Lead Score Case Study

Created By:
Aakash Agarwal

Problem Statement

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. 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%.

Business Goal

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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%

Strategy

1. Source the data for analysis.
2. Clean and prepare the data.
3. Exploratory Data Analysis.
4. Splitting the data into Test and Train dataset.
5. Feature Scaling.
6. Building a logistic Regression model and calculate Lead Score.
7. Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
8. Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Problem Solving Flow

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 Numeric data
- Splitting data into train and test set

Model Building

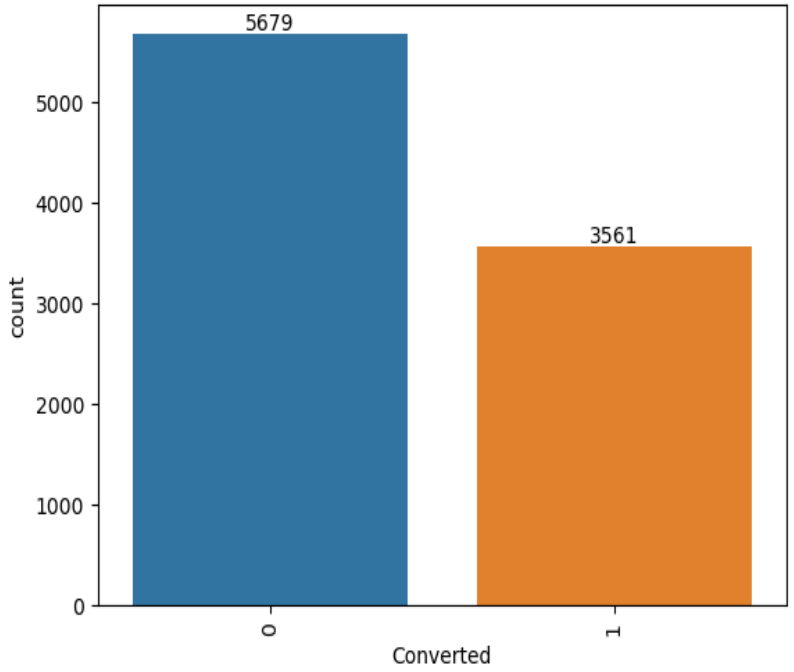
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model

