Lead Score Case Study

By Srinivas Shenoy, Sri Harsha and Srushti Neoge

Problem Statement

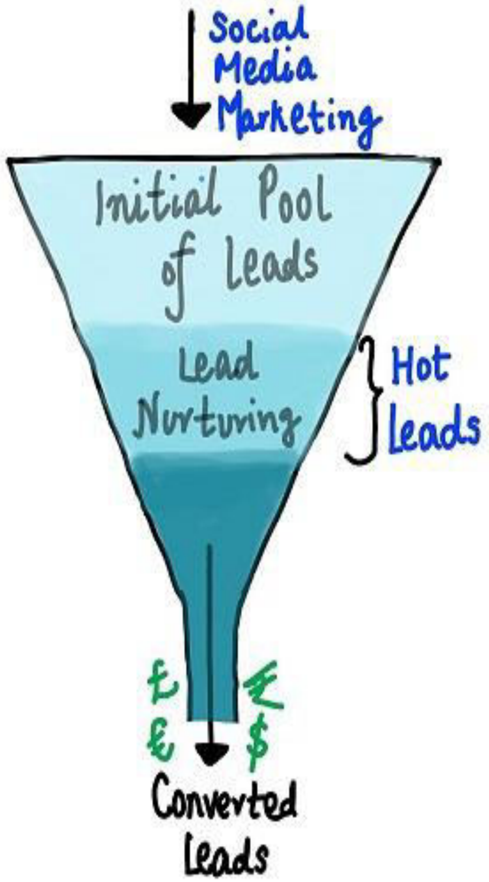
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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.

Problem Statement

conversion process can be represented using the following funnel:



Business Objective

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 a higher lead score have a higher conversion chance and the customers with a 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%.

Dataset

- The Case study was done using dataset from the past which has around 9000 data points

- Target Variable in this data set is the column Converted which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

- The data set also consits of many categorical variables

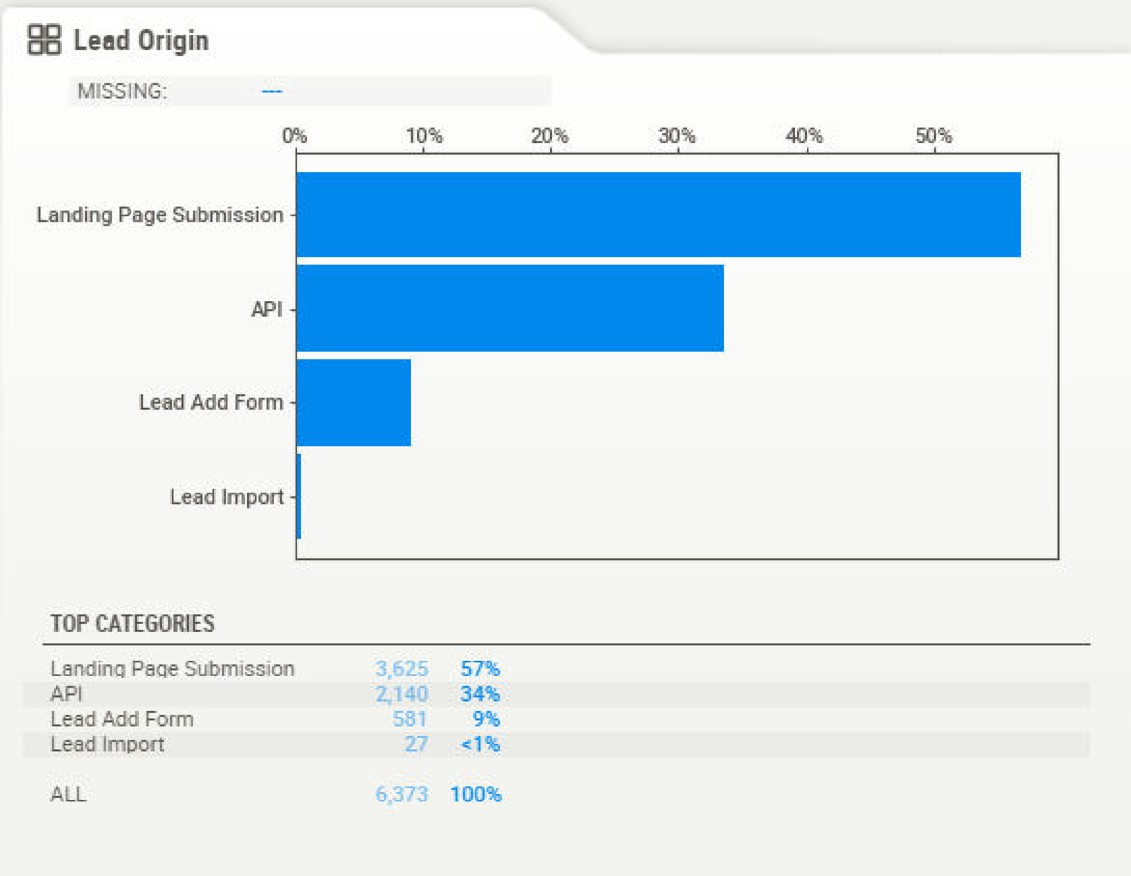
Steps followed

- Understanding the domain/variables.
- Import/Load the data.
- Check the structure/metadata of the data.
- Perform Exploratory Data Analysis
- Data cleaning and Preparation
- Train - Test split of data
- Scaling of variables
- Building a model using Logistic Regression and calculate Lead Score
- Model Evaluation on train data using metrics like Accuracy, Sensitivity, Specificity or Precision and Recall
- Using the final model on the test data
- Model Evaluation on Test data using metrics like Accuracy, Sensitivity, Specificity

Exploratory Data Analysis

- Sweet Viz was used to perform the EDA
- In the inital round of EDA, the colums with zero variance were removed
- Rows with missing values were removed
- Columns having Select which means blank selection as highest counts of values were removed
- Finally data frame was created with required variables and removal of missing values
- Post this another round of EDA was conducted and the same is presented in the following slides

