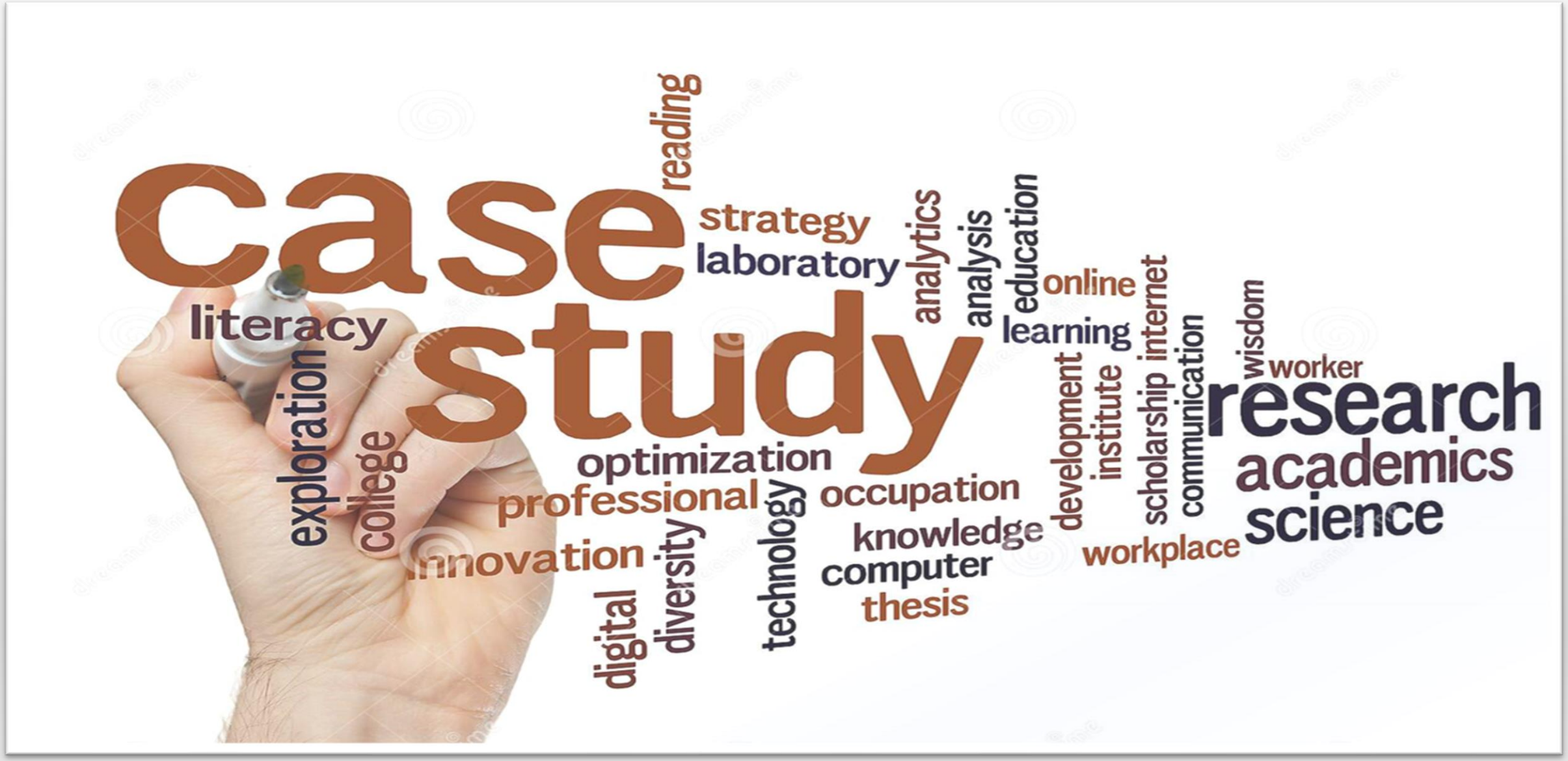
Lead Score Case Study



Lead Score Case Study for X Education

Problem Statement :

- X Education sells online courses to industry professionals.
- 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.

Business Goal:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Strategy

- Inspecting the data for analysis
- Cleaning and preparing the data
- Exploratory Data Analysis.
- Feature Scaling.
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics and majorly by Accuracy.

Problem solving methodology

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Uniform data
- Splitting data into train and test set.