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

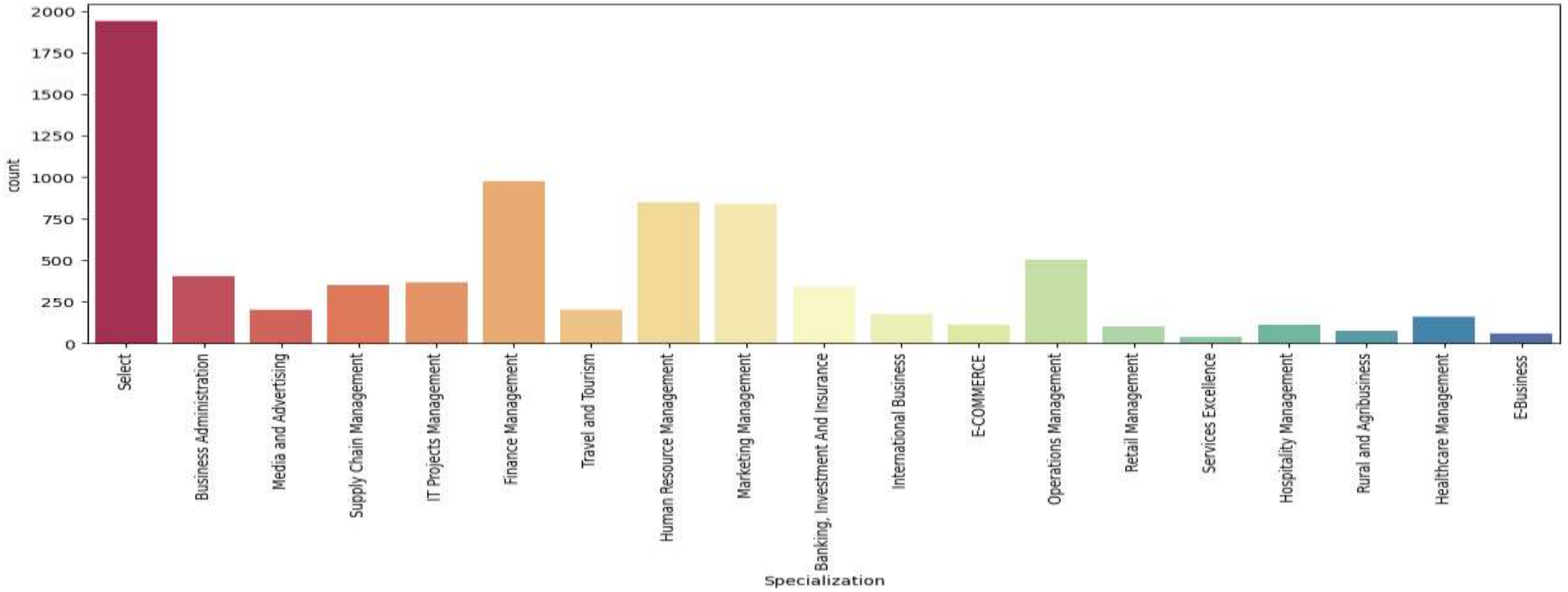
Exploratory Data Analysis

We have around **39%** Conversion Rate

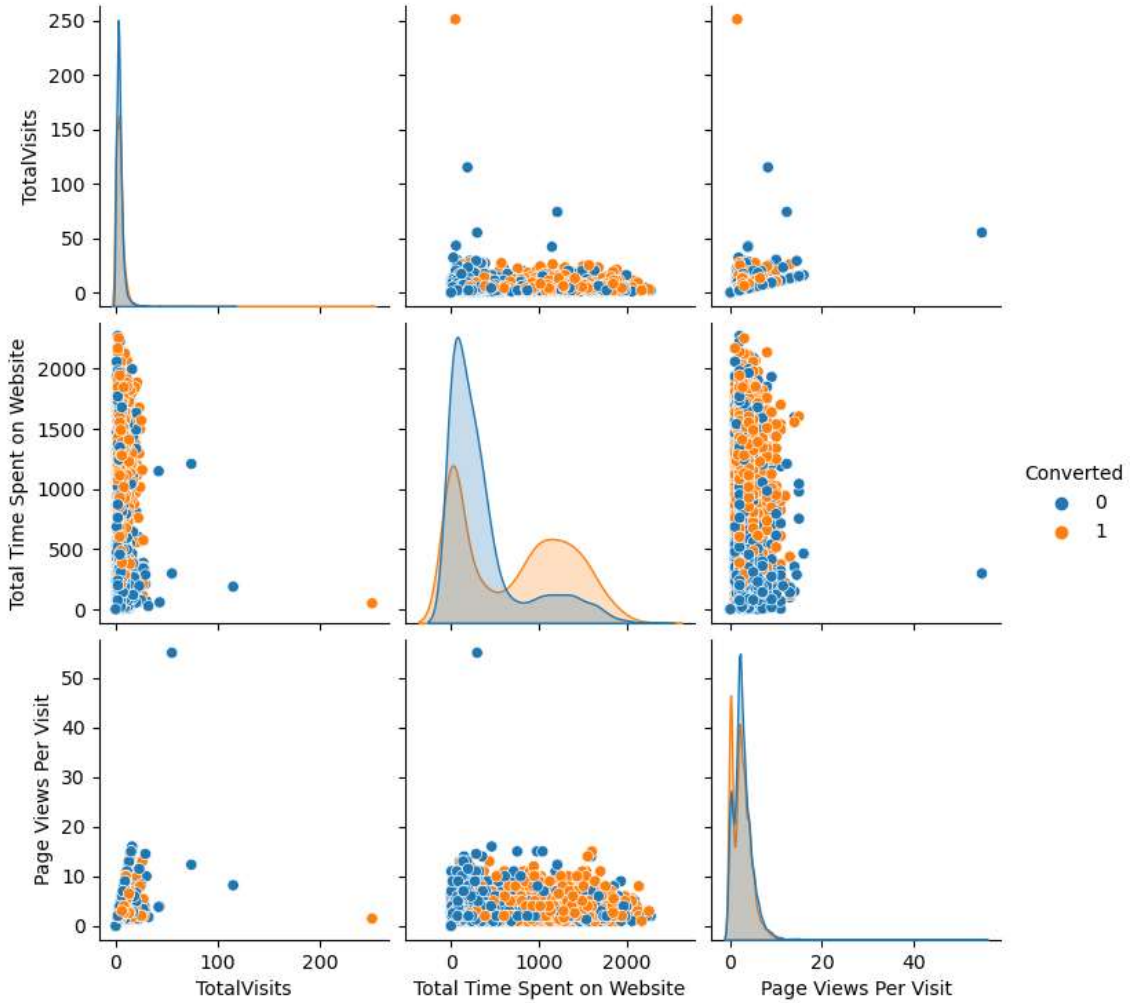


Plotting count plot for Specialization

Analyzing the column **Specialization** as it seems to be an important feature for the analysis.



