Lead Scoring Case Study

By

Team

Deepthi Yalavarthi

and

Muthuraja Sivanantham

Agenda:

- Problem Statement and Objective
- Analysis Approach
- Results and Business Recommendation

Problem Statement:

- An education company named X Education sells online courses to industry professionals. 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. 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.

Objective:

- Build a model wherein we 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:

Step 1: Import required libraries

Step 2: Read and Understand the data

Step 3: Data cleaning

Step 4: Outlier Analysis and Treatment

Step 5: Data Pre-Processing

- Creating Dummy Variables
- Train-Test Split
- Scaling
- Looking at Correlation

Step 6: Model Building

Step 7: Feature Selection using RFE – Coarse Tuning

Step 8: Checking P-Value and VIF – Manual Tuning

Step 9: Plotting ROC Curve

Step 10: Finding optimal probability cut-off

Step 11: Model Evaluation and Model Performance

Step 12: Generate Score Variables

Key Classification Analysis:

1. Data Cleaning

- Dropping Variables having more missing values
- Imputing Variables having less missing values
- Dropping Skewed variables
- Checking null values after dropping and imputing the values in the dataset
- Finally verifying the Percentage of null values in the variables and here is the result

Percentage of null values in the variables

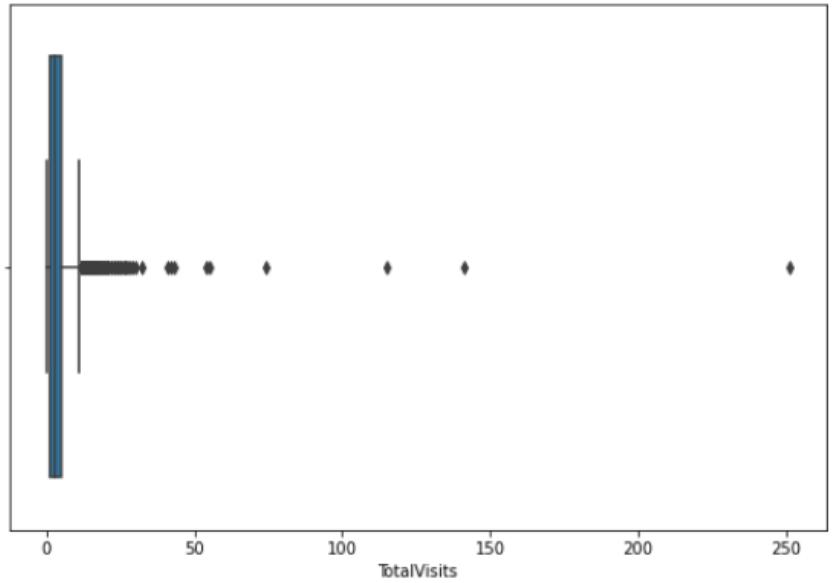
```
In [61]:  1  lead_df.isnull().mean()*100
```

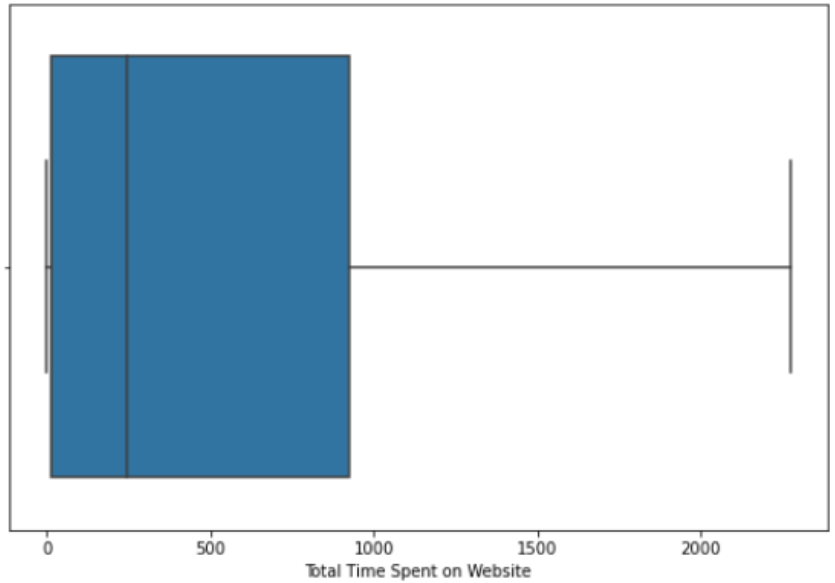
```
Out[61]: Lead Number                                  0.0
         Lead Origin                                  0.0
         Lead Source                                  0.0
         Do Not Email                                 0.0
         Converted                                    0.0
         TotalVisits                                  0.0
         Total Time Spent on Website                  0.0
         Last Activity                                0.0
         Country                                      0.0
         Specialization                               0.0
         What is your current occupation              0.0
         City                                         0.0
         A free copy of Mastering The Interview       0.0
         Last Notable Activity                        0.0
         dtype: float64
```

