



# LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

SUBMITTED BY:




UJWAL.K

SHASHWAT GAONKAR

ZEESHAN MAINDARGI



# CONTENTS:

1. Problem statement
  2. Problem approach
  3. EDA (Data cleaning and Data preparation)
  4. Model Evaluation
  5. Observations
  6. Conclusion
- 
- 
- 

# PROBLEM STATEMENT :

1. 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. They have process of form filling on their website after which the company that individual as a lead.
2. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
3. The typical lead conversion rate at X education is around 30%. Now, this means 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.
4. 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:

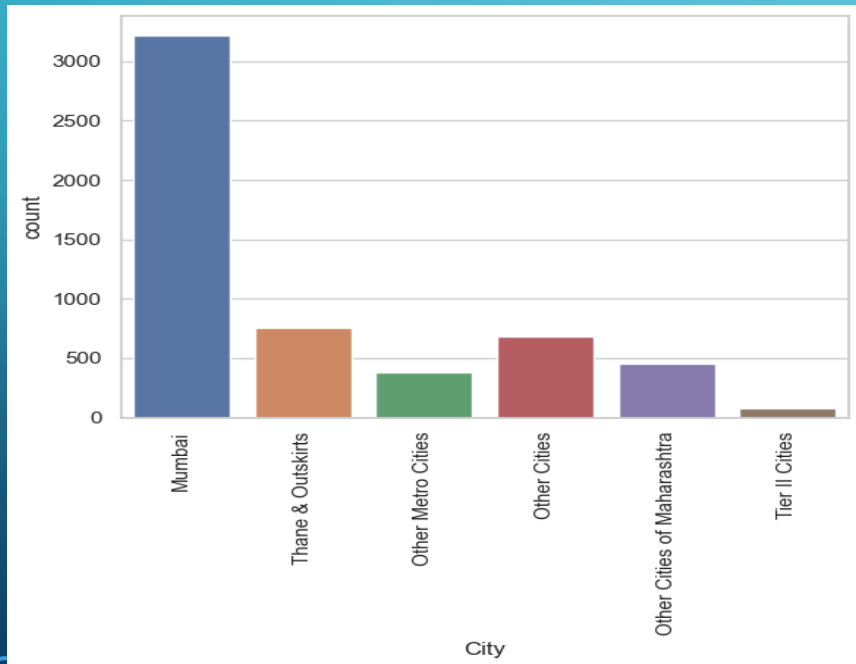
- Lead X wants us to build a model to give every lead a lead score between 0 -100 .So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

# PROBLEM APPROACH:

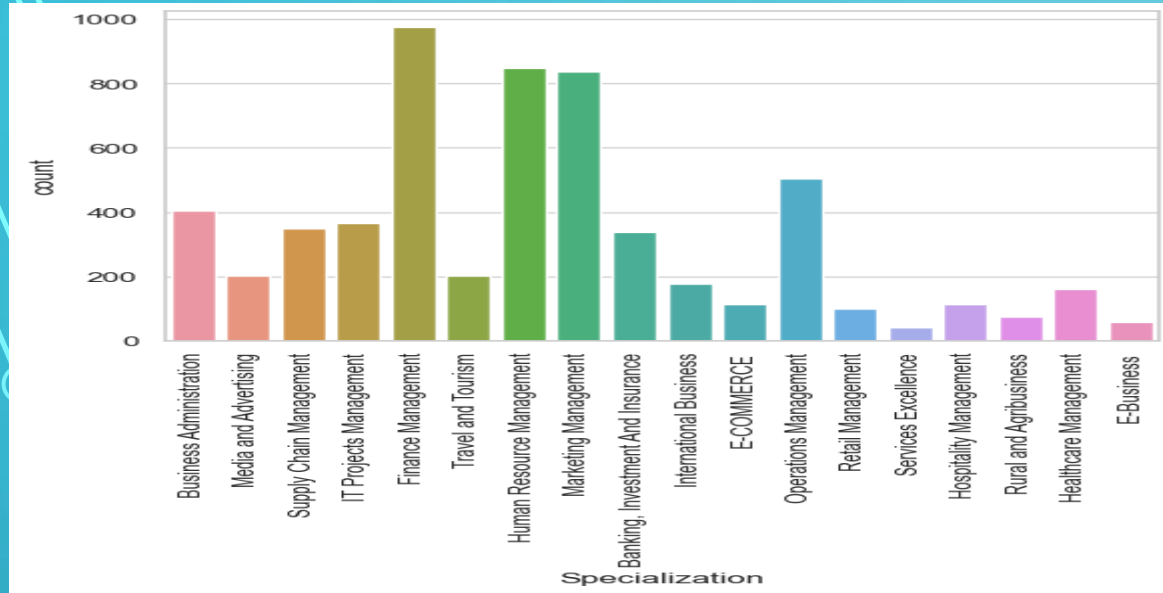
- Importing the data and inspecting the data frame
- Data preparation
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Model Building (RFE (feature selection), VIF and p values)
- Model Evaluation
- Making predictions on test set

