ASSIGNMENT SUMMARY

Pratima Rathore
Apoorv Dubey

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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%..

Business Goal

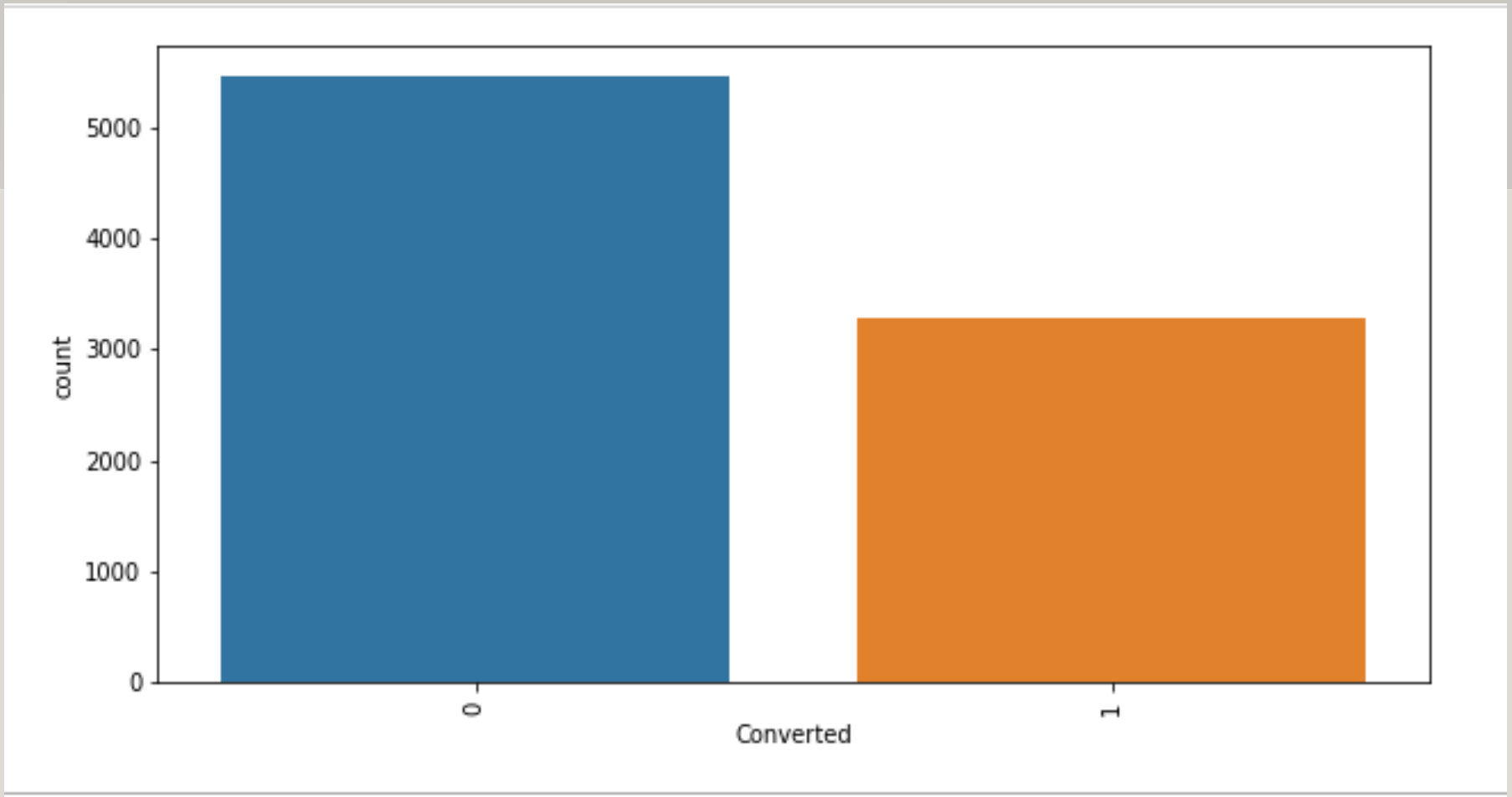
To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Data Understanding