Observation: Percentage of Null Values - Finally we have cleared null values in all the variables and removed the skewed, unique identifier variables as well. After doing data cleaning we are ended up with 13 variables

Key Classification Analysis:

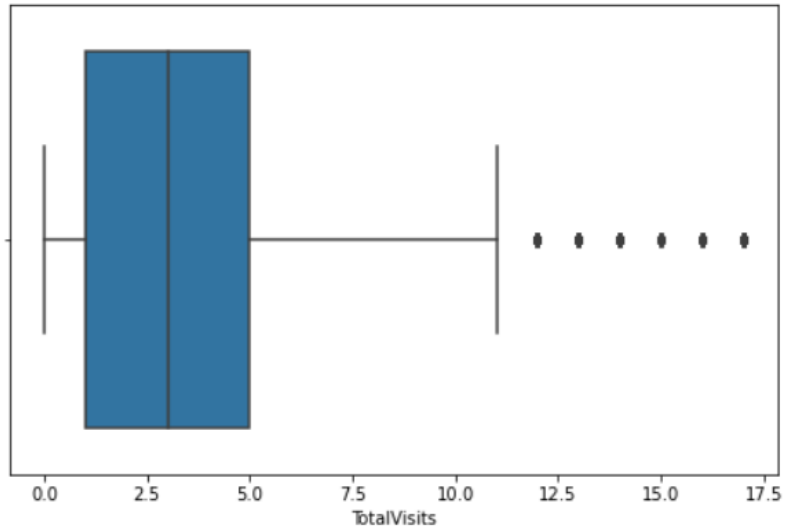
2. Outlier Analysis and Treatment





Observation:

- We don't have any outliers in 'Total Time Spent on Website' variable but we have in 'TotalVisits'. So we need to remove the outliers using soft capping method

