Lead Scoring Case Study

Participants:

- Hiral Salvi
- Tejas Barhate
- Ramakrishna Beeraveli

Problem Statement

- X Education, an education company sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google. Interested candidates browse for courses and submit the forms with personal details, referred to as leads.

- Currently the lead conversion ratio, meaning the proportion of leads that actually enrol in the course, is 30%.

- Due to thousands of leads coming through multiple channels, it is difficult to follow up on all the leads.

- Expectation is to identify hot leads to narrow down and follow up only with them to increase the lead conversion ratio to 80%.

Objective

- X Education would like to know promising leads or hot leads
- They would need a model that assigns lead score of 0 to 100 to the leads where higher the score, higher the conversion chances
- Lead conversion ratio target is 80% (currently 30% leads are getting converted)
- Deploy the model considering future changes

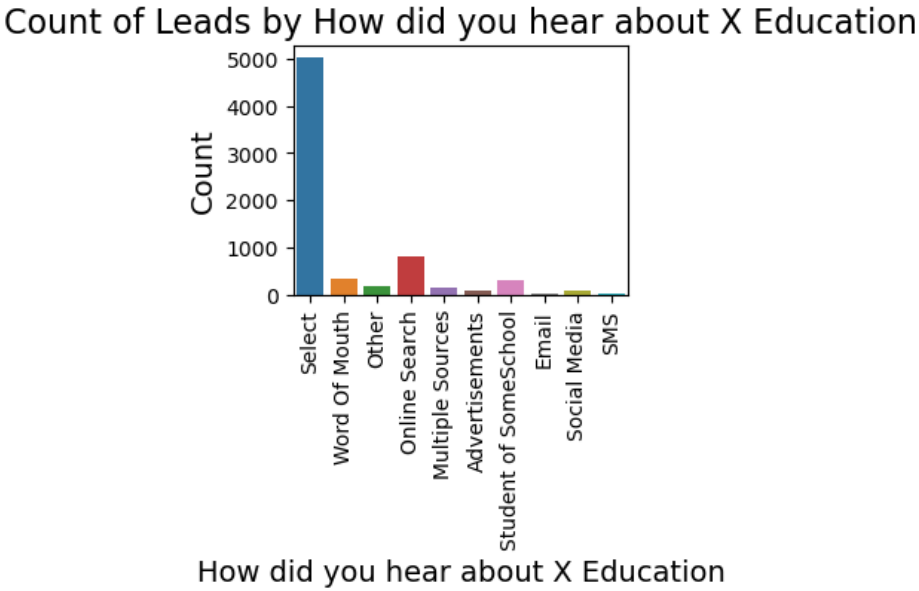
Data Cleaning – Missing values

- Many attributes have more than 30% missing value. Examples are Tags, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score and Asymmetrique Profile Score. These columns are dropped.

Through Recommendations	0.000000
Receive More Updates About Our Courses	0.000000
Tags	36.287879
Lead Quality	51.590909
Update me on Supply Chain Content	0.000000
Get updates on DM Content	0.000000
Lead Profile	29.318182
City	15.367965
Asymmetrique Activity Index	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Score	45.649351
I agree to pay the amount through cheque	0.000000
A free copy of Mastering The Interview	0.000000
Last Notable Activity	0.000000

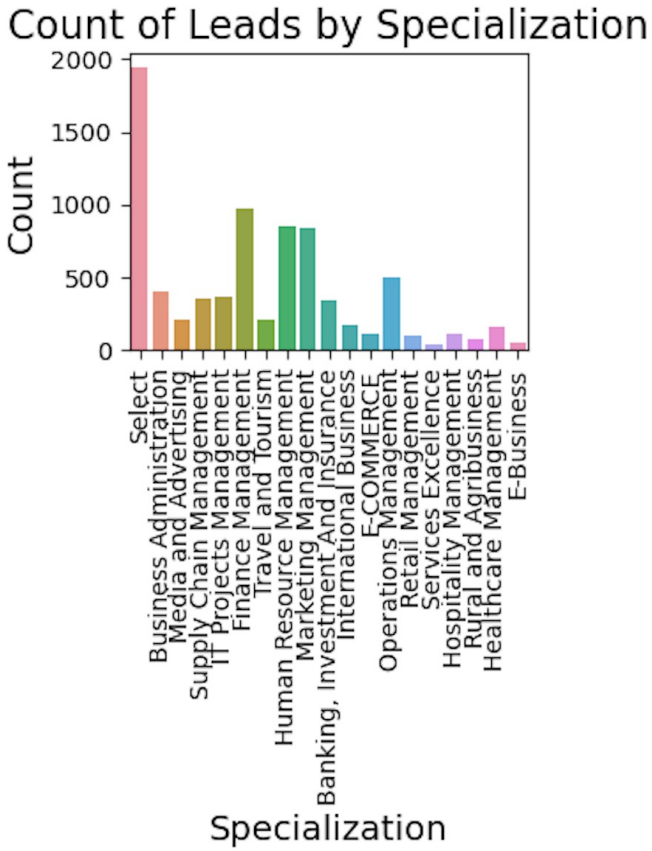
Data Cleaning – Special value like ‘Select’