# EDA

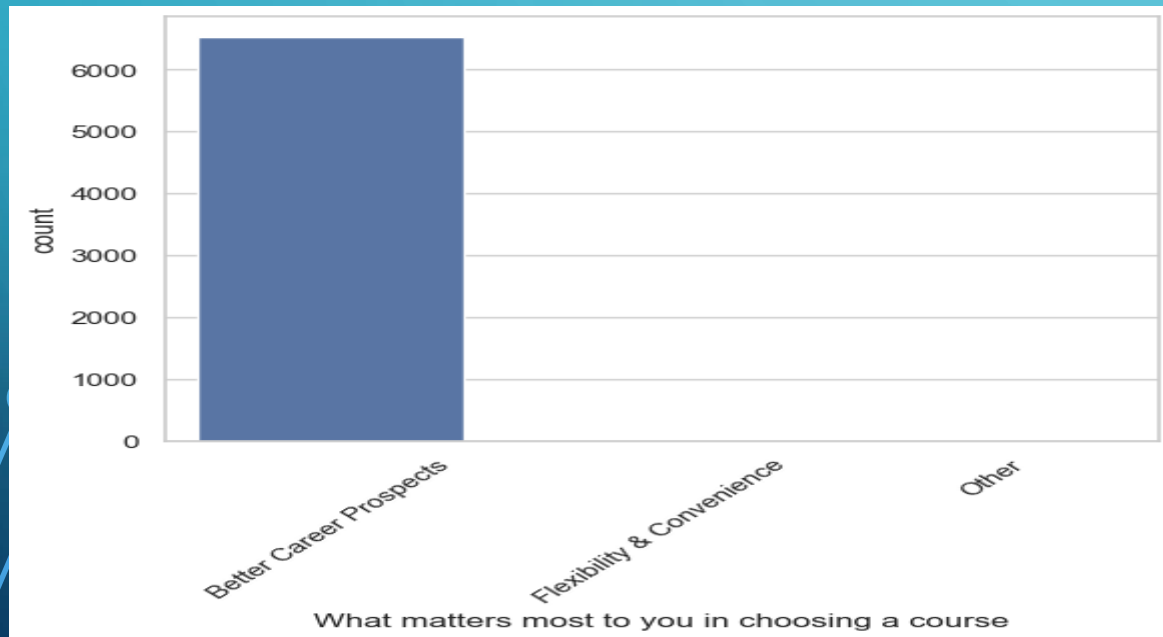
- Columns that has more than 40% missing values and that are dropped: 'How did you hear about X Education','Lead Quality','Lead Profile','Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score'.



- Missing values in city , are replaced with Mumbai, since Mumbai has highest count.