Exploratory Data Analysis

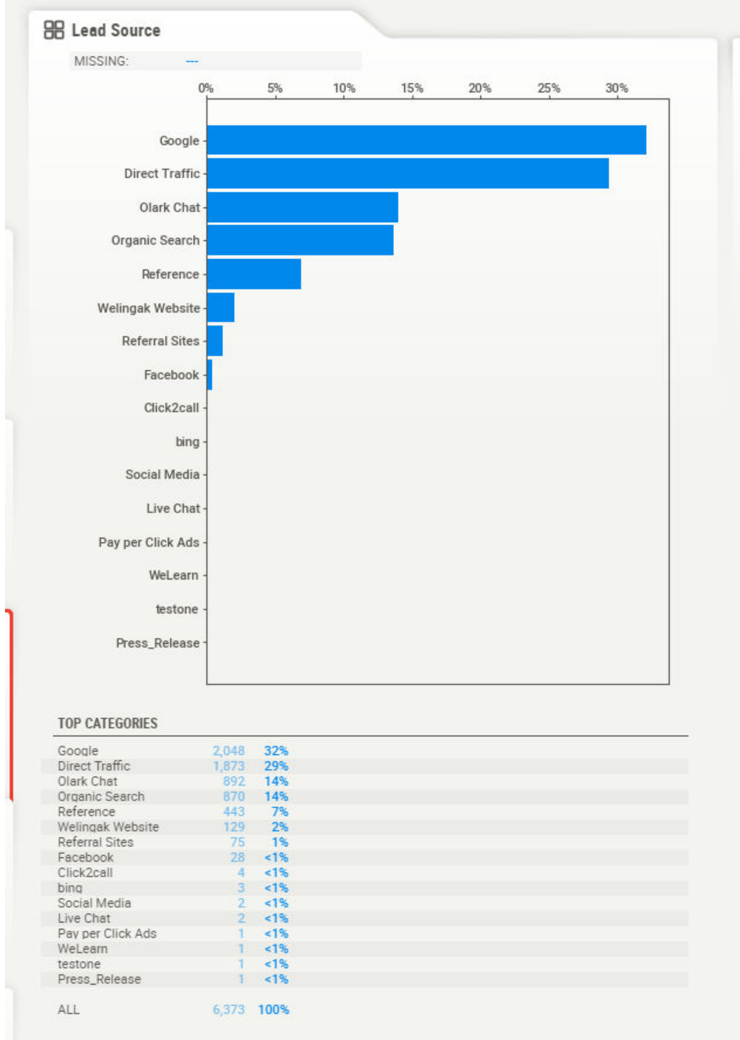


Lead Origin

MISSING: ---

TOP CATEGORIES

Landing Page Submission	3,625	57%
API	2,140	34%
Lead Add Form	581	9%
Lead Import	27	<1%
ALL	6,373	100%



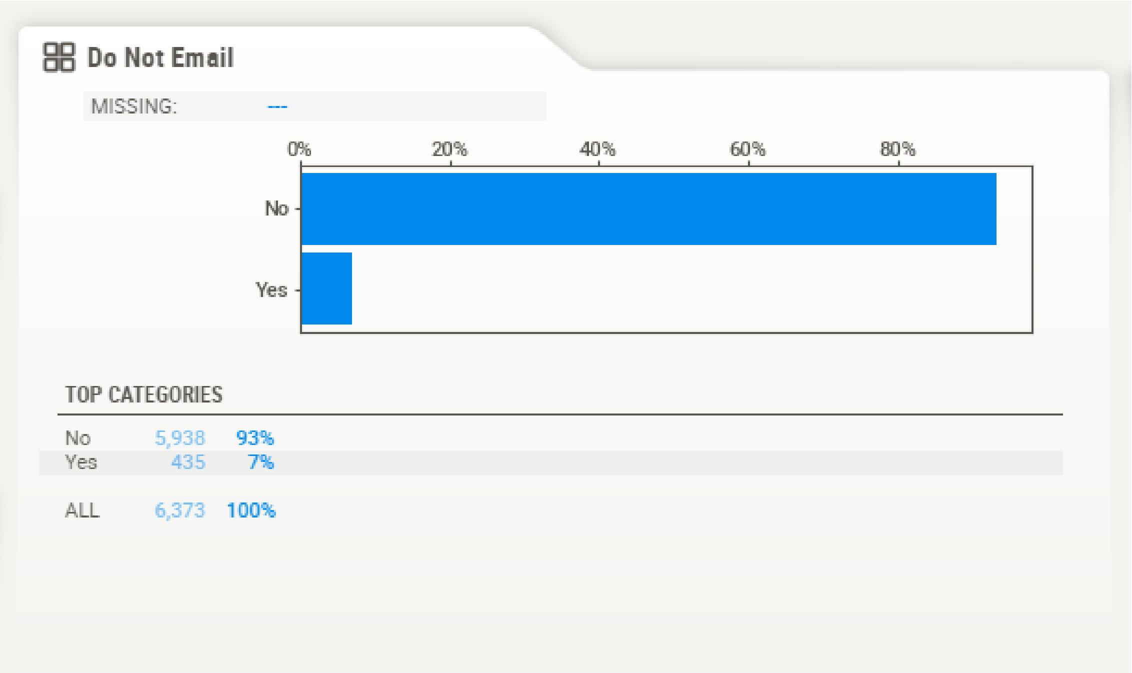
Lead Source

MISSING: ---

TOP CATEGORIES

Google	2,048	32%
Direct Traffic	1,873	29%
Olark Chat	892	14%
Organic Search	870	14%
Reference	443	7%
Welingak Website	129	2%
Referral Sites	75	1%
Facebook	28	<1%
Click2call	4	<1%
bing	3	<1%
Social Media	2	<1%
Live Chat	2	<1%
Pay per Click Ads	1	<1%
WeLearn	1	<1%
testone	1	<1%
Press_Release	1	<1%
ALL	6,373	100%

