LEAD SCORING CASE STUDY

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 markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, 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'. 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.

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%.

Logistic Regression Modelling Steps

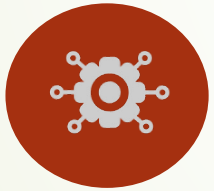
1. Import necessary libraries

2. Read & understand data

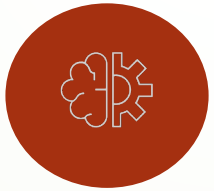
3. Data Cleaning

4. Exploratory Data Analysis

5. Data Pre-processing



6. Data Preparation



7. Modelling

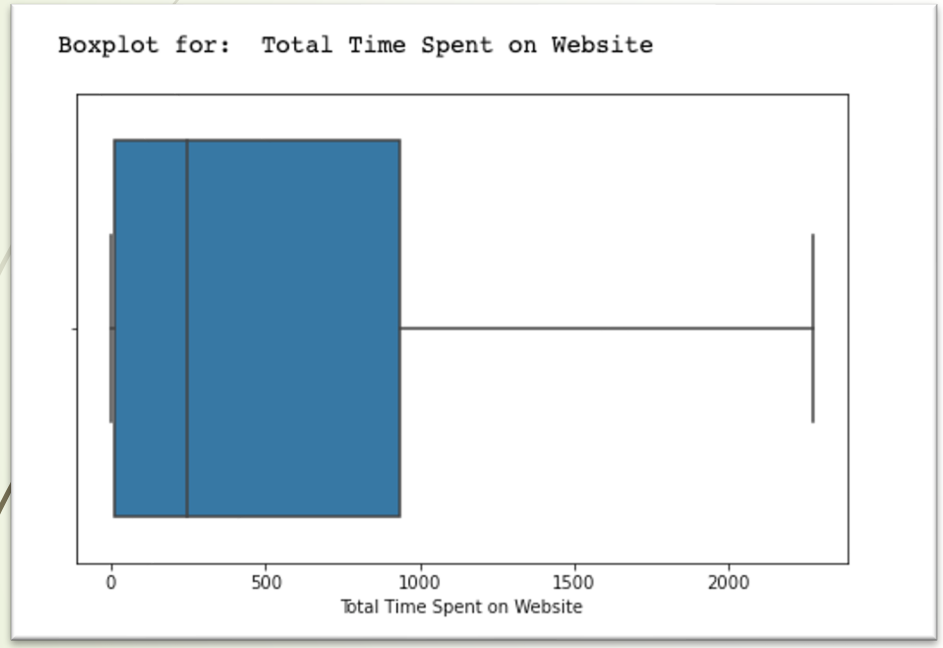
8. Model Evaluation

9. Calculate Lead Score

Data Cleaning

- Replaced ‘Select’ with Nan
- Dropped columns that have more than 40% null values
- Check number of unique categories in categorical column & taken necessary step to either drop column that has variance in data or merge <1% category-wise values to one category
- Dropped id columns