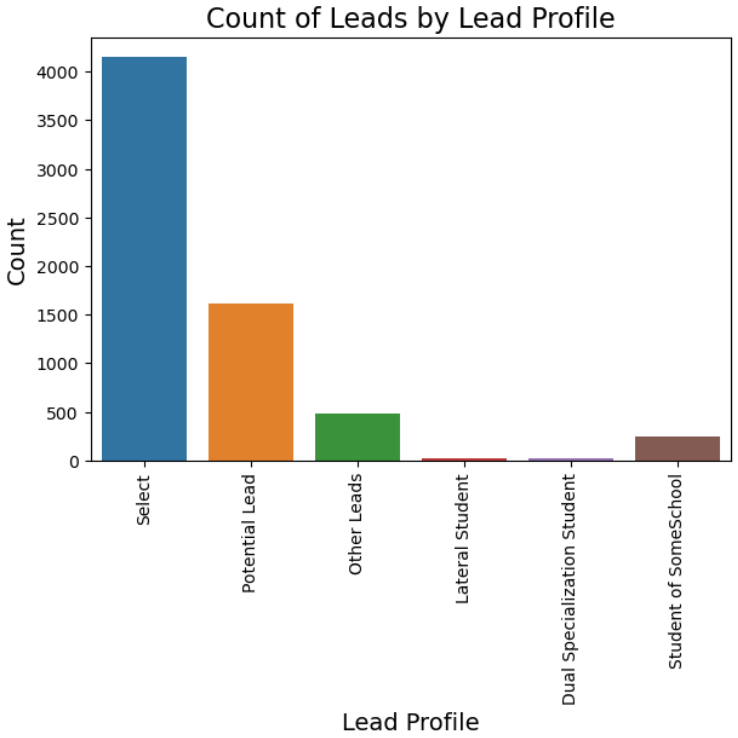
- Many columns 'Select’ as their value. Which means they don't actually hold useful values and can be treated same as null.
- EDA is used to check overall distribution of the values of ’Select’ compared to other values.
- For example, How did you hear about X Education has 70%+ ‘Select’ so we can drop this attribute.



Data Cleaning – EDA for special value like ‘Select’

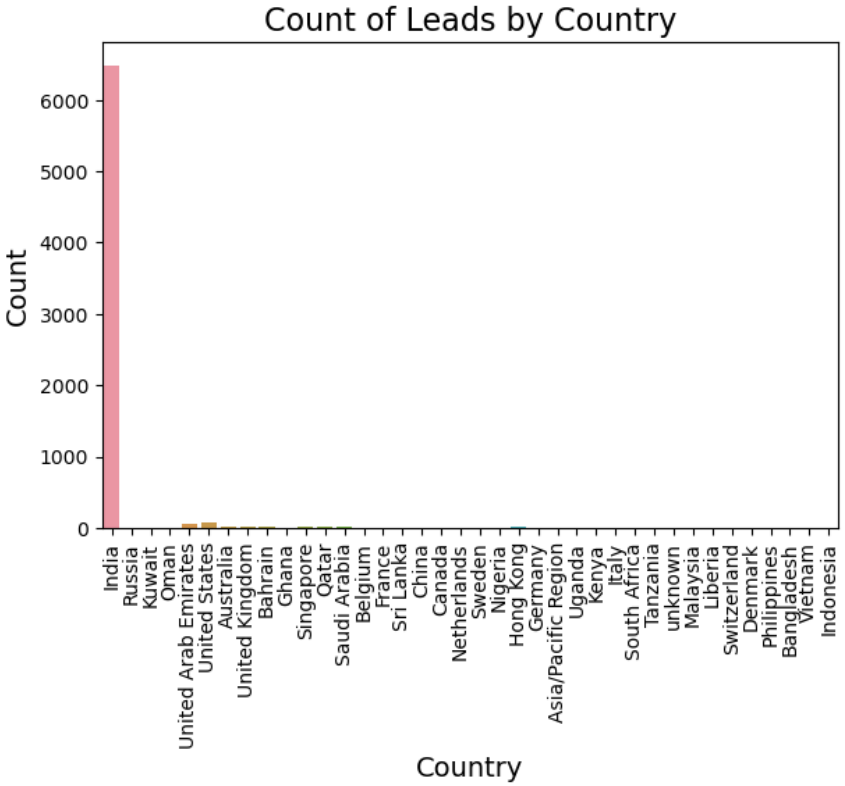
- Specialization and Lead Profile also have very high ‘Select’ which are similar to null values so both attributes can be dropped.