Exploratory Data Analysis

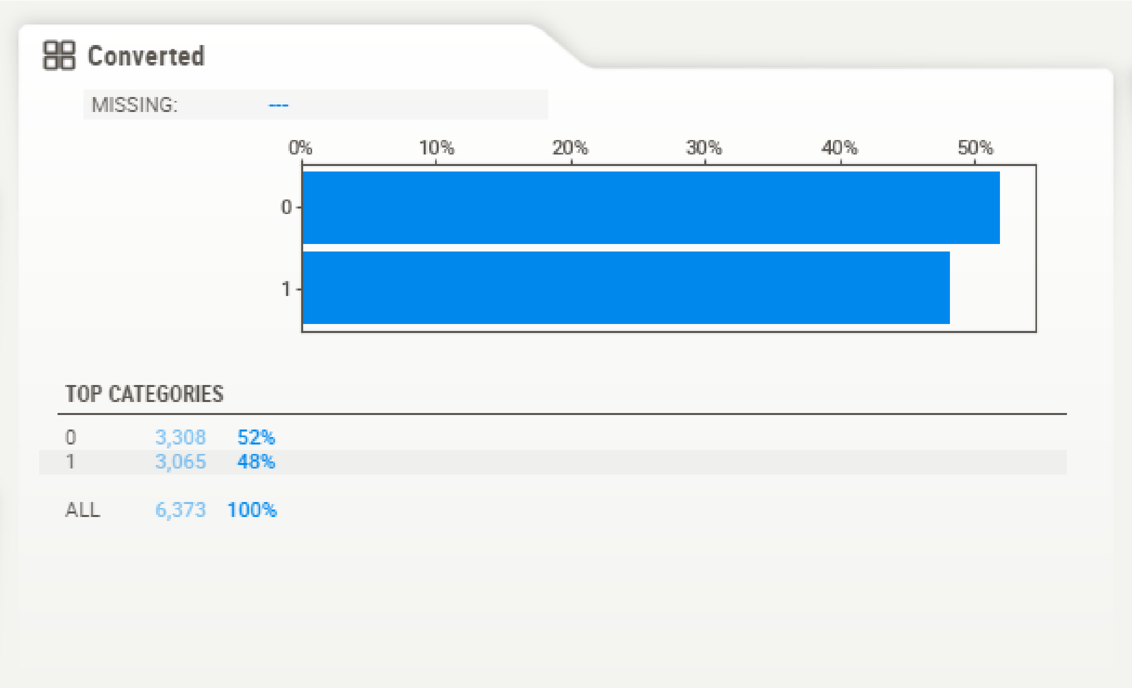


Do Not Email

MISSING: ---

TOP CATEGORIES

No	5,938	93%
Yes	435	7%
ALL	6,373	100%



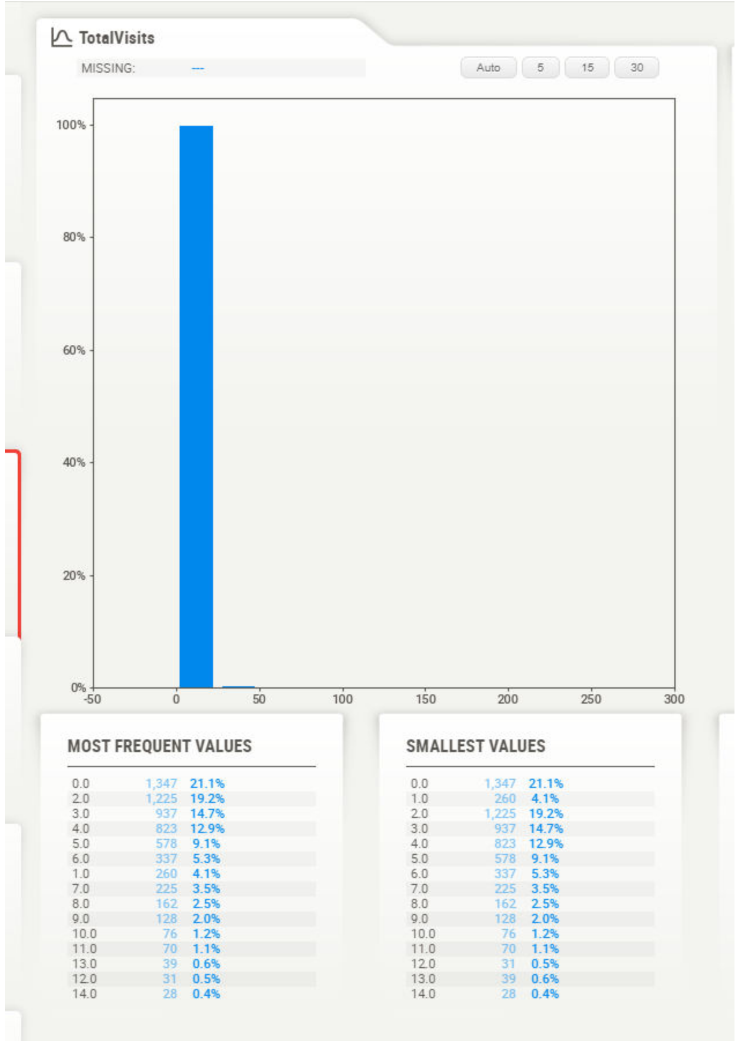
Converted

MISSING: ---

TOP CATEGORIES

0	3,308	52%
1	3,065	48%
ALL	6,373	100%

Exploratory Data Analysis

