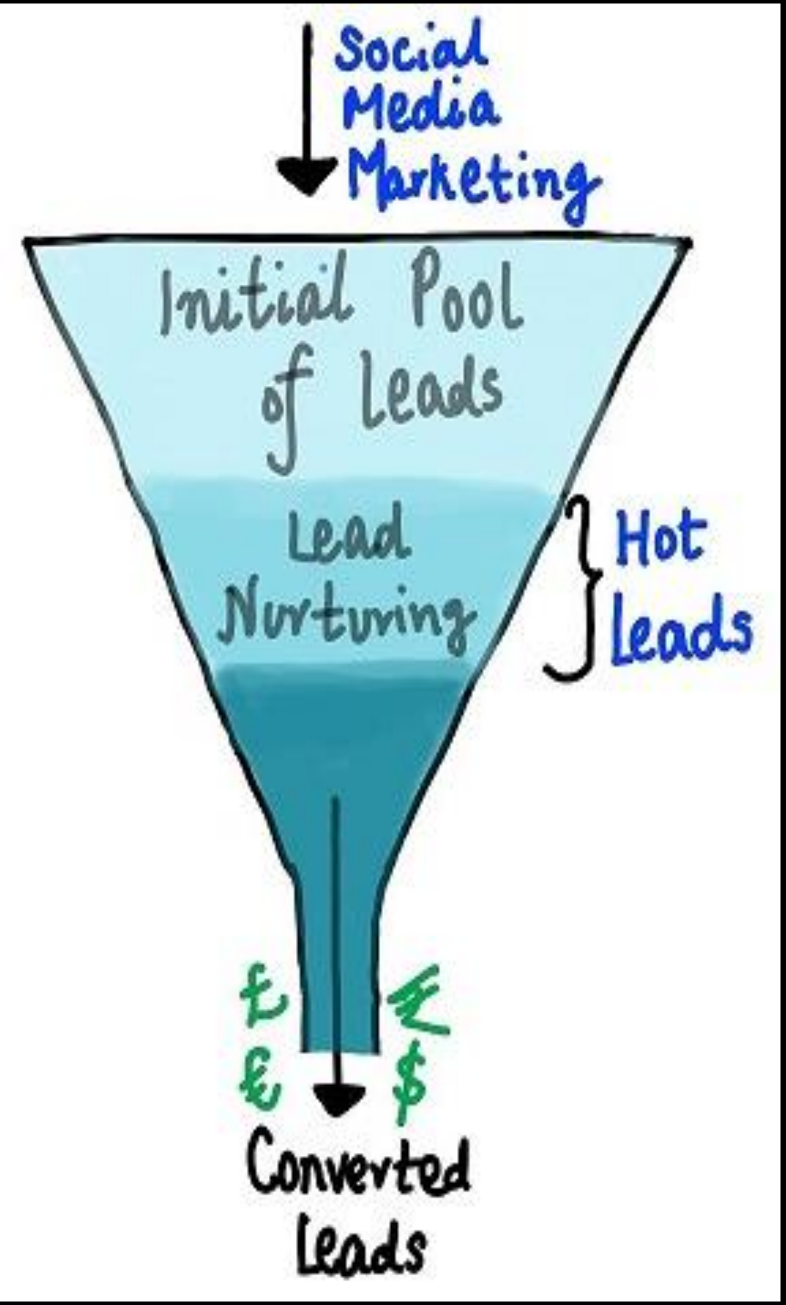
Lead Scoring Case study

By

Rahul

Srinivasan

Vijaysampath



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. A typical lead conversion process can be represented using the following funnel:

- As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

- 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%.

Analysis Approach

Data understanding, preparation and EDA

- ❖ Importing, Inspecting & Data Preparation

Model building

- ❖ Test-Train Split & Feature Scaling
- ❖ Initial Model Building
- ❖ Feature Selection Using RFE
- ❖ Final Model Building

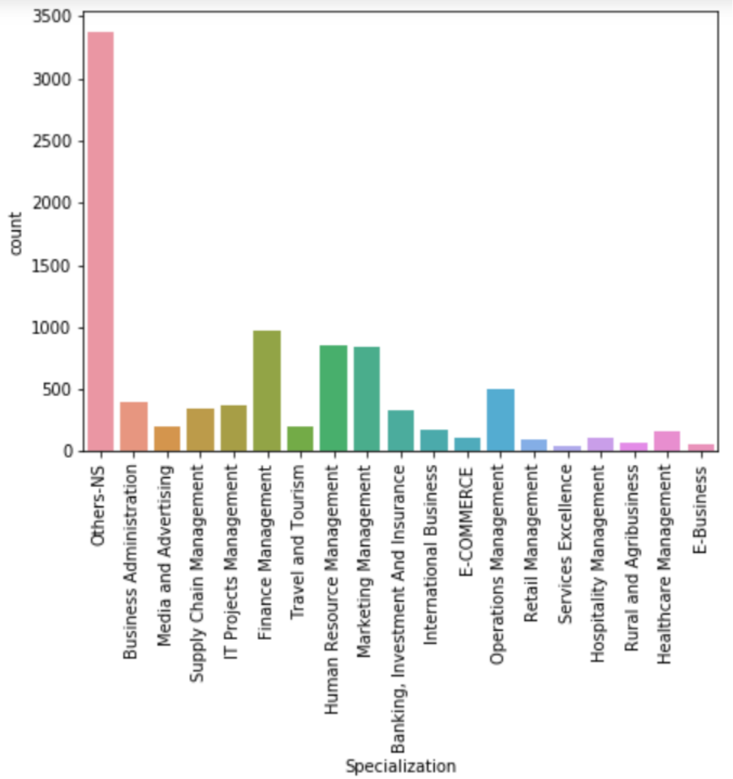
Evaluation

- ❖ Plotting the ROC Curve
- ❖ Finding Optimal Cut off Point
- ❖ Making Predictions on the test set
- ❖ Comparing Accuracy, Sensitivity & Specificity between Train and Test data
- ❖ Assign Lead score based on converted Probability