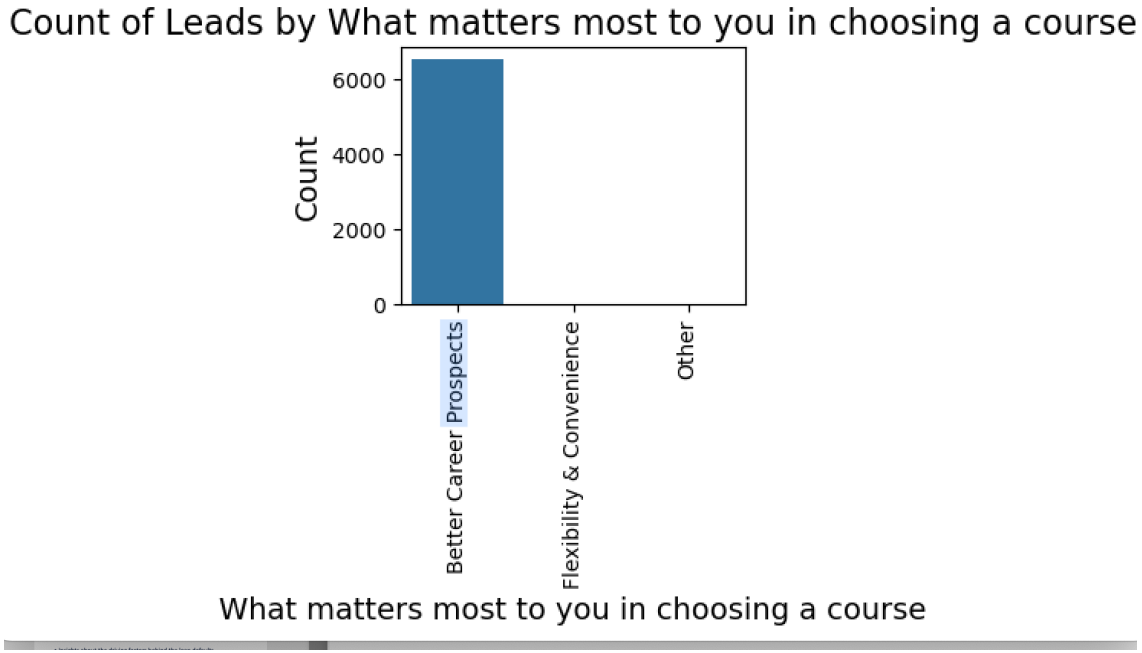




Data Cleaning – Dominating values

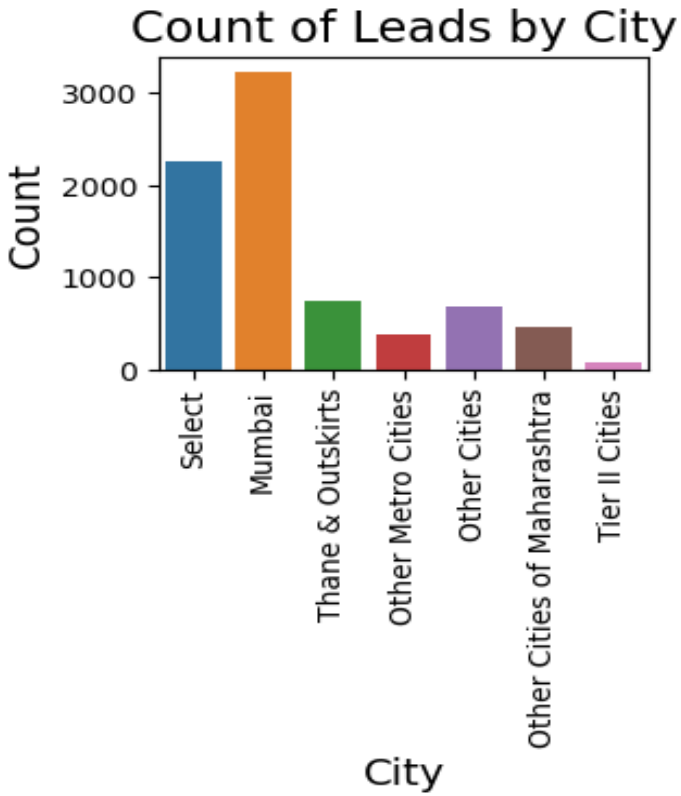
- Country has most values as ‘India’, similarly What matters most to you in choosing a course has dominating value as Better career prospect. Due to this, we may not get correct predictions to we may choose to drop such attributes as well.





Data Cleaning – Looking at meaningless values

- Values like Other Cities, Other Metro Cities, Tier II cities are not useful while making predictions. Even if we find some relation with City values and predicted lead conversion score, these are not meaningful or explainable to the client. We can choose to drop these as well.