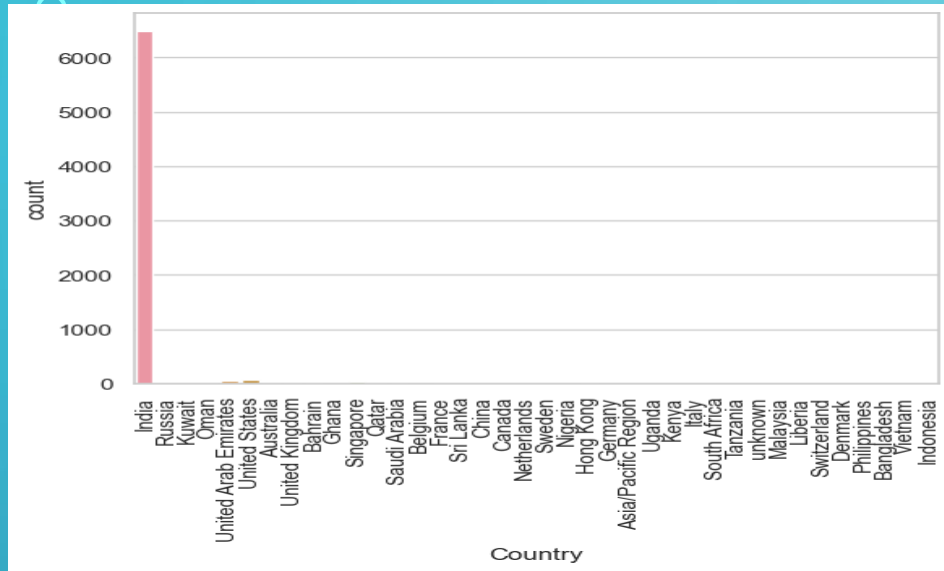


Column specialization has 37% missing values, so missing values are imputed with “other”.

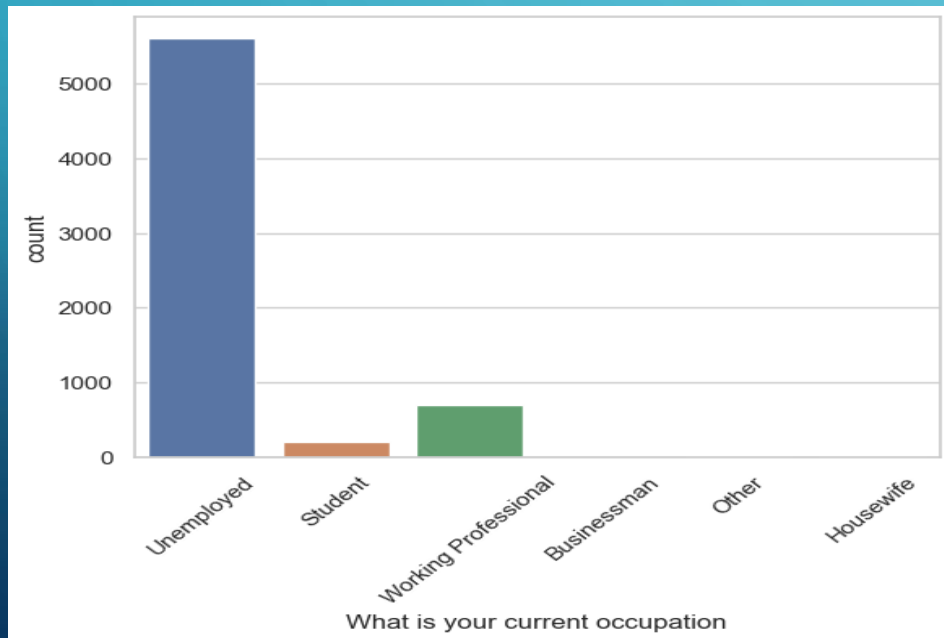


Column “What matters most to you in choosing a course” has all the values in “Better career prospects” category so which will not be usefull for the model so the column has been dropped.





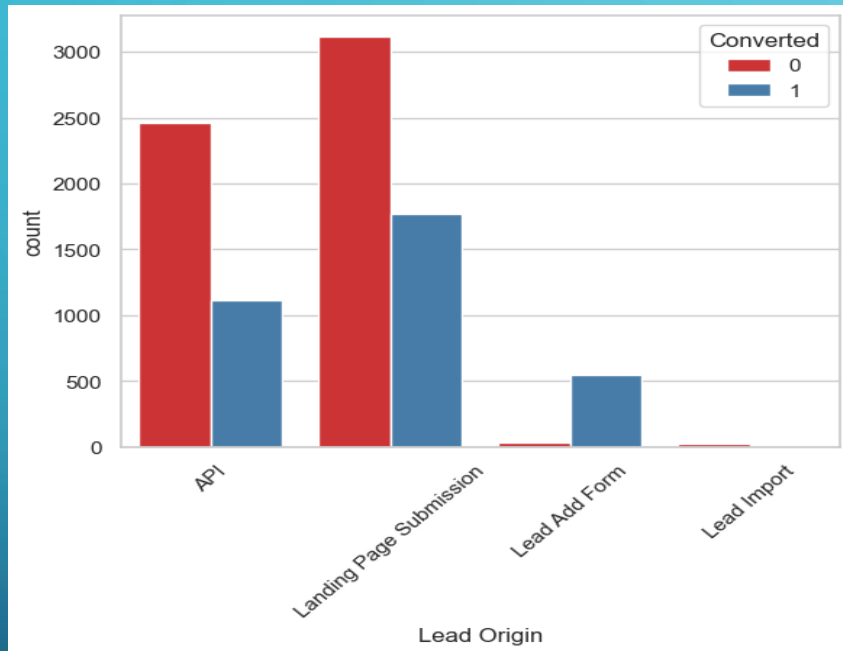
The missing values in Country column is imputed with “India” since “India” has highest count.