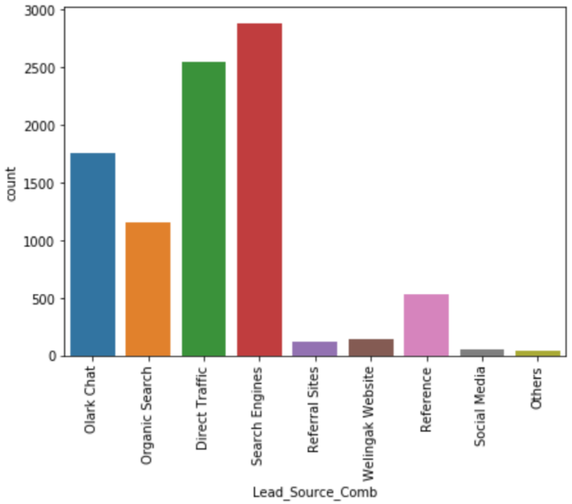
Importing, Inspecting & Data Preparation

- **Converting Binary Variables(Yes, No) to (0, 1).**
- **Prospect ID is unique for each customers.**
- **Converting “Select” to np.NaN as Select is the default value from front end system.**
- **Columns with Null values greater than 45% are dropped from dataset as they will not provide useful information.**
- **Impute mean(if no outlier) and median(if outlier) values for continuous variables and mode for Categorical variables.**
- **Dropping fields where % of value for single category is greater than 99%.**
- **Remove outliers from continuous variables field as they will disturb the results of model.**
- **Tags Field has been removed as it is not adding much value to business.**

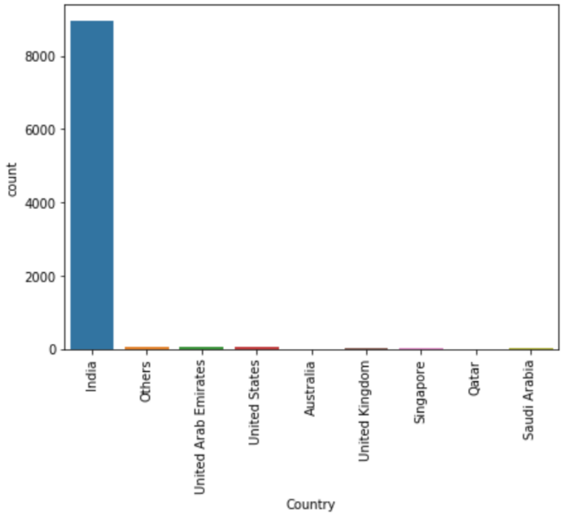
Importing, Inspecting & Data Preparation