This assignment has 2 files as explained below:

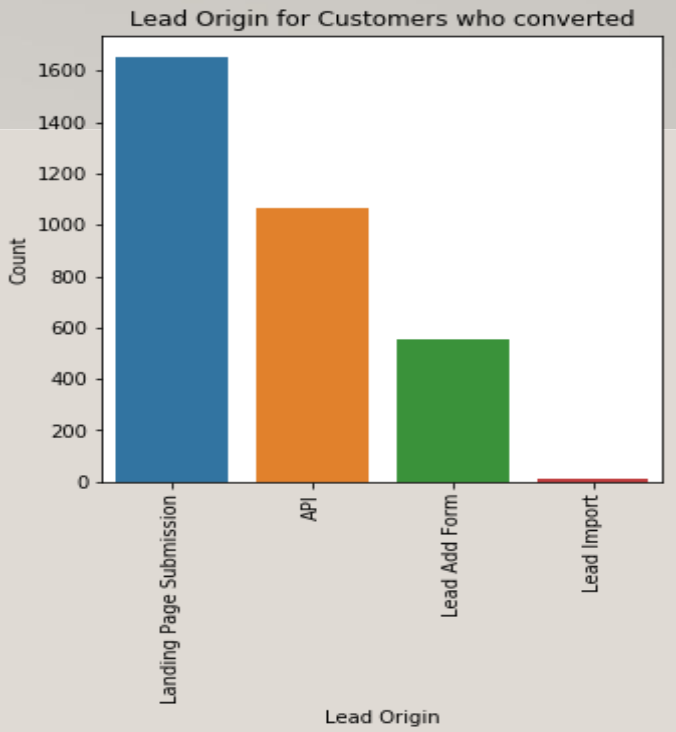
- ‘Leads.csv' contains all features related to customer,leads and conversion.
- ‘Leads_data_dictionary.xlxs' is data dictionary which describes the meaning of the features.

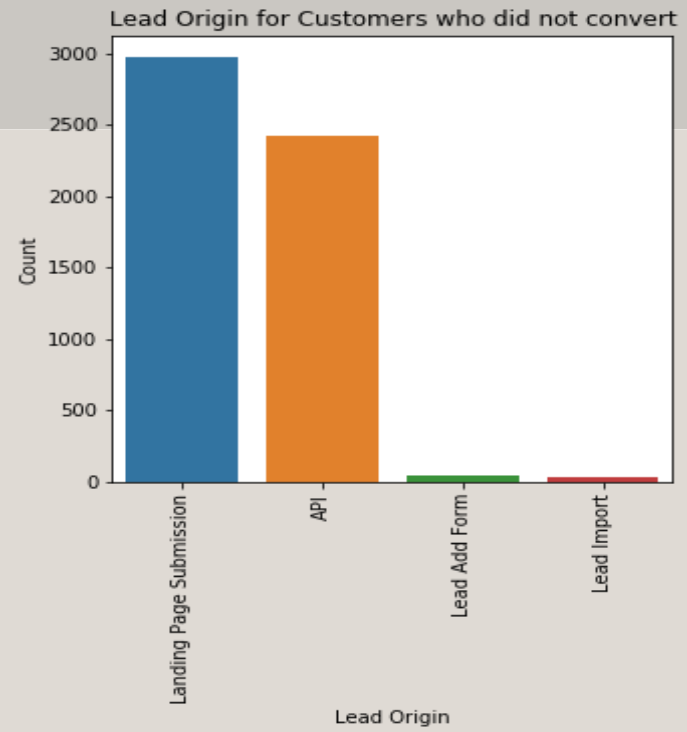
Subsetting the dataset on ‘Converted’ column for better analysis on Conversion