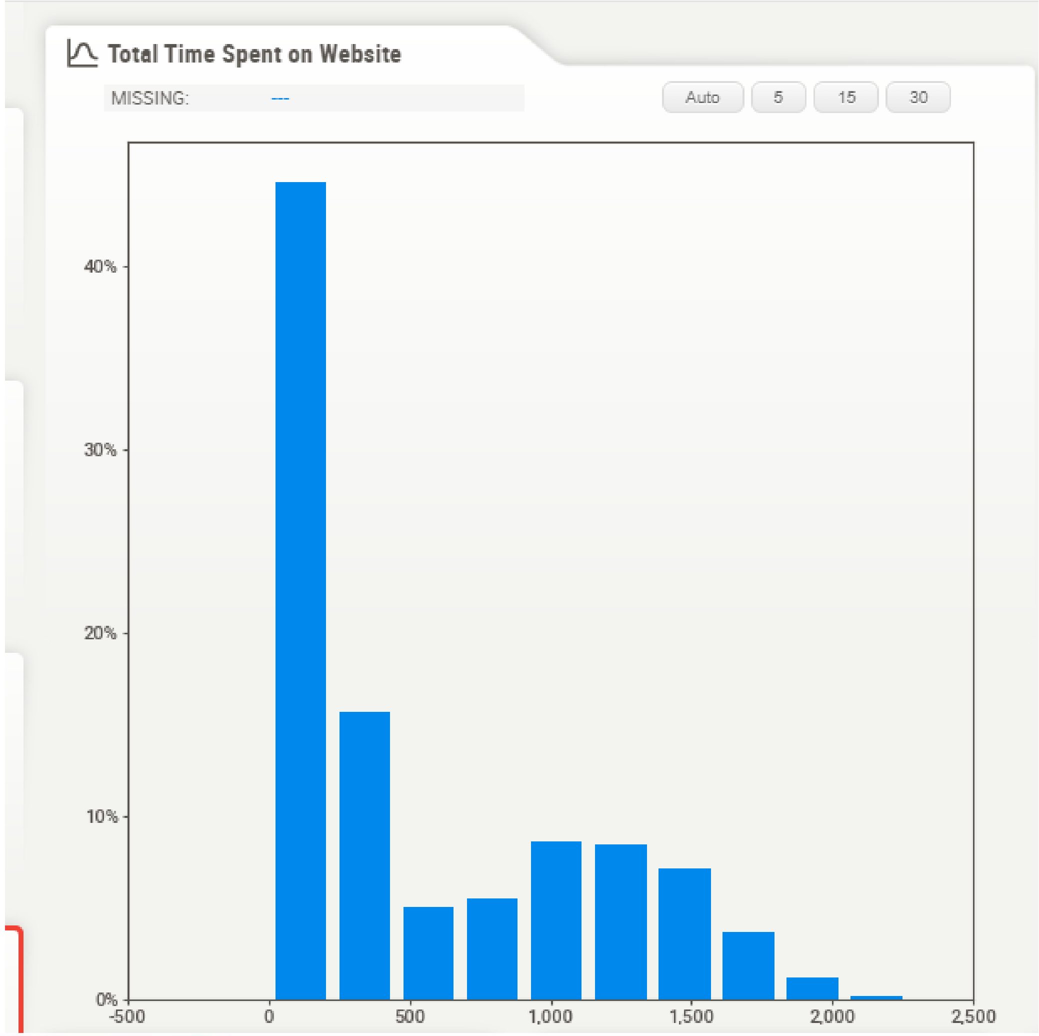


MOST FREQUENT VALUES

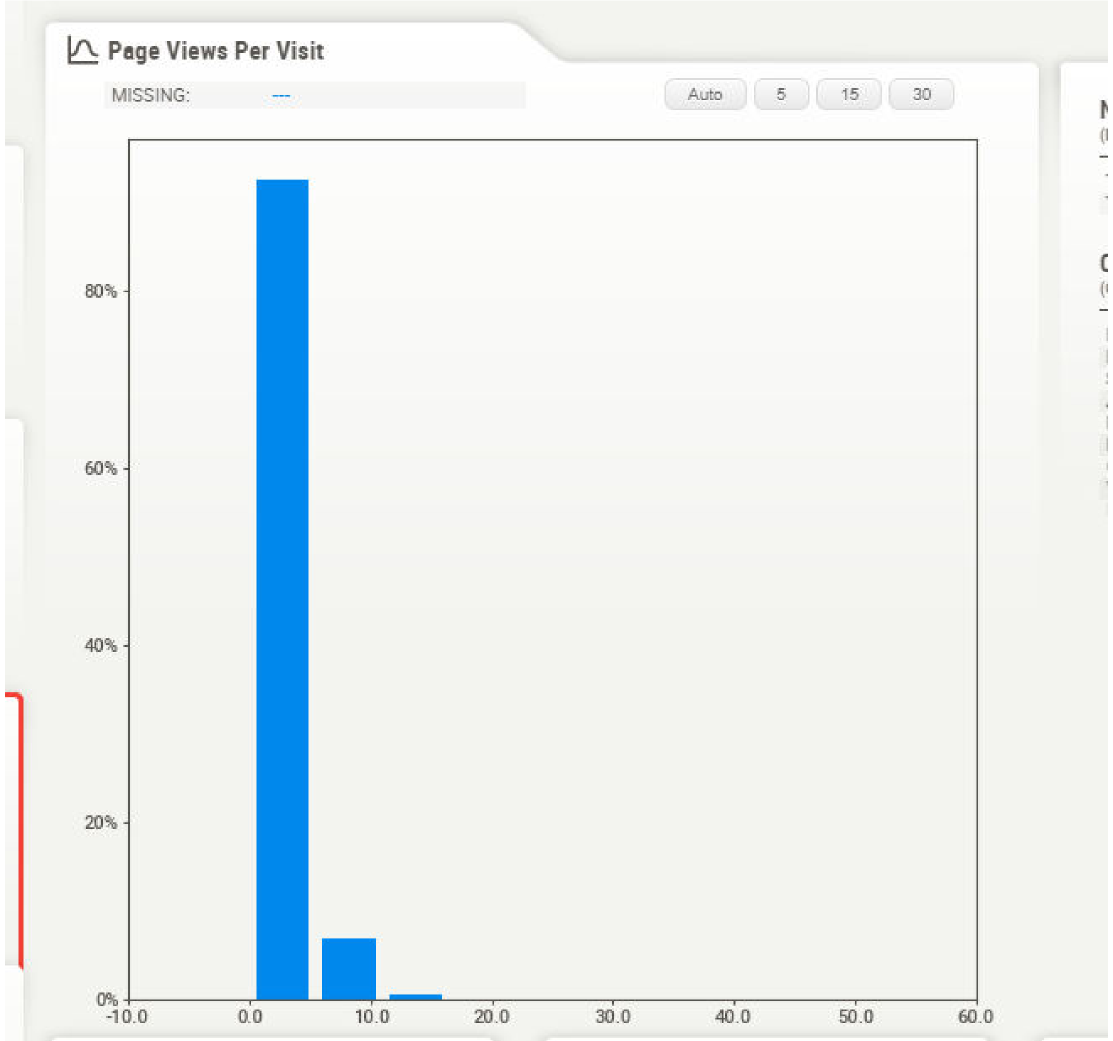
0.0	1,347	21.1%
2.0	1,225	19.2%
3.0	937	14.7%
4.0	823	12.9%
5.0	578	9.1%
6.0	337	5.3%
1.0	260	4.1%
7.0	225	3.5%
8.0	162	2.5%
9.0	128	2.0%
10.0	76	1.2%
11.0	70	1.1%
13.0	39	0.6%
12.0	31	0.5%
14.0	28	0.4%