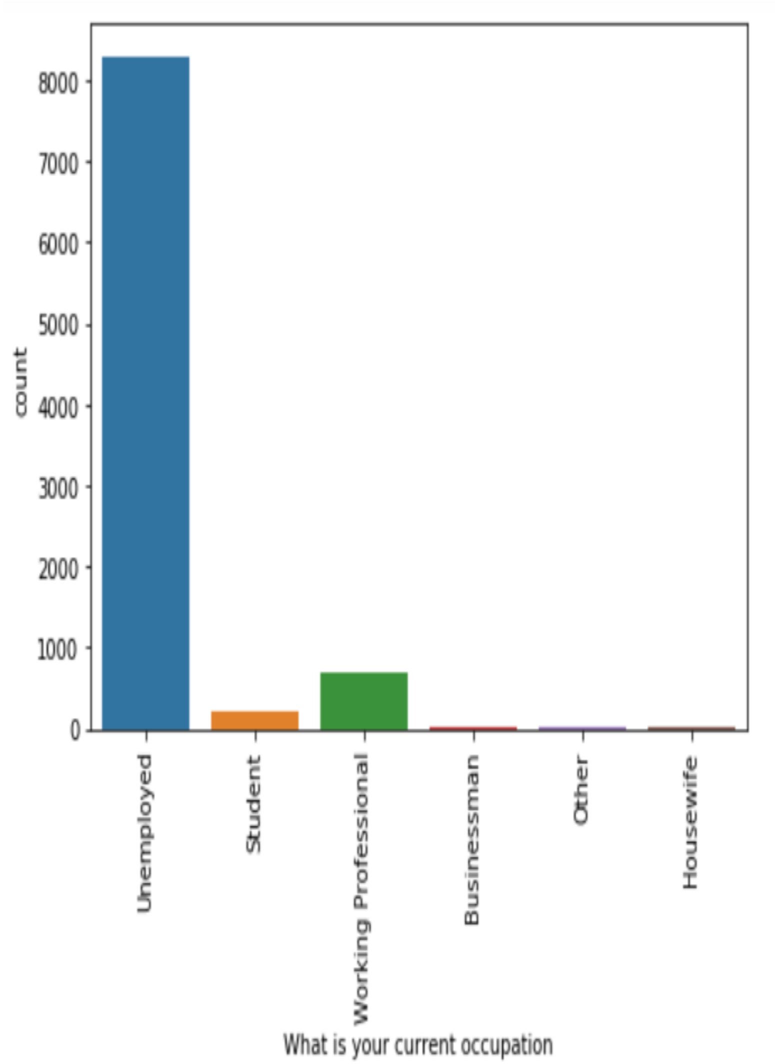
Specialization



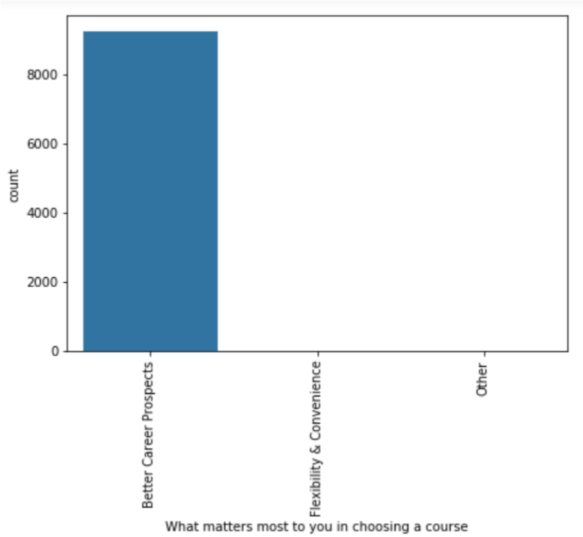
Lead Source



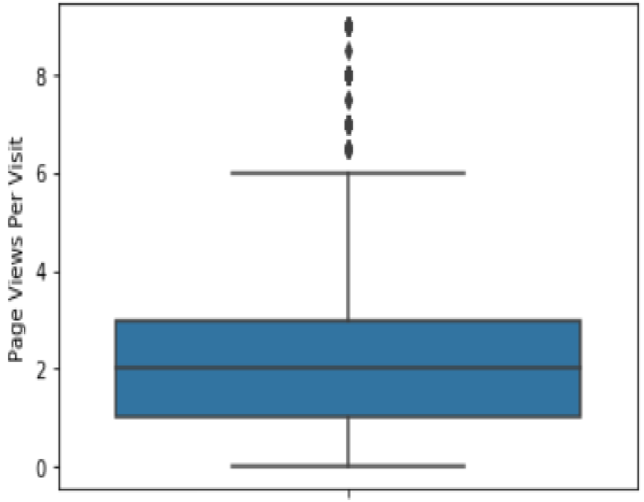
Country



What is your current occupation

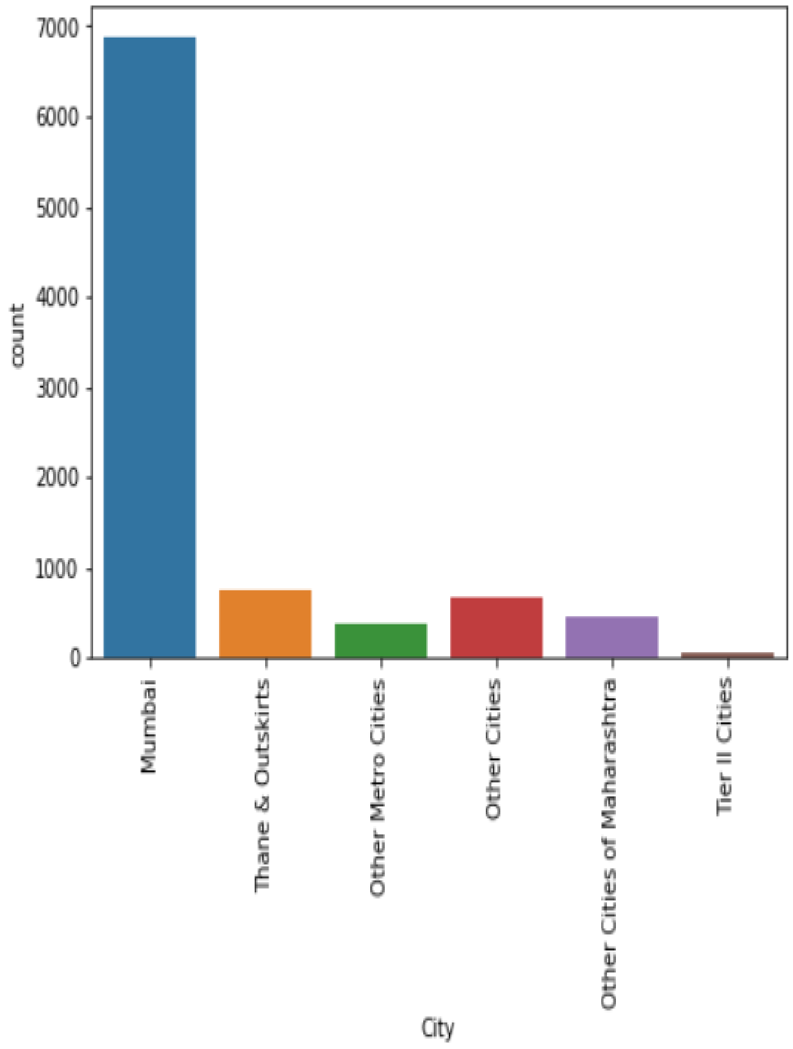


What matters most to you in choosing a course

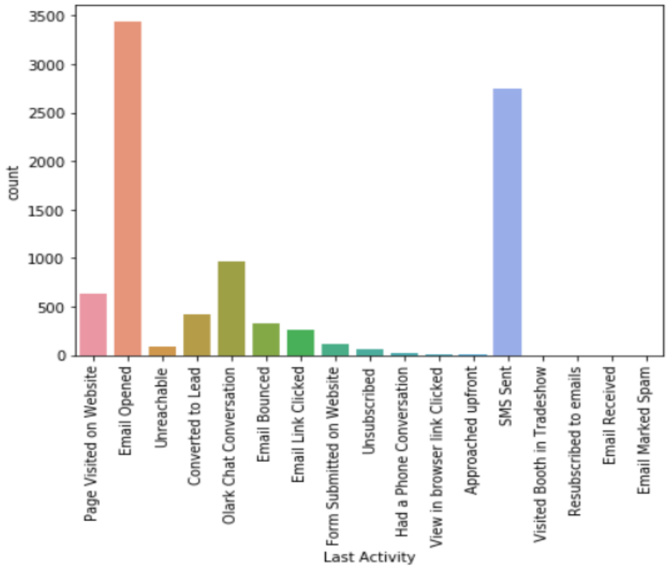


Pages Views Per Visit

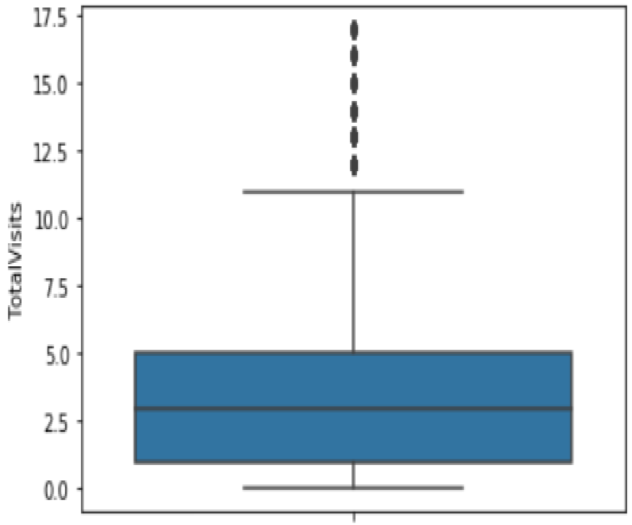
Importing, Inspecting & Data Preparation



