Lead Score Case Study

Prepared by:

Nivethitha P

Problem Statement:

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google
- Although X Education gets a lot of leads to the course, 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’. We need 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.

Business Goal:

- To select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- To successfully identify this set of leads, so that 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
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology:

- **Data Cleansing and Manipulation**

  - Read data from source and check for duplicates

  - Handle the “Select” level that is present in many of the categorical variables.

  - Drop columns that are having high percentage of missing values.

  - Check and handle NA values and missing values in other columns. Use Imputation technique wherever necessary.

  - Check and handle outliers in data.

- **Exploratory Data Analysis and Data Preparation**

  - Check for skewed categorical columns and drop them

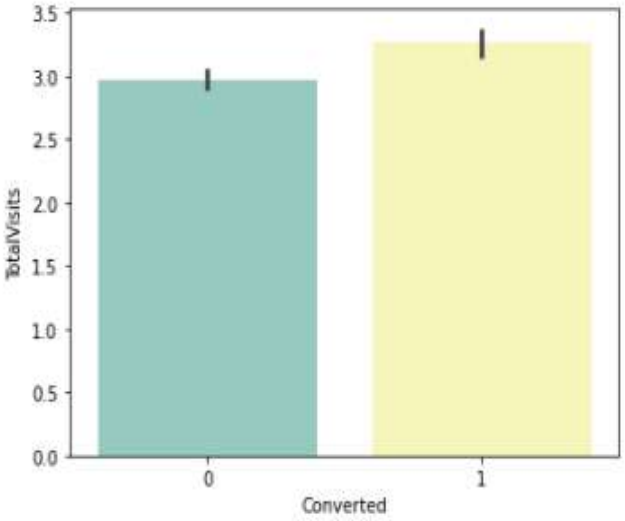
  - Analyze the various variables with respect to the Target variable and visualize them to find insights. (Find visualizations on the next slide)

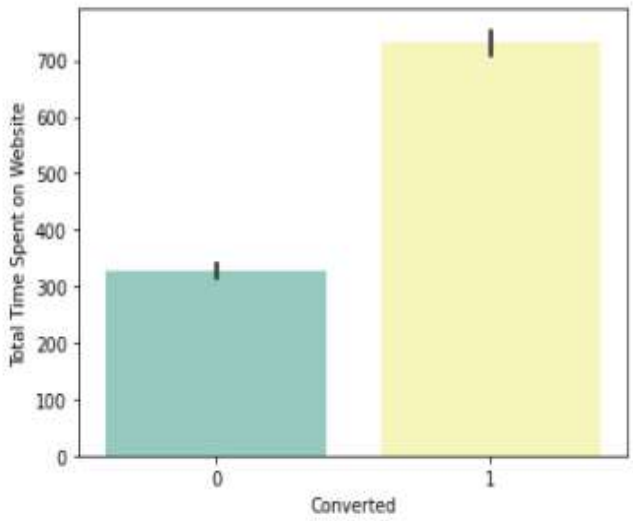
  - Create dummies for all categorical columns