SMALLEST VALUES

0.0	1,347	21.1%
1.0	260	4.1%
2.0	1,225	19.2%
3.0	937	14.7%
4.0	823	12.9%
5.0	578	9.1%
6.0	337	5.3%
7.0	225	3.5%
8.0	162	2.5%
9.0	128	2.0%
10.0	76	1.2%
11.0	70	1.1%
12.0	31	0.5%
13.0	39	0.6%
14.0	28	0.4%

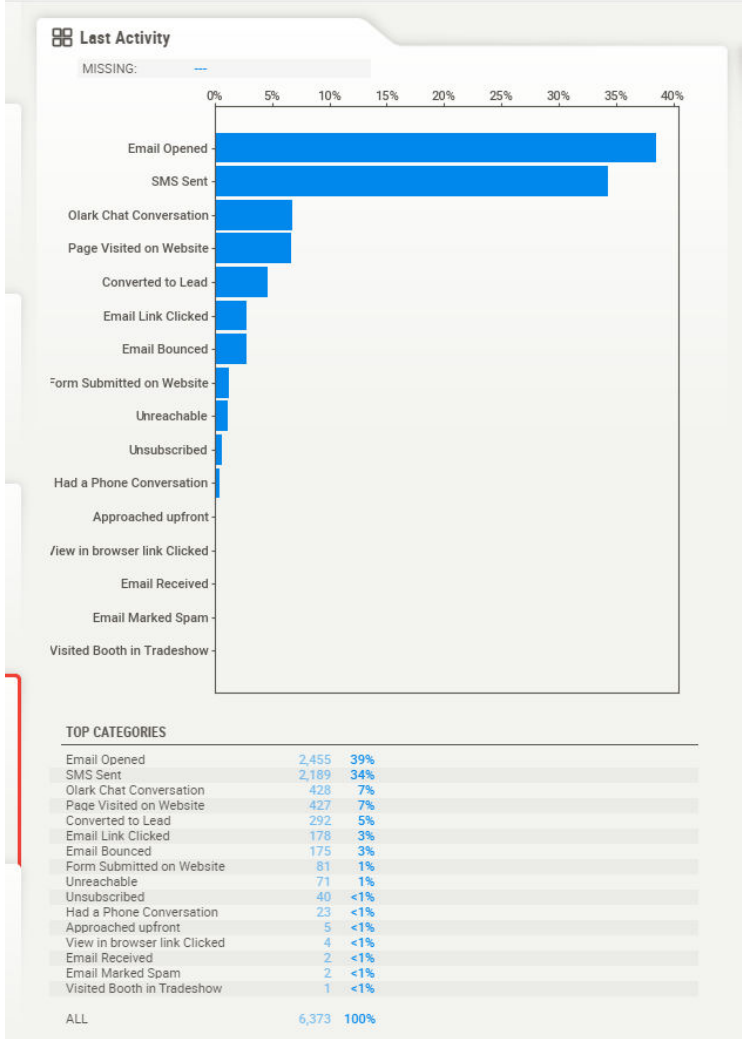


Exploratory Data Analysis



Page Views Per Visit

MISSING: ---



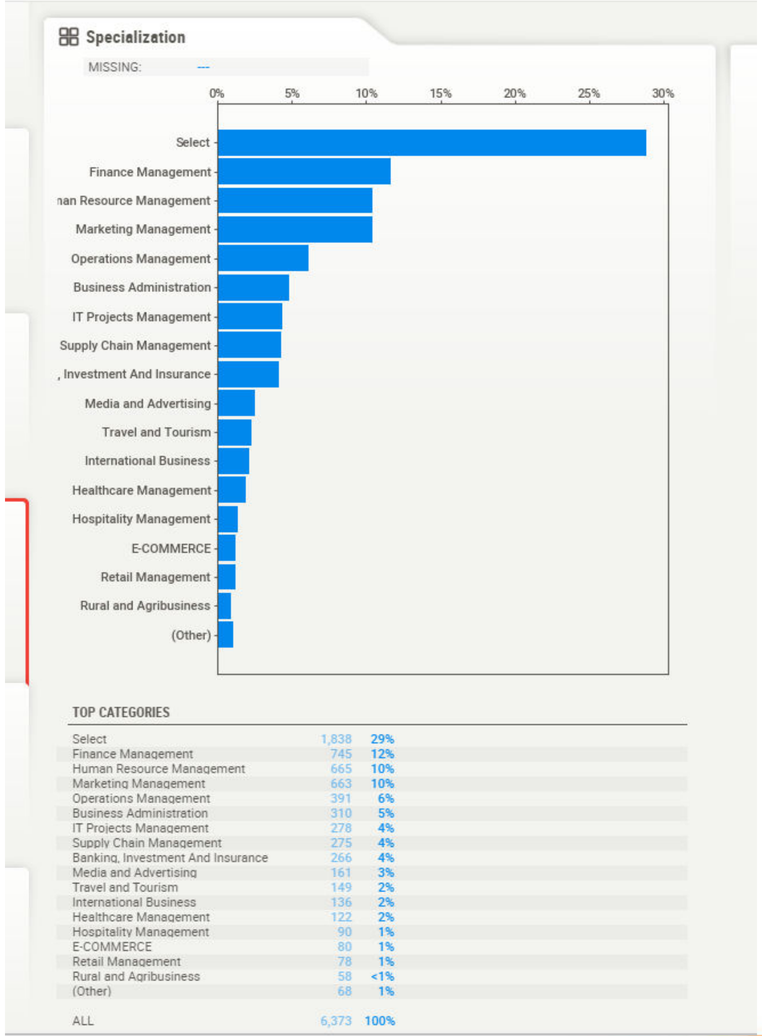
Last Activity

MISSING: ---

TOP CATEGORIES

Email Opened	2,455	39%
SMS Sent	2,189	34%
Olark Chat Conversation	428	7%
Page Visited on Website	427	7%
Converted to Lead	292	5%
Email Link Clicked	178	3%
Email Bounced	175	3%
Form Submitted on Website	81	1%
Unreachable	71	1%
Unsubscribed	40	<1%
Had a Phone Conversation	23	<1%
Approached upfront	5	<1%
View in browser link Clicked	4	<1%
Email Received	2	<1%
Email Marked Spam	2	<1%
Visited Booth in Tradeshow	1	<1%
ALL	6,373	100%