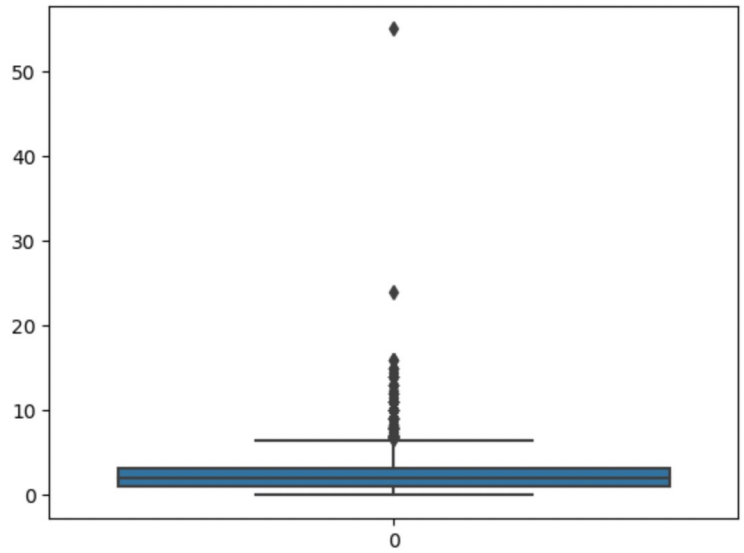


Data Preparation

- Mapping Yes/No values to 1 and 0 : Many attributes including Do Not Email, Do Not Call, Search, Magazine, Newspaper Article and others have Yes and No values, these are converted to 1 and 0.
- Dummy features for categorical variables with multiple levels : Lead Origin, Lead Source, Last Activity and others with multiple levels can be converted into dummy attributes.

```
2
3 sns.boxplot(leads_data['Page Views Per Visit'])
4
```

```
Out[389]: <Axes: >
```