Major factors impacting Conversion

- Lead Origin

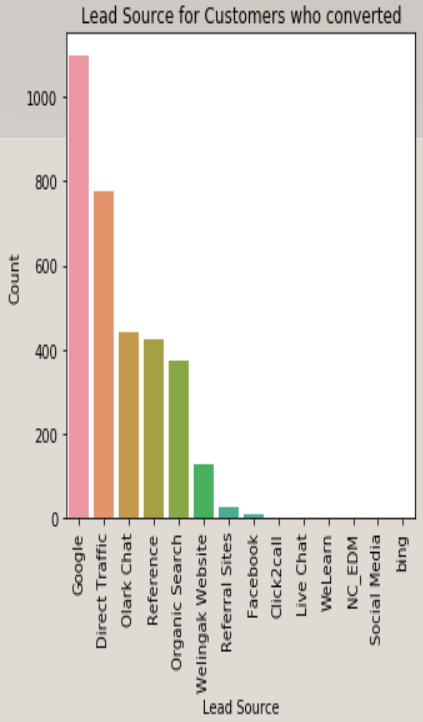


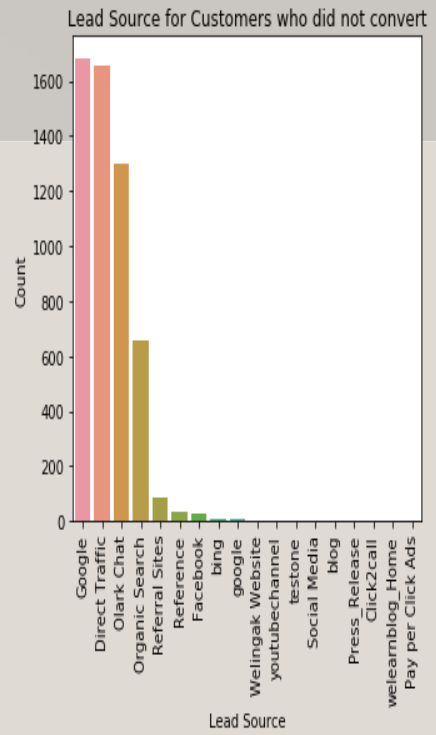


➢ The origin identifier with which the customer was identified to be a lead. These are plot are converted vs not converted