Exploratory Data Analysis

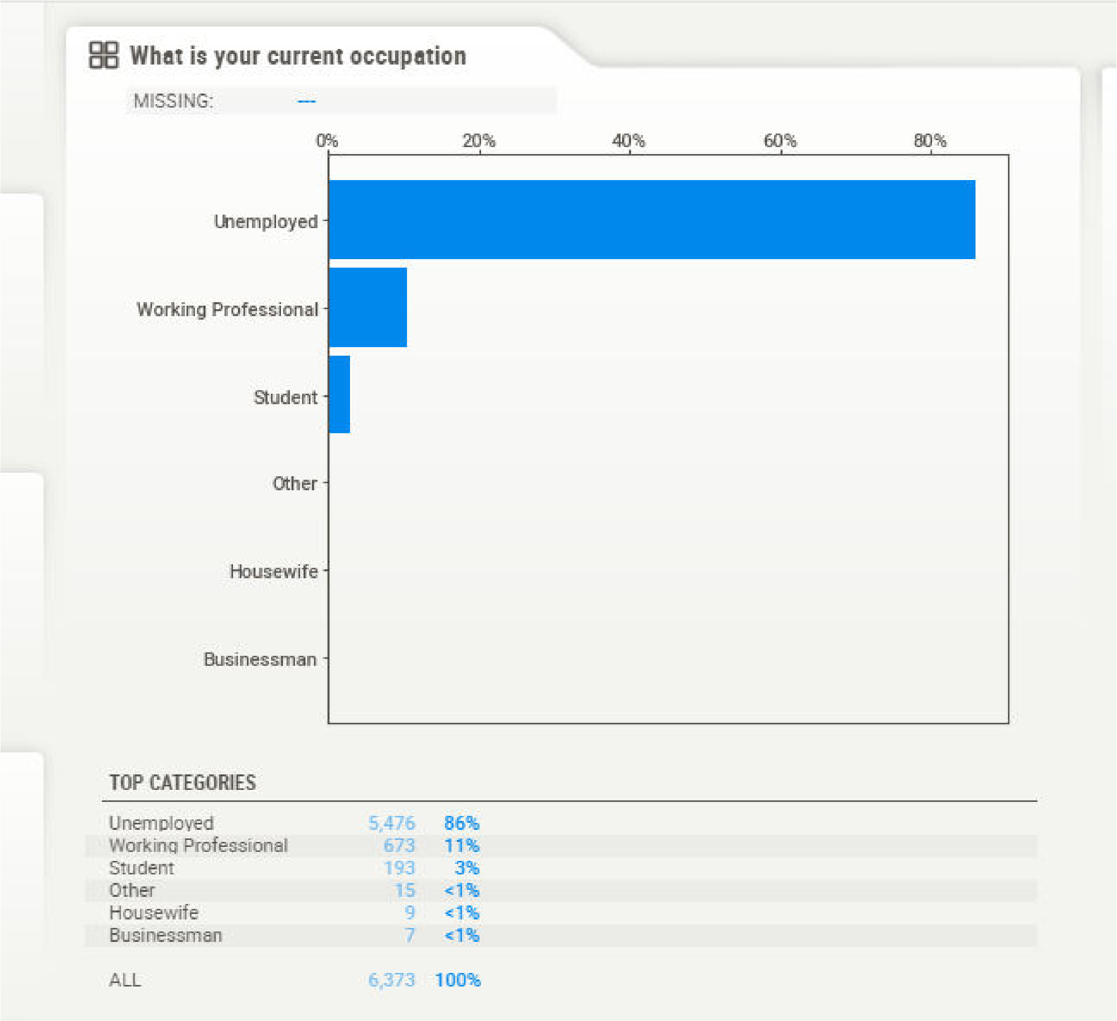


Specialization

MISSING: ---

TOP CATEGORIES

Select	1,838	29%
Finance Management	745	12%
Human Resource Management	665	10%
Marketing Management	663	10%
Operations Management	391	6%
Business Administration	310	5%
IT Projects Management	278	4%
Supply Chain Management	275	4%
Banking, Investment And Insurance	266	4%
Media and Advertising	161	3%
Travel and Tourism	149	2%
International Business	136	2%
Healthcare Management	122	2%
Hospitality Management	90	1%
E-COMMERCE	80	1%
Retail Management	78	1%
Rural and Agribusiness	58	<1%
(Other)	68	1%
ALL	6,373	100%