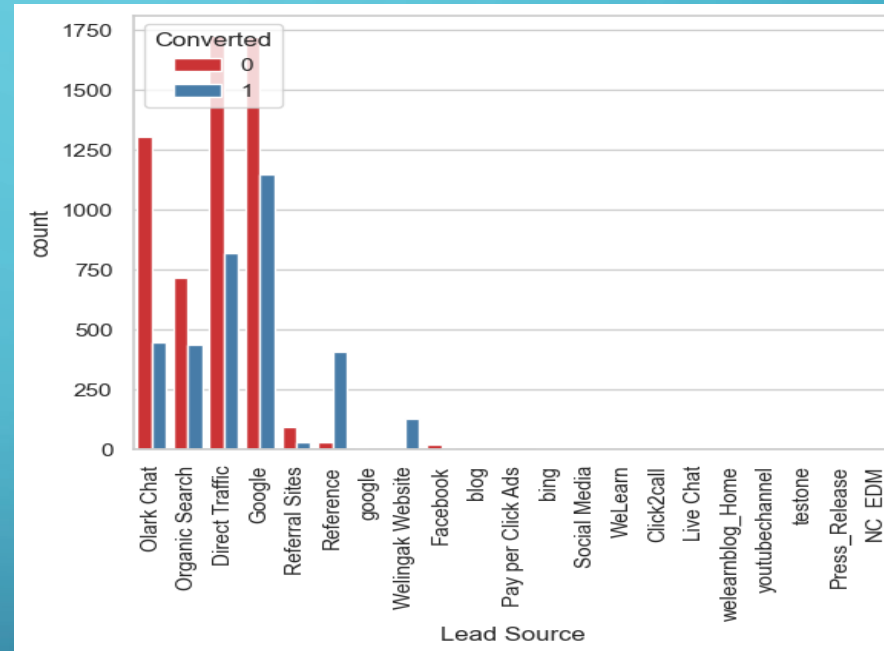
The missing values in column “What is your current occupation” is replaced with “Unemployed”.