City



Last Activity



Total Visits

Importing, Inspecting & Data Preparation

Test-Train Split & Feature Scaling

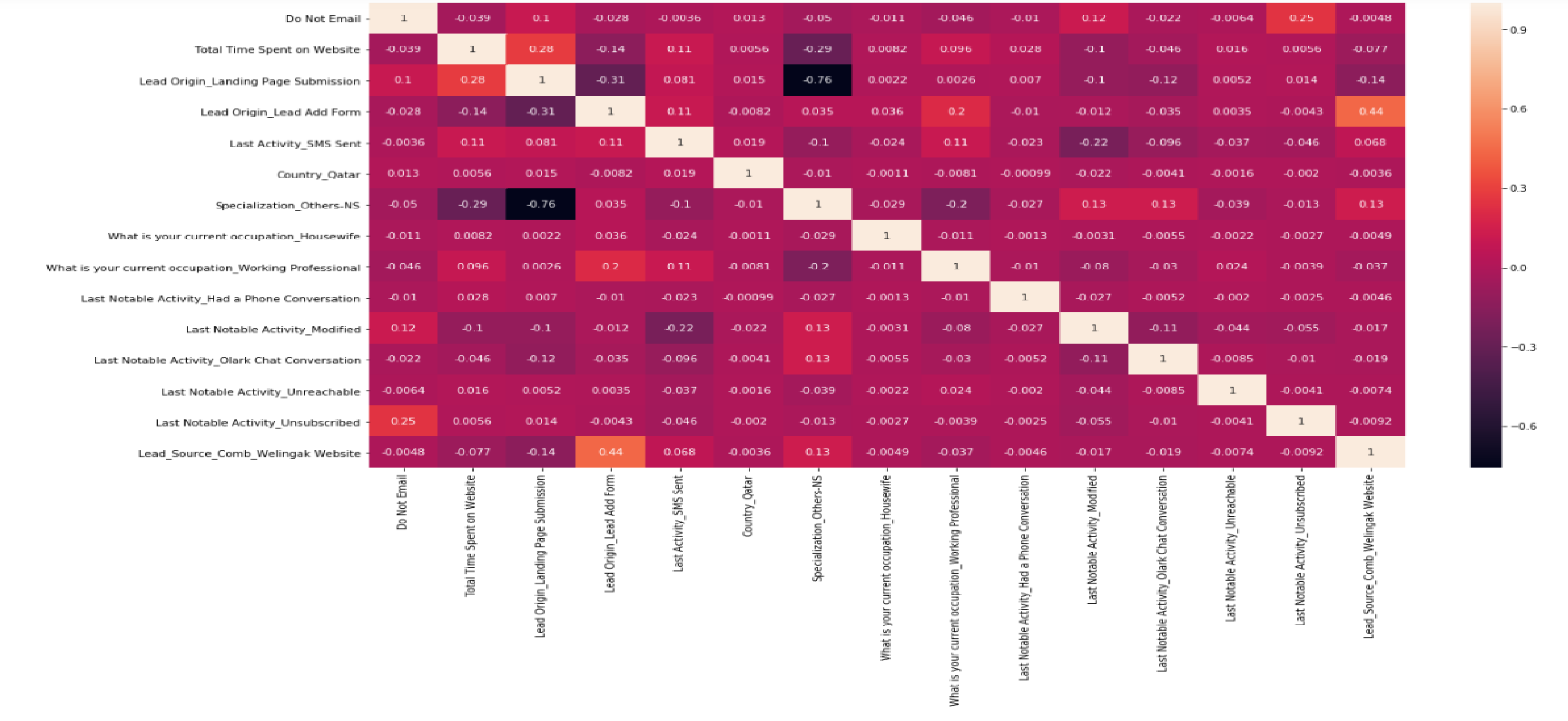
- **Splitting Dataset into Train and Test as 70% and 30%**
- **Fit Transform on training data using standard scaler**

Initial Model Building

- **Most of the features have high probability. Hence going to use RFE to eliminate not required features**

