Lead Score Case Study

Created By:

Nayanshi Sahu

Sameer Kumar

Sakshi Bhasin

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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%.

Goals of Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps Followed:

- Importing Libraries and Data
- Inspecting the Dataframe
- Data Cleaning
- Data Preparation
- Test-Train Split
- Feature Scaling
- Model Building using Stats Model & RFE
- Predicting a Train model
- Predictions on the test set
- Conclusion

Library Used:

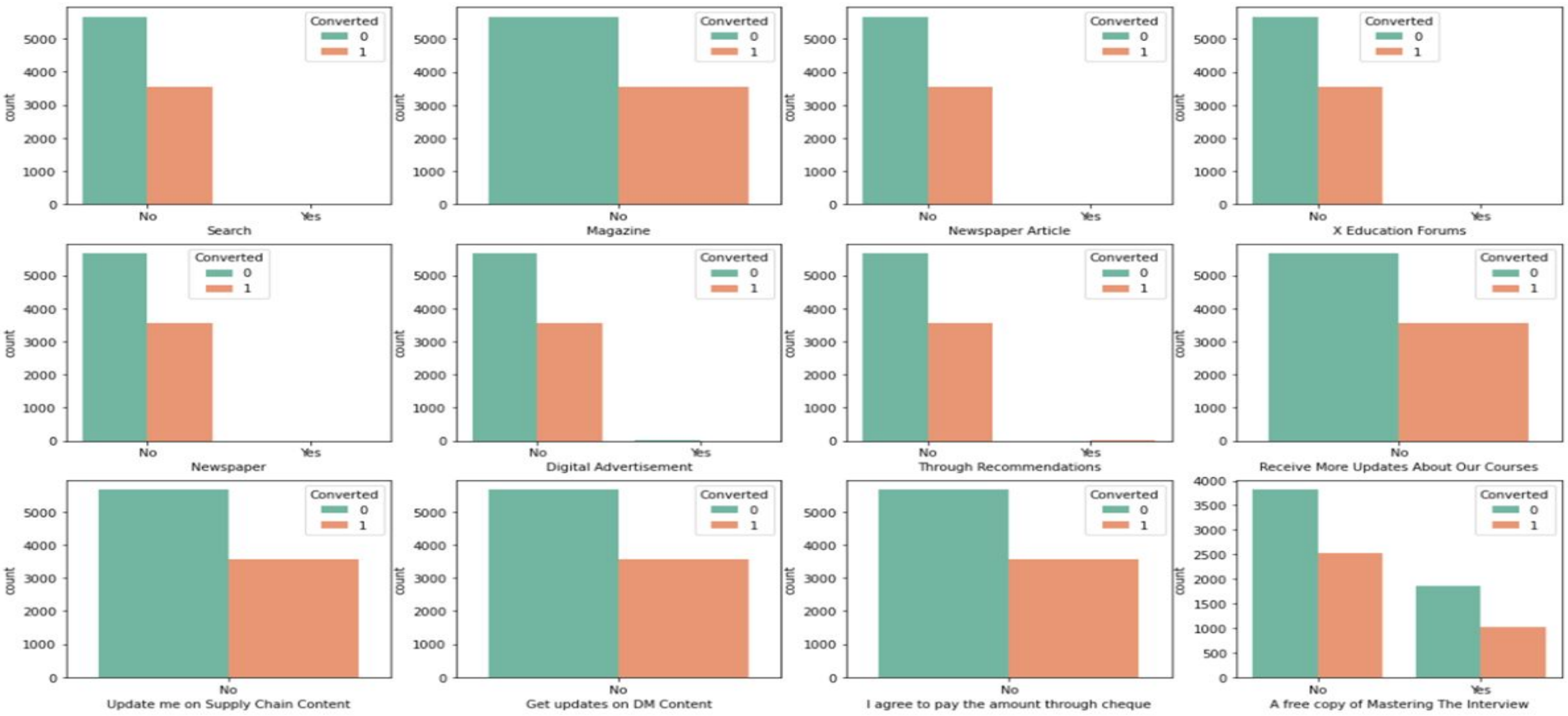
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn

Data Inspection:

- Checked data Dimension size and description
- Checked for Duplicity
- Deleted Unique keys columns

Data Cleaning:

- Dropping Unique Keys columns and columns with more than 35% null value
- Dropped few column whose data are highly imbalanced shown in the below graph



Examined every Categorical column and if found any null values, we have replaced it with Mode in most of the cases, In some cases we replace null values with text like not provided.

Some columns that looked screwed and had redundant information we have dropped them.