# UNIVARIATE AND BIVARIATE ANALYSIS

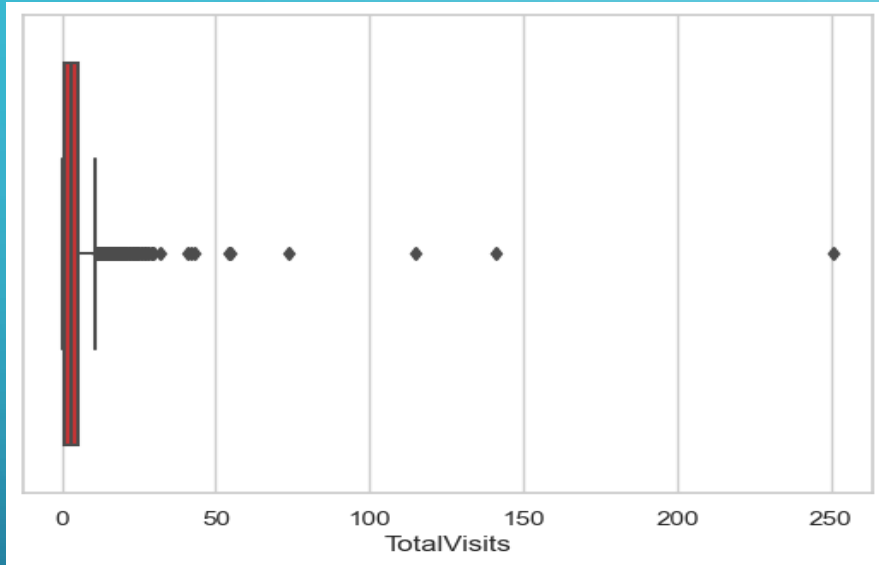


Lead origin most number of leads are landing on submission

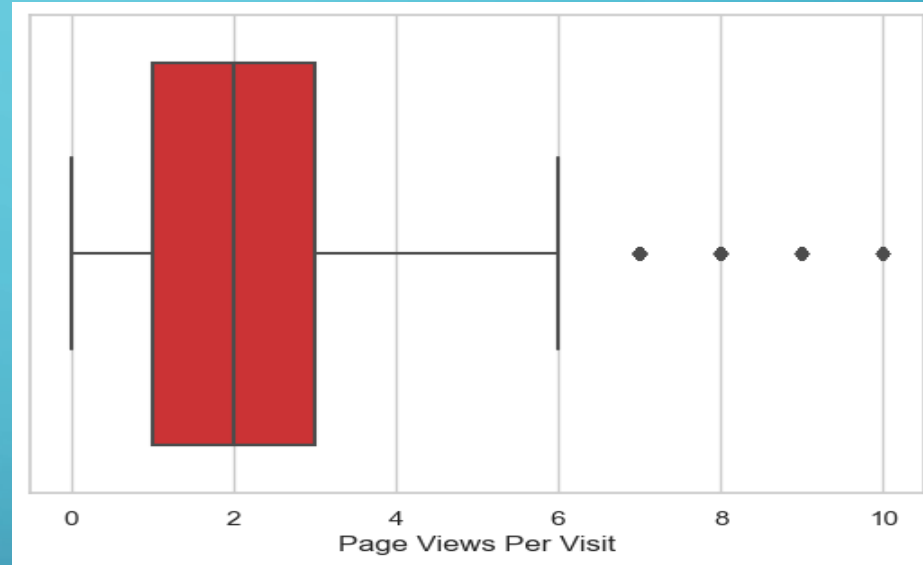


The lead source welingak website and reference has highest conversion rate.

# OUTLIER CHECK



There are outliers in “Total visits” we can cap it to 95%



There are outliers we can cap it to 95%.