Feature Selection Using RFE

- **Heat Map of RFE selected 15 variables**

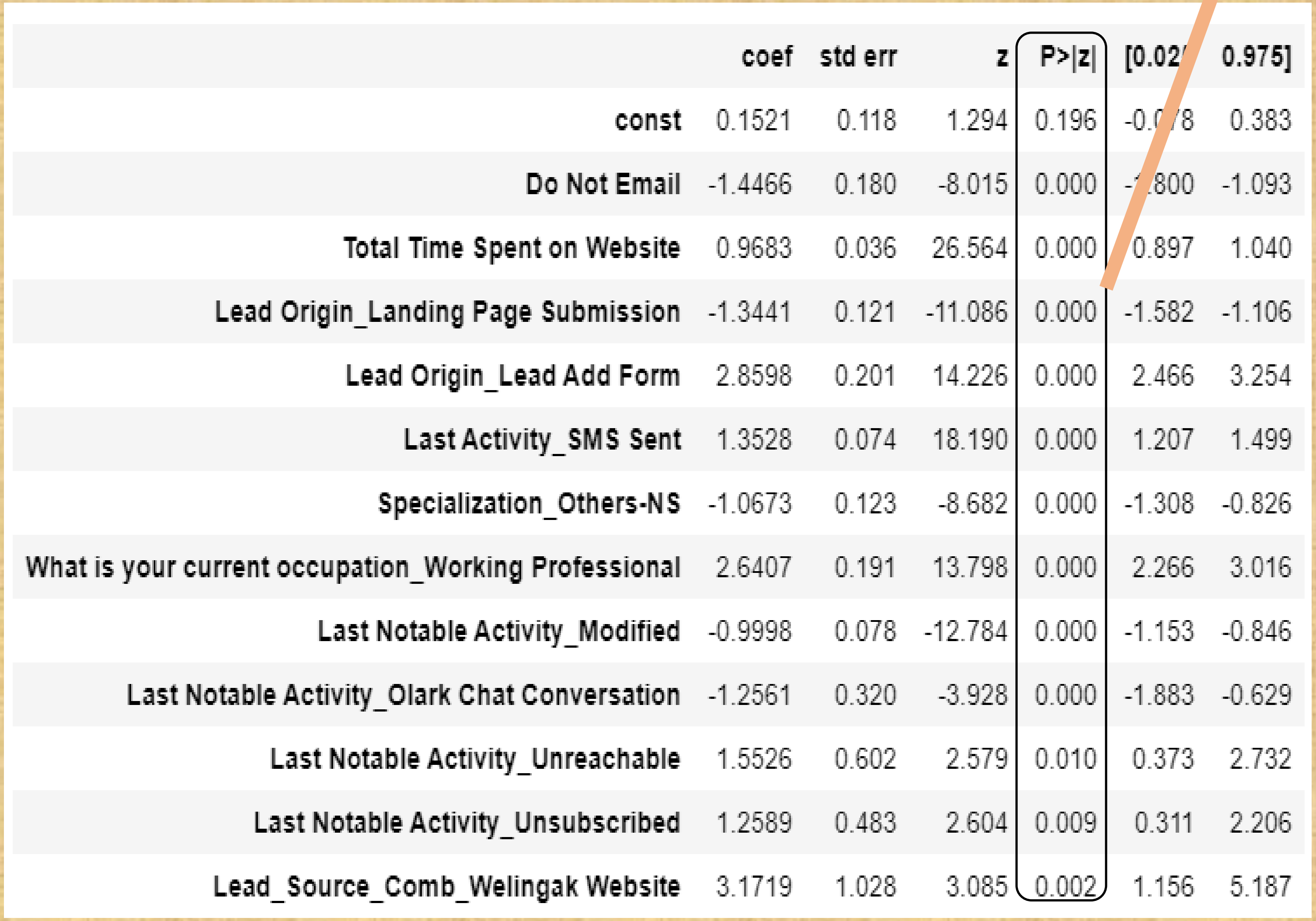


Final Model Building

Probability values are less than 5%

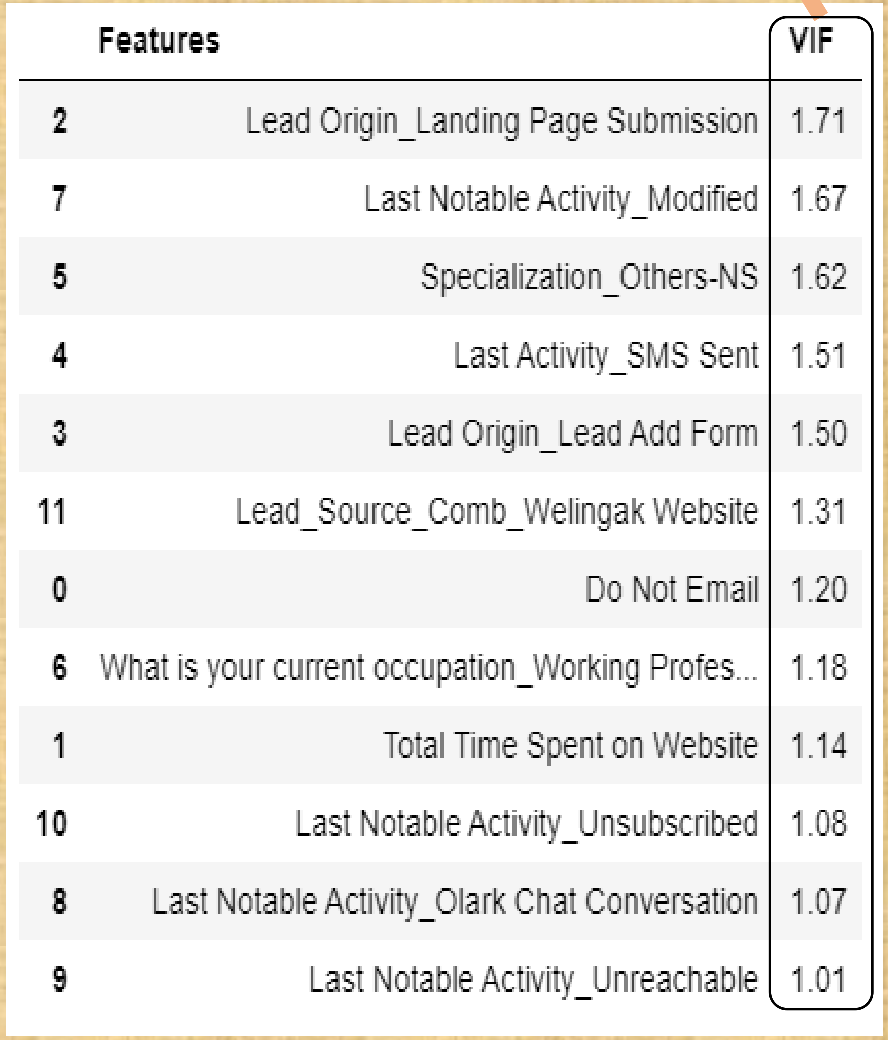
VIF values are less than 5

❖ **Final Model features with Probability Values**



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1521	0.118	1.294	0.196	-0.078	0.383
Do Not Email	-1.4466	0.180	-8.015	0.000	-1.800	-1.093
Total Time Spent on Website	0.9683	0.036	26.564	0.000	0.897	1.040
Lead Origin_Landing Page Submission	-1.3441	0.121	-11.086	0.000	-1.582	-1.106
Lead Origin_Lead Add Form	2.8598	0.201	14.226	0.000	2.466	3.254
Last Activity_SMS Sent	1.3528	0.074	18.190	0.000	1.207	1.499
Specialization_Others-NS	-1.0673	0.123	-8.682	0.000	-1.308	-0.826
What is your current occupation_Working Professional	2.6407	0.191	13.798	0.000	2.266	3.016
Last Notable Activity_Modified	-0.9998	0.078	-12.784	0.000	-1.153	-0.846