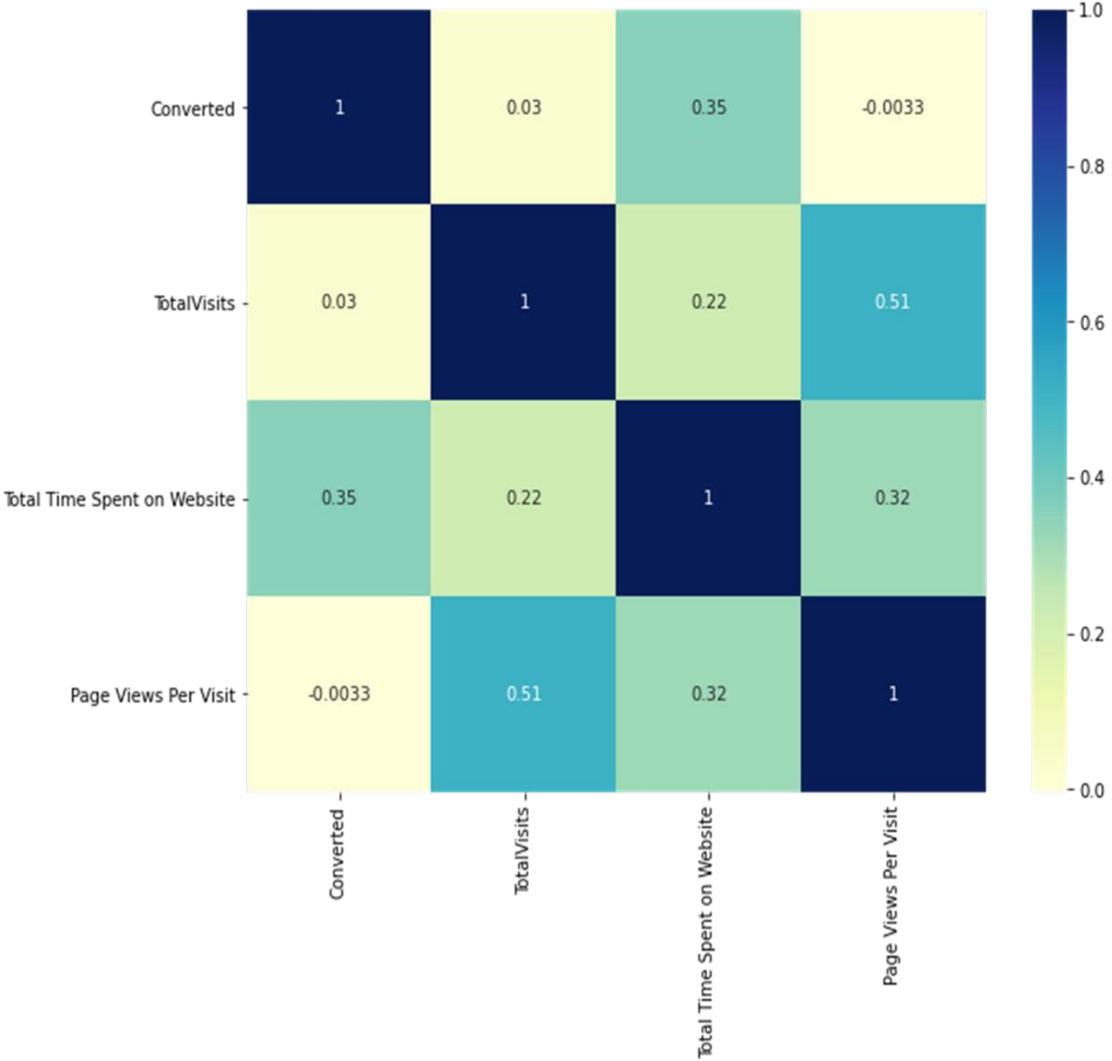
We have noticed Below 2 point from above exercise:

1. Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
2. SMS sent as last activity has high conversion rate.

Correlation:

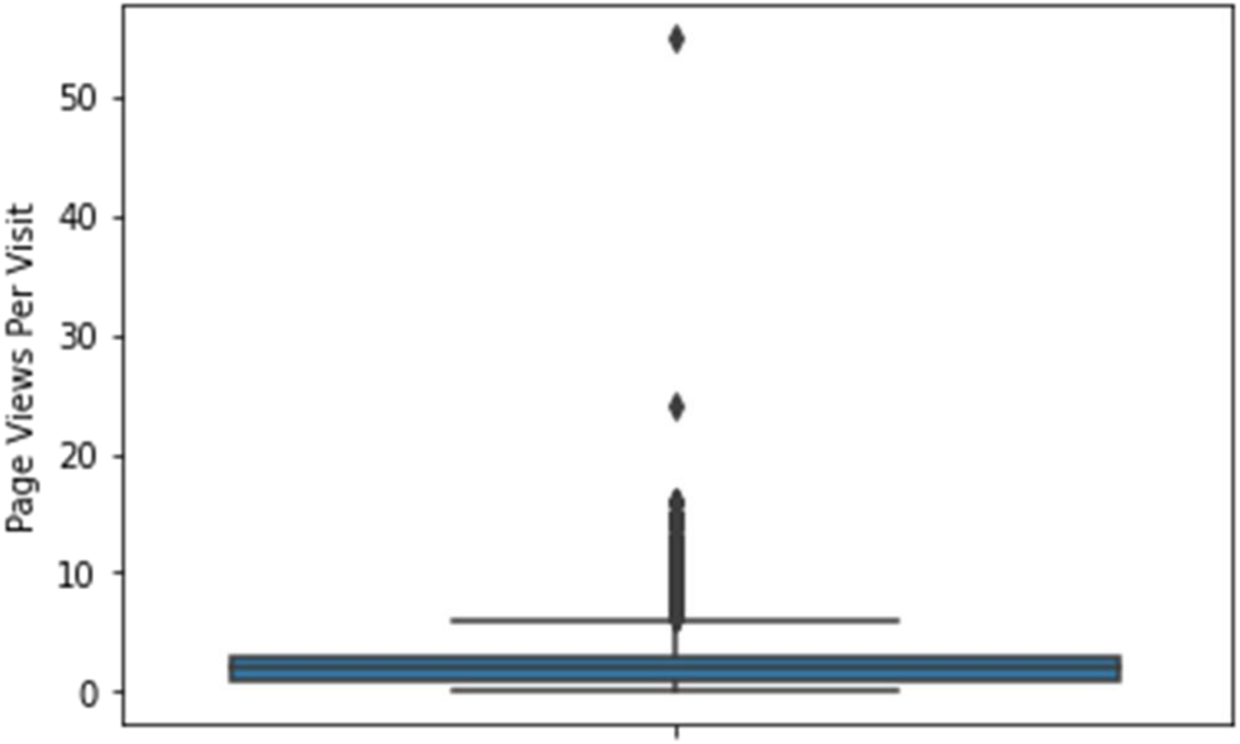
- Numerical Columns examin by using correlation Matrix as shown in the right side image.

- By Correlation we have figured out that mostly People who has spend more time on website has higher chance to get converted.

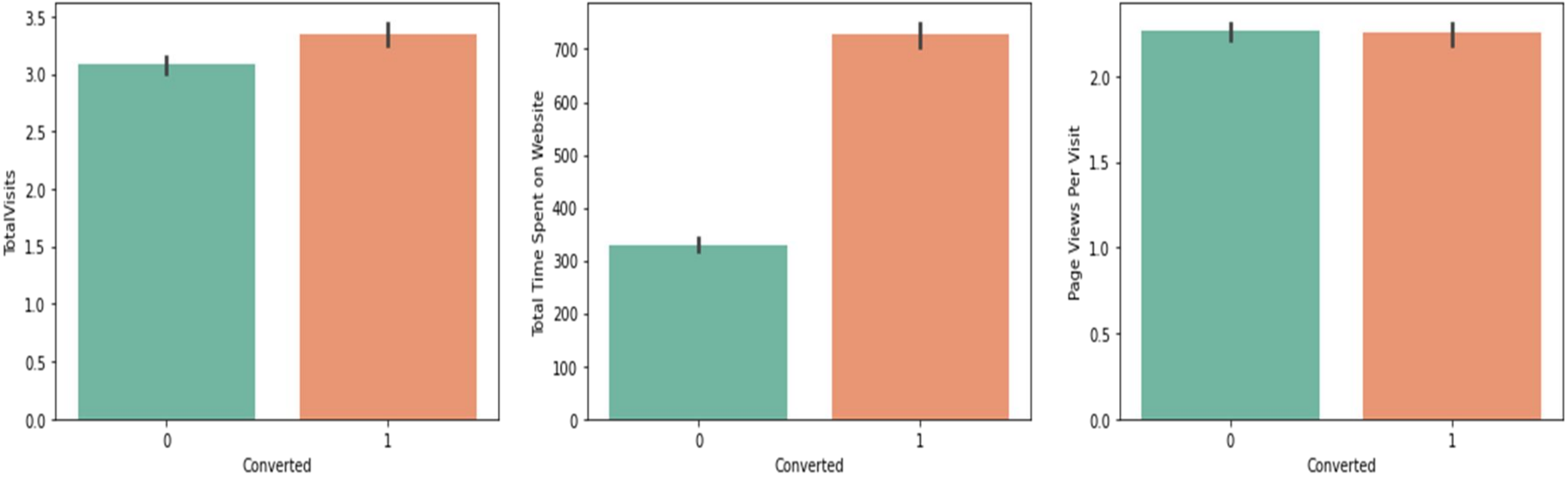


Outlier:

Outlier found only in “Page Views per Visit” Column and we have treated them by replacing all the values which are higher than one percentile with one percentile.



