LEAD SCORING CASE STUDY

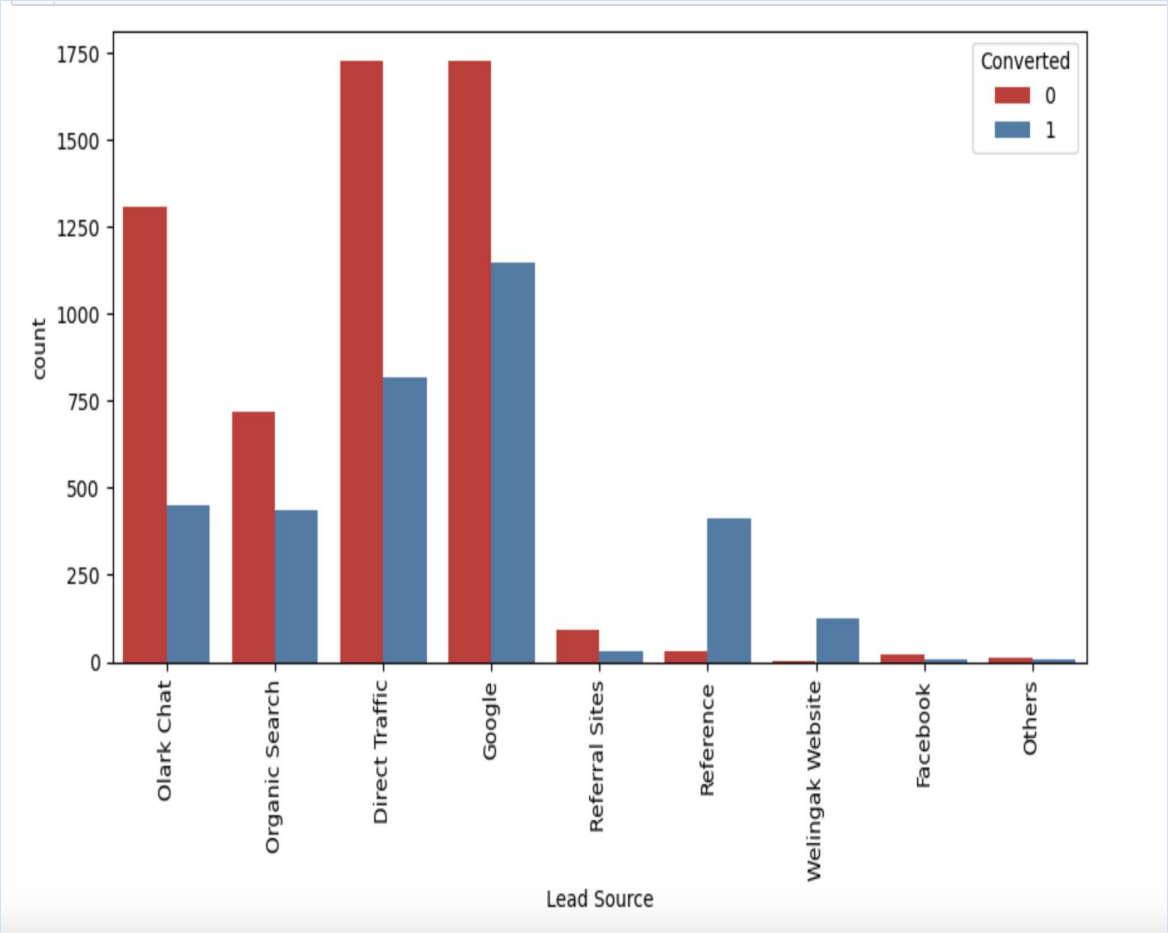
PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The leads are acquired through various mediums and the sales team start making calls, writing emails, etc. to convert the leads.
- The current conversion rate is 30 %. This indicates that most of the leads are not getting converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’
- The CEO wants the conversion rate to increase to 80 %

APPROACH FOR ANALYSIS AND MODELLING.