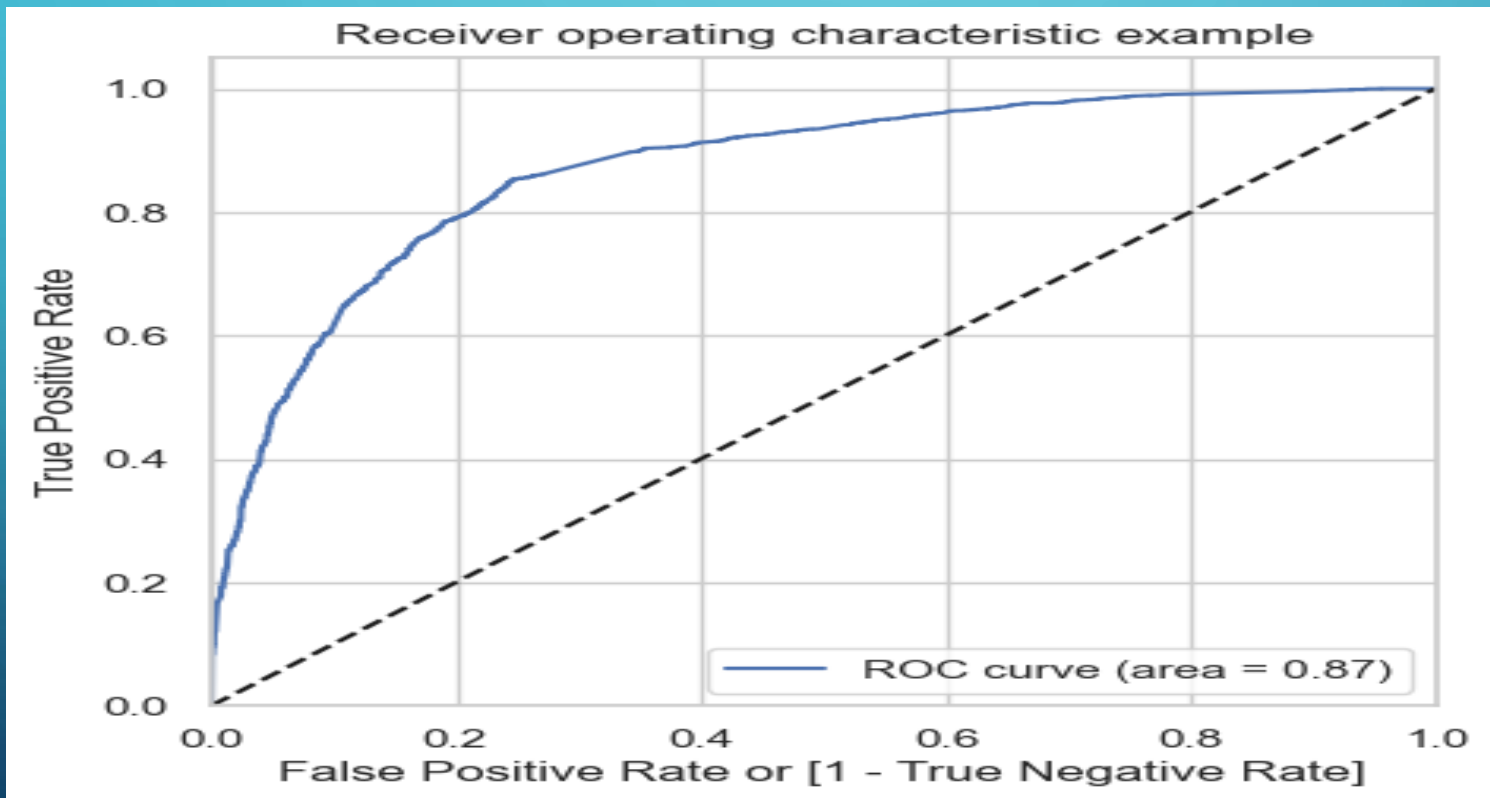
# FINAL MODEL DETAILS

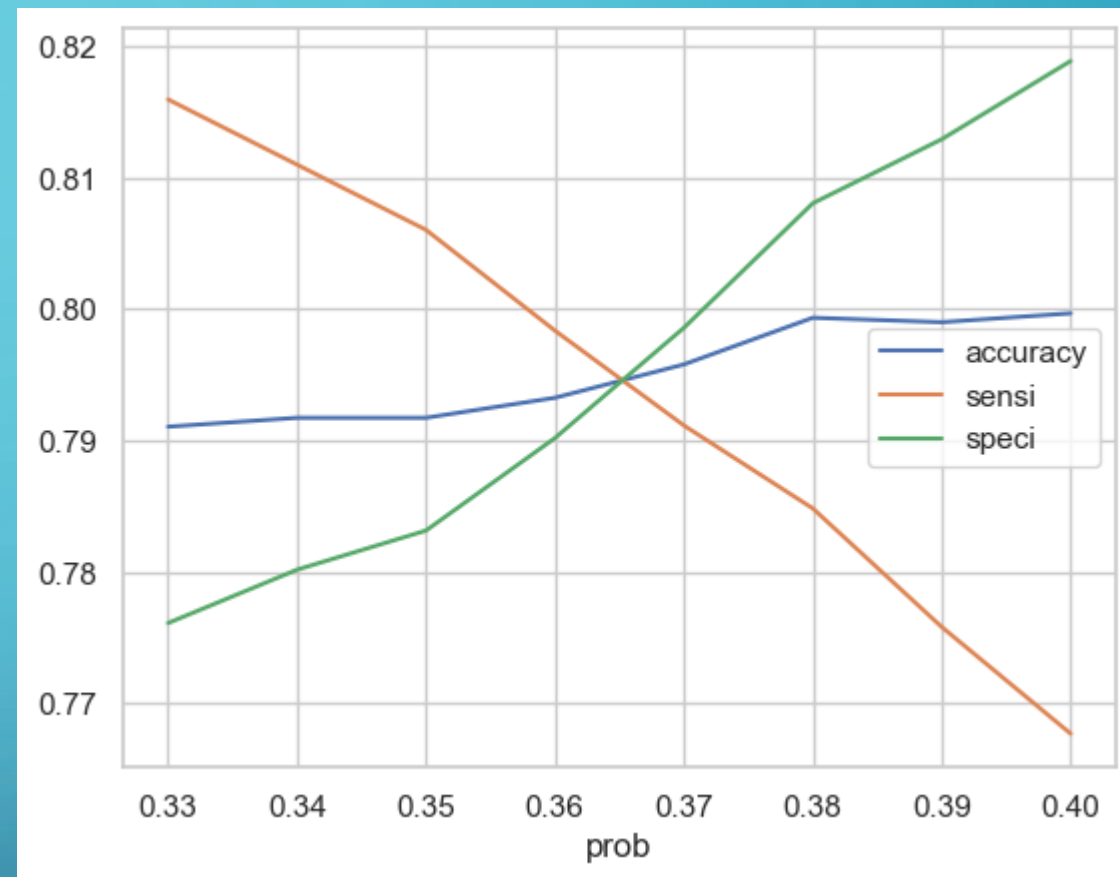
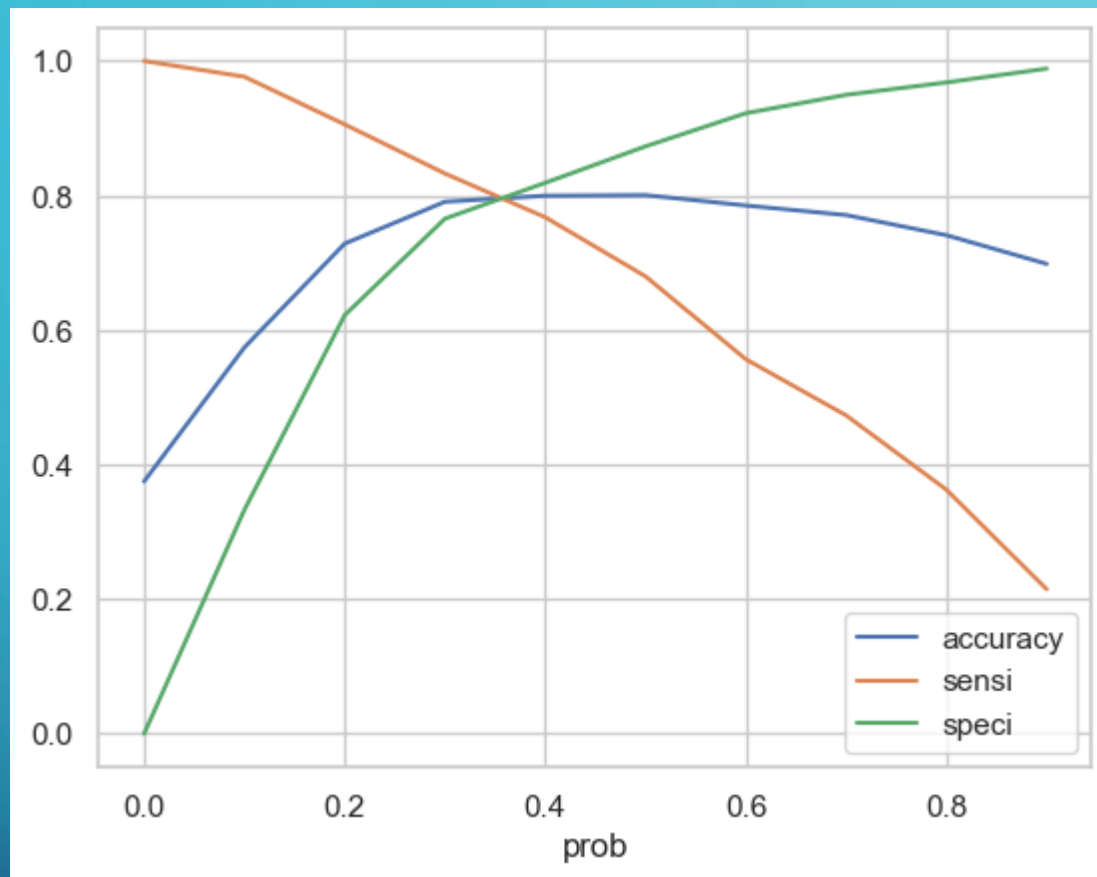
	coef	std err	z	P> z	[0.025	0.975]
<b>const</b>	-0.4084	0.136	-3.009	0.003	-0.675	-0.142
<b>Do Not Email</b>	-1.4525	0.198	-7.353	0.000	-1.840	-1.065
<b>Total Time Spent on Website</b>	1.0460	0.040	25.946	0.000	0.967	1.125
<b>Lead Origin_Landing Page Submission</b>	-1.2789	0.128	-9.989	0.000	-1.530	-1.028
<b>Lead Source_Olark Chat</b>	1.0007	0.123	8.166	0.000	0.761	1.241
<b>Lead Source_Reference</b>	3.4156	0.241	14.191	0.000	2.944	3.887
<b>Lead Source_Welingak Website</b>	6.3704	1.021	6.237	0.000	4.369	8.372
<b>Last Activity_Email Opened</b>	0.5520	0.093	5.951	0.000	0.370	0.734
<b>Last Activity_Olark Chat Conversation</b>	-1.1150	0.185	-6.035	0.000	-1.477	-0.753
<b>Last Activity_Other_Activity</b>	2.5932	0.634	4.089	0.000	1.350	3.836
<b>Last Activity_Unsubscribed</b>	1.5807	0.455	3.474	0.001	0.689	2.472
<b>Specialization_Others</b>	-1.4589	0.125	-11.716	0.000	-1.703	-1.215
<b>Last Notable Activity_Email Bounced</b>	1.4783	0.497	2.976	0.003	0.505	2.452
<b>Last Notable Activity_SMS Sent</b>	1.9652	0.101	19.379	0.000	1.766	2.164
<b>Last Notable Activity_Unreachable</b>	2.2357	0.534	4.184	0.000	1.188	3.283