Checking Numerical value conversion rate:



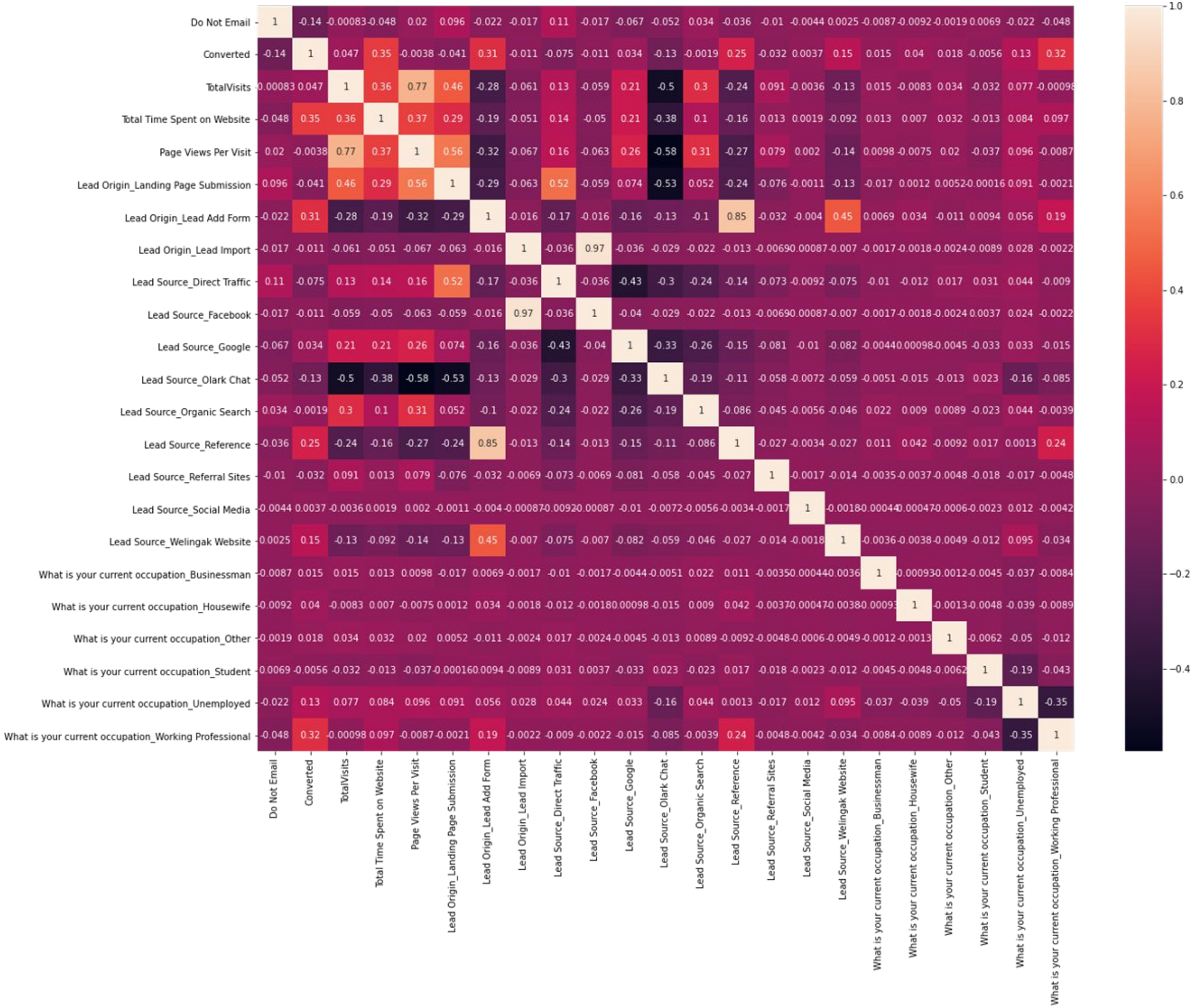
The conversion rate is high for all of the above three Columns.