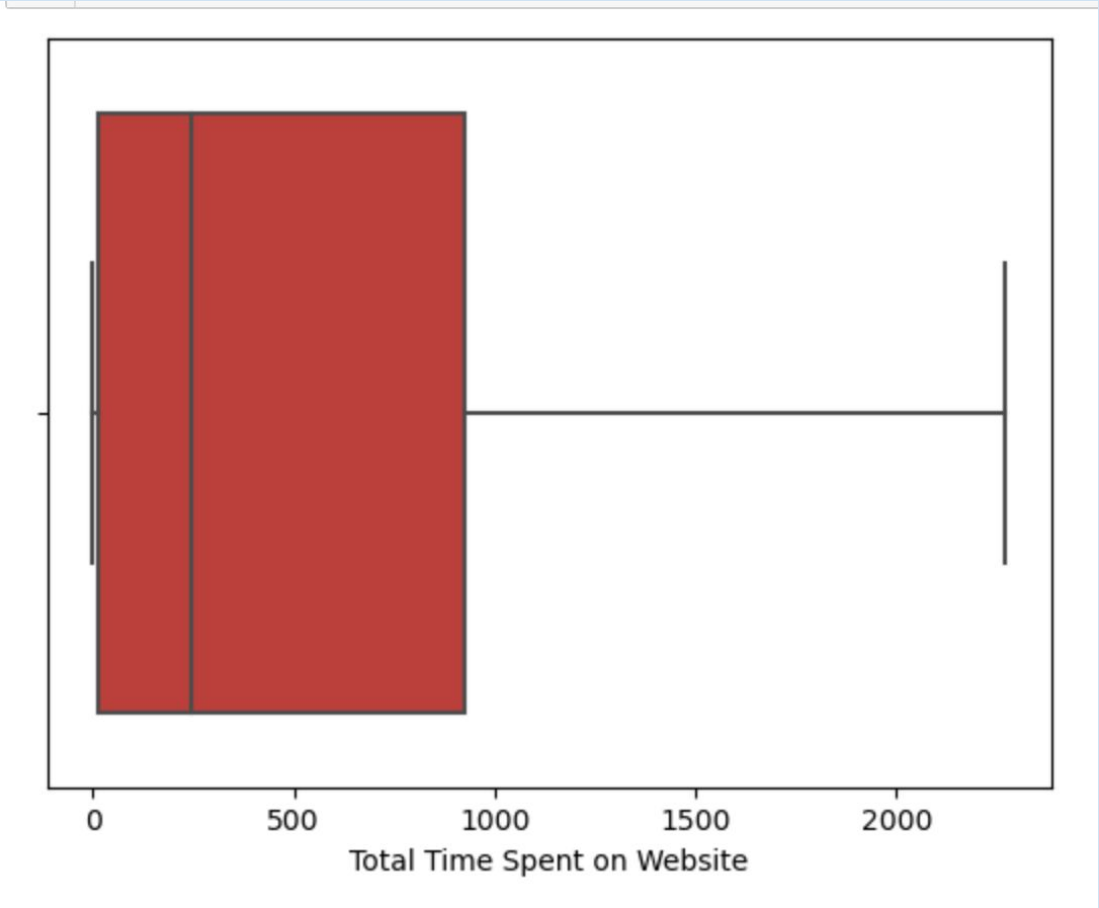
- Cleaning and Understanding the data
- (EDA) Exploratory Data Analysis for finding out most useful variable for conversion
- Preparing the data for model building
- Build the logistic Regression model
- Test the model on train and test dataset
- Evaluate the model with different measures and matrices
- Interpret the model and its parameters

EDA- LEAD SOURCE VS CONVERTED



- In lead source, categories such as “Wellingak website” and “reference” have high higher conversion numbers.
- Also the absolute value of count is also considerable.
- “Click2call” and “live chat” also have higher conversion numbers,
however the count value is very less.