Plotting pair plot for Continuous Variables

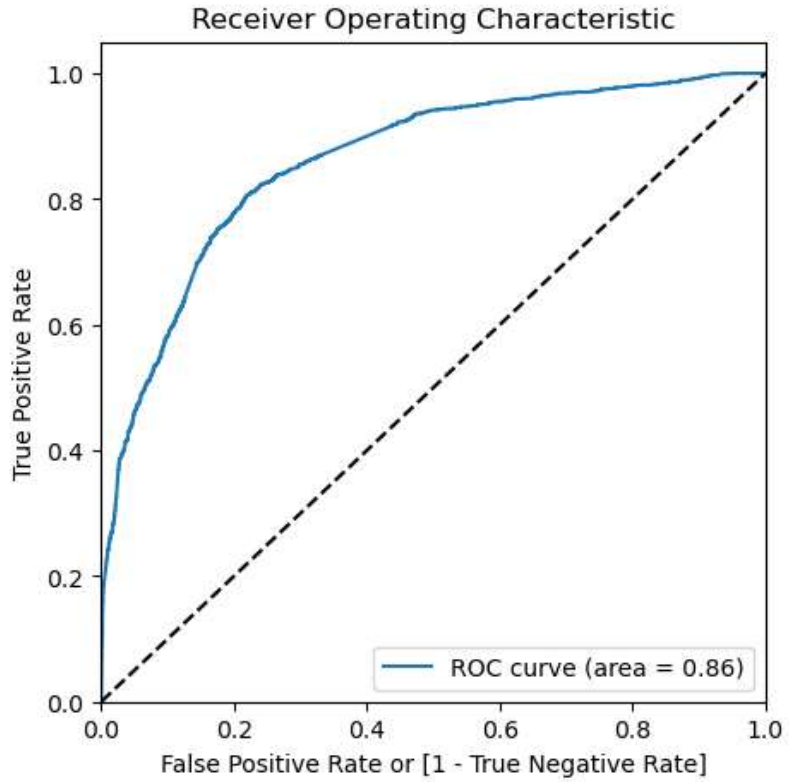


Features Impacting the Conversion Rate

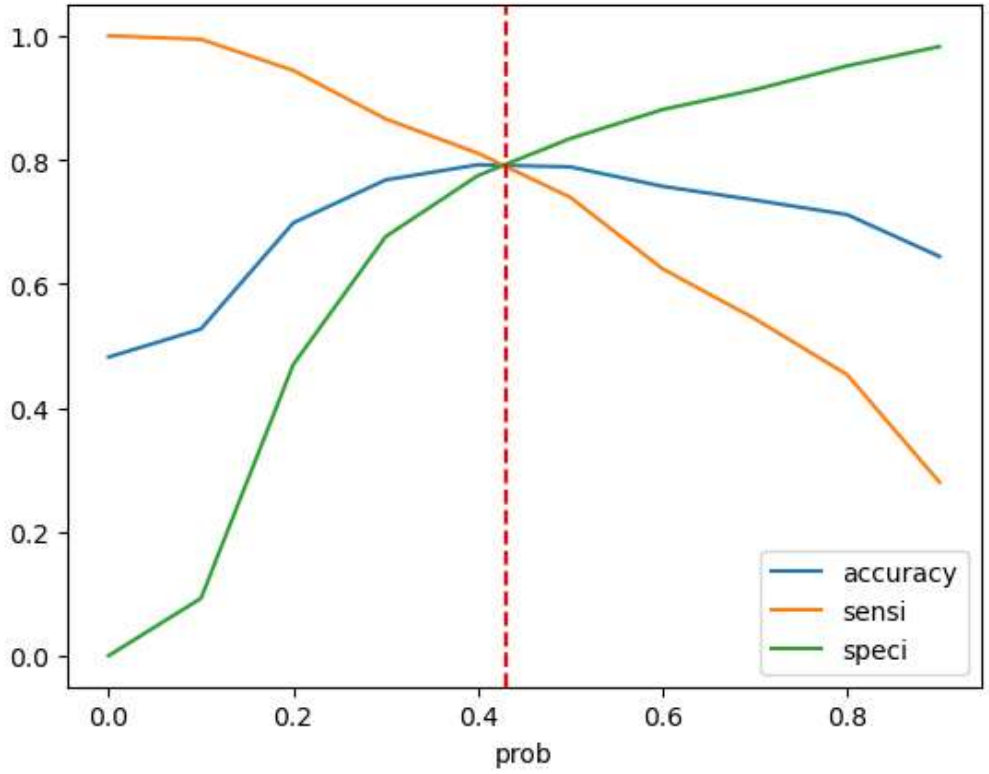
1. TotalVisits
2. Total Time Spent on Website
3. Lead Origin_Lead Add Form
4. Last Notable Activity_Unreachable
5. Last Activity_Had a Phone Conversation
6. Lead Source_Welingak Website
7. Lead Source_Olark Chat
8. Last Activity_SMS Sent
9. Do Not Email_Yes
10. What is your current occupation_Student
11. What is your current occupation_Unemployed

Model Evaluation - Sensitivity and Specificity on Train Data Set

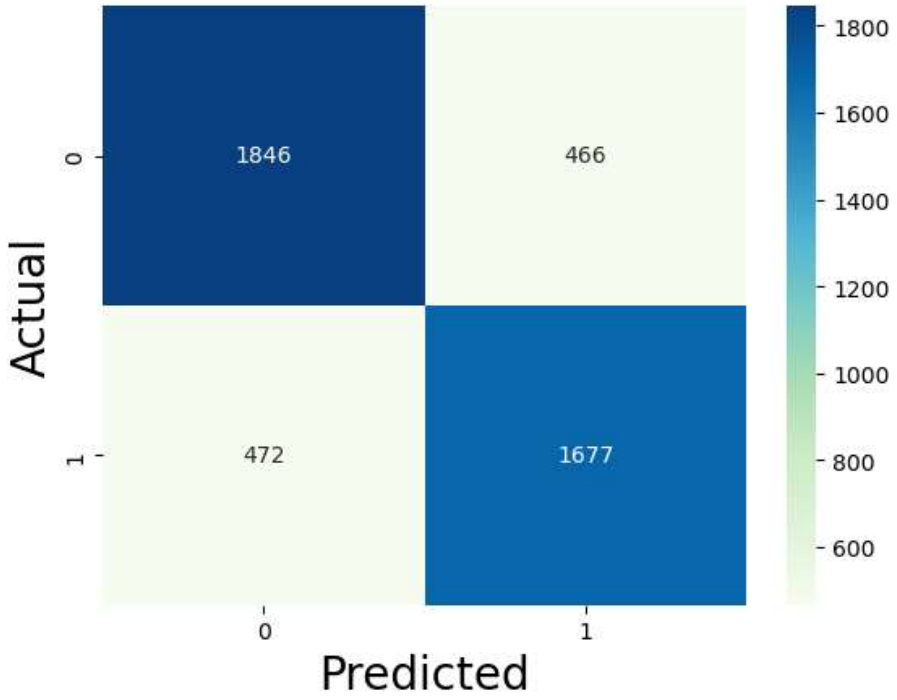
Receiver Operating Characteristic (ROC) Curve



Finding Optimal Cutoff by plotting Accuracy, Sensitivity and Specificity = **0.43**



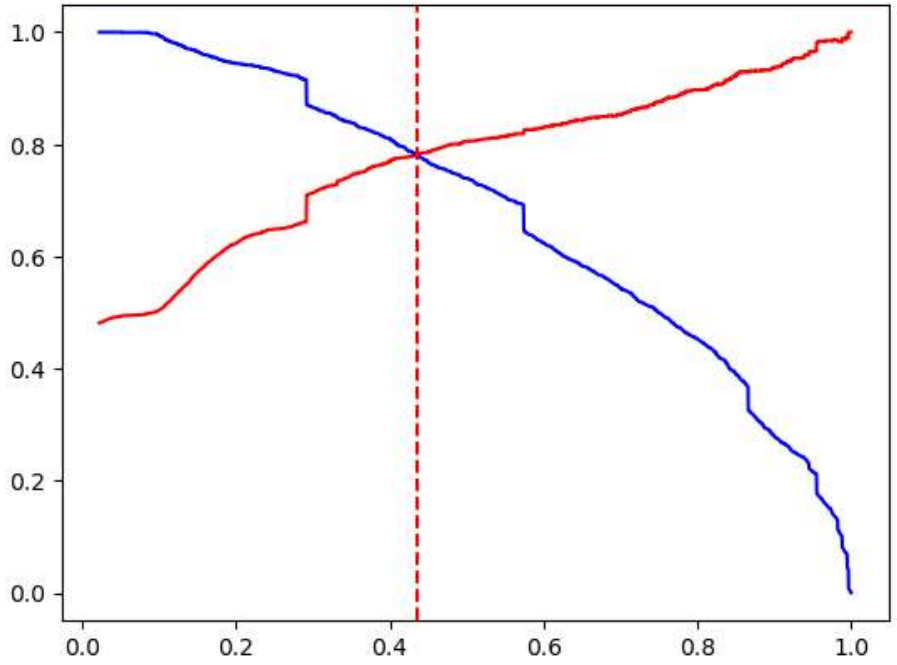
Confusion Matrix of Train Data



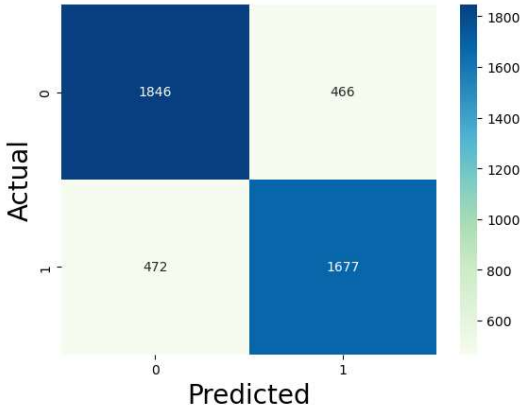
- **Accuracy: 0.79**
- **Sensitivity: 0.78**
- **Specificity: 0.79**

Model Evaluation - Precision and Recall on Train Data Set

Finding Optimal Cutoff by plotting Precision and Recall =
0.435



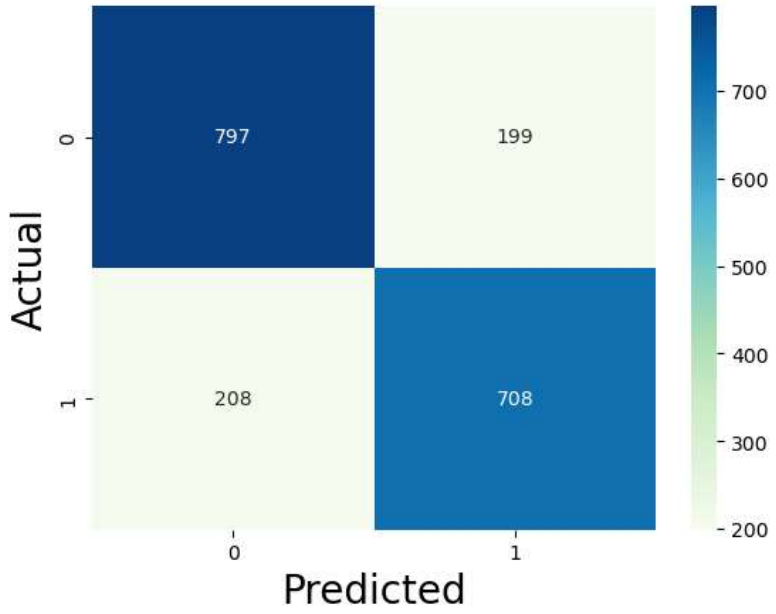
Confusion Matrix



- Precision: 0.78
- Recall: 0.78

Confusion Matrix is same as of Sensitivity and Specificity as the cutoff is almost the same.

Model Evaluation - Sensitivity and Specificity on Test Data Set



- **Accuracy: 0.78**
- **Sensitivity: 0.77**
- **Specificity: 0.79**
- **Precision: 0.78**
- **Recall: 0.77**

Final Model Line Equation

Converted = 0.204 + 11.1489 X TotalVisits + 4.4223 X Total Time Spent on Website + 4.2051 X Lead Origin_Lead Add Form + 1.4526 X Lead Source_Olark Chat + 2.1526 X Lead Source_Welingak Website - 1.5037 X Do Not Email_Yes + 2.7552 X Last Activity_Had a Phone Conversation + 1.1856 X Last Activity_SMS Sent - 2.3578 X What is your current occupation_Student - 2.5445 X What is your current occupation_Unemployed + 2.7846 X Last Notable Activity_Unreachable

Summary

- While we have checked both Sensitivity-Specificity as well as Precision-Recall Metrics, we have considered the optimal cut off to be **0.435** for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 78%, 77% and 79% which are approximately closer to the respective values calculated using trained set.
- Precision and Recall values of test set are around 78% and 77% which are approximately closer to the respective values calculated using trained set.
- The top 3 variables that contribute for lead getting converted in the model are:
    1. TotalVisits
    2. Total Time Spent on Website
    3. Lead Origin_Lead Add Form
- Hence overall this model seems to be accurate.

Thank you