	Features	VIF
<b>10</b>	Specialization_Others	2.39
<b>3</b>	Lead Source_Olark Chat	2.29
<b>2</b>	Lead Origin_Landing Page Submission	2.28
<b>6</b>	Last Activity_Email Opened	2.17
<b>12</b>	Last Notable Activity_SMS Sent	1.77
<b>7</b>	Last Activity_Olark Chat Conversation	1.75
<b>0</b>	Do Not Email	1.37
<b>1</b>	Total Time Spent on Website	1.31
<b>4</b>	Lead Source_Reference	1.21
<b>11</b>	Last Notable Activity_Email Bounced	1.11
<b>9</b>	Last Activity_Unsubscribed	1.10
<b>5</b>	Lead Source_Welingak Website	1.08
<b>8</b>	Last Activity_Other_Activity	1.01
<b>13</b>	Last Notable Activity_Unreachable	1.01

# MODEL EVALUATION

## ROC Curve





Trade off between accuracy, sensitivity and specificity which comes out to be 0.365.

# OBSERVATIONS:

## Train data:

Sensitivity :79.4%

Specificity :79.4%

Accuracy :80.0%

## Test data:

Sensitivity :81.8%

Specificity :79.7%

Accuracy : 80.5%

## FINAL FEATURE LIST WITH THEIR COEFFICIENTS

Lead Source_Welingak Website	6.370434
Lead Source_Reference	3.415614
Last Activity_Other_Activity	2.593184
Last Notable Activity_Unreachable	2.235664
Last Notable Activity_SMS Sent	1.965153
Last Activity_Unsubscribed	1.580658
Last Notable Activity_Email Bounced	1.478256
Total Time Spent on Website	1.045983
Lead Source_Olark Chat	1.000745
Last Activity_Email Opened	0.552007
const	-0.408436
Last Activity_Olark Chat Conversation	-1.115020
Lead Origin_Landing Page Submission	-1.278882
Do Not Email	-1.452472
Specialization_Others	-1.458854
dtype: float64	



# CONCLUSION:

- We can see that maximum conversions happens in Lead Source “Welingak Website” and “Reference”.
- Leads who spent more time on website, more likely to convert.
- Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with working professional.
- Company should not make calls to leads coming from "Lead Source\_Olark Chat","Last Activity\_Email Opened","Last Activity\_Olark Chat Conversation","Lead Origin\_Landing Page Submission","Do Not Email" and "Specialization\_Others ".
- Company should make calls to leads coming from "Lead Source\_Welingak Website ","Lead Source\_Reference","Last Activity\_Other\_Activity","Last Notable Activity\_Unreachable ","Last Notable Activity\_SMS Sent ","Last Activity\_Unsubscribed","Last Notable Activity\_Email Bounced" and "Total Time Spent on Website".