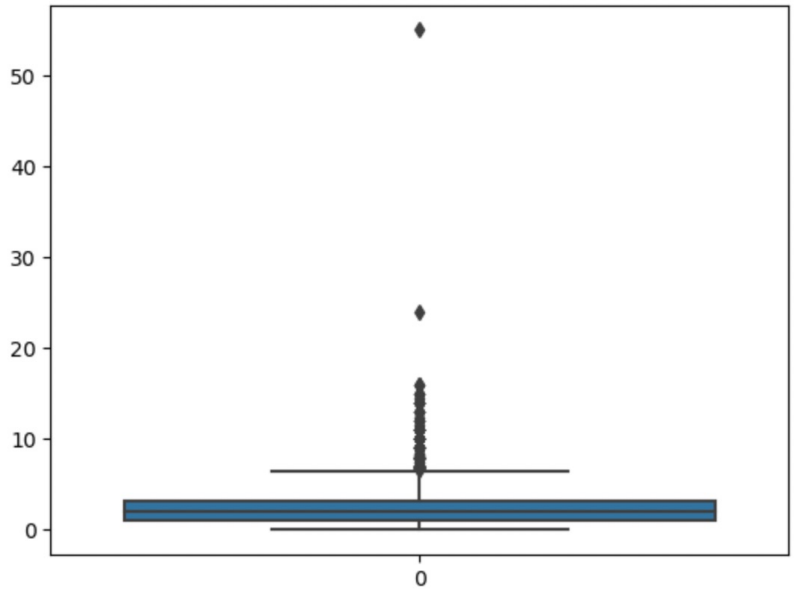


Outlier detection through EDA

- EDA is used to detect outlier in the data.
- For example - Page Views Per Visit, Total Visits , Page Views Per Visit etc.

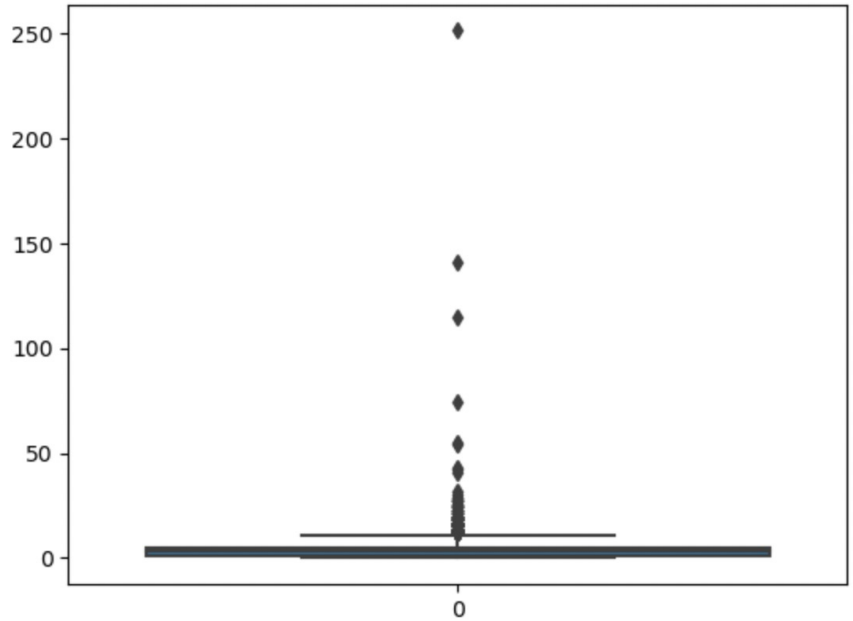
```
2
3 sns.boxplot(leads_data['Page Views Per Visit'])
4
```