EDA: Univariate Analysis of Continuous Variable

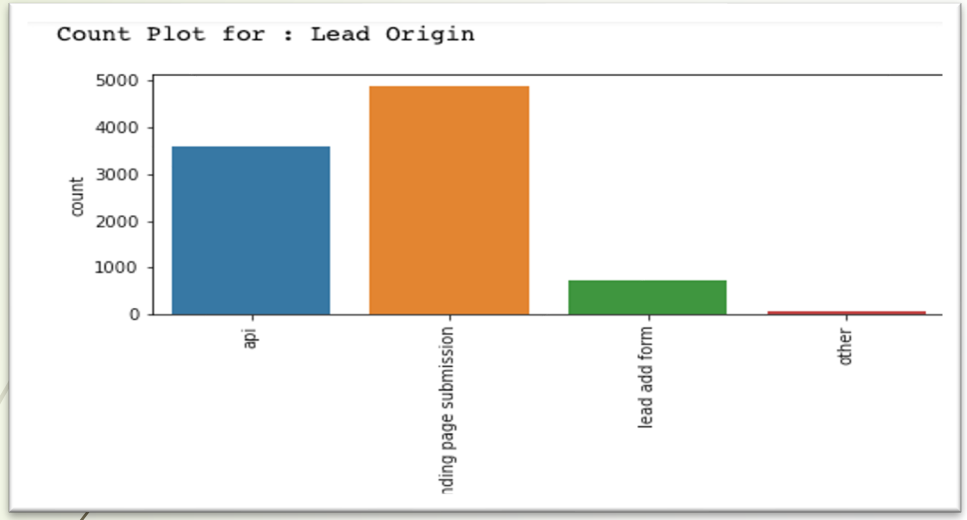


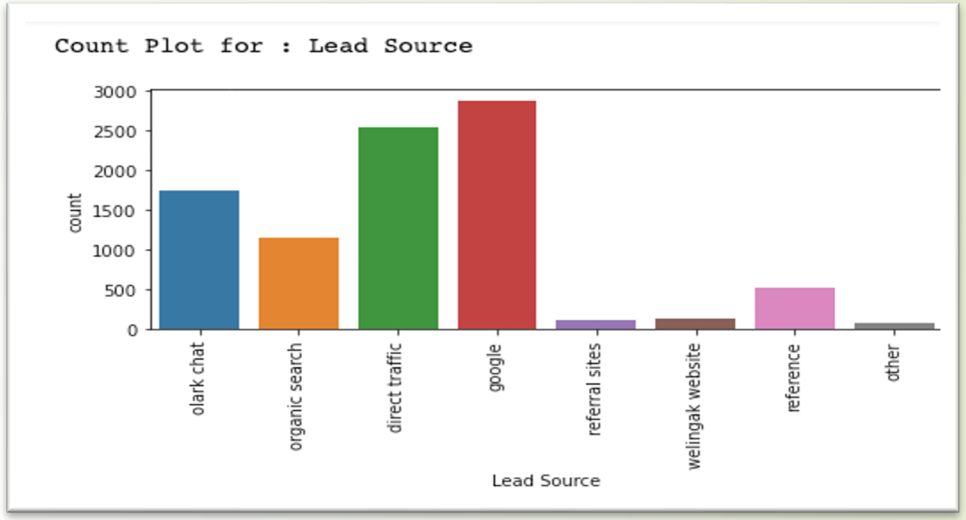
3 Average lead visits in the website