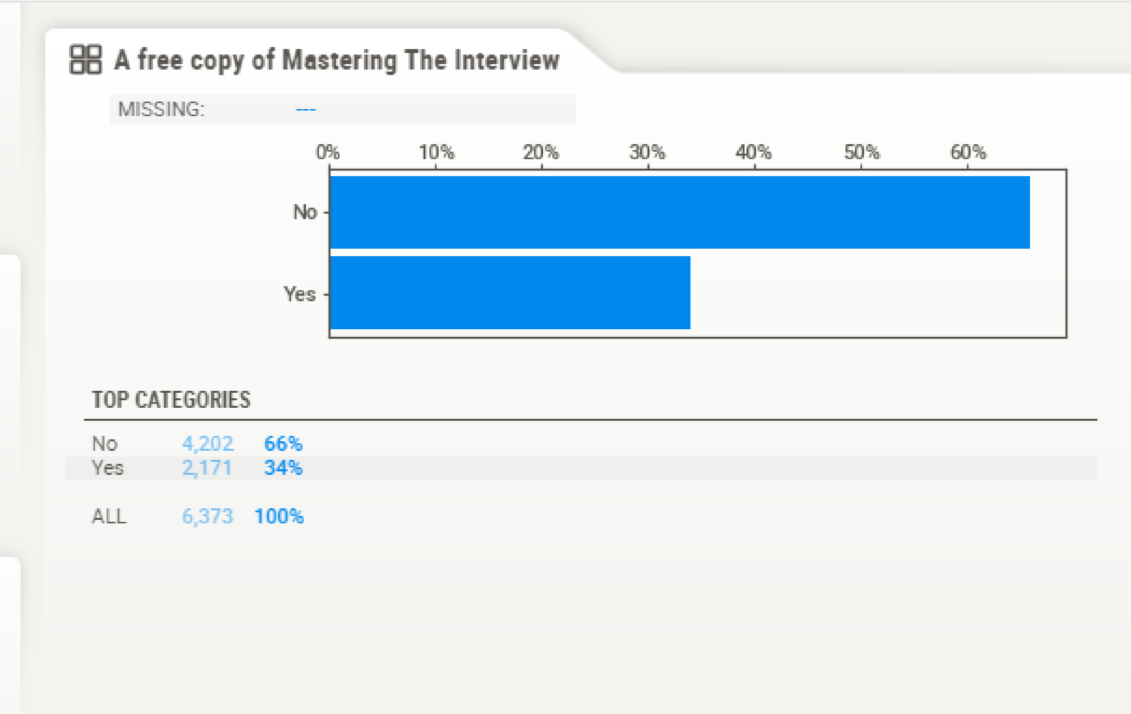
What is your current occupation

MISSING: ---

TOP CATEGORIES

Unemployed	5,476	86%
Working Professional	673	11%
Student	193	3%
Other	15	<1%
Housewife	9	<1%
Businessman	7	<1%
ALL	6,373	100%

Exploratory Data Analysis

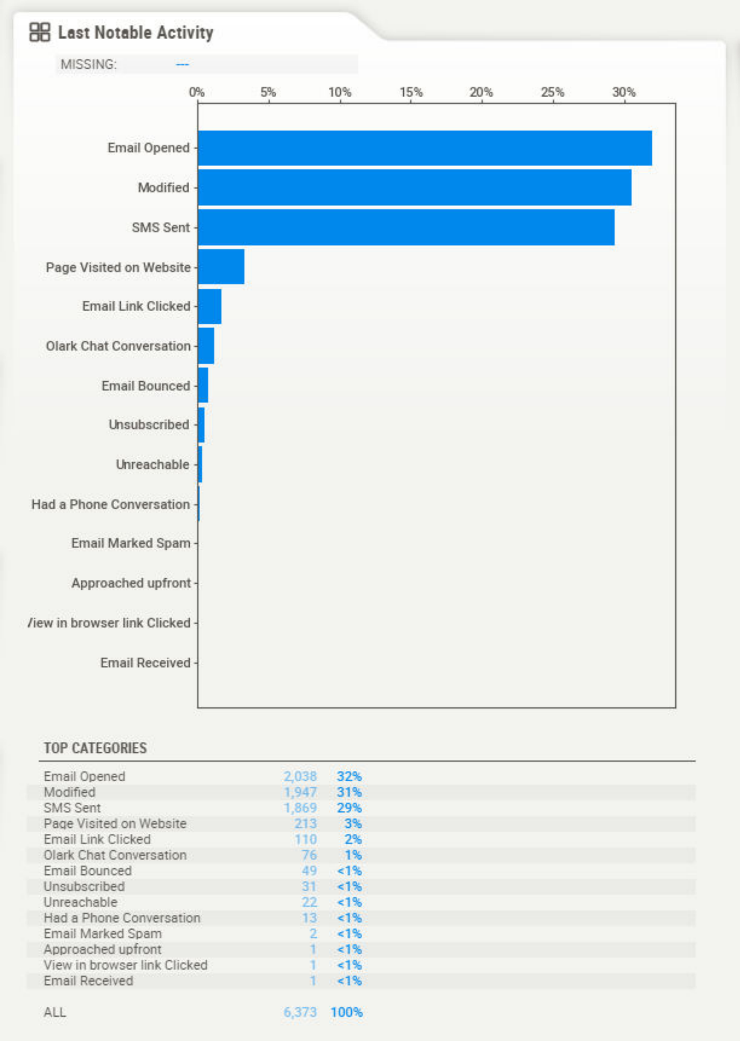


A free copy of Mastering The Interview

MISSING: ---

TOP CATEGORIES

No	4,202	66%
Yes	2,171	34%
ALL	6,373	100%



Last Notable Activity

MISSING: ---

TOP CATEGORIES

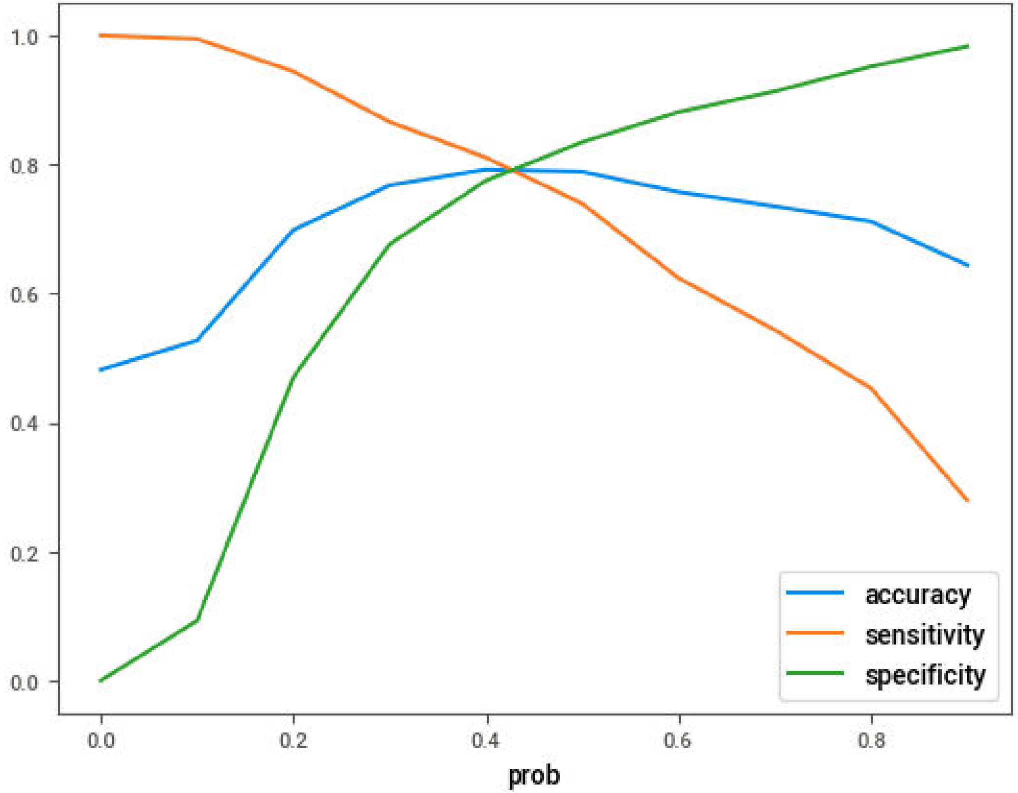
Email Opened	2,038	32%
Modified	1,947	31%
SMS Sent	1,869	29%
Page Visited on Website	213	3%
Email Link Clicked	110	2%
Olark Chat Conversation	76	1%
Email Bounced	49	<1%
Unsubscribed	31	<1%
Unreachable	22	<1%
Had a Phone Conversation	13	<1%
Email Marked Spam	2	<1%
Approached upfront	1	<1%
View in browser link Clicked	1	<1%
Email Received	1	<1%
ALL	6,373	100%