Model Building

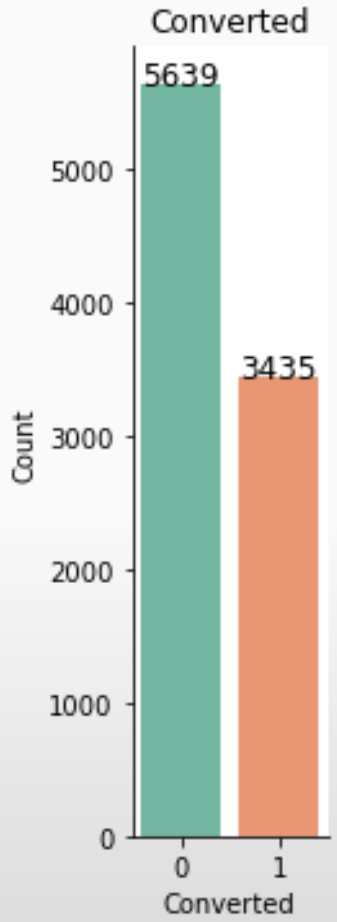
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

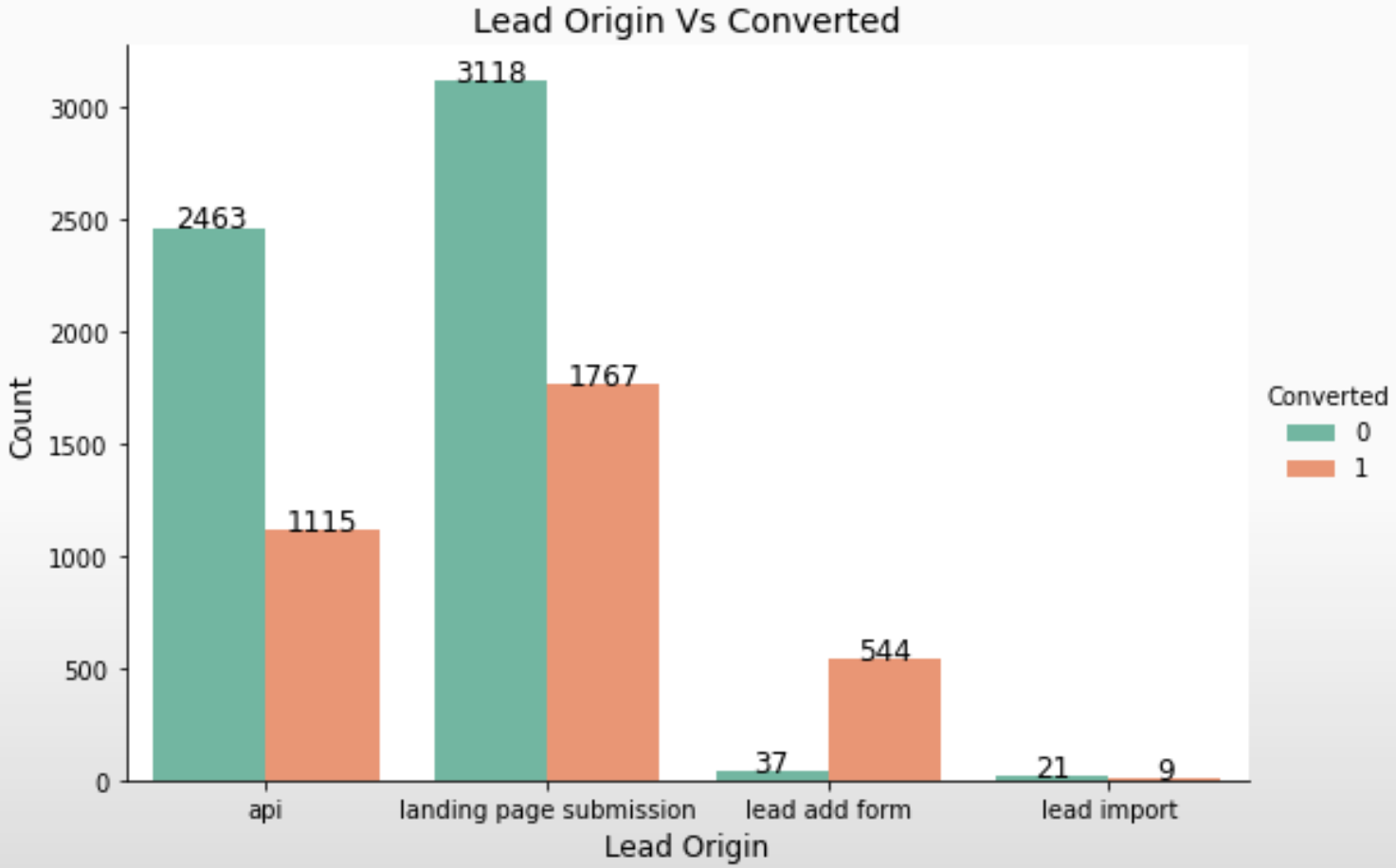