Data Preparation:

- Converting some binary variables (Yes/No) to 0/1
- Created Dummy Variables
- Test Train Split

Feature Scaling:

- Using Correlation matrix after creating Dummy Variables and dropped highly correlated matrix



	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Reference	Lead Source_Referral Sites	Lead Source_Social Media	Lead Source_Welingak Website	What is your current occupation_Businessman	What is your current occupation_Housewife	What is your current occupation_Other	What is your current occupation_Student	What is your current occupation_Unemployed	What is your current occupation_Working Professional
Do Not Email	1	-0.14	-0.00083	-0.048	0.02	0.096	-0.022	-0.017	0.11	-0.017	-0.067	-0.052	0.034	-0.036	-0.01	-0.0044	0.0025	-0.0087	-0.0092	-0.0019	0.0069	-0.022	-0.048
Converted	-0.14	1	0.047	0.35	-0.0038	-0.041	0.31	-0.011	-0.075	-0.011	0.034	-0.13	-0.0019	0.25	-0.032	0.0037	0.15	0.015	0.04	0.018	-0.0056	0.13	0.32
TotalVisits	-0.00083	0.047	1	0.36	0.77	0.46	-0.28	-0.061	0.13	-0.059	0.21	-0.5	0.3	-0.24	0.091	-0.0036	-0.13	0.015	-0.0083	0.034	-0.032	0.077	-0.00098
Total Time Spent on Website	-0.048	0.35	0.36	1	0.37	0.29	-0.19	-0.051	0.14	-0.05	0.21	-0.38	0.1	-0.16	0.013	0.0019	-0.092	0.013	0.007	0.032	-0.013	0.084	0.097
Page Views Per Visit	0.02	-0.0038	0.77	0.37	1	0.56	-0.32	-0.067	0.16	-0.063	0.26	-0.58	0.31	-0.27	0.079	0.002	-0.14	0.0098	-0.0075	0.02	-0.037	0.096	-0.0087
Lead Origin_Landing Page Submission	0.096	-0.041	0.46	0.29	0.56	1	-0.29	-0.063	0.52	-0.059	0.074	-0.53	0.052	-0.24	-0.076	-0.0011	-0.13	-0.017	0.0012	0.0052	-0.00016	0.091	-0.0021
Lead Origin_Lead Add Form	-0.022	0.31	-0.28	-0.19	-0.32	-0.29	1	-0.016	-0.17	-0.016	-0.16	-0.13	-0.1	0.85	-0.032	-0.004	0.45	0.0069	0.034	-0.011	0.0094	0.056	0.19
Lead Origin_Lead Import	-0.017	-0.011	-0.061	-0.051	-0.067	-0.063	-0.016	1	-0.036	0.97	-0.036	-0.029	-0.022	-0.013	-0.0069	-0.00087	-0.007	-0.0017	-0.0018	-0.0024	-0.0089	0.028	-0.0022
Lead Source_Direct Traffic	0.11	-0.075	0.13	0.14	0.16	0.52	-0.17	-0.036	1	-0.036	-0.43	-0.3	-0.24	-0.14	-0.073	-0.0092	-0.075	-0.01	-0.012	0.017	0.031	0.044	-0.009
Lead Source_Facebook	-0.017	-0.011	-0.059	-0.05	-0.063	-0.059	-0.016	0.97	-0.036	1	-0.04	-0.029	-0.022	-0.013	-0.0069	-0.00087	-0.007	-0.0017	-0.0018	-0.0024	0.0037	0.024	-0.0022