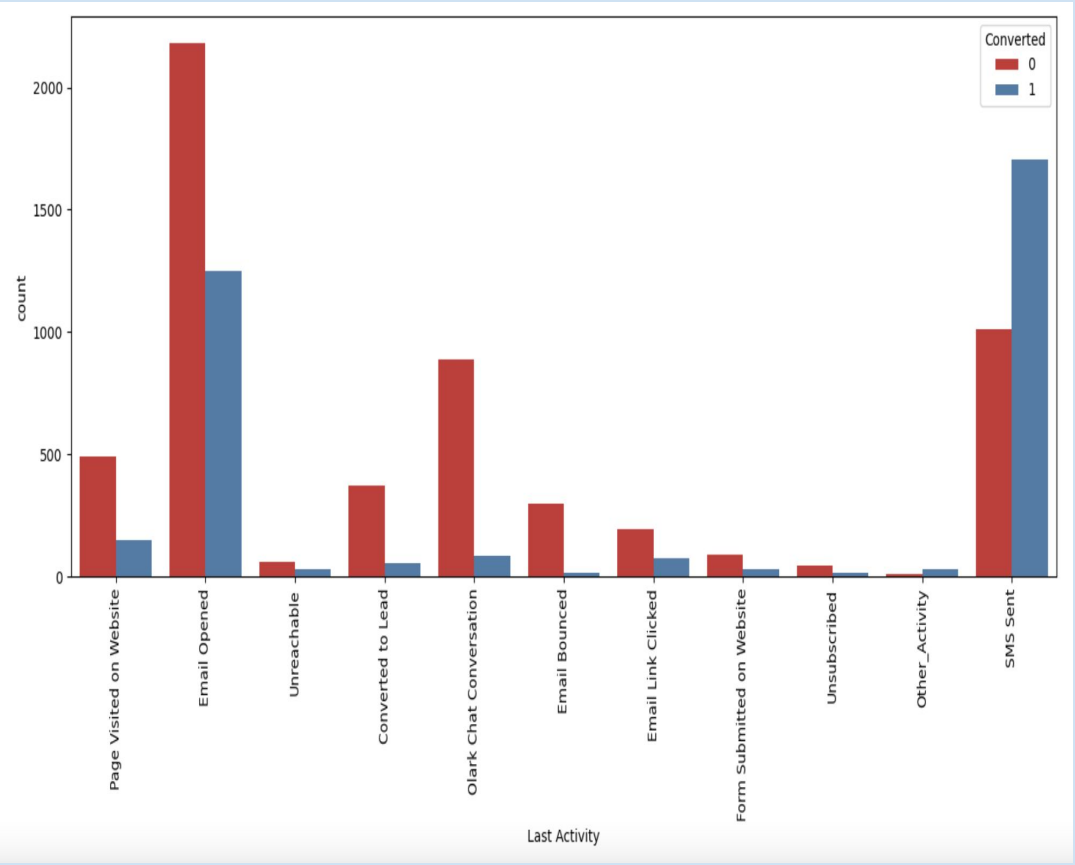
EDA-TOTAL TIME SPENT ON WEBSITE



Leads spending more time on the weblise are more likely to be converted.

Website should be made more engaging to make leads spend more time.