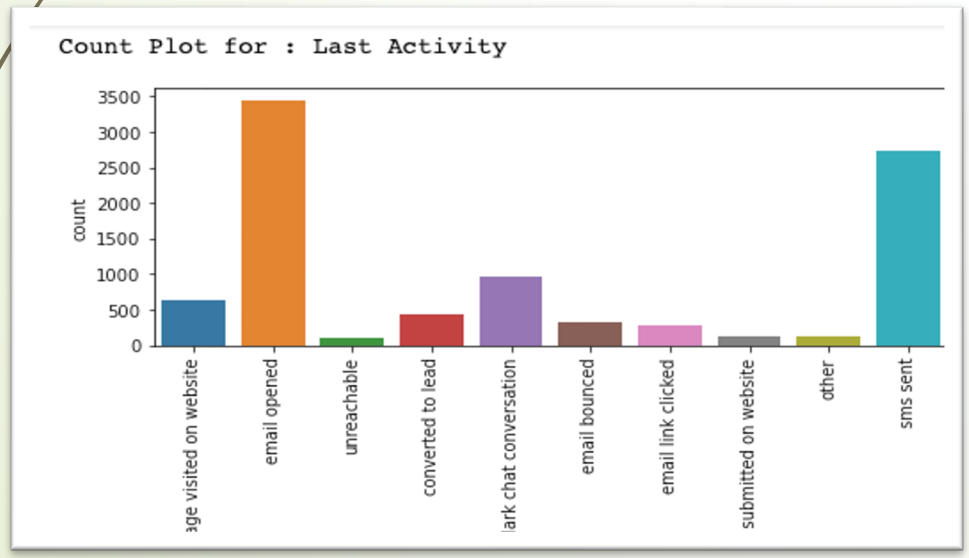
500 Average minutes spent on website

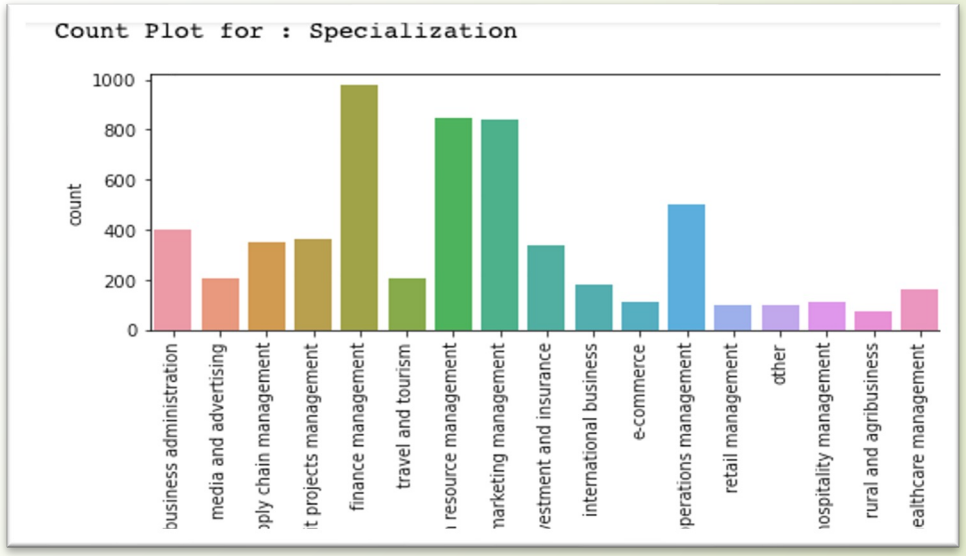
2 Average number of pages viewed on every visit