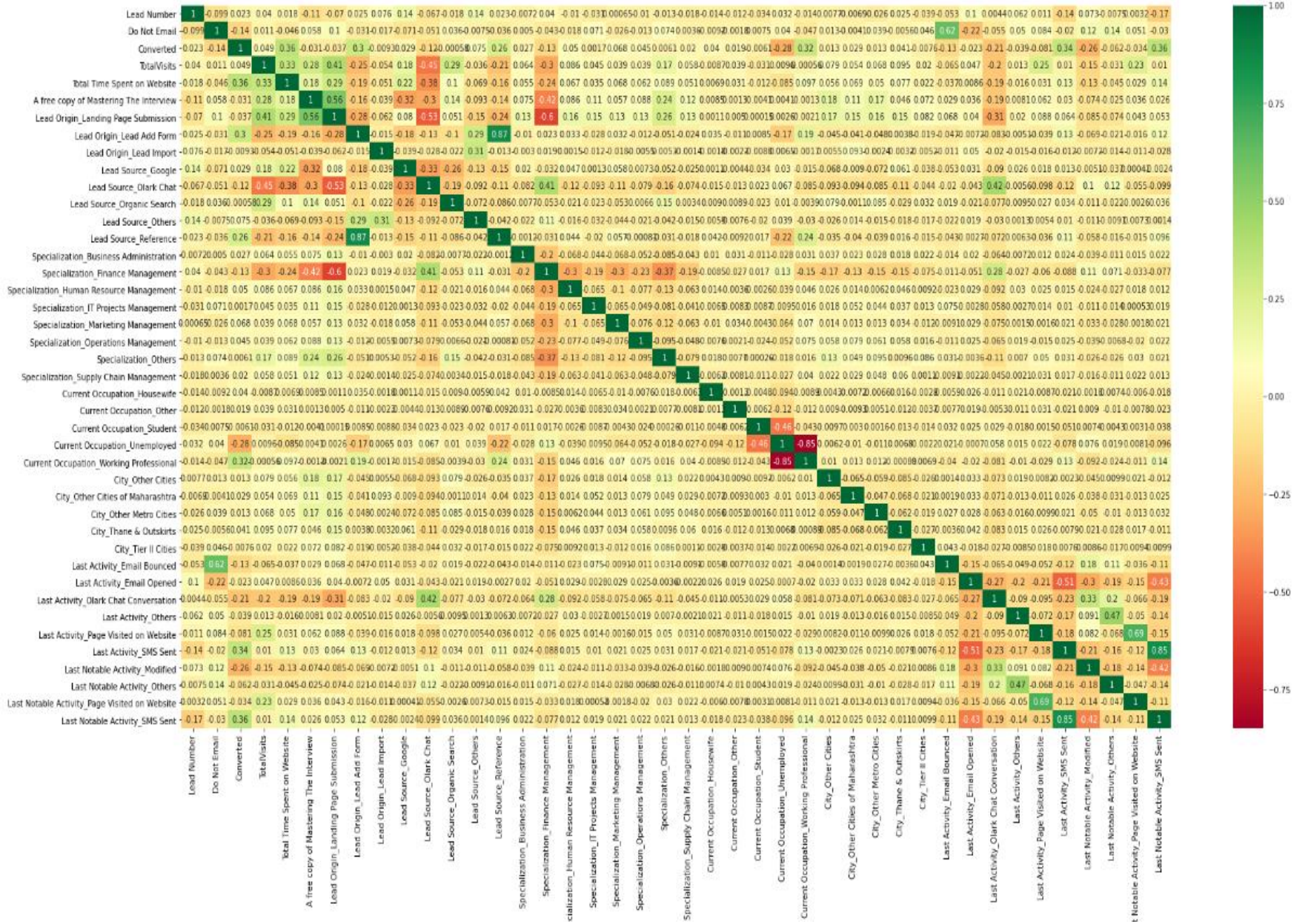


Observation:

- The outliers which is there in 'TotalVisits' variable has been reduced drastically after outlier treatment by applying the soft capping method

Key Classification Analysis:

3. Looking at Correlation

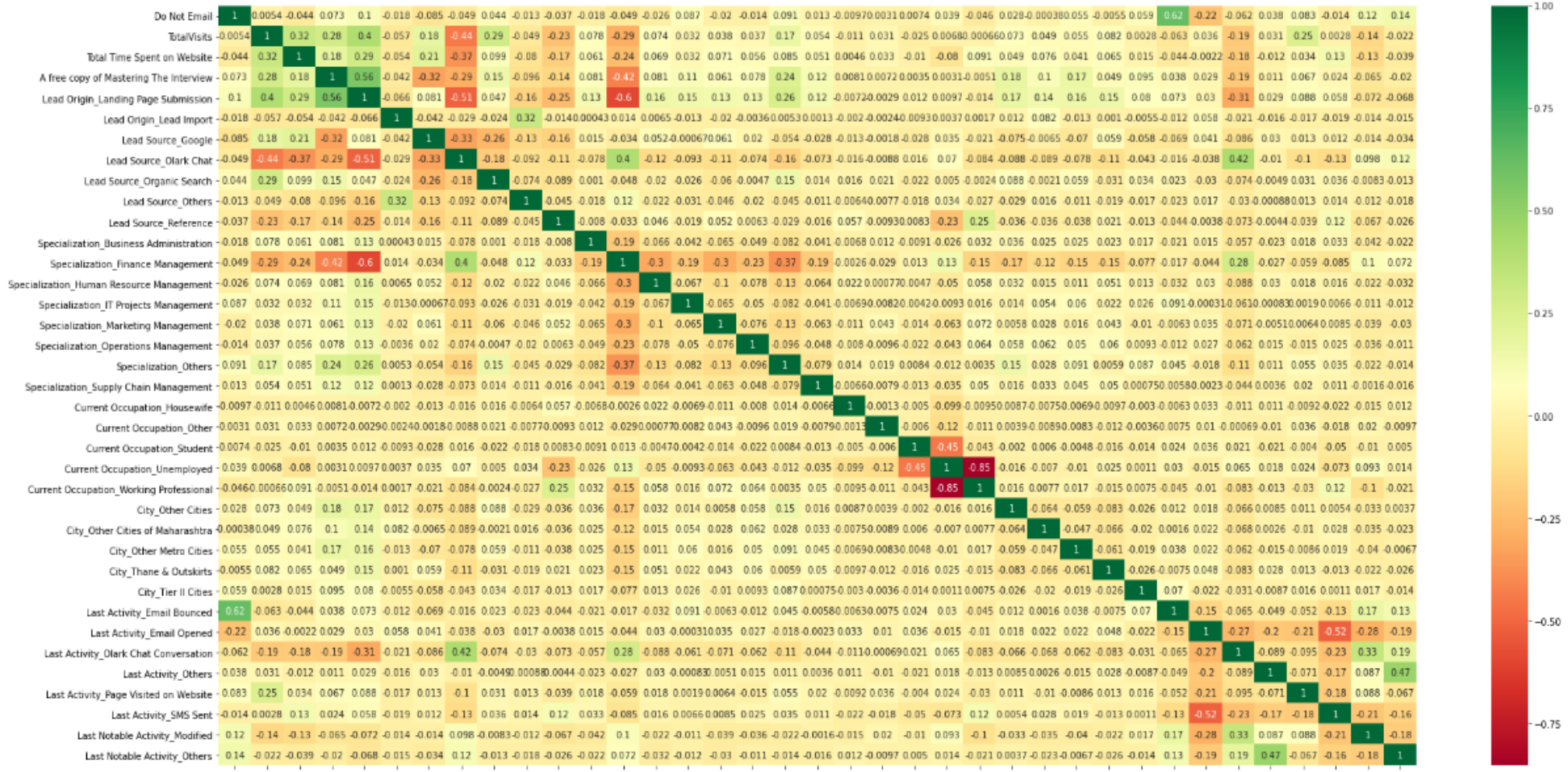


Observation: Based on the heatmap we found that the high correlation is between the following variables:

- Last Notable Activity_SMS Sent
- Last Notable Activity_Page Visited on Website
- Lead Origin_Lead Add Form

So these variables can be dropped

Key Classification Analysis:



Observation: After dropping the high correlated variables, we now have all the variables are correlated as expected. So we can go ahead with these variables for further processing

Key Classification Analysis:

4. Final Model Results

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6340
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2742.2
Date:	Mon, 08 Mar 2021	**Deviance:**	5484.4
Time:	16:10:00	**Pearson chi2:**	6.55e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8938	0.084	-10.617	0.000	-1.059	-0.729
Do Not Email	-1.4108	0.165	-8.540	0.000	-1.735	-1.087
Total Time Spent on Website	1.0946	0.040	27.616	0.000	1.017	1.172
Lead Origin_Landing Page Submission	-0.3690	0.088	-4.216	0.000	-0.540	-0.197
Lead Source_Olark Chat	0.9463	0.119	7.942	0.000	0.713	1.180
Lead Source_Others	1.8674	0.169	11.068	0.000	1.537	2.198
Lead Source_Reference	3.6793	0.238	15.480	0.000	3.213	4.145
Current Occupation_Working Professional	2.7358	0.188	14.543	0.000	2.367	3.105
Last Activity_Olark Chat Conversation	-1.0946	0.167	-6.569	0.000	-1.421	-0.768
Last Activity_SMS Sent	1.2929	0.073	17.792	0.000	1.150	1.435
Last Notable Activity_Modified	-0.9113	0.078	-11.644	0.000	-1.065	-0.758

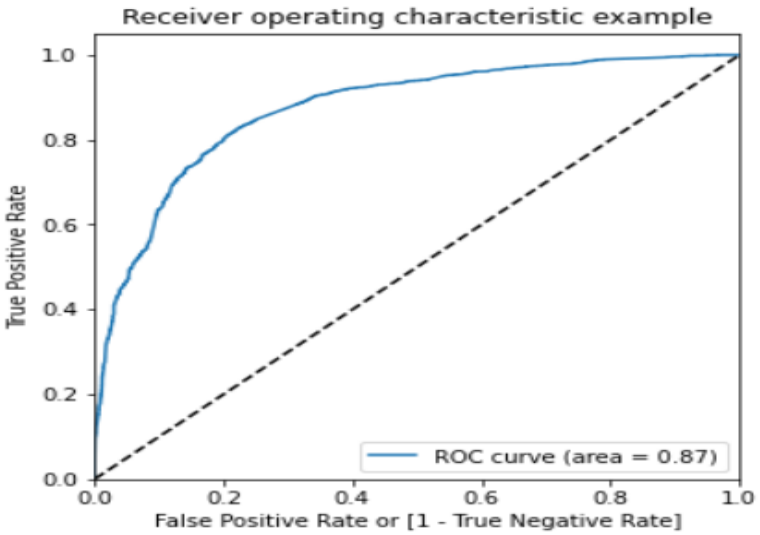
	Features	VIF
9	Last Notable Activity_Modified	1.65
2	Lead Origin_Landing Page Submission	1.63
3	Lead Source_Olark Chat	1.60
7	Last Activity_Olark Chat Conversation	1.55
8	Last Activity_SMS Sent	1.45
1	Total Time Spent on Website	1.29
5	Lead Source_Reference	1.23
6	Current Occupation_Working Professional	1.18
0	Do Not Email	1.13
4	Lead Source_Others	1.04