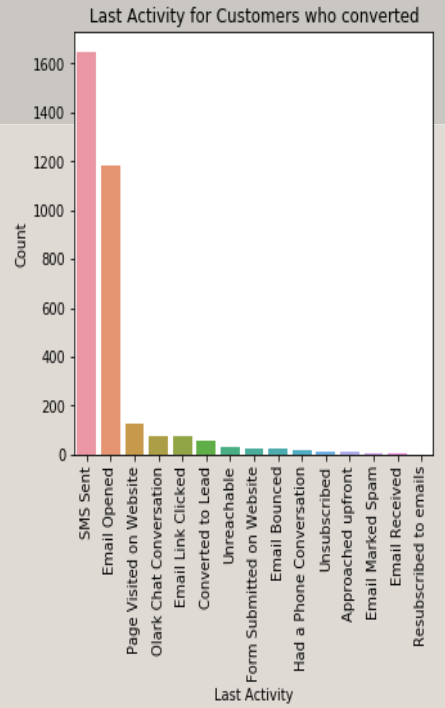
Major factors impacting Conversion

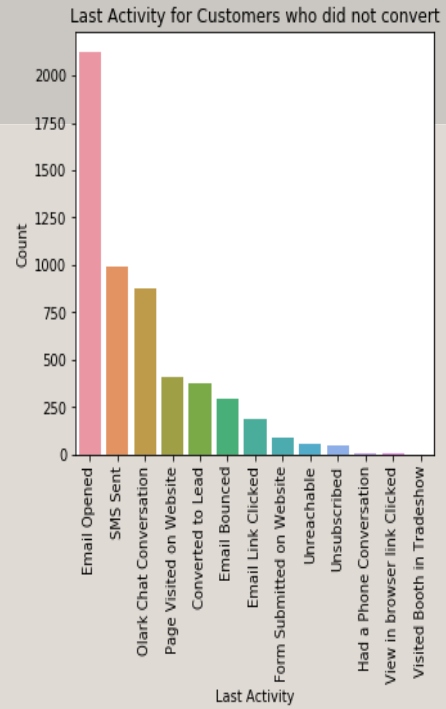
- Lead Source





- Last Activity

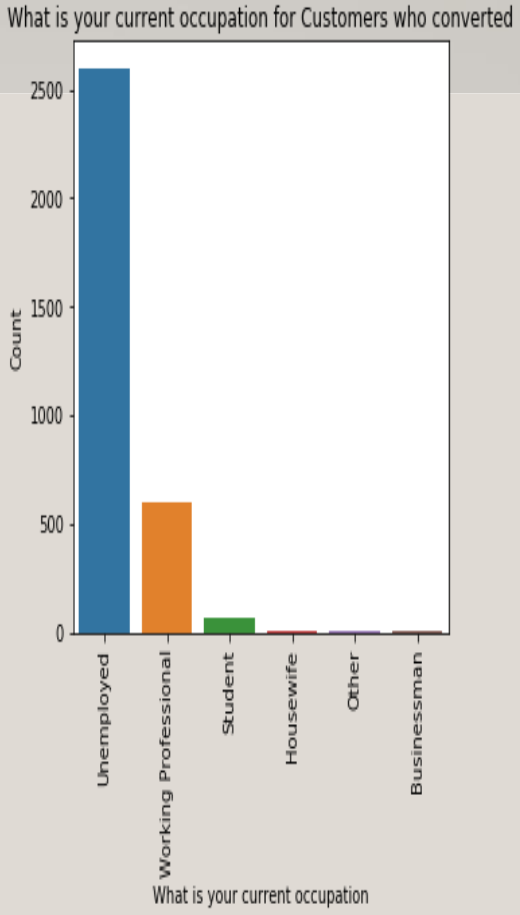


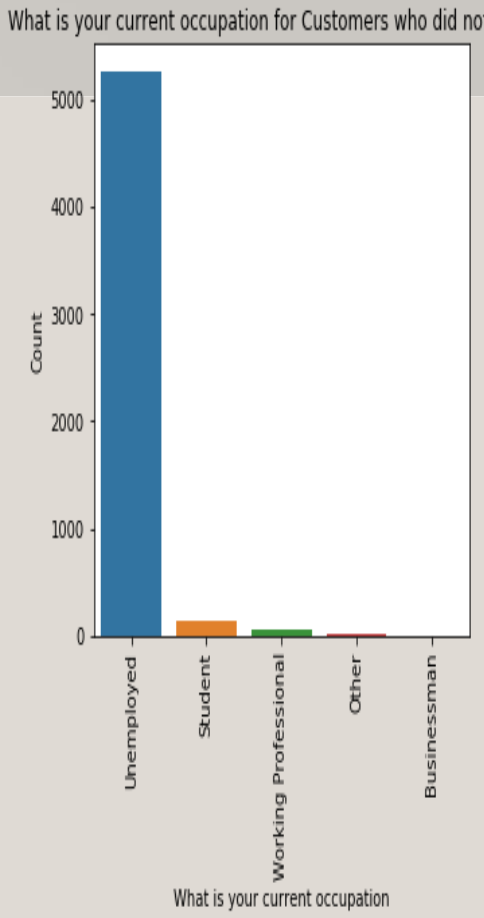


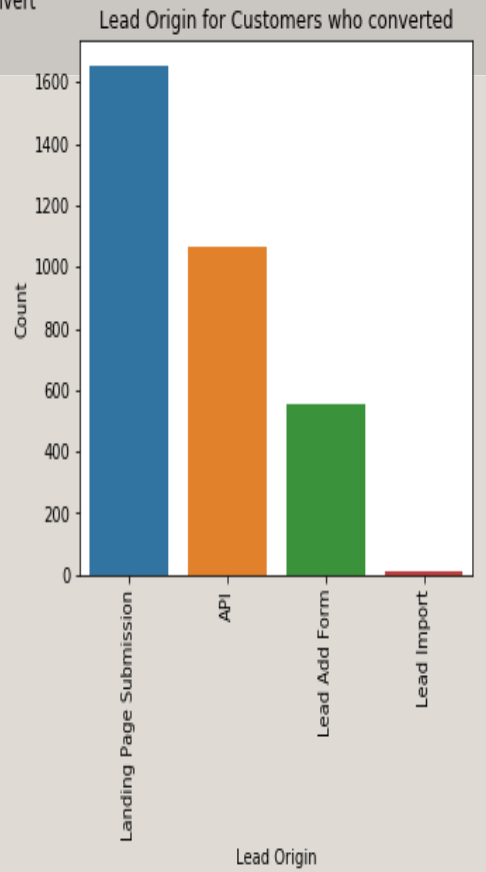
- The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
- Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.

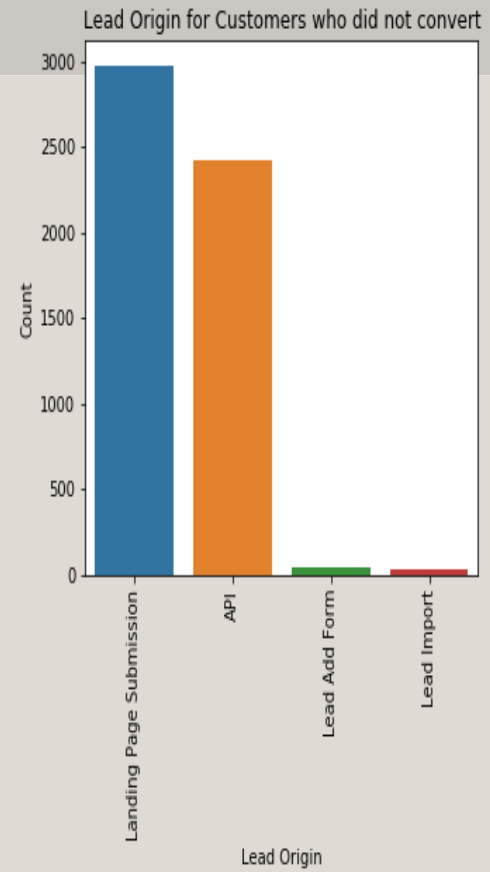
Major factors impacting Conversion

- What is your current occupation
- Lead Origin



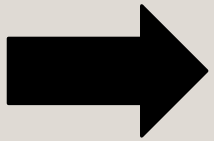






Final Features after RFE and multiple model evaluation

```
Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Lead Source_Welingak Website', 'Last Activity_Converted to Lead',
       'Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation',
       'Last Activity_Olark Chat Conversation', 'Specialization_Others',
       'What is your current occupation_Housewife',
       'What is your current occupation_Working Professional',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unreachable'],
      dtype='object')
```