Features Selected

- Features were selected using RFE and was applied to a logistic regression model
- The final set of features after p value < 0.05 and VIF < 5 were
    1. TotalVisits
    2. Total Time Spent on Website
    3. Lead Origin - Lead Add Form
    4. Last Notable Activity - Unreachable
    5. Last Activity - Had a Phone Conversation
    6. Lead Source - Welingak Website
    7. Lead Source - Olark Chat
    8. Last Activity - SMS Sent
    9. Do Not Email - Yes
    10. What is your current occupation - Student
    11. What is your current occupation - Unemployed

Model Evaluation On Train Data Set - Sensitivity, Specificity and Accuracy

- Below graph shows optimal cutoff as 0.42



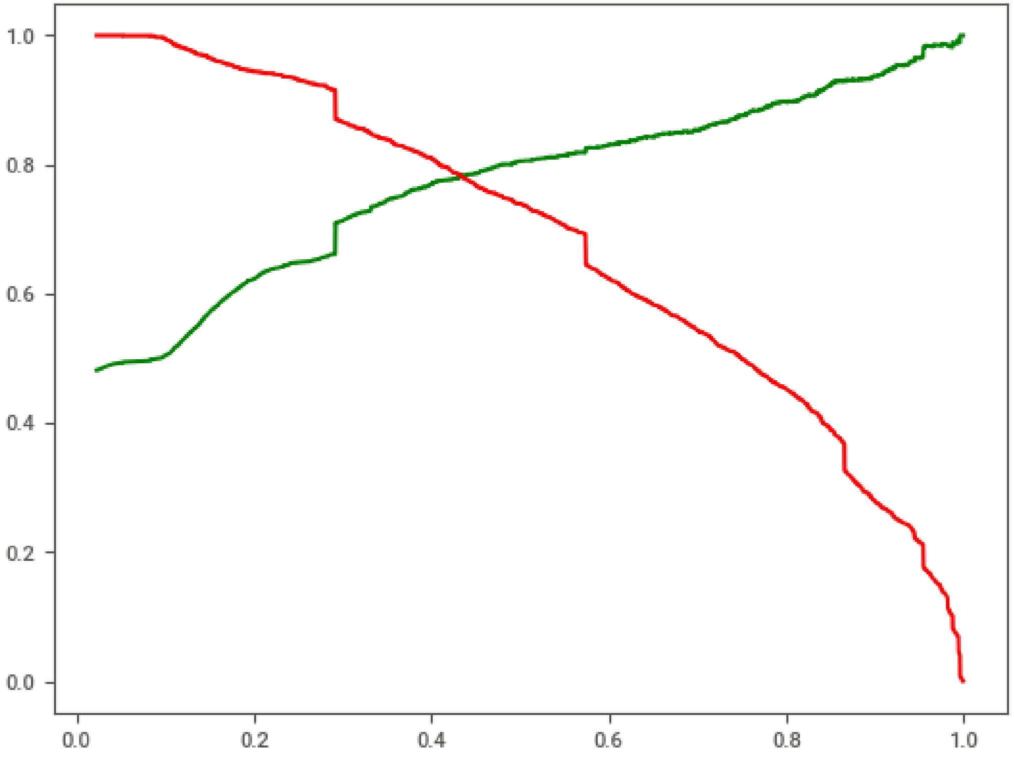
Confusion Matrix

1823	**489**
444	1705

- accuracy - 79
- Sesitivity - 79.3
- Specitficity - 79

Model Evaluation On Train Data Set - Recall and Precision

- Below graph shows optimal cutoff as 0.44



Confusion Matrix

1852	**460**
479	1670

- Precision - 78

- Recall - 77.8

Model Evaluation On Test Data Set - Sensitivity, Specificity and Accuracy

Confusion Matrix

786	**210**
202	714

- accuracy - 78.5
- Sesitivity - 78
- Specitficity - 79

Conclusion and Reccomendations

- Optimal cutoff selected as 0.42 based on Accuracy, Sensitivity and Specificity
- The model are accuracy - 79 ,Sesitivity - 79.3 and Specitficity - 79 on the train data set which is close to the test data set
- THe lead score calculated on the final prediction model is close to 80% in both the train and test data set
- The final list of variables positvely affecting the conversion are as below
  1. TotalVisits
  2. Total Time Spent on Website
  3. Lead Origin - Lead Add Form
  4. Last Notable Activity - Unreachable
  5. Last Activity - Had a Phone Conversation
  6. Lead Source - Welingak Website
  7. Lead Source - Olark Chat
  8. Last Activity - SMS Sent
- The final list of variables negatively affecting the conversion are as below
  1. Do Not Email - Yes
  2. What is your current occupation - Student
  3. What is your current occupation - Unemployed

Conclusion and Reccomendations

- The company should make calls to the leads Having high 'Total Visits'
- The company should make calls to the leads 'Who spend more time on websites'
- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was 'SMS sent'
- The company should make calls to the leads whose last activity was 'Had a Phone Conversation '
- The company should make calls to the leads who are in process of "Add Form"
- The company should not make calls to the leads who selected 'Do Not Email'
- The company should not make calls to the leads who are 'Students'
- The company should not make calls to the leads who are 'Unemployed'