EDA: Univariate Analysis of Categorical Variables





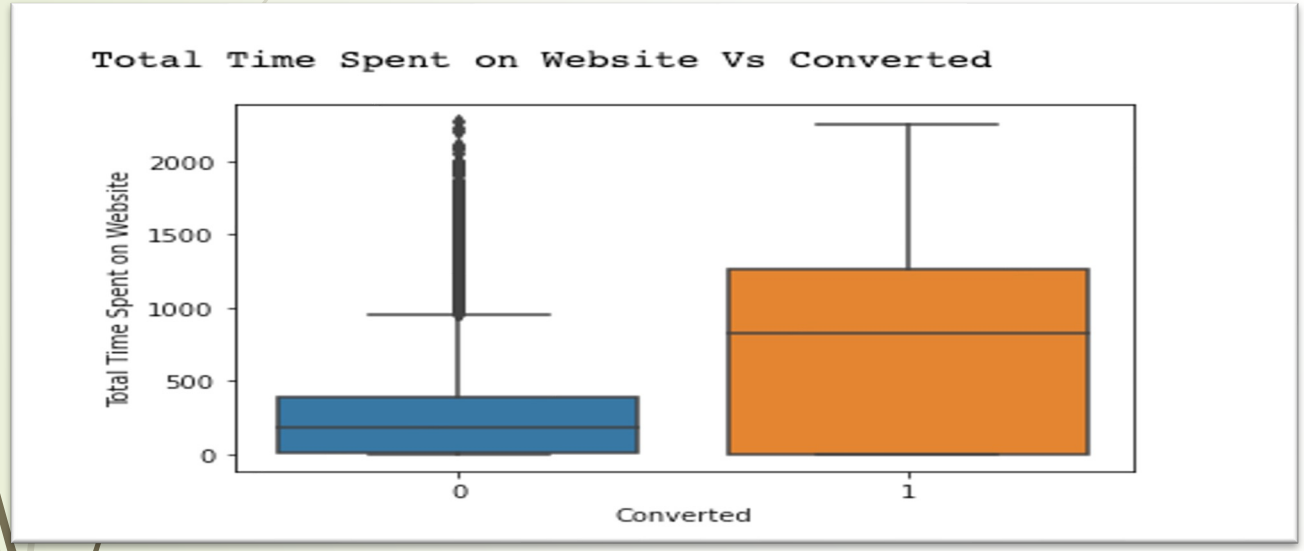




EDA: Insights based on univariate analysis of categorical variables

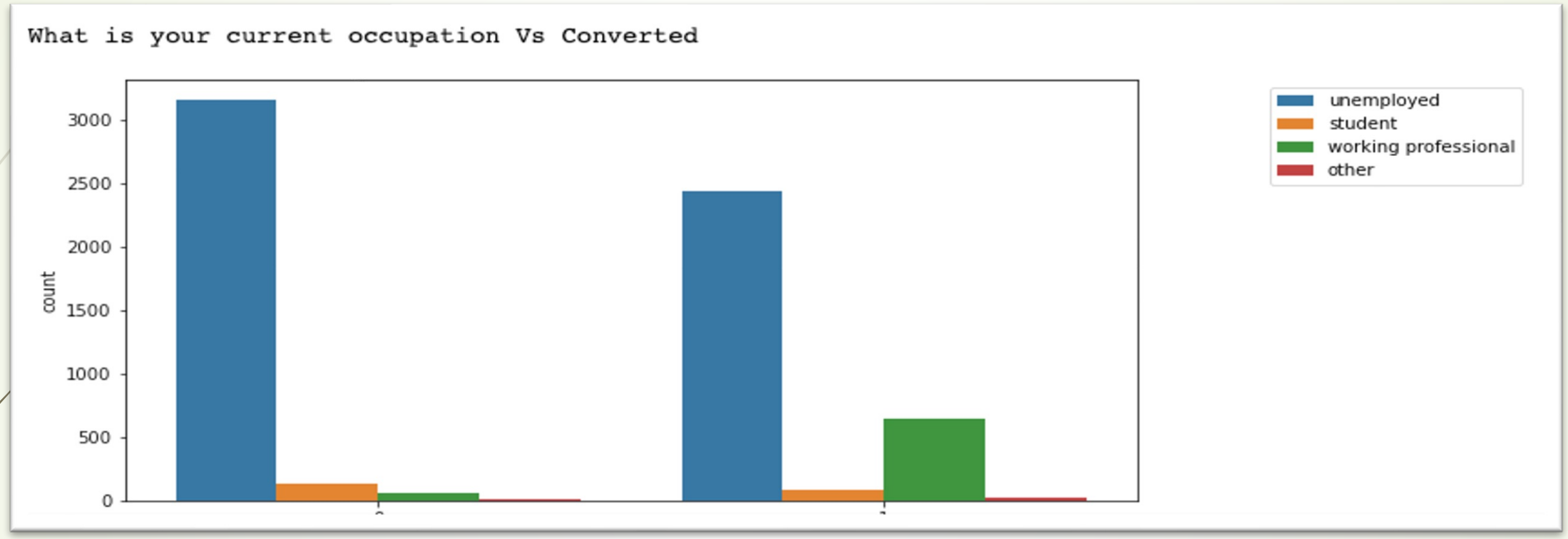
- Most of the Lead origin is from 'Landing Page Submission' followed by 'API' and 'Lead Add form'.
- The large source of Lead is Google, followed by Direct Traffic and Olark Chat.
- Most of the Leads do not want to receive any mail.
- Most of the Leads opened mail, sent SMS and did olark chat conversation
- Most of the customers are from Finance Management domain followed by Human Resource Management and Marketing Management domain.
- Most of the Leads are currently Unemployed.
- Most of the customer responded that they will revert after reading the mail.
- Most of the Leads are from Mumbai followed by Thane and outskirts.
- Most of the Leads don't want free copy of Mastering the Interview.
- Last notable activity of most of the Leads is that they done modification in their application