Last Notable Activity_Olark Chat Conversation	-1.2561	0.320	-3.928	0.000	-1.883	-0.629
Last Notable Activity_Unreachable	1.5526	0.602	2.579	0.010	0.373	2.732
Last Notable Activity_Unsubscribed	1.2589	0.483	2.604	0.009	0.311	2.206
Lead_Source_Comb_Welingak Website	3.1719	1.028	3.085	0.002	1.156	5.187

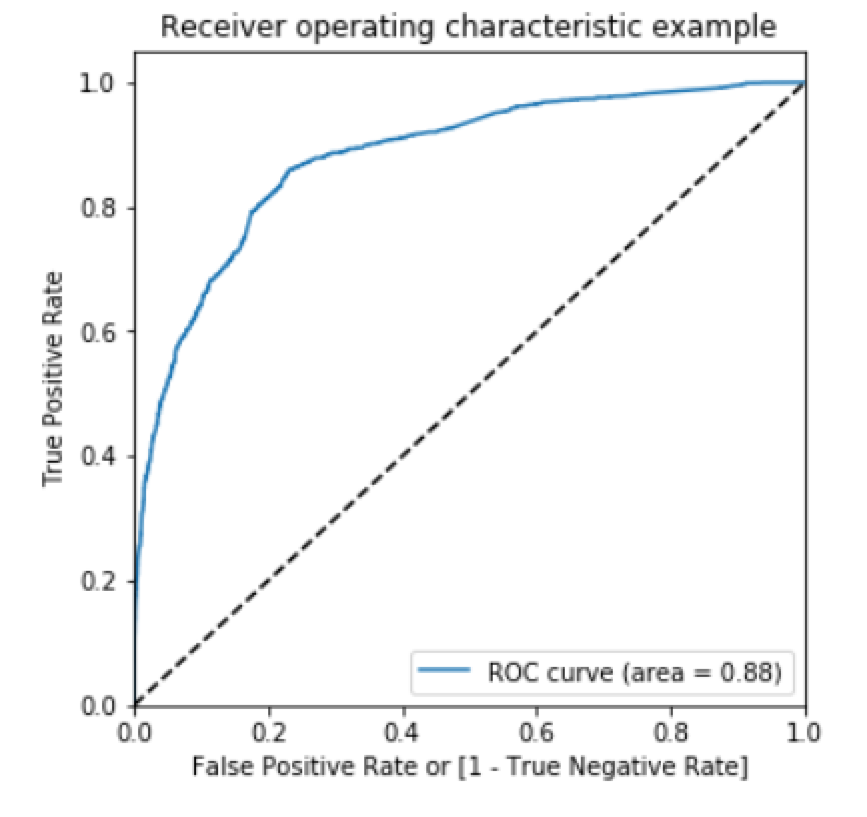
❖ **VIF values of final list of Features**



	Features	VIF
2	Lead Origin_Landing Page Submission	1.71
7	Last Notable Activity_Modified	1.67
5	Specialization_Others-NS	1.62
4	Last Activity_SMS Sent	1.51
3	Lead Origin_Lead Add Form	1.50
11	Lead_Source_Comb_Welingak Website	1.31
0	Do Not Email	1.20
6	What is your current occupation_Working Profes...	1.18
1	Total Time Spent on Website	1.14
10	Last Notable Activity_Unsubscribed	1.08
8	Last Notable Activity_Olark Chat Conversation	1.07
9	Last Notable Activity_Unreachable	1.01

Plotting the ROC Curve

- ROC Curve of the Model

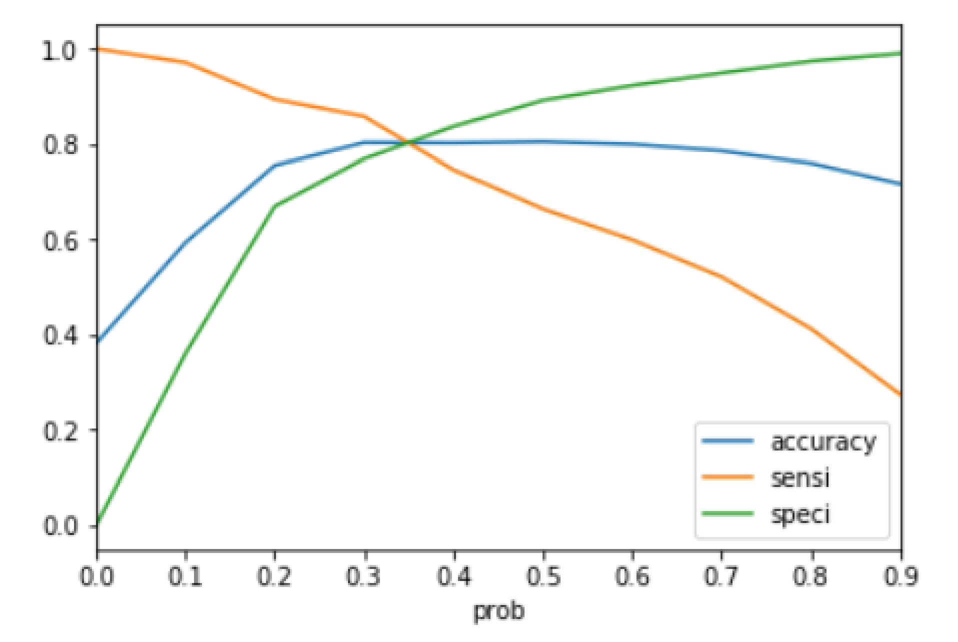


Observations:

- **ROC Value is 0.88 close to 1**
- **Area under the curve is maximum.**
- **Model is a good predictive model**

Finding Optimal Cut off Point

- **Accuracy, Sensitivity and Specificity plot**



Observations:

- **From the curve 0.33 is the optimum point taken as cut off Probability.**

Comparing Accuracy, Sensitivity & Specificity between Train and Test data

Train Data:

- Accuracy: 80.34%
- Sensitivity: 84.31%
- Specificity: 77.89%
- Precision: 70.17%
- Recall: 84.31%

Test Data:

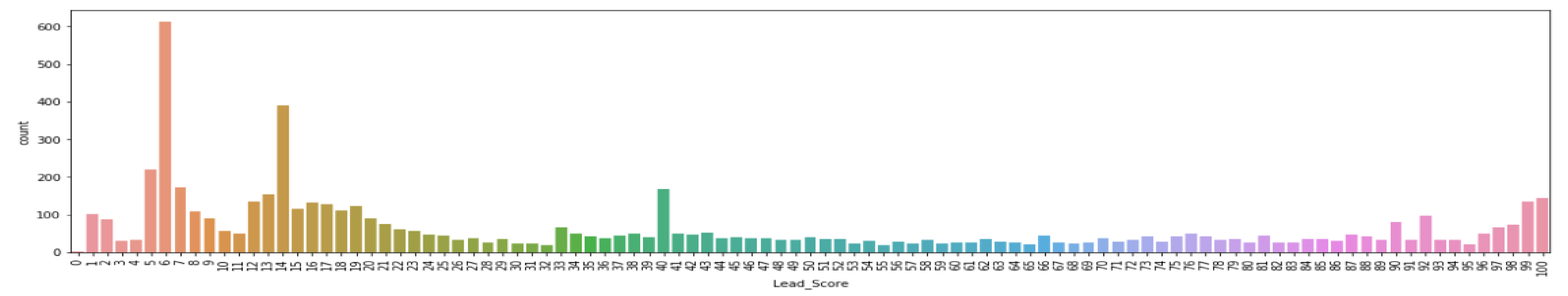
- Accuracy: 79.76% (~80%)
- Sensitivity: 84.07%
- Specificity: 76.99%
- Precision: 70.13%
- Recall: 84.07 %

Observations:

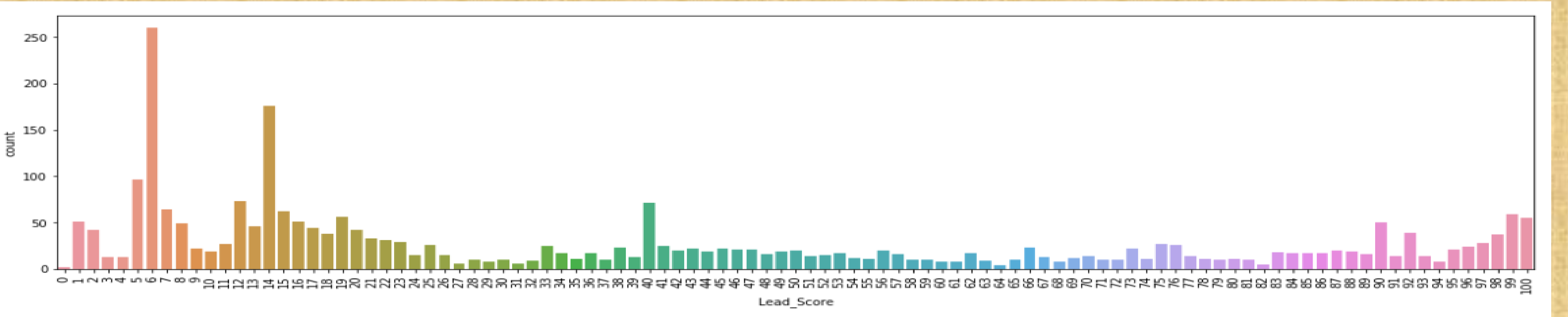
- Sensitivity value of both Train and Test data are greater than 80%
- Difference between Train and Test data in terms of Accuracy, Sensitivity & Specificity are less than 5%.

Assign Lead score based on converted Probability

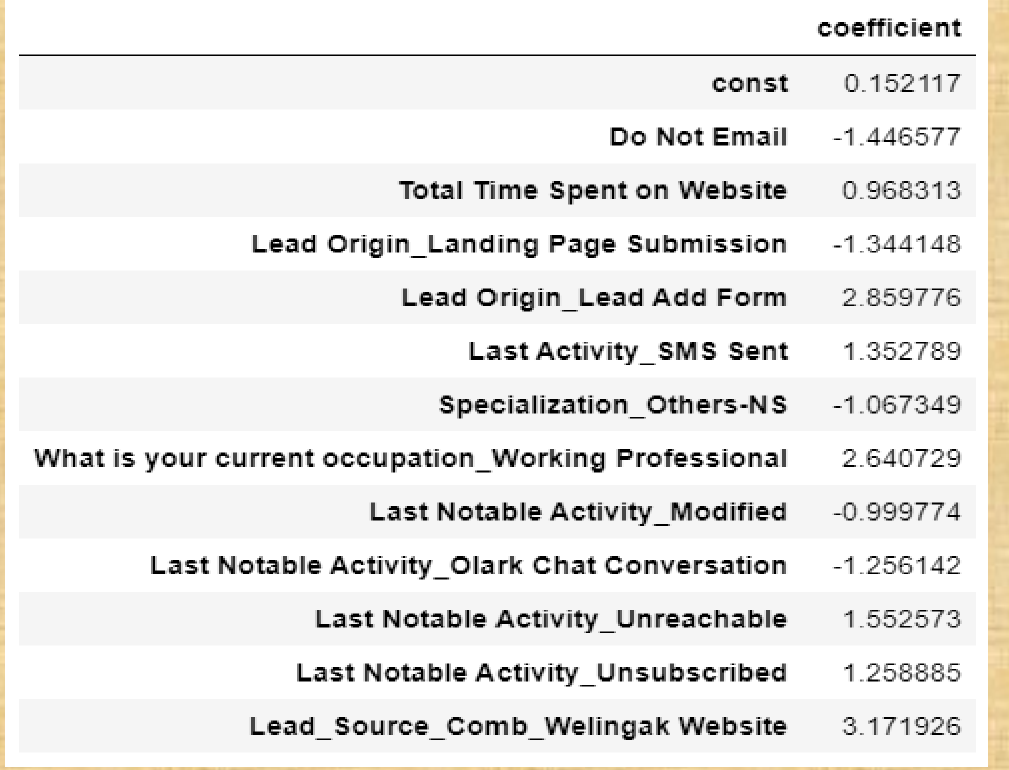
- ❖ **Lead score count plot on train dataset**