Exploratory Data Analysis

We have around 65% Conversion rate in Total

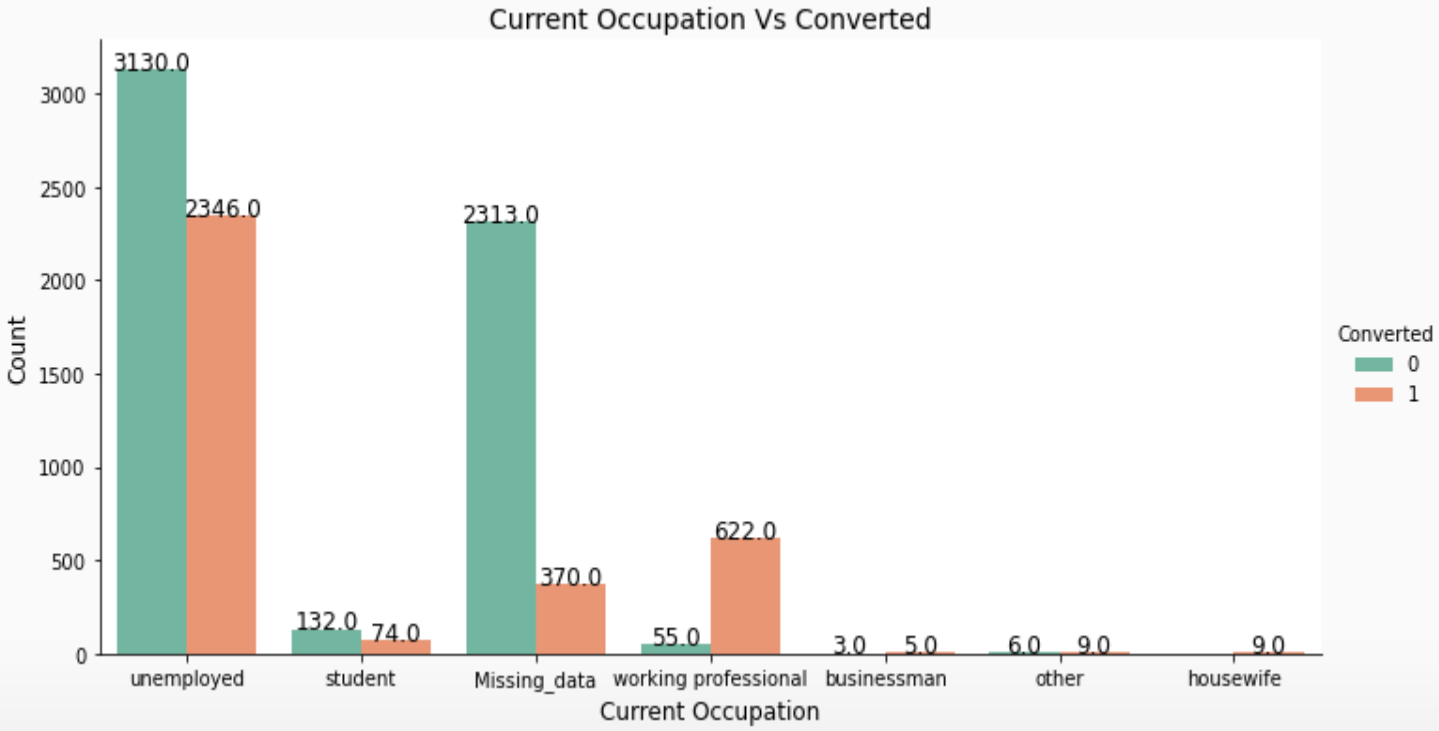


.