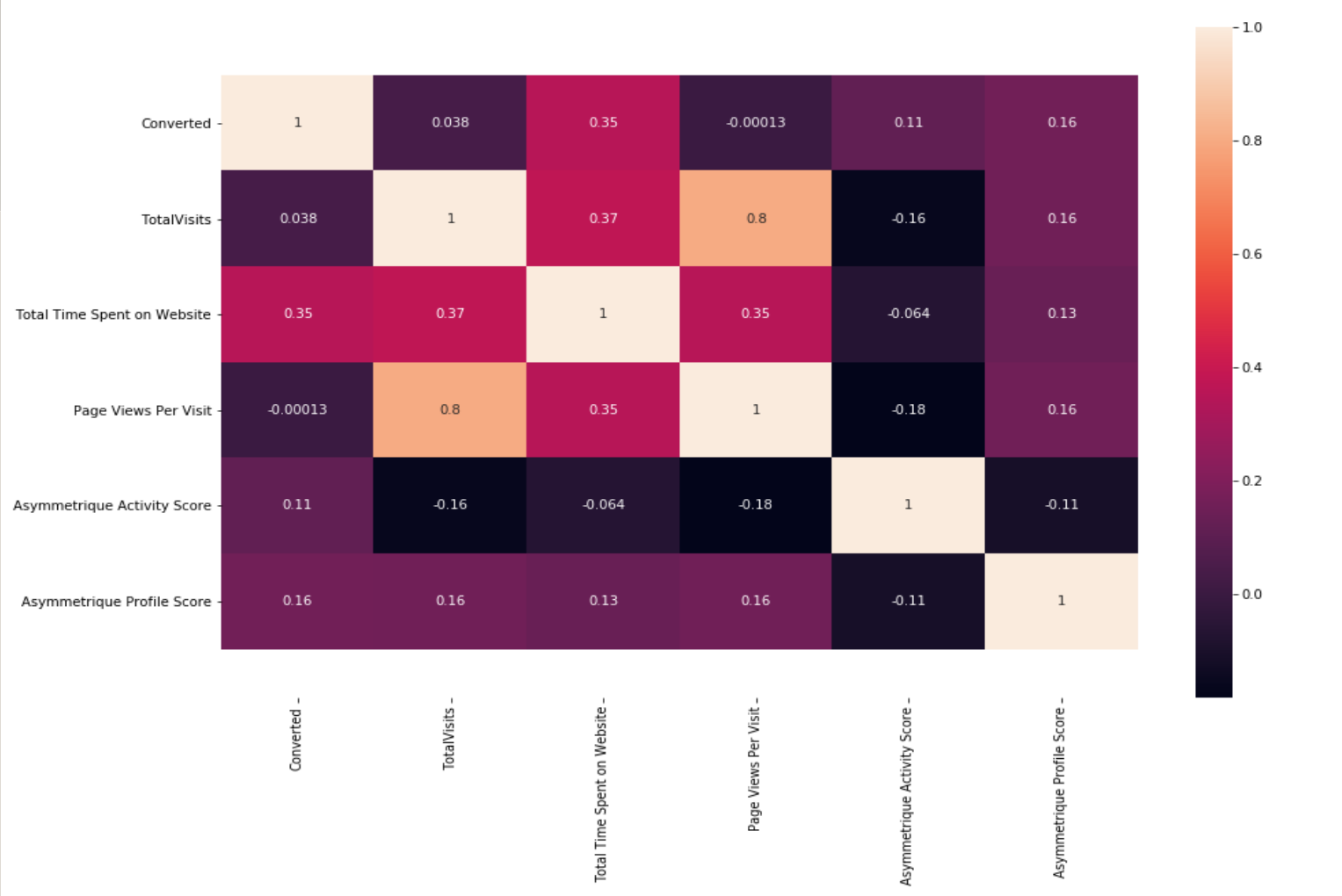


- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Last Activity_Converted to Lead
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation
- Specialization_Others
- What is your current occupation_Working Professional
- Last Notable Activity_SMS Sent
- Last Notable Activity_Unreachable