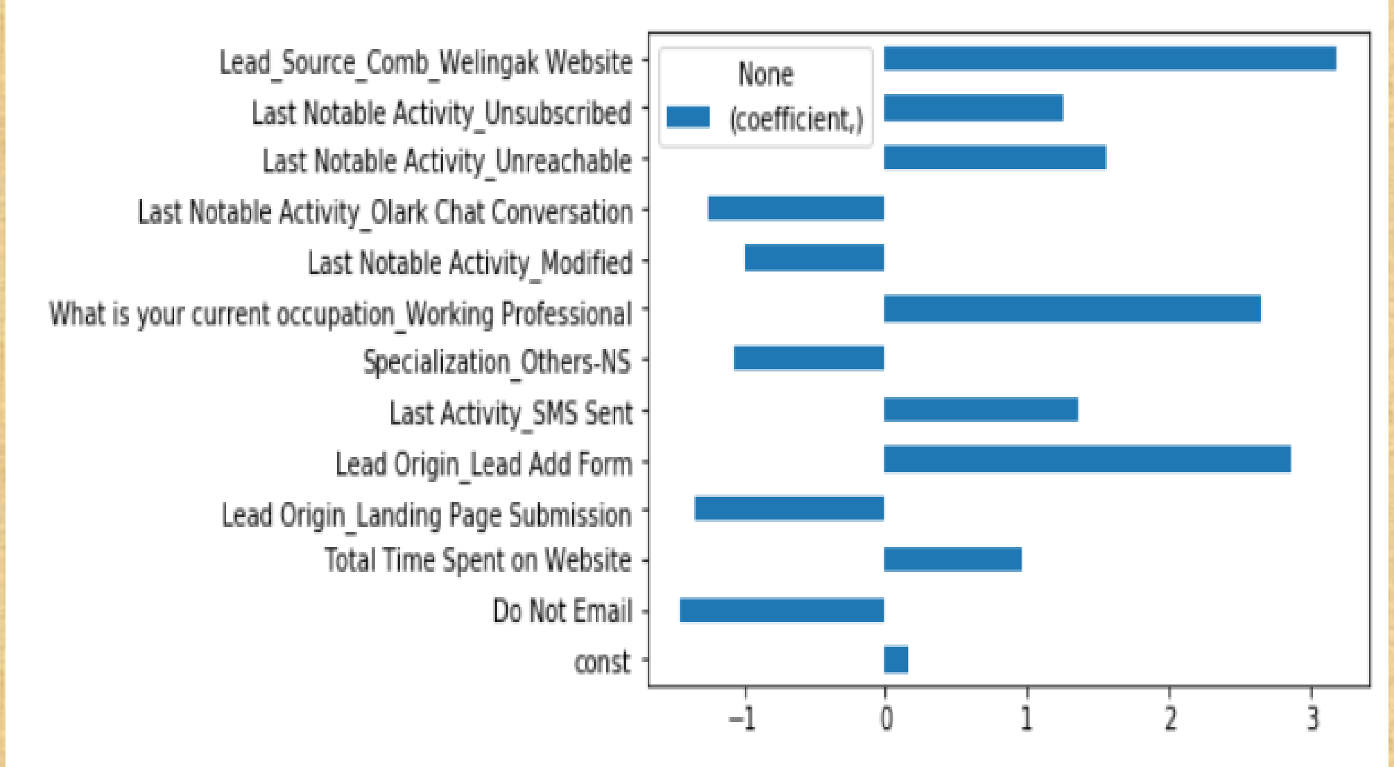
- ❖ **Lead score count plot on final predicted dataset**



Top three variables towards the probability of a lead getting converted



	coefficient
const	0.152117
Do Not Email	-1.446577
Total Time Spent on Website	0.968313
Lead Origin_Landing Page Submission	-1.344148
Lead Origin_Lead Add Form	2.859776
Last Activity_SMS Sent	1.352789
Specialization_Others-NS	-1.067349
What is your current occupation_Working Professional	2.640729
Last Notable Activity_Modified	-0.999774
Last Notable Activity_Olark Chat Conversation	-1.256142
Last Notable Activity_Unreachable	1.552573
Last Notable Activity_Unsubscribed	1.258885
Lead_Source_Comb_Welingak Website	3.171926



Observation:

❖ **Lead Source (Welingak Website), Lead Origin (Lead Add Form) & What is your current occupation (Working Professional)**

Recommendations

1) Lead Source (Welingak Website), Lead Origin (Lead Add Form) & What is your current occupation (Working Professional), Last Notable Activity (Unreachable) are the top 4 features that drive towards potential lead conversion.

2) Apply the model to real time data get the sales Team to reach out to hot leads based on their lead score which will result in achieving the quarterly targets early. This will help company to prioritize and convert the most probable leads to paying customers.

3) CEO can advise the sales Team to target as much as Working professionals as possible. Discounts or aggressive campaigns towards corporate and more corporate B2B tie-ups.

4) Invest more on channel partner websites such as Welingak by giving incentives or some revenue sharing model for each successful conversion.