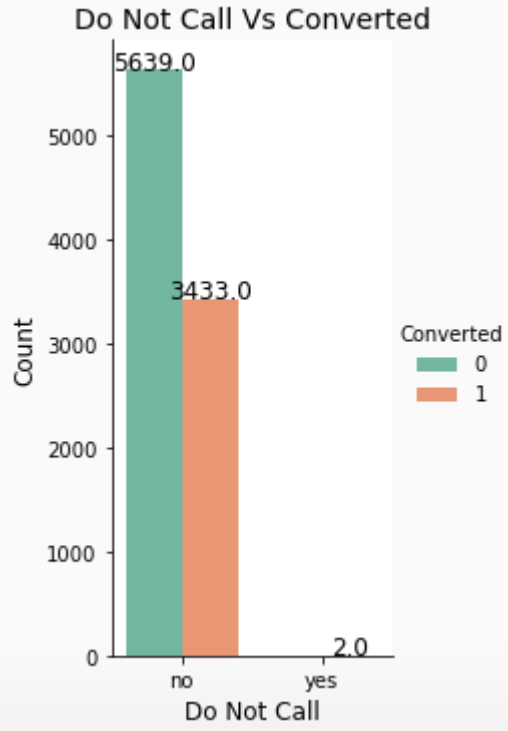
In Lead Origin, maximum conversion occurred during Landing Page Lucida Fax Submission

