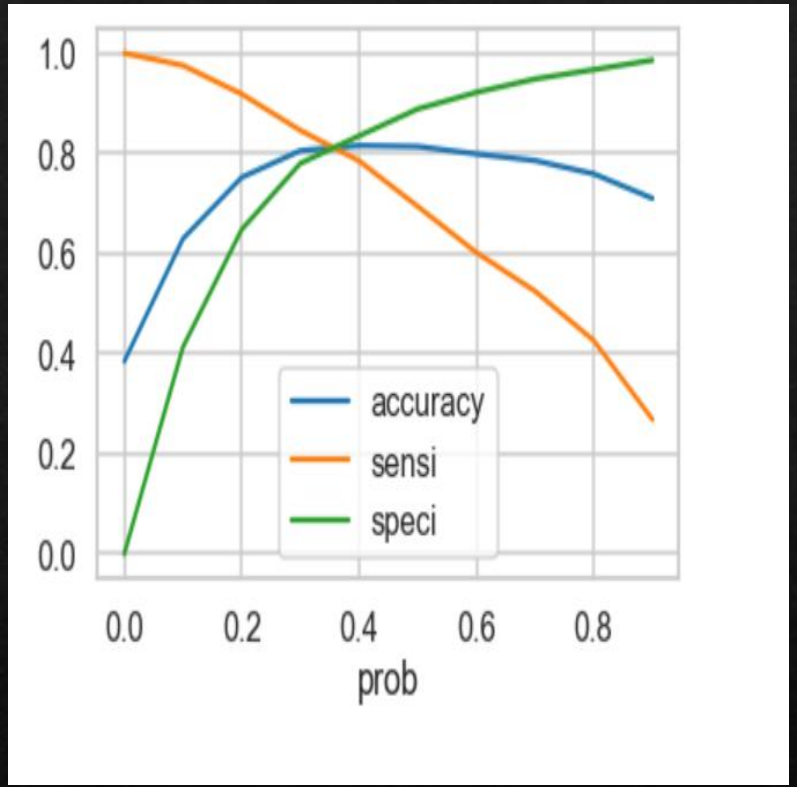
	Features	VIF
4	Country_Not Provided	2.56
5	Specialization_Not Provided	2.44
2	Lead Origin_Landing Page Submission	2.21
1	Total Time Spent on Website	2.04
3	Lead Origin_Lead Add Form	1.86
6	What is your current occupation_Not Provided	1.61
8	Last Notable Activity_SMS Sent	1.40
9	Lead Source_Welingak Website	1.34
7	What is your current occupation_Working Profes...	1.21
0	Do Not Email	1.11

Optimization & Accuracy



ROC Curve





accuracy sensitivity and specificity

Points to be concluded from above curves -

- The curve is closer to the left side of the border than to the right side hence our model is having great accuracy.
- The area under the curve is 89% of the total area.
- optimal cut off is at 0.38

Lead Score Prediction

	Lead ID	Converted	Converted_Prob	final_predicted	Lead Score
0	3497	1	0.888899	1	89
1	4044	1	0.828010	1	83
2	7200	0	0.115272	0	12
3	1183	0	0.221400	0	22
4	8216	0	0.032840	0	3
5	8746	1	0.221400	0	22
6	9047	0	0.058941	0	6
7	6526	1	0.798730	1	80
8	7694	0	0.085717	0	9
9	8184	1	0.902206	1	90

Conclusion

- With Cut Off of 0.41 , we got Accuracy Score of 81.3 % on Train data and 80.9% on Test Data set
- We got Precision – 75% and Recall – 77% on Train dataset and Precision – 73% , Recall – 76% on Test dataset
- Feature impacting positively : Total time spent on website , Lead Source – Welingak website, Current Occupation – Working Professional , Lead Origin – Lead Add Form , and Last Notable Activity – SMS Sent
- Feature Negatively Impacting : Do not Email , Specialization – Not provided, Current Occupation – Not provided , Lead Origin – Landing Page Submission

Thank You