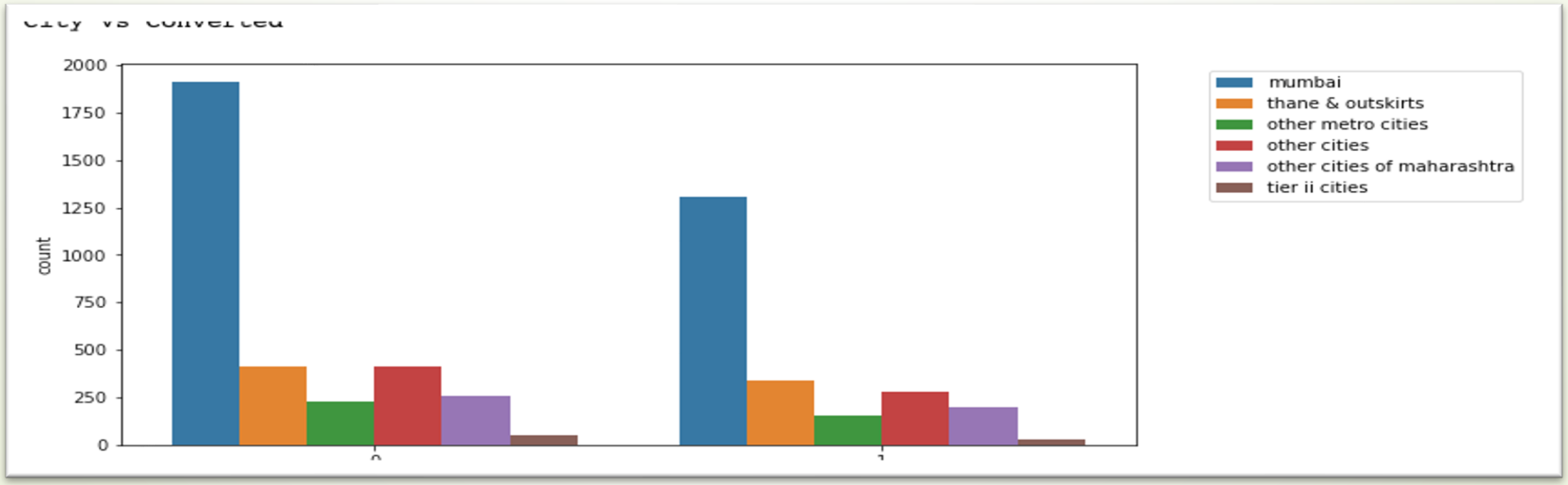
EDA: Bivariate Analysis of Continuous Variables



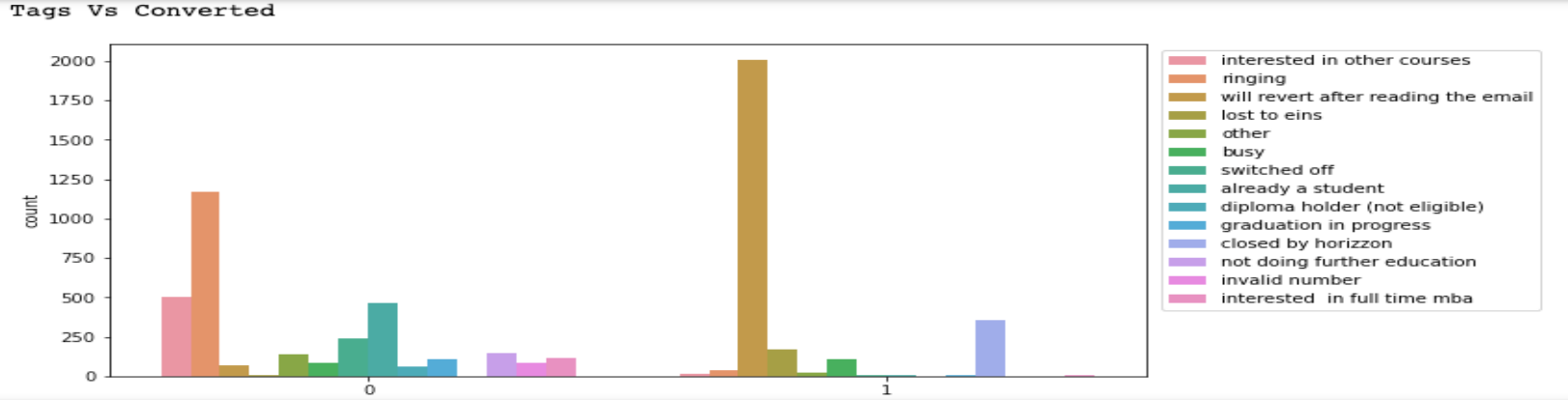
Leads who visited more on the website and spent more time on the website have high chances to get converted.