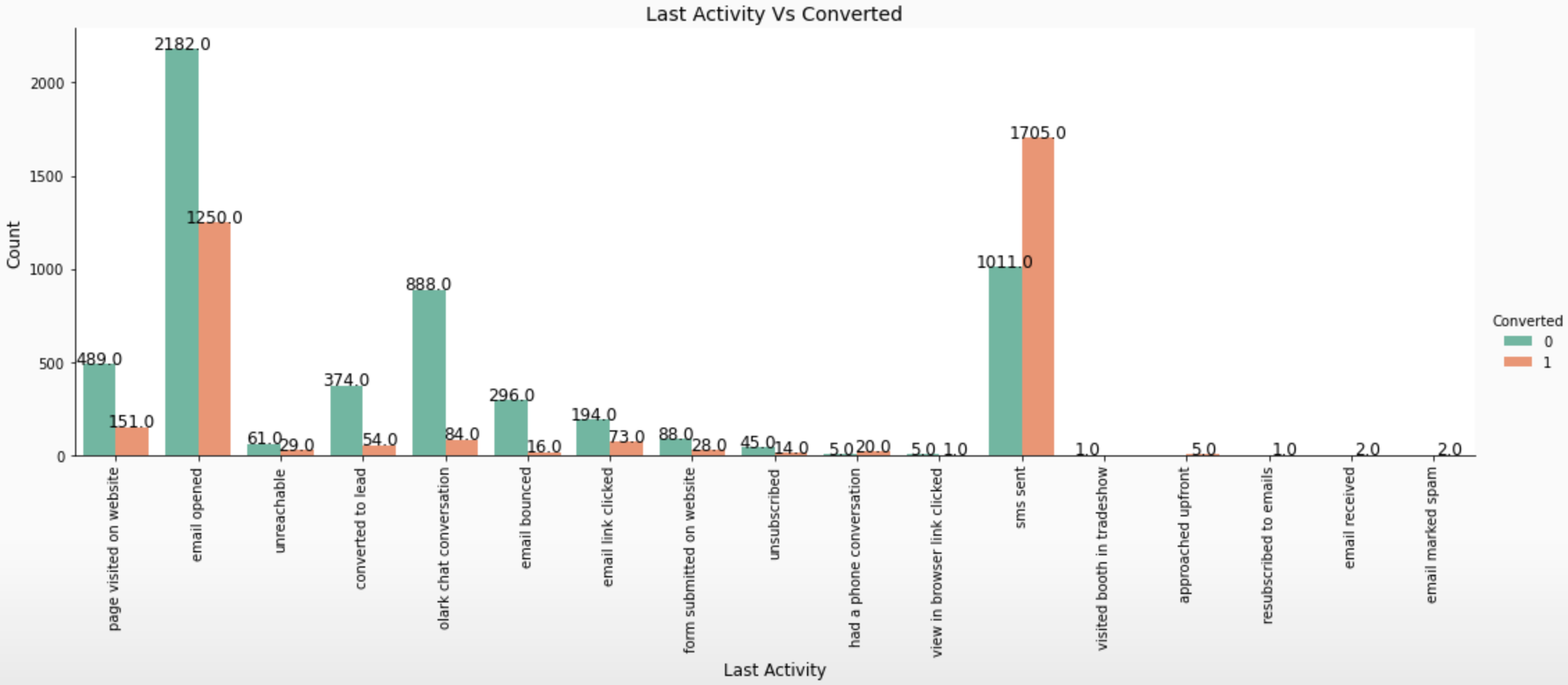


More conversion happened with people who are unemployed

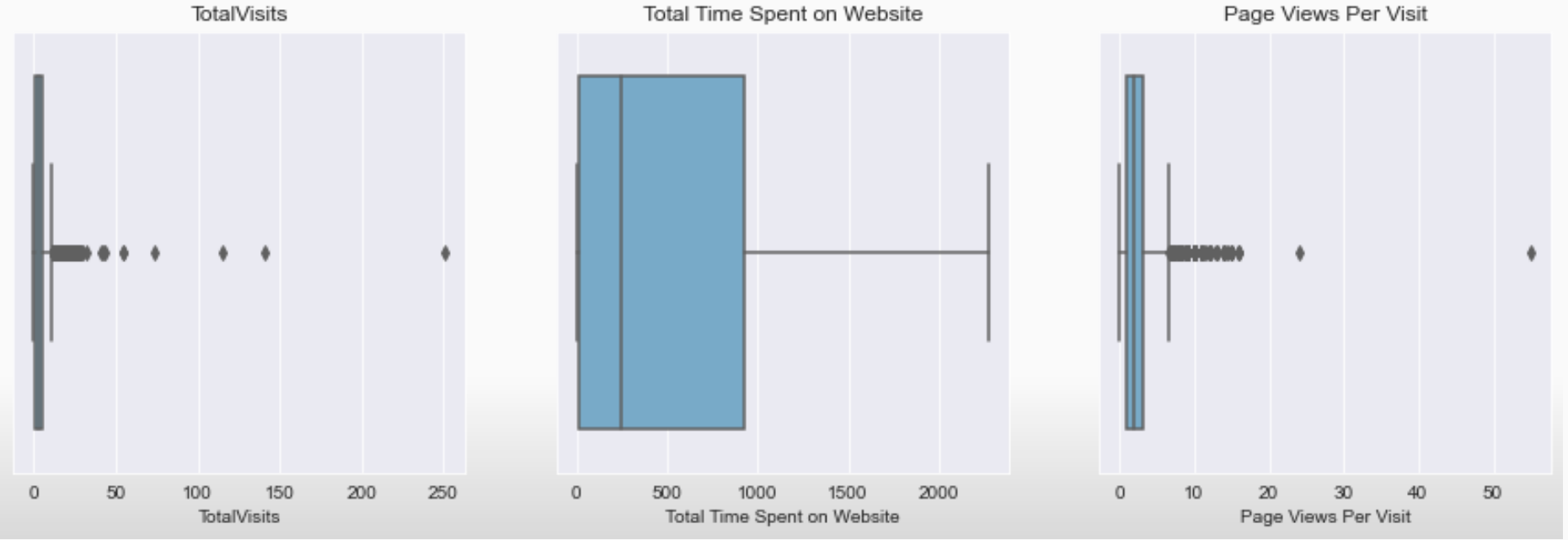


Major conversion Calls made





Last Activity value of SMS Sent' had more conversion.



Outlier Treatment for all three numerical variables.

Fitting data between IQR

DATA PREPARATION:

- Getting Dummies variable for categorical columns
- Dropping columns with redundant values

Test Train split (70:30)

Feature Scaling for uniform scale of data (using MinMaxScaler)

Feature Selection using RFE with 15 variable as output

Model building

Prediction on test data

Obtaining accuracy 80%

- **Creating a dataframe with the actual converted and the converted probabilities**
- keeping a threshold of P >= 0.5 for conversion