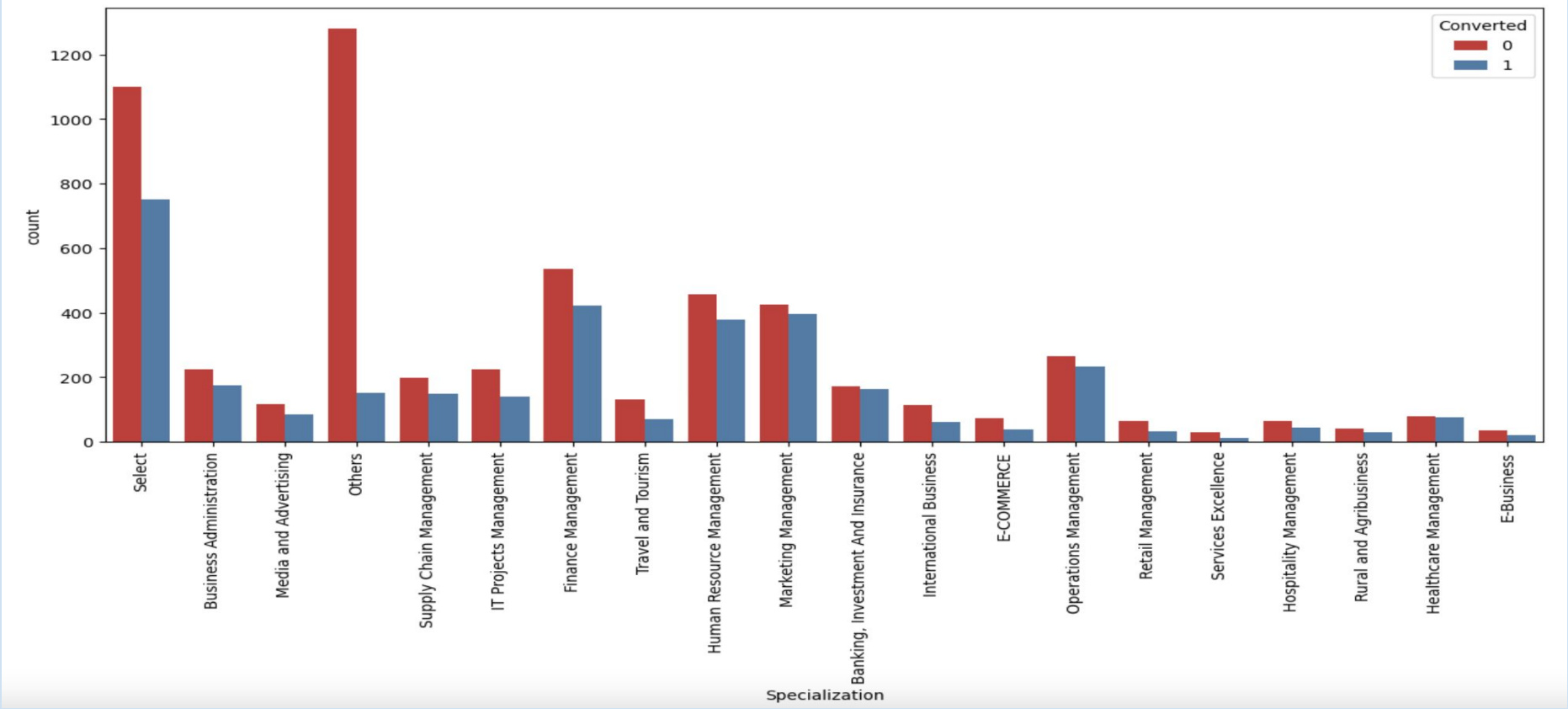
EDA- LAST ACTIVITY VS CONVERTED



Most of the lead have their Email opened as their last activity.

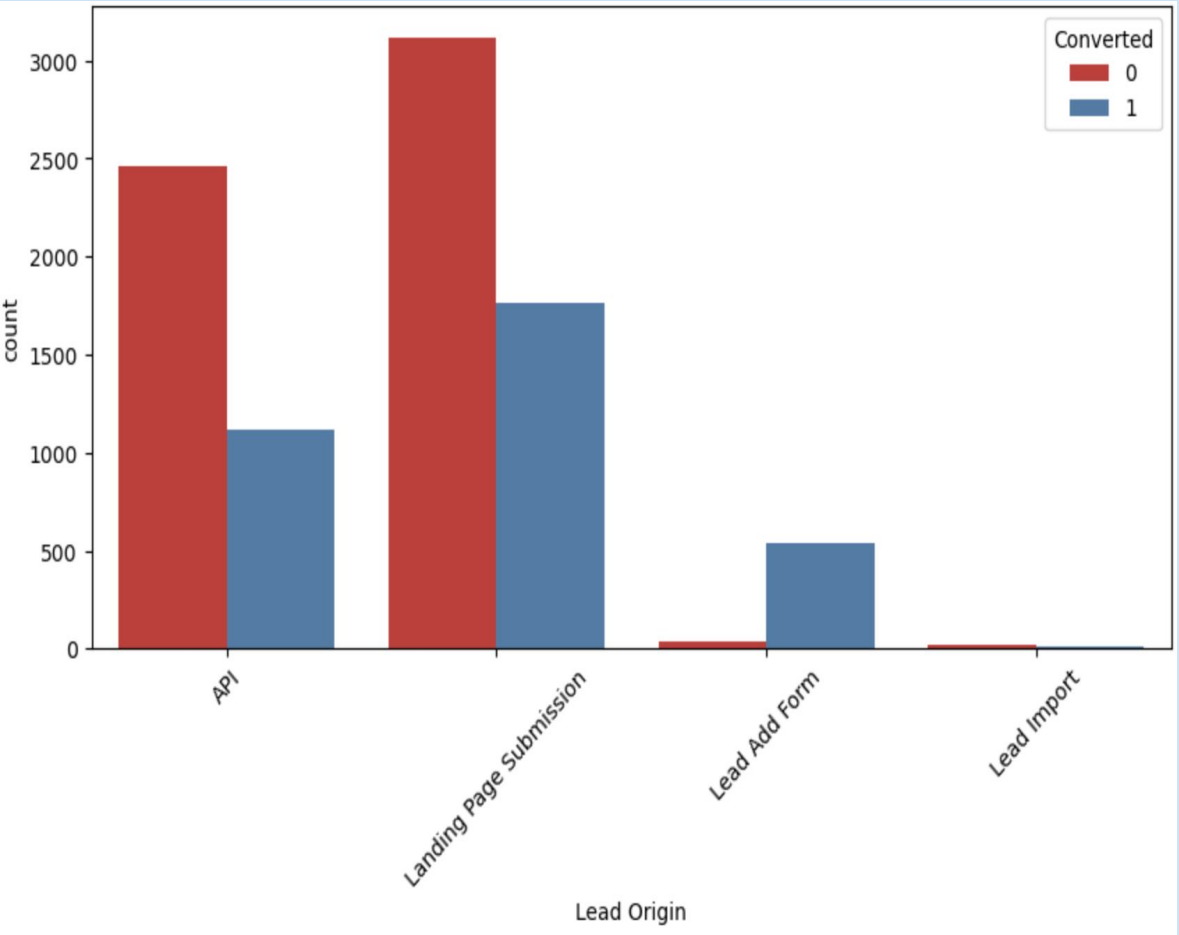
Conversion rate for leads with last activity as SMS Sent is almost 60%.