Observation: After doing coarse tuning (RFE) and manual tuning (p-value and VIF), all the variables now have a VIF and p-values are within a range

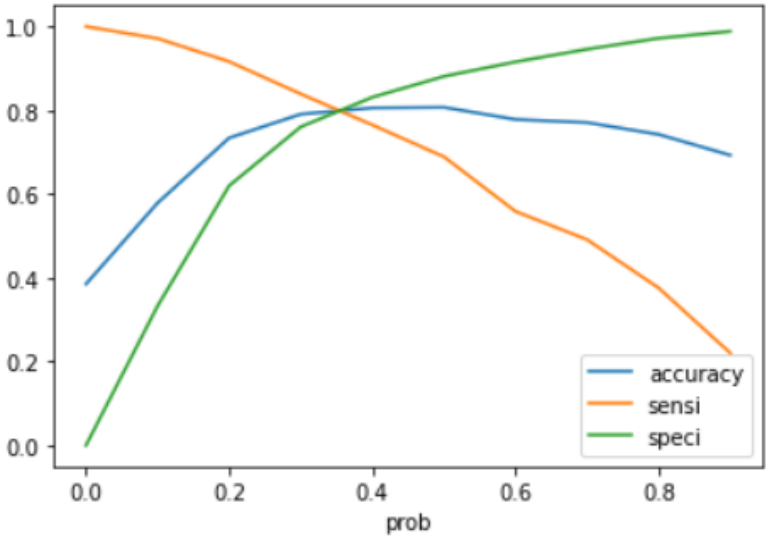
Key Classification Analysis:

5. ROC Curve and Optimal Cut-off

ROC Curve



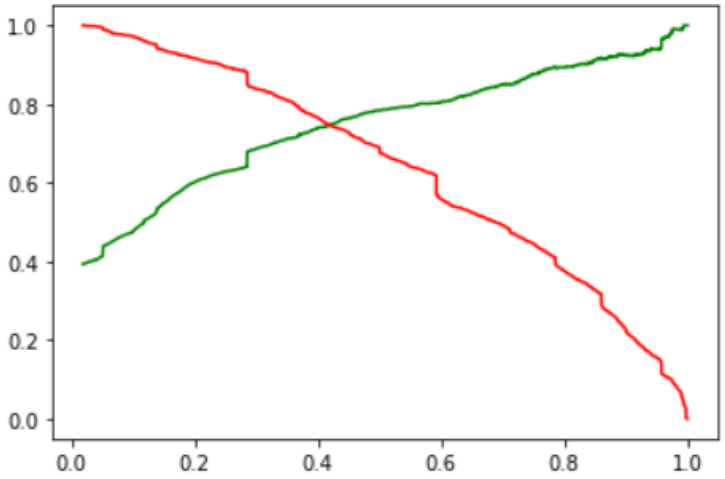
Optimal Cut-off (for Lead Score)



Observation:

- ROC Area under curve is **0.87**
- Optimal cut-off point based on the graph would be **0.38**. So leads having probability conversion more than **38%** are considered as **hot leads**

6. Precision and Recall – Trade Off



Key Classification Analysis:

7. Final Observation

The following are the model performance metrics for Train Data set and Test Data Set

	Train Data	Test Data
Accuracy	79.05%	80.35%
Sensitivity	83.89%	76.13%
Specificity	76.03%	82.75%
Precision	78.37%	71.57%

Final Conclusion: The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model performance

Thank You !!!