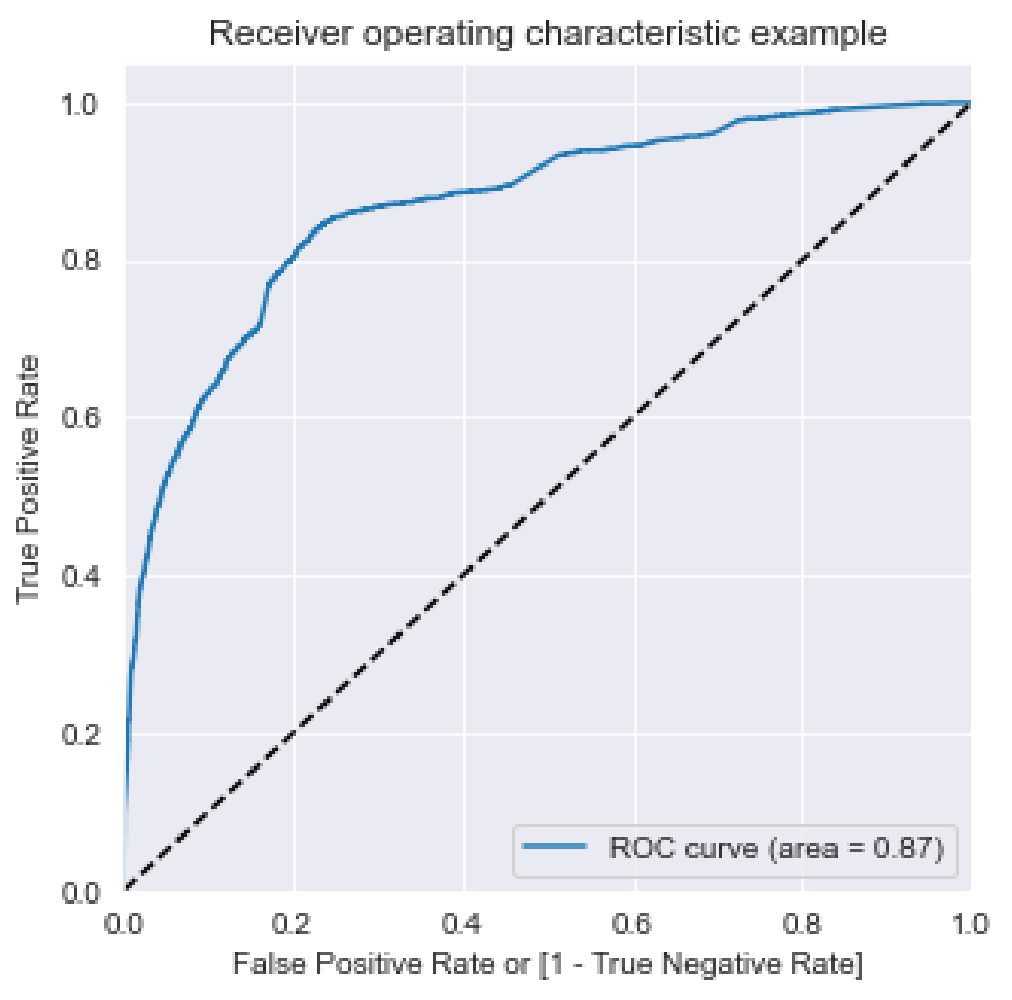
- **Optimizing the model for P- values and VIF**