EDA -SPECIALIZATION VS CONVERTED



Focus should be more on the Specialization with high conversion rate.

EDA – LEAD ORIGIN VS CONVERTED



- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead converSion of API and Landing Page
- Submission origin and generate more leads from Lead Add Form.

DATA CONSIDERATION FOR MODEL BUILDING USING RFE

```
['Do Not Email', 'Total Time Spent on Website',
 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
 'Lead Source_Olark Chat', 'Lead Source_Reference',
 'Lead Source_Welingak Website', 'Last Activity_Email Opened',
 'Last Activity_Other_Activity', 'Last Activity_SMS Sent',
 'Last Activity_Unsubscribed', 'Specialization_Others',
 'What is your current occupation_Housewife',
 'What is your current occupation_Student',
 'What is your current occupation_Unemployed',
 'What is your current occupation_Working Professional', 'City_Select',
 'Last Notable Activity_Modified',
 'Last Notable Activity_Olark Chat Conversation',
 'Last Notable Activity_Unreachable'],
dtype='object')
```

- These are the columns which we get after RFE , we can proceed with these columns for building model

MODEL PARAMETERS AND EVALUATION METRICS

	Features	VIF
2	Lead Origin_Landing Page Submission	3.26
6	Last Activity_Email Opened	2.24
8	Last Activity_SMS Sent	2.21
11	City_Select	2.09
3	Lead Source_Olark Chat	1.98
9	Specialization_Others	1.87
12	Last Notable Activity_Modified	1.81
4	Lead Source_Reference	1.36
1	Total Time Spent on Website	1.30
10	What is your current occupation_Working Profes...	1.19
0	Do Not Email	1.18
5	Lead Source_Welingak Website	1.11
7	Last Activity_Other_Activity	1.02
13	Last Notable Activity_Unreachable	1.01

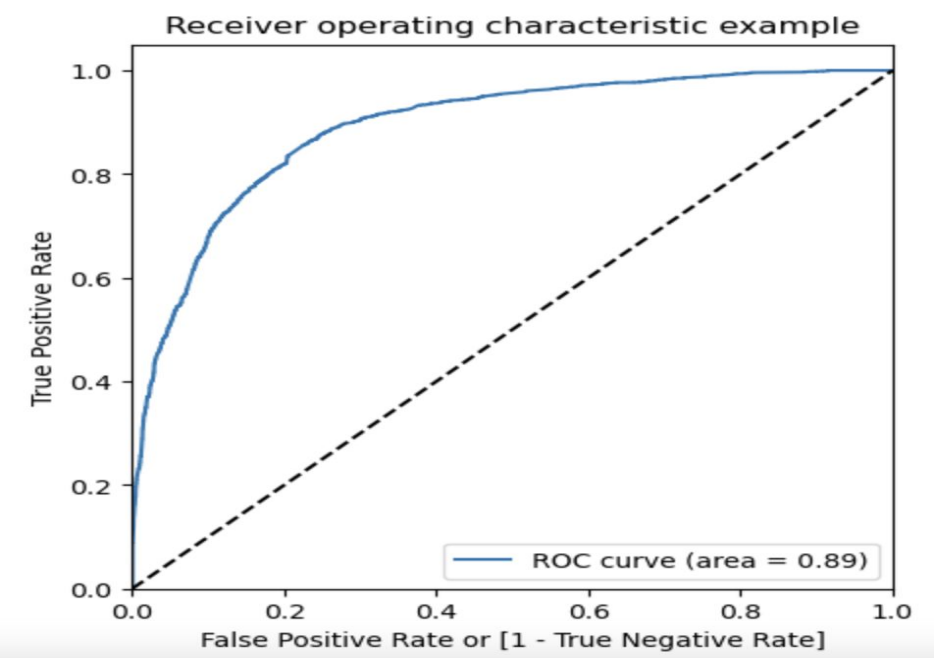
Since the Pvalues of all variables is 0 and VIF values are low for all the variables, model-7 is our final model. We have 14 variables in our final model.