Out[389]: <Axes: >




```
1 sns.boxplot(leads_data['TotalVisits'])
```

<Axes: >

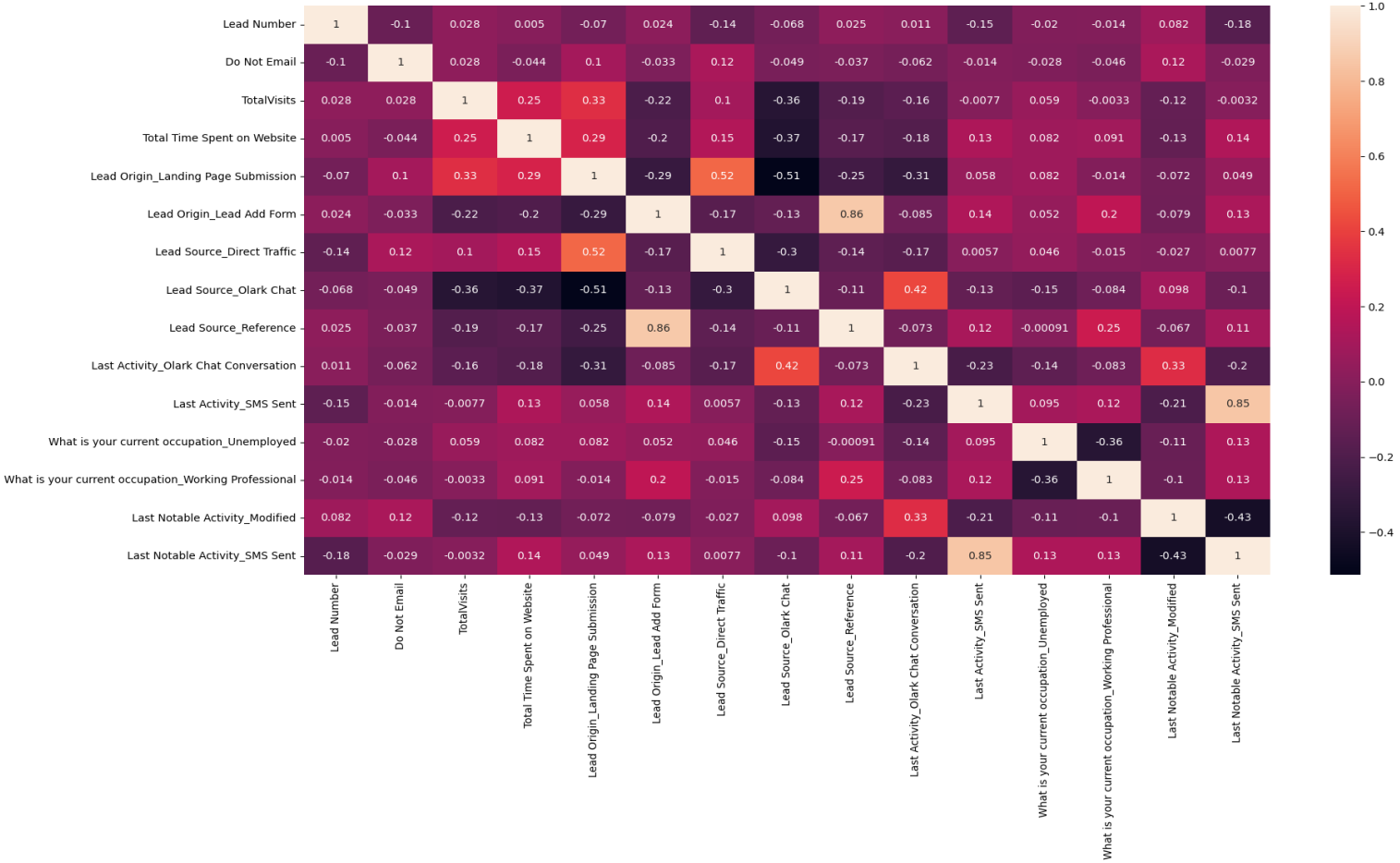


Model Building – Considerations

- Using 70% as train and 30% as test data
- Feature scaling for continuous attributes like Total Time Spent on Website, Total Visits and Page Views Per Visit
- Current conversation rate as per the data is 30%
- FRE is used to identify most important 15 attributes, rest are not needed for the modeling.
- Checking correlation between the attributes through EDA
- Acceptable VFA values are < 5
- Sequentially dropping the attributes having VIF > 5 and adjusting the model until all VIF value are below 5

Model Building – Correlation using EDA

- Many attributes having high correlation as per the heatmap below.