EDA: Bivariate Analysis of Categorical Variables





EDA: Insights based on Bivariate analysis of categorical variables



- ✓ Both in Converted and potential Leads, most of the Leads are originated from 'Landing Page Submission' followed by 'API' and 'Lead Add form'
- ✓ Google, Direct Traffic and Olark chat are the main sources of Leads, but the customers who sourced from Google get converted most
- ✓ Most of the converted and non-converted customers do not opt to receive any mails
- ✓ Leads last activity was that they opened email and sent SMS, but the one who sent SMS are converted most.
- ✓ Most of the converted and not converted Leads are from Finance Management domain.
- ✓ Most of the converted and not converted Leads are currently Unemployed.
- ✓ Most of the Leads who tagged as they will revert after reading the mail get converted most.
- ✓ Most of the Leads are from Mumbai.
- ✓ Most of the Leads don't want free copy of Mastering the Interview.
- ✓ Most of the Leads who have done modification in their application not get converted but the Leads who sent SMS get converted most.

Data Pre-processing & Data Preparation

- Data Pre-processing: Map ‘Yes’ & ‘No’ to 1’s & 0’s
- Data Preparation:
  - Data is split to 70% train data & 30% test data
  - Missing Value Imputation with median in continuous variable & mode in categorical variable
  - Outlier Treatment with upper and lower bound values
  - Dummy Variable creation
  - Scaling of continuous variables

Modelling

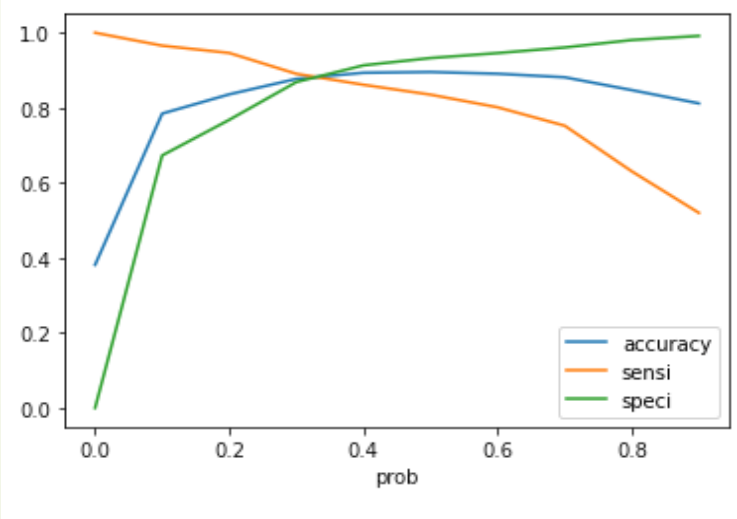
- Feature selection using RFE followed by manual feature selection based on p-value<0.05 & VIF's<5.
- Final Model:

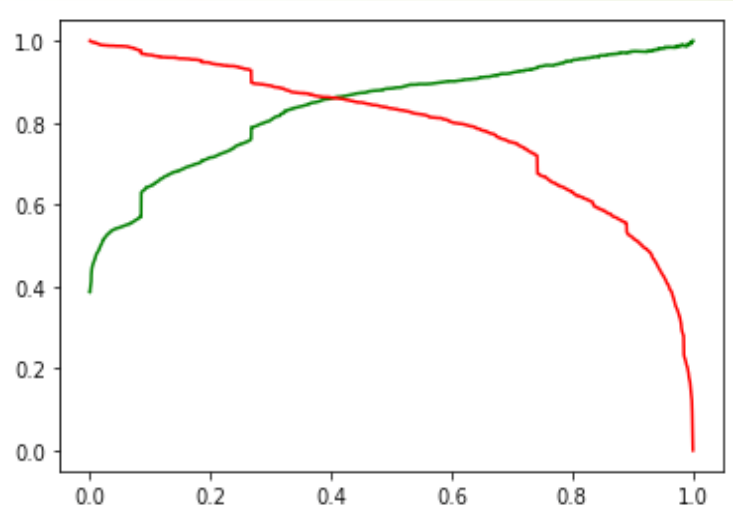
	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.6956	0.364	-12.903	0.000	-5.409	-3.982
Total Time Spent on Website	4.5205	0.204	22.185	0.000	4.121	4.920
Page Views Per Visit	-0.6243	0.164	-3.805	0.000	-0.946	-0.303
Lead Origin_lead add form	3.0962	0.232	13.346	0.000	2.641	3.551
Last Activity_email bounced	-1.7555	0.350	-5.022	0.000	-2.441	-1.070
Last Activity_sms sent	0.8921	0.170	5.252	0.000	0.559	1.225
What is your current occupation_unemployed	-1.2592	0.321	-3.921	0.000	-1.889	-0.630
What is your current occupation_working professional	1.6008	0.399	4.016	0.000	0.820	2.382
Tags_busy	2.9100	0.306	9.515	0.000	2.311	3.509
Tags_closed by horizzon	9.1496	1.037	8.824	0.000	7.117	11.182
Tags_graduation in progress	1.9406	0.559	3.469	0.001	0.844	3.037
Tags_lost to eins	8.5710	0.759	11.287	0.000	7.083	10.059
Tags_ringing	-1.3097	0.309	-4.245	0.000	-1.914	-0.705
Tags_switched off	-1.2880	0.565	-2.278	0.023	-2.396	-0.180
Tags_will revert after reading the email	3.5834	0.212	16.914	0.000	3.168	3.999
Last Notable Activity_email opened	1.3653	0.109	12.480	0.000	1.151	1.580
Last Notable Activity_other	2.1162	0.308	6.879	0.000	1.513	2.719
Last Notable Activity_sms sent	2.5348	0.180	14.089	0.000	2.182	2.887

	Features	VIF
0	const	45.76
5	Last Activity_sms sent	4.05
7	What is your current occupation_working profes...	3.99
6	What is your current occupation_unemployed	3.93
17	Last Notable Activity_sms sent	3.78
14	Tags_will revert after reading the email	2.03
12	Tags_ringing	1.68
9	Tags_closed by horizzon	1.42
15	Last Notable Activity_email opened	1.39
3	Lead Origin_lead add form	1.35
2	Page Views Per Visit	1.30
1	Total Time Spent on Website	1.25
13	Tags_switched off	1.17
8	Tags_busy	1.13
4	Last Activity_email bounced	1.11
11	Tags_lost to eins	1.11
16	Last Notable Activity_other	1.08
10	Tags_graduation in progress	1.05

Model Evaluation

- Initially, predicted probability based 0.5 as cut-off & results were as below,
  - Accuracy: 89.53 %, Recall: 83.5 %, Precision: 88.41 %, F1 Score: 85.88 %
- Found out new Optimal cut-off as 0.4 using ROC curve and Precision-Recall Trade-off & results were as below,
  - Accuracy: 89.33 %, Recall: 86.13 %, Precision: 85.92 %, F1 Score: 86.03 %
  - Applied final model on test data: Accuracy: 89.21 %, Recall: 86.94 %, Precision: 85.92 %, F1 Score: 86.43 %





Thank You!