Lead Scoring Case Study

Presented by,

Vishal Batra,

Prashanth Reddy,

Hemalatha H

Lead Score Case Study for X Education

Problem Statement:

X-Education is an education company sells online Education courses to professionals and marketing through online advertisements.

Company gets information through different channels and if candidates enquiring with certain education level it calls lead.

Typically lead conversion is 30% of certain education. Company identifying Hot Leads on certain criteria also.

Lead conversion ratio is lesser than number of enrollment. Company given Target to achieve 80% of total enrollment.

Business Goal:

Building logistics regression model to finding leads for Company and help to achieve potential targets.

Alternative approach should be ready in case Company's requirement changes in futures should be flexible.

Approach

- Source the data for analysis
- Reading and Understanding the data
- Data cleaning
- Exploratory Data analysis
- Data Preparation
- Splitting the data into Test and Train dataset
- Feature Scaling
- Model Building
- Model Evaluation
- Plotting the ROC curve and Finding the optical cutoff
- Evaluating the model by using different metrics – Specificity and Sensitivity or Precision and Recall
- Making predictions

Problem solving methodology

1. Data sourcing, cleaning and preparation:

- Read the data from CSV file
- Outlier treatment
- Data cleaning – Handling Null values and removing higher null values data
- Removing Redundant columns in the data
- Imputing Null values
- Exploratory data analysis
- Feature standardization

2. Feature scaling and splitting train and test sets:

- Feature scaling of Numeric data
- Splitting data into train and test set.

3. Model building:

- Feature selection using RFE
- Determine the optimal model using logistic regression
- Calculating the model by using different metrics – Specificity and Sensitivity or Precision and Recall

4. Result or conclusion

- Determine the lead score and check if target final predictions amount
- Evaluating the model by using different metrics – Specificity and Sensitivity or Precision and Recall

Data Sourcing, Cleaning and Preparation

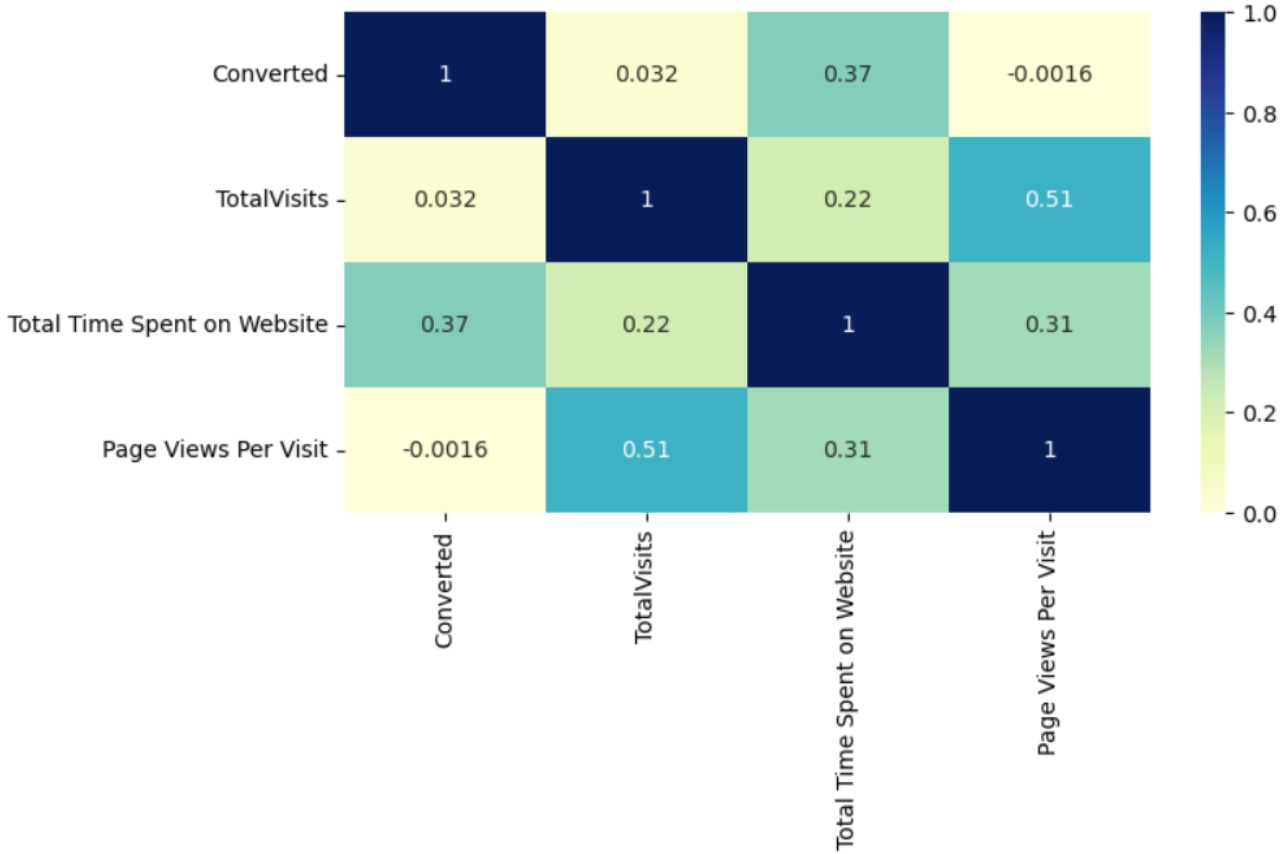
- Read the data from CSV file
- Outlier treatment
- Data cleaning – Handling Null values and removing higher null values data
- Removing Redundant columns in the data
- Imputing Null values
- Exploratory data analysis
- Feature standardization

- Read the data from source
- Convert data into clean format suitable for anlysis
- Remove duplicate data
- Outlier treatment
- Exploratory data analysis
- Feature standardization

Exploratory Data Analysis

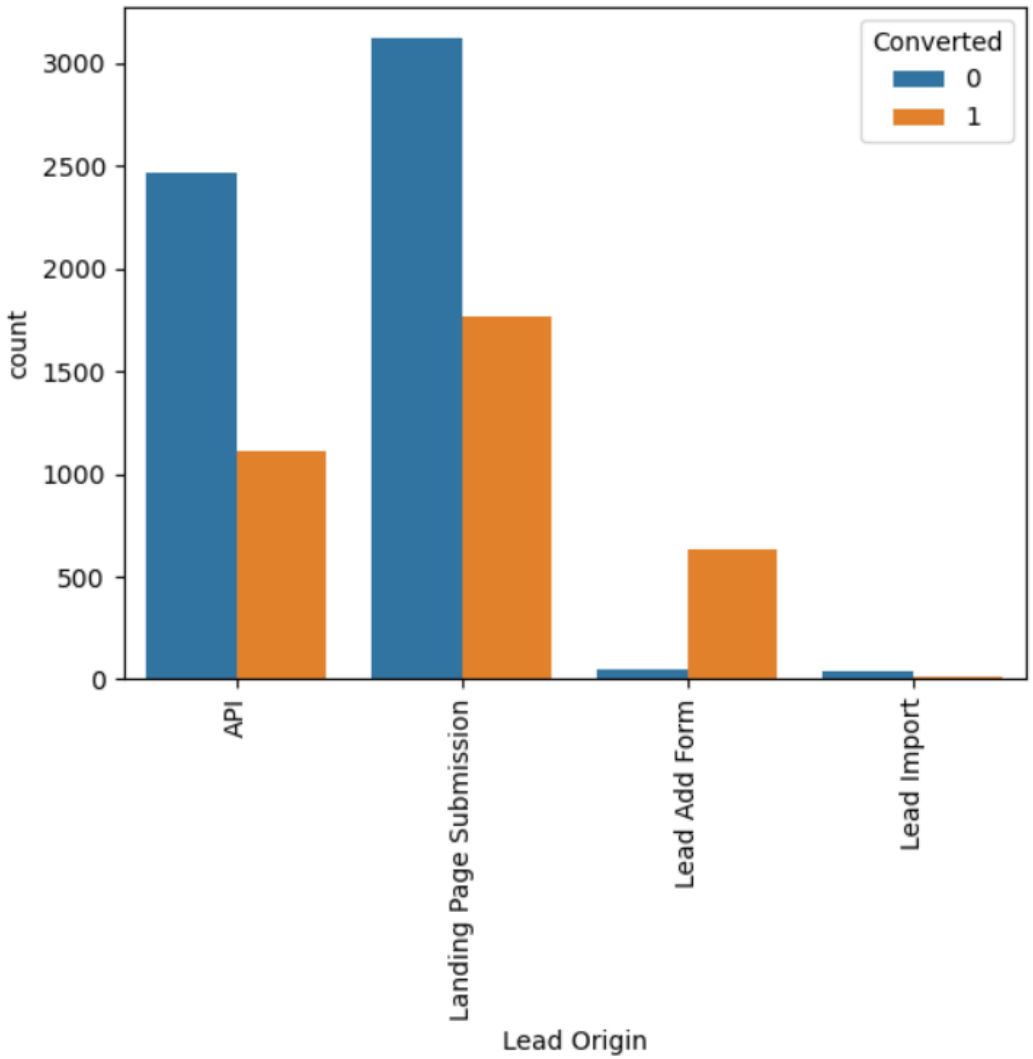
Numerical Analysis:

Checking correlations of numeric values



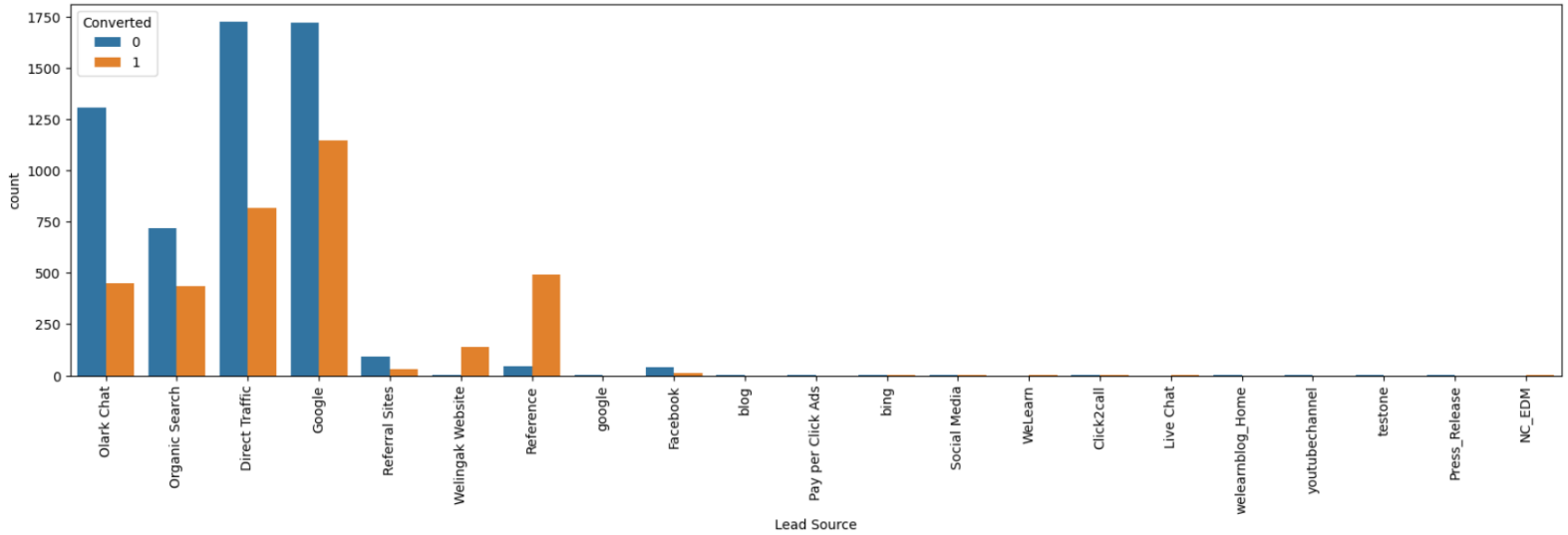
Univariate Analysis:

Plotting count plot of 'Lead Origin' for both 'Converted' 0 and 1 :

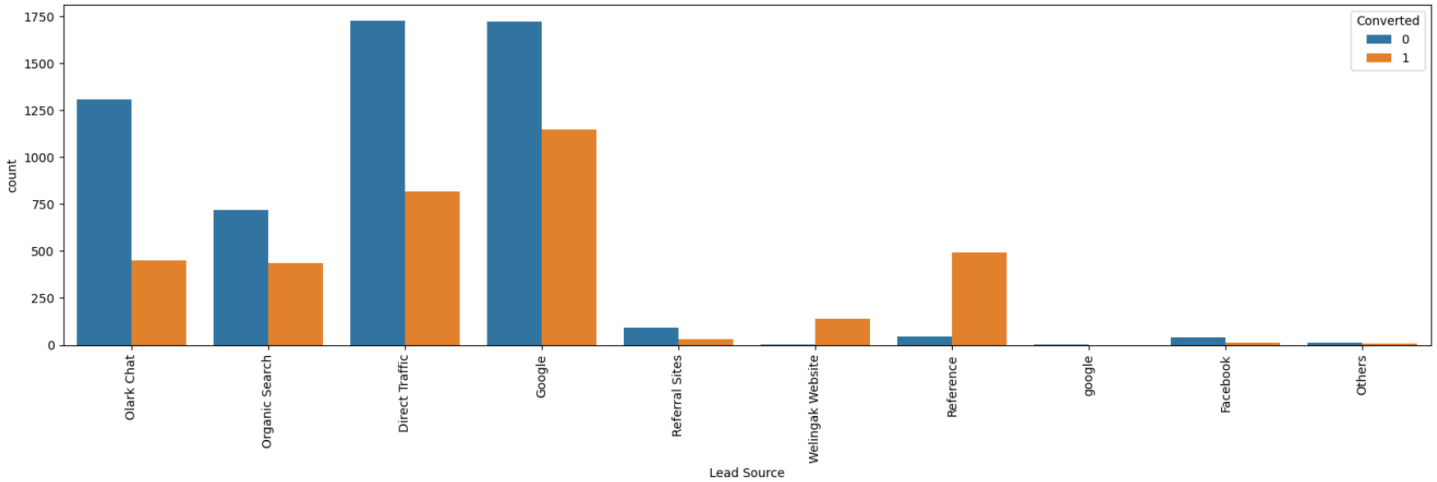


Univariate Analysis:

Plotting count plot of 'Lead Source' based on 'Converted' value 0 and 1

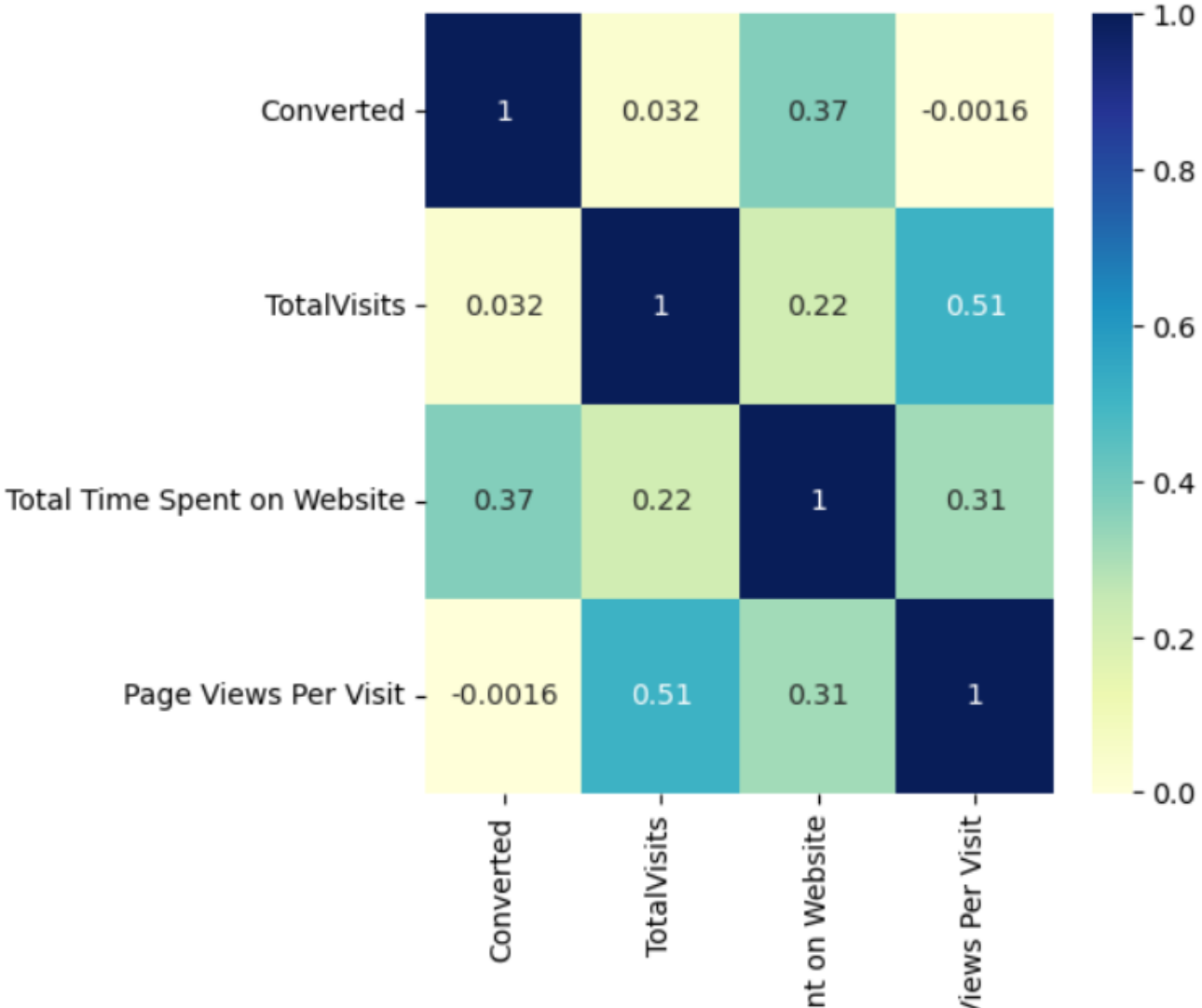


Again Plotting count plot of 'Lead Source' based on 'Converted' value 0 and 1



Bivariate Analysis:

Visualizing the correlation between all set of usable columns

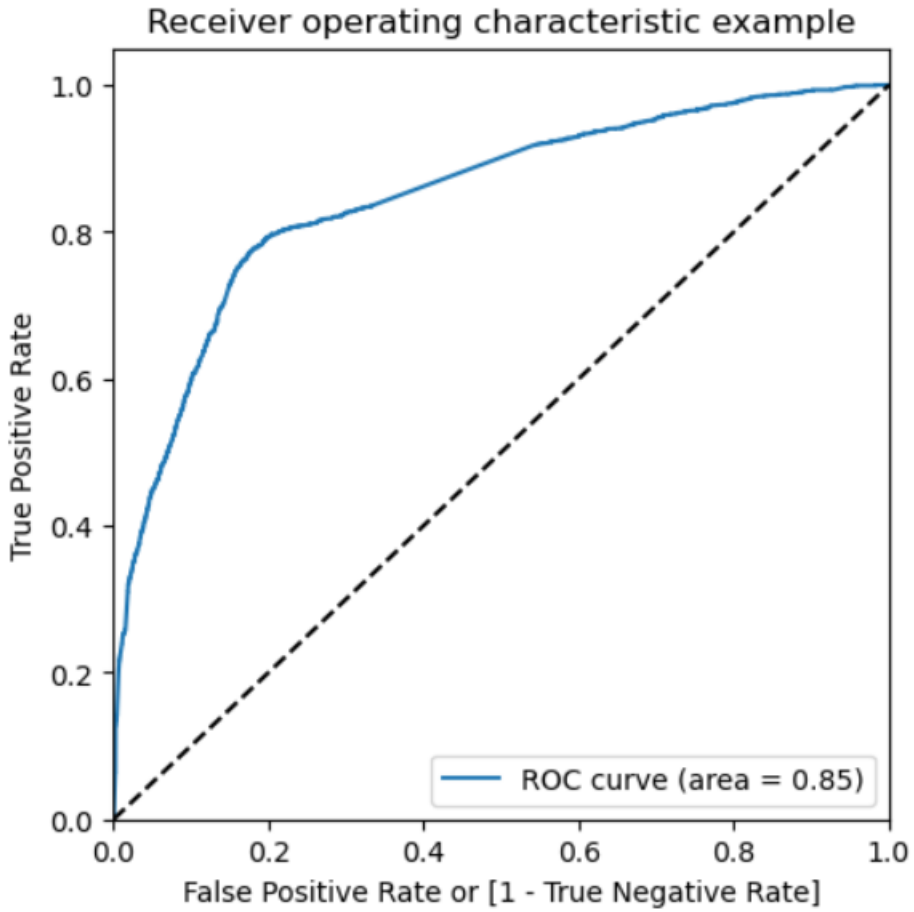


Variables impacting the conversion rate

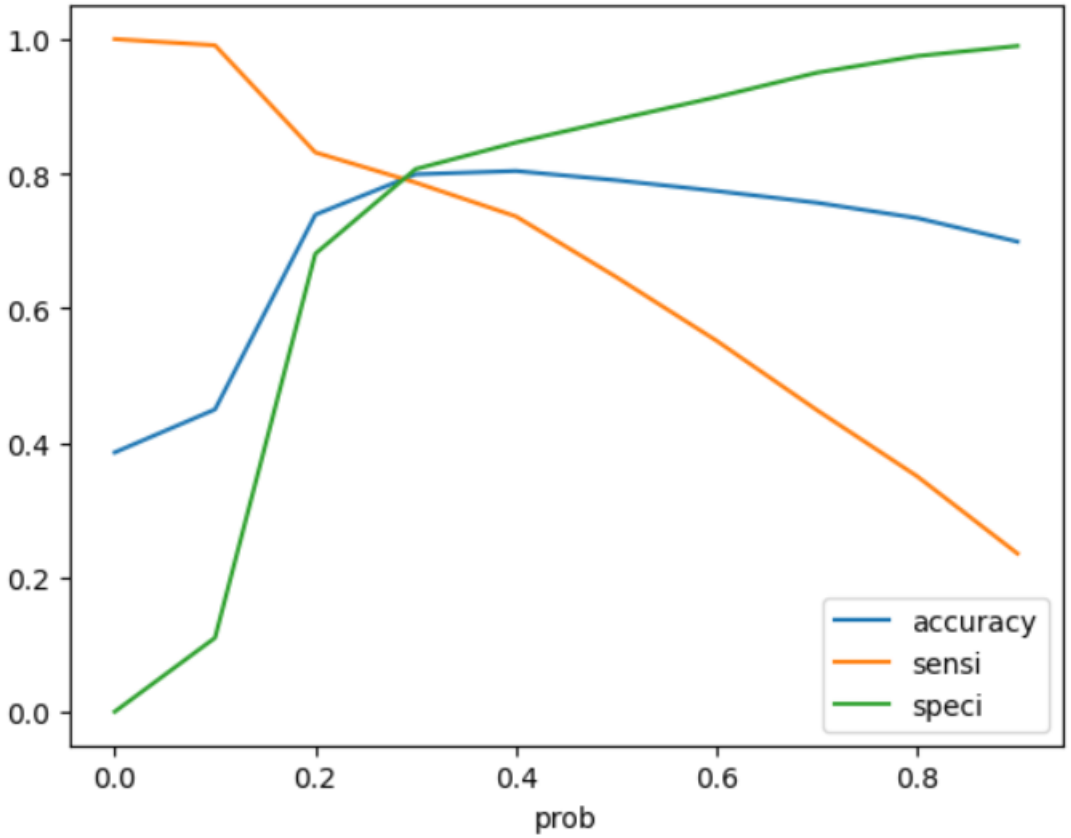
- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source – Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had A Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation

The area under the ROC curve is 0.85, which is pretty good. So it looks like we have a pretty good model. To find the cutoff, we also examine the sensitivity and the specificity of change.



As we can see, you get the best value of the three parameters around 0.3. Now let's choose 0.3 as our breakpoint.



Confusion Matrix

3069	735
509	1877

Accuracy: 80%
Sensitivity: 78%
Specificity : 80%

Making Predictions

Confusion Matrix

1311	335
239	769

Sensitivity: 78%
Specificity : 80%

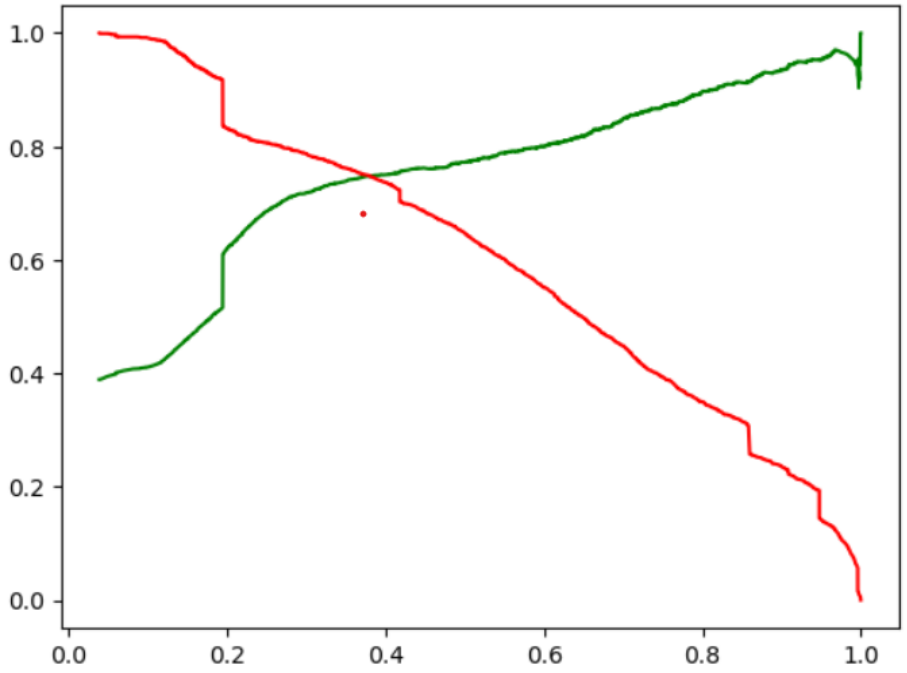
Precision and Recall View

Confusion Matrix

3349	455
844	1542

Precision: 77%
Recall : 64%

Precision and Recall tradeoff



Confusion Matrix

3069	735
509	1877

Accuracy: 79%
Precision: 72%
Recall: 79%

Final overall Predictions on the test set

Confusion Matrix

1311	335
239	769

Accuracy: 78%
Precision : 70%
Recall: 76%

Conclusion: The train and test data shows similar results

Evaluation Metrics for the train Dataset:-

- Accuracy :0.79
- Sensitivity:~0.78
- Specificity:0.80
- Precision: 0.72
- Recall: 0.78

Evaluation Metrics for the test Dataset:-

- Accuracy : 0.78
- Sensitivity: ~ 0.78
- Specificity: 0.80
- Precision: 0.70
- Recall: 0.76

Recommendations

- Lead Origin_Lead Add Form: Leads who participate in the "Lead Add Form" have a higher conversion rate, so companies can focus on getting more leads because there is a higher chance of change.
- What is your current occupation_Working Professional: Managers with "Employee Professional" job have more flexibility, companies should focus on working professionals and try to get more leads.
- Total Time Spent on Website: Administrators who spend a lot of time on the site can do it for us.

****The Model predicts the Conversion Rate very well and we can give the CEO confidence in making good calls based on this.**