Confusion Matrix

3342	403
795	1417

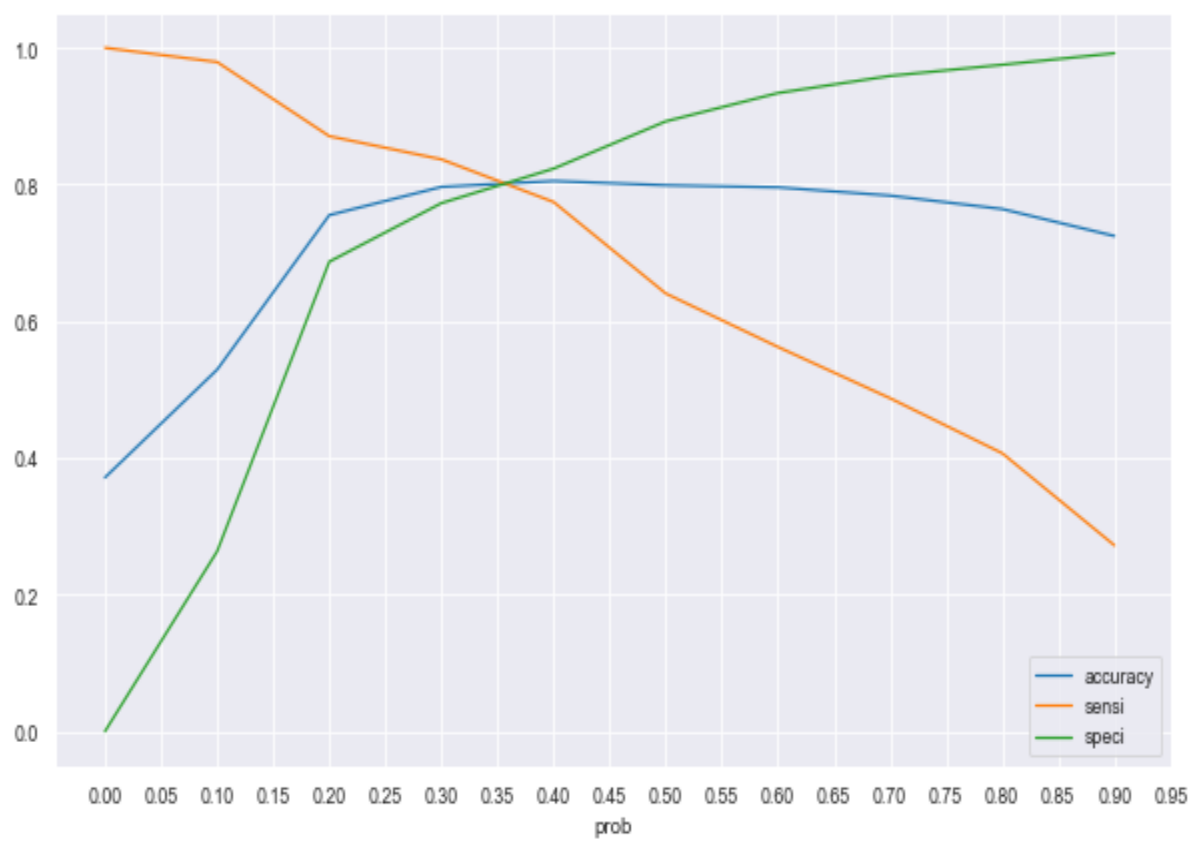
- Accuracy - 80%
- Sensitivity - 64 %
- Specificity - 89 %
- False Positive Rate - 18 %
- Positive Predictive Value - 85%
- Positive Predictive Value – 79%

ROC Curve for Train set



- Area under the curve is : 0.87

The graph depicts an optimal cut off of 0.36 based on Accuracy, Sensitivity and Specificity



	prob	accuracy	sensi	speci
0.0	0.0	0.371328	1.000000	0.000000
0.1	0.1	0.529125	0.979204	0.263284
0.2	0.2	0.755246	0.870705	0.687049
0.3	0.3	0.796542	0.836799	0.772764
0.4	0.4	0.805103	0.774412	0.823231
0.5	0.5	0.798892	0.640597	0.892390
0.6	0.6	0.795870	0.562387	0.933778
0.7	0.7	0.783784	0.487342	0.958879
0.8	0.8	0.764143	0.406872	0.975167
0.9	0.9	0.724694	0.272152	0.991989

Model Evaluation – using cut-off 3.6

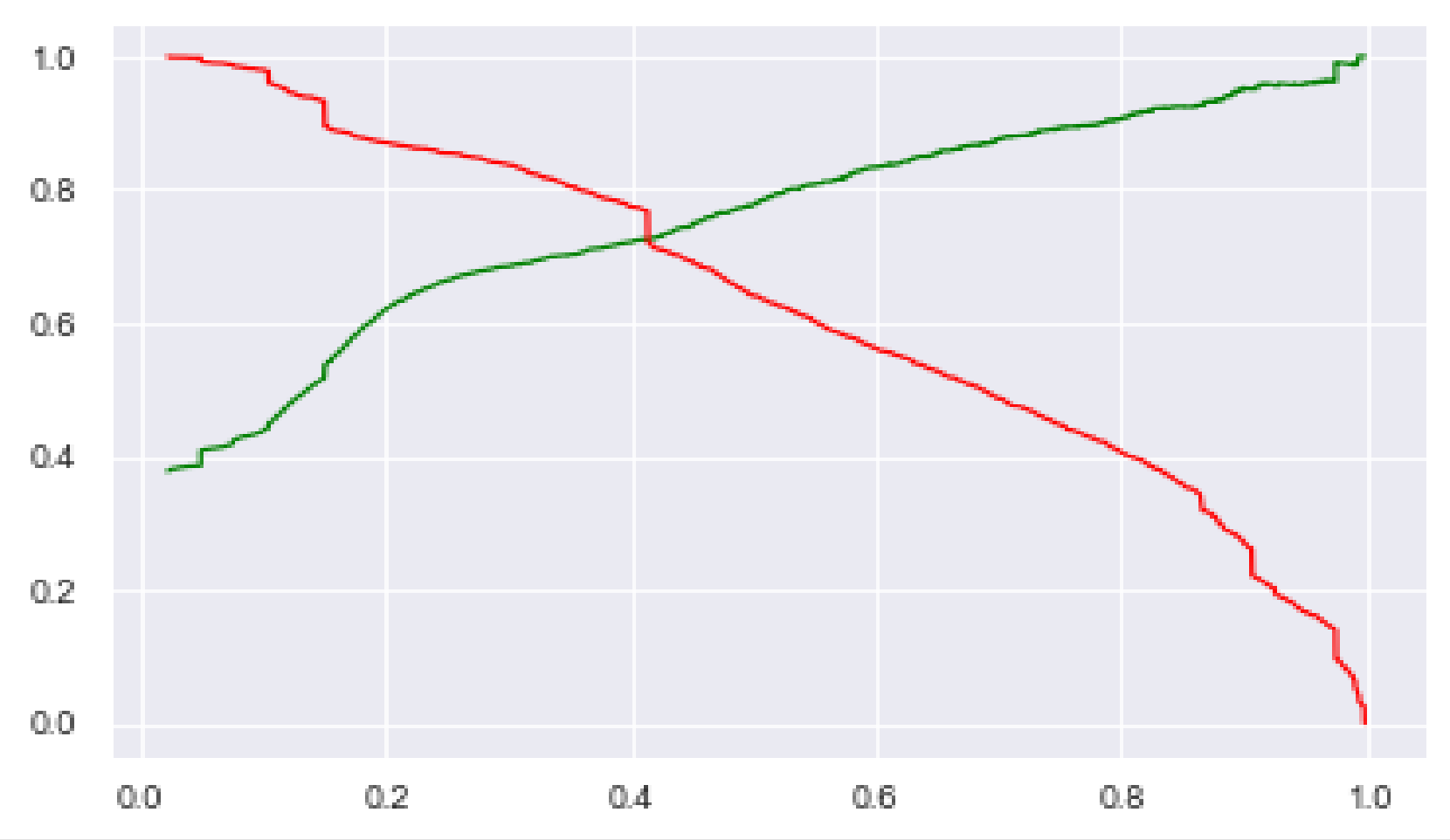
Confusion Matrix for test set

- Accuracy - 80%
- Sensitivity - 80 %
- Specificity - 80%
- Precision - 78 %
- Recall - 64 %

Confusion Matrix for train set

- Accuracy - 80%
- Sensitivity - 80 %
- Specificity - 81 %
- Precision - 78 %
- Recall - 64 %

Graph above shows that, we get the optical threshold value as close to .41. However our target is to have Lead Conversion Rate around 80%



F1 Score:
78 %

Model Evaluation – using cut-off 0.41(save to go with)

Confusion Matrix for test set

- Accuracy - 80%
- Precision - 72 %
- Recall - 77 %

Confusion Matrix for train set

- Accuracy - 81%
- Precision - 75%
- Recall - 78 %

Determining Feature Importance

It was found that the variables that mattered the most in the potential buyers are (In descending order)

- ✓ The total time spend on the Website.
- ✓ Lead Origin- lead add form
- ✓ Lead Source_direct traffic
- ✓ Do Not Email_yes
- ✓ Last Activity_olark chat conversation
- ✓ What is your current occupation_working profession
- ✓ Last Notable Activity_email link clicked
- ✓ Last Notable Activity_email opened
- ✓ Last Notable Activity_modified
- ✓ Last Notable Activity_olark chat conversation
- ✓ Last Notable Activity_page visited on website

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

Conclusion

- Accuracy, Sensitivity and Specificity values of test set are around 81%, 81.5% and 81% which are approximately closer to the respective values calculated using trained set.

- Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 81% in test set

- The top 3 variables that contribute for lead getting converted in the model are
  - ✓ The total time spend on the Website.
  - ✓ Lead Origin- lead add form
  - ✓ Lead Source_direct traffic

- Hence overall this model seems to be good.