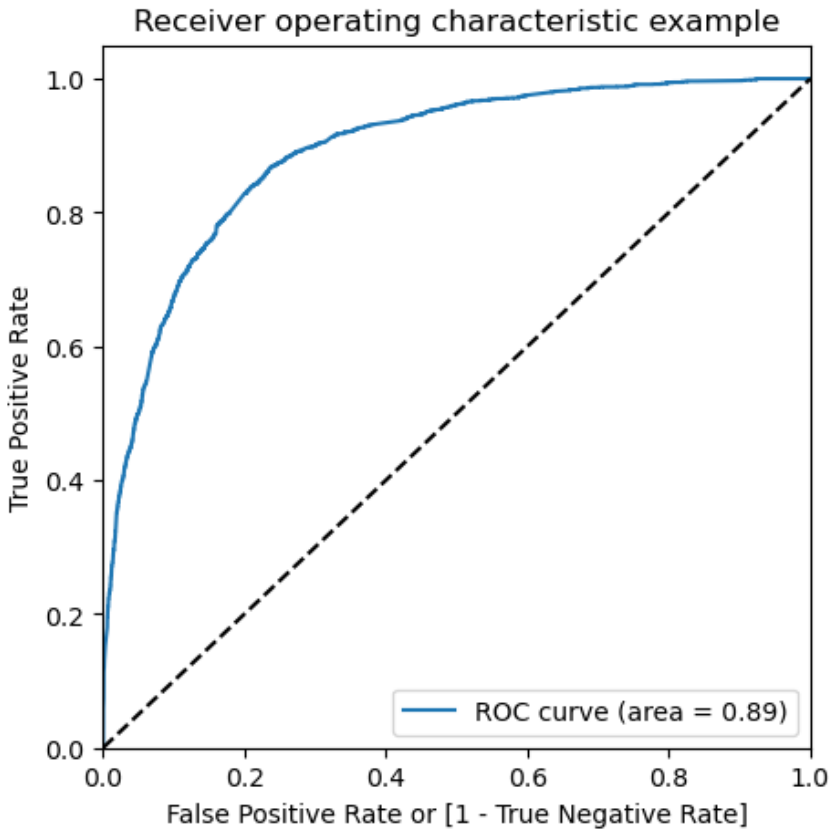
Model Evaluation – Train Data

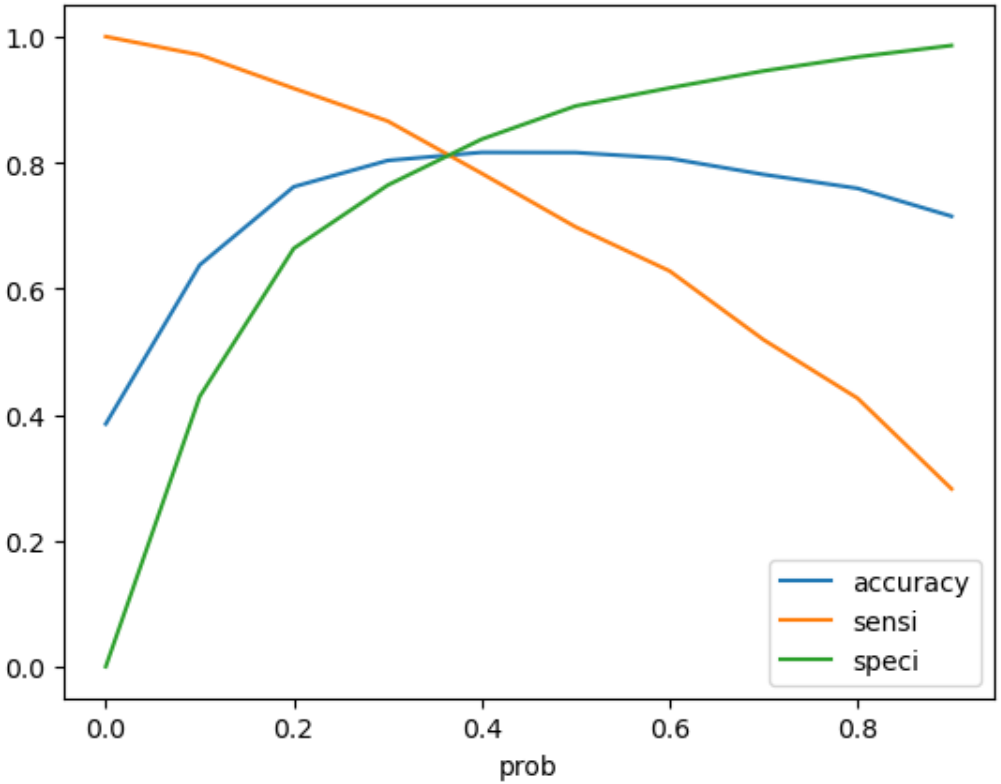
- Accuracy of the model : ~81%
- Specificity of the model : ~70%
- Sensitivity of the model : ~89%

- Other Matrix:
  - False positive rate : ~11%
  - Positive predictive value : ~80%
  - Negative predictive value : ~82%

Model Evaluation – ROC Curve

- Optimum cut off point for accuracy, sensitivity and specificity is 0.4





Model Evaluation – Test Data

- Accuracy of train data and test data are 81% which is acceptable

- Overall matrix for the test data :
  - ✓ Accuracy of the model : ~81%
  - ✓ Specificity of the model : ~83%
  - ✓ Sensitivity of the model : ~77%

Conclusion

- Lead score is assigned by multiplying the probability with 100 to get the expected lead score.
- High lead score is an indication of hot deals, 100 is highest and 0 is lowest.
- Leads who use the website more are identified as hot leads. Attributes such as Total Time spend on the Website matter the most.
- Most leads getting converted as Lead source such as Olark Chat and Lead Add Form.
- Additionally, when current occupation is a working profession, there is a high chance of conversion.

Thank You!