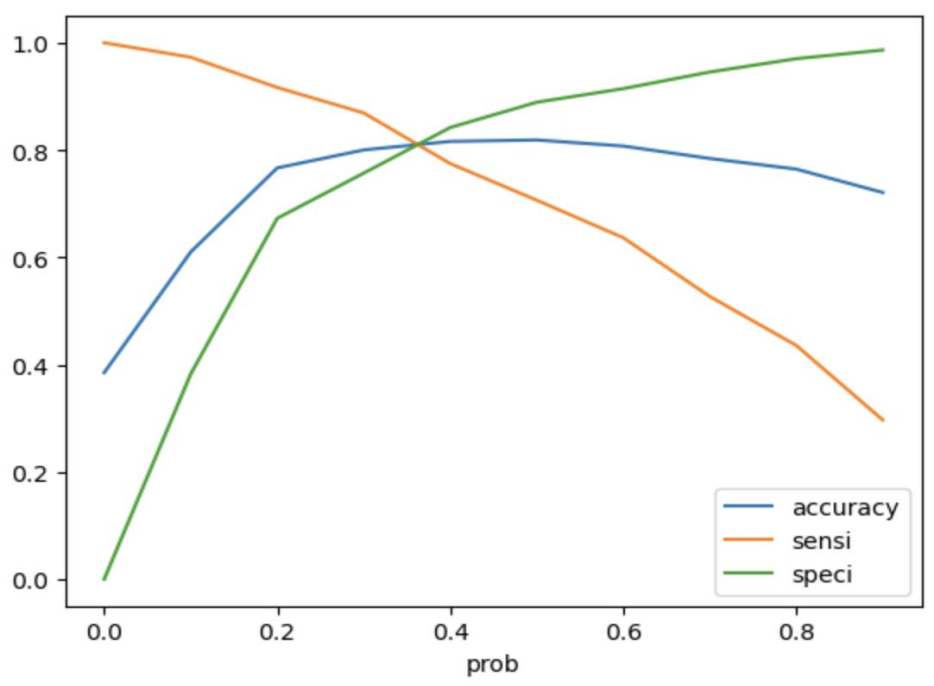
We found out that our specificity was good (~88%) but our sensitivity was only 70%. Hence, this needed to be taken care of.

We have got sensitivity of 70% and this was mainly because of the cut-off point of 0.5 that we had arbitrarily chosen. Now,

this cut-off point had to be optimised in order to get a decent value of sensitivity and for this we will use the ROC curve

ROC CURVE AND OPTIMUM CUT-OFF





Since we have higher (0.89) area under the ROC curve , therefore our model is a good one.

From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Finding out the Important Features from our final model

```
Lead Source_Welingak Website                                  5.446004
Lead Source_Reference                                         3.249778
What is your current occupation_Working Professional          2.685798
Last Activity_Other_Activity                                  2.596923
Last Notable Activity_Unreachable                             2.034601
Last Activity_SMS Sent                                        1.655729
Lead Source_Olark Chat                                        1.134335
Total Time Spent on Website                                   1.112072
Last Activity_Email Opened                                    0.424606
const                                                        -0.396090
Last Notable Activity_Modified                               -0.829726
City_Select                                                  -0.887802
Lead Origin_Landing Page Submission                          -1.213959
Do Not Email                                                 -1.368053
Specialization_Others                                        -2.109764
dtype: float64
```

RECOMMENDATIONS

- The company should make calls to the leads coming from the lead sources **"Welingak Websites" and "Reference"** as these are more likely to get converted.
- The company should make calls to the leads who are the **"working professionals"** as they are more likely to get converted.
- The company should make calls to the leads who spent **"more time on the websites"** as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources **"Olark Chat"** as these are more likely to get converted.
- The company should make calls to the leads whose last activity was **SMS Sent** as they are more likely to get converted.

- The company should not make calls to the leads whose last activity was **"Olark Chat Conversation"** as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is **"Landing Page Submission"** as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was **"Others"** as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of **"Do not Email" as "yes"** as they are not likely to get converted.

CONCLUSIONS

- The final model has Sensitivity of 0.809 on the test data set, this means the model is able to predict 80% customers out of all the converted customers, (Positive conversion) correctly.

- The accuracy on training data set and test data set is almost similar, proving that the model is stable.

- We can go ahead with the final model and use it for improving the conversion rate of the leads for X Educations.


Thank
you