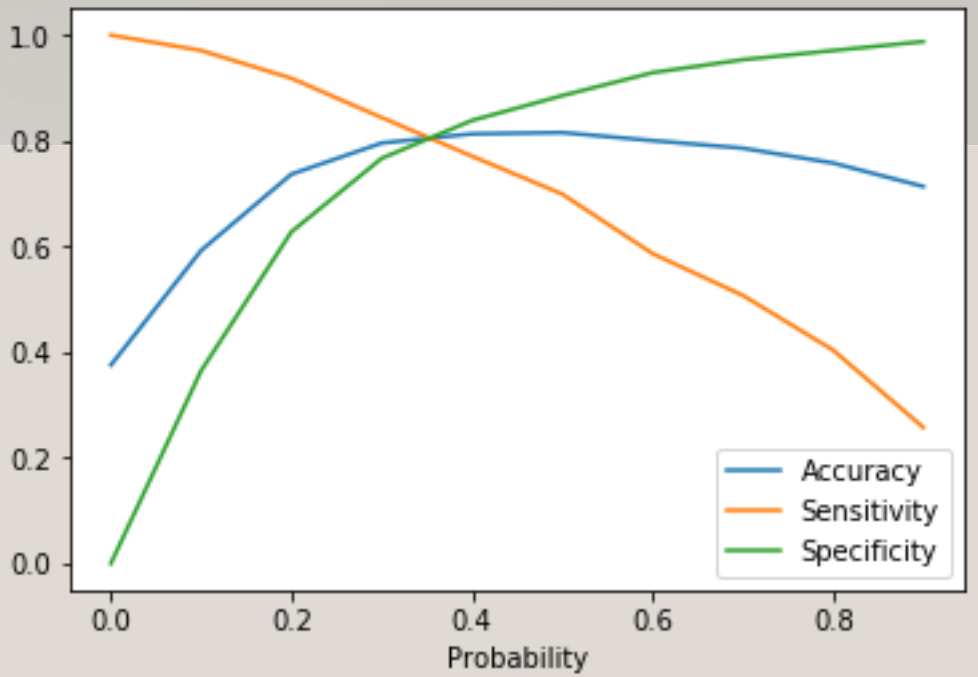
RFE selected features

Final model features

Heatmap showing relationship btw various features



Selecting right threshold to attain 80% conversion



	Probability	Accuracy	Sensitivity	Specificity
0.0	0.0	0.375510	1.000000	0.000000
0.1	0.1	0.592490	0.970870	0.364967
0.2	0.2	0.736816	0.918261	0.627712
0.3	0.3	0.795755	0.843913	0.766797
0.4	0.4	0.813061	0.770870	0.838431
0.5	0.5	0.815510	0.699130	0.885490
0.6	0.6	0.800327	0.586522	0.928889
0.7	0.7	0.785959	0.507391	0.953464
0.8	0.8	0.757878	0.403913	0.970719
0.9	0.9	0.713469	0.257391	0.987712

- We came on conclusion to –

As we want 80% lead conversion of the potential leads, we can **choose an optimal threshold value for Conversion Probability ie 0.36**

Assigning Lead Score

Lead Score = 100 * Conversion Probability

	Converted	Prospect ID	Prediction Probability	Final Prediction	Lead Score
0	1	3667	0.319164	0	31.92
1	0	1952	0.087698	0	8.77
2	0	7355	0.559689	1	55.97
3	1	8865	0.843473	1	84.35
4	1	4272	0.537565	1	53.76

Model Metrics

Training Data
Accuracy 80.88%
Sensitivity 80.17%
Specificity 81.69%

Testing Data
Accuracy 80.68%
Sensitivity 78.82%
Specificity 81.8%

As the Business statement says 80% conversion which means 80% of constomers to me converted.The best metric to evaluate that is Sensitivity.
Sensitivity (also called the **true positive rate**) measures the proportion of actual positives that are correctly identified .

So we achieved our objective .

Top features that are impacting conversion

- Lead Add Form (from **Lead Origin**)

It implies that the origin identifier 'Lead Form' has a big role to play in lead conversion.

➢ Welingak Website (from **Lead Source**)

It implies customer that visit this website has strong impact on conversion.

➢ Working Professional (from **What is your current occupation**)

It implies customers who were Working Professionals are the ones that converted the most or have major impact on conversion .

- Email Bounced (**Last Activity**)

It has negative impact on conversion implying when ever last activity of customer is 'email bounced' , he is more likely to not get converted.