Lead Source_Google	-0.067	0.034	0.21	0.21	0.26	0.074	-0.16	-0.036	-0.43	-0.04	1	-0.33	-0.26	-0.15	-0.081	-0.01	-0.082	-0.0044	0.00098	-0.0045	-0.033	0.033	-0.015
Lead Source_Olark Chat	-0.052	-0.13	-0.5	-0.38	-0.58	-0.53	-0.13	-0.029	-0.3	-0.029	-0.33	1	-0.19	-0.11	-0.058	-0.0072	-0.059	-0.0051	-0.015	-0.013	0.023	-0.16	-0.085
Lead Source_Organic Search	0.034	-0.0019	0.3	0.1	0.31	0.052	-0.1	-0.022	-0.24	-0.022	-0.26	-0.19	1	-0.086	-0.045	-0.0056	-0.046	0.022	0.009	0.0089	-0.023	0.044	-0.0039
Lead Source_Reference	-0.036	0.25	-0.24	-0.16	-0.27	-0.24	0.85	-0.013	-0.14	-0.013	-0.15	-0.11	-0.086	1	-0.027	-0.0034	-0.027	0.011	0.042	-0.0092	0.017	0.0013	0.24
Lead Source_Referral Sites	-0.01	-0.032	0.091	0.013	0.079	-0.076	-0.032	-0.0069	-0.073	-0.0069	-0.081	-0.058	-0.045	-0.027	1	-0.0017	-0.014	-0.0035	-0.0037	-0.0048	-0.018	-0.017	-0.0048
Lead Source_Social Media	-0.0044	0.0037	-0.0036	0.0019	0.002	-0.0011	-0.004	-0.00087	-0.0092	-0.00087	-0.01	-0.0072	-0.0056	-0.0034	-0.0017	1	-0.0018	-0.00044	-0.00047	-0.0006	-0.0023	0.012	-0.0042
Lead Source_Welingak Website	0.0025	0.15	-0.13	-0.092	-0.14	-0.13	0.45	-0.007	-0.075	-0.007	-0.082	-0.059	-0.046	-0.027	-0.014	-0.0018	1	-0.0036	-0.0038	-0.0049	-0.012	0.095	-0.034
What is your current occupation_Businessman	-0.0087	0.015	0.015	0.013	0.0098	-0.017	0.0069	-0.0017	-0.01	-0.0017	-0.0044	-0.0051	0.022	0.011	-0.0035	-0.00044	-0.0036	1	-0.00093	-0.0012	-0.0045	-0.037	-0.0084
What is your current occupation_Housewife	-0.0092	0.04	-0.0083	0.007	-0.0075	0.0012	0.034	-0.0018	-0.012	-0.0018	0.00098	-0.015	0.009	0.042	-0.0037	-0.00047	-0.0038	-0.00093	1	-0.0013	-0.0048	-0.039	-0.0089
What is your current occupation_Other	-0.0019	0.018	0.034	0.032	0.02	0.0052	-0.011	-0.0024	0.017	-0.0024	-0.0045	-0.013	0.0089	-0.0092	-0.0048	-0.0006	-0.0049	-0.0012	-0.0013	1	-0.0062	-0.05	-0.012
What is your current occupation_Student	0.0069	-0.0056	-0.032	-0.013	-0.037	-0.00016	0.0094	-0.0089	0.031	0.0037	-0.033	0.023	-0.023	0.017	-0.018	-0.0023	-0.012	-0.0045	-0.0048	-0.0062	1	-0.19	-0.043
What is your current occupation_Unemployed	-0.022	0.13	0.077	0.084	0.096	0.091	0.056	0.028	0.044	0.024	0.033	-0.16	0.044	0.0013	-0.017	0.012	0.095	-0.037	-0.039	-0.05	-0.19	1	-0.35
What is your current occupation_Working Professional	-0.048	0.32	-0.00098	0.097	-0.0087	-0.0021	0.19	-0.0022	-0.009	-0.0022	-0.015	-0.085	-0.0039	0.24	-0.0048	-0.0042	-0.034	-0.0084	-0.0089	-0.012	-0.043	-0.35	1

Model Building:

- Used Stats Model & RFE
- By optimizing p value we have rebuilt our model 5 times.
- After 5th time rebuilding the model the all features VIF So we need to further optimize our model and we can proceed with making predictions using this model only
- Predicting a Train Model and calculated Metrics -Accuracy, Sensitivity, Specificity, False Positive Rate, Positive Predictive Value and Negative Predictive Value.

Confusion Matrix:

2905	**1048**
414	**2005**

Accuracy: 77.05%

Sensitivity: 82.89%

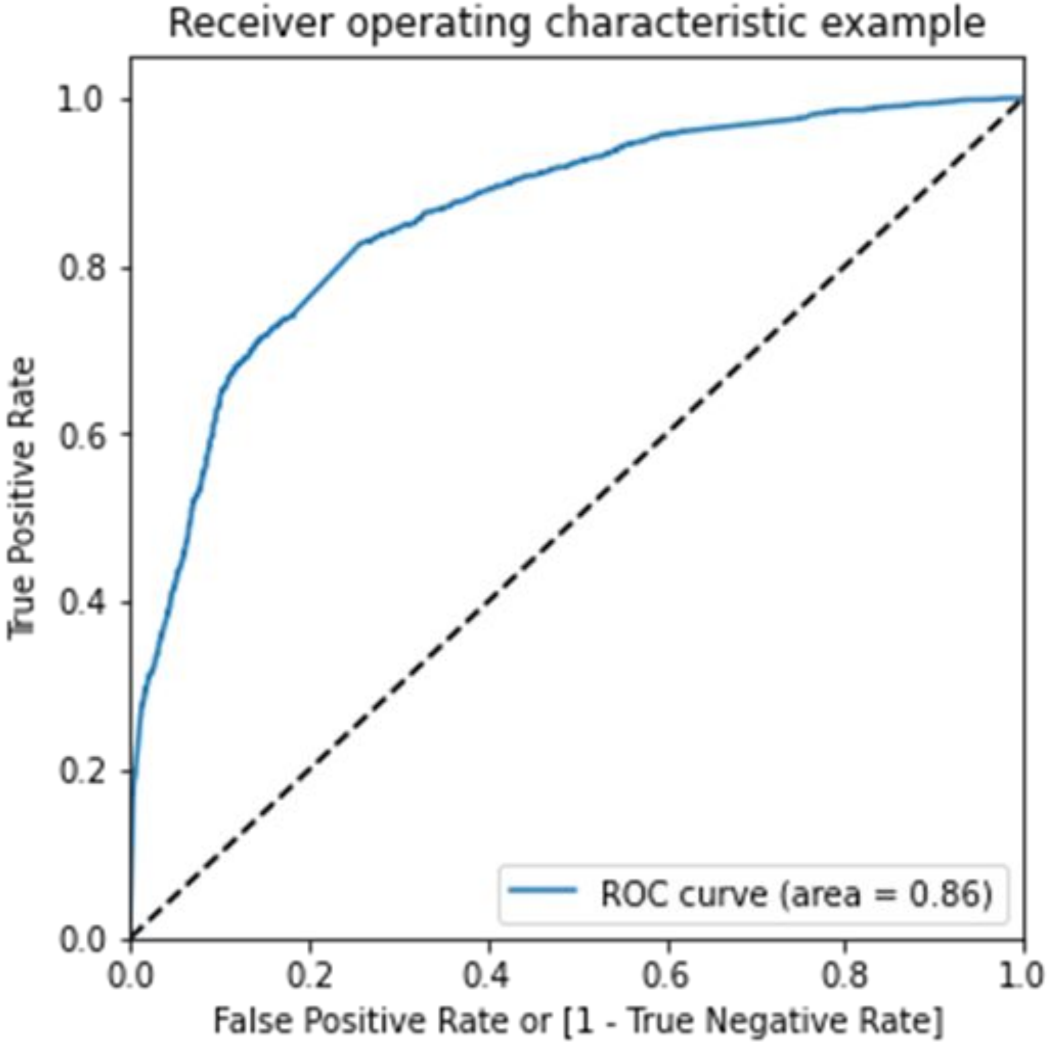
Specificity: 73.9%

Precision: 65.67%

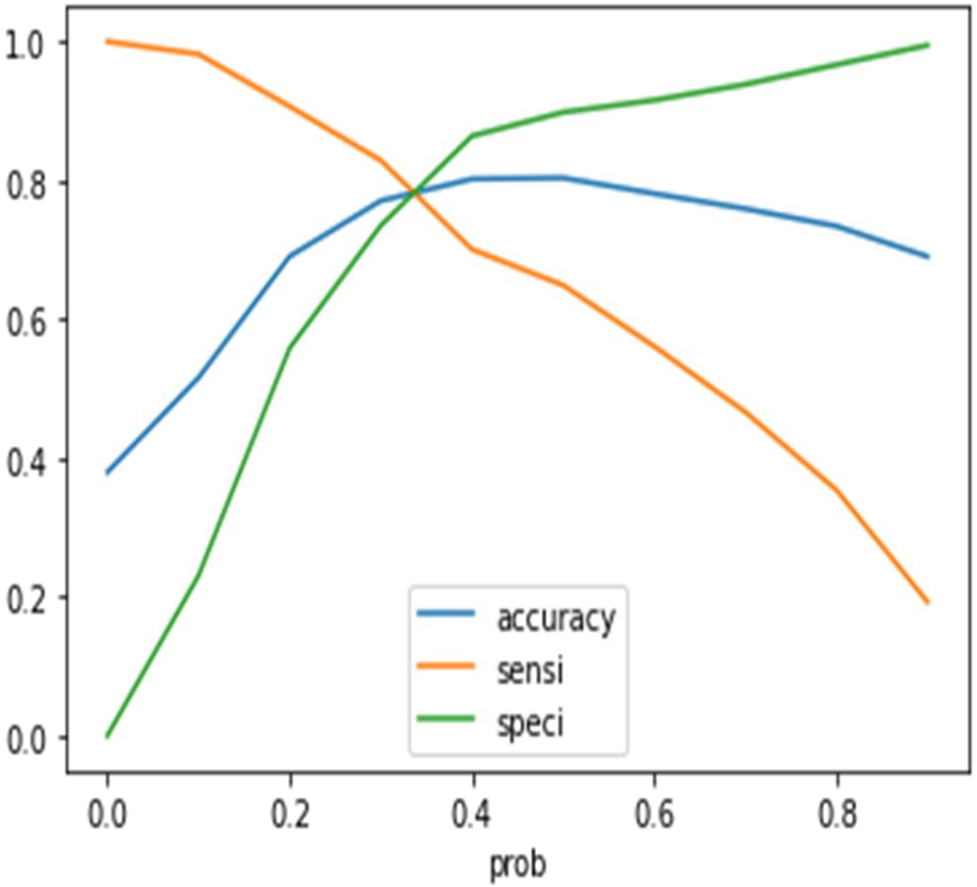
Recall: 82.88

PLOTTING ROC CURVE

The ROC Curve should be a value close to 1. We are getting a good value of 0.86 indicating a good predictive model

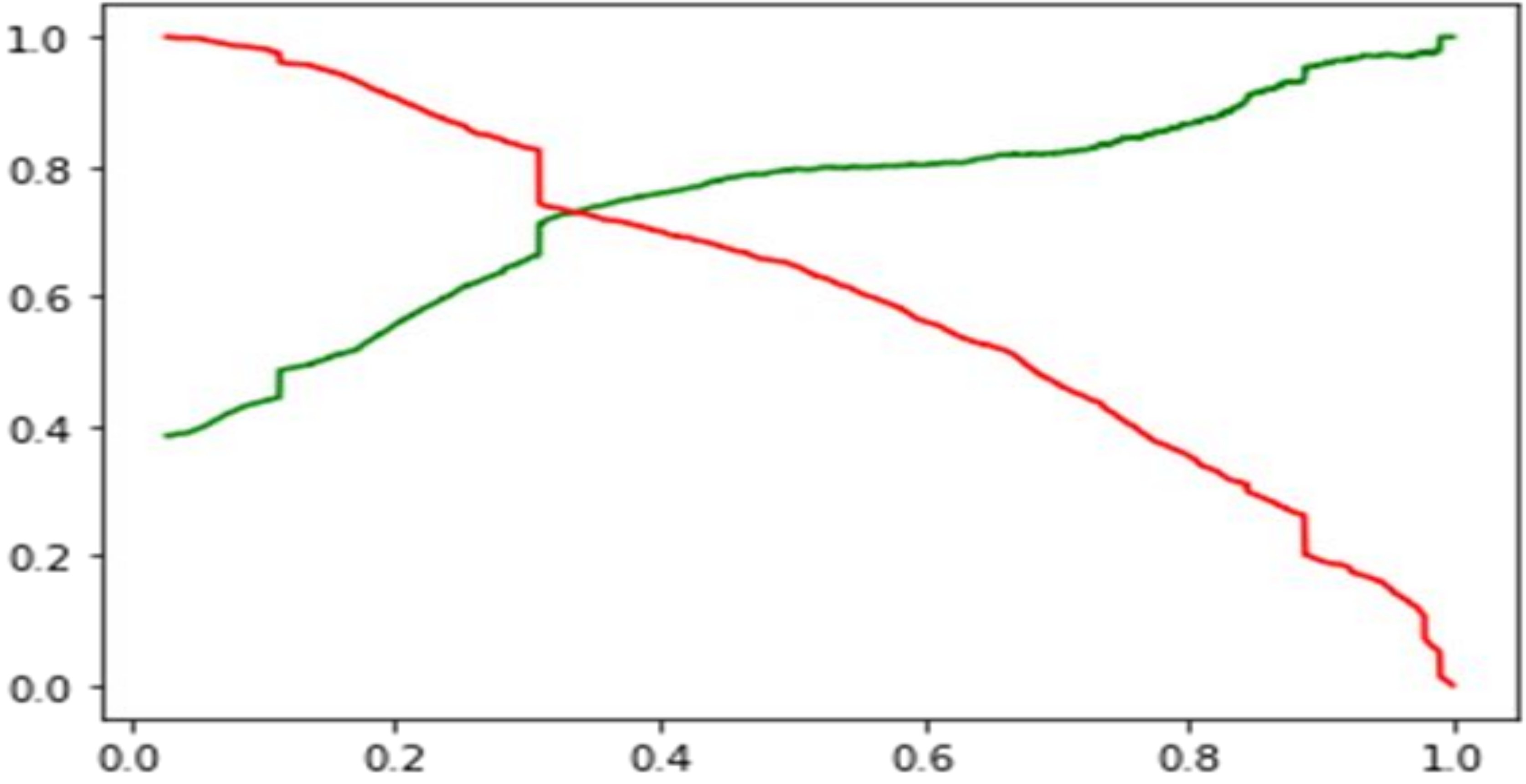


Accuracy sensitivity and specificity Plot



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Precision and Recall Trade-off



Prediction on the Test Set

- The final prediction of conversions have a target rate of 83% (same as predictions made on training data set)
- Overall Metrics - Accuracy, Confusion Metrices, Sensitivity, Specificity, False Positive Rate, Positive Predictive Value, Negative Predictive Value on final prediction on test set
- After running the model on the Test Data these are the figures we obtain:
  - Accuracy : 77.52%
  - Sensitivity :83.01%
  - Specificity : 74.13%

Conclusion:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%.
- Hence overall this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Total Time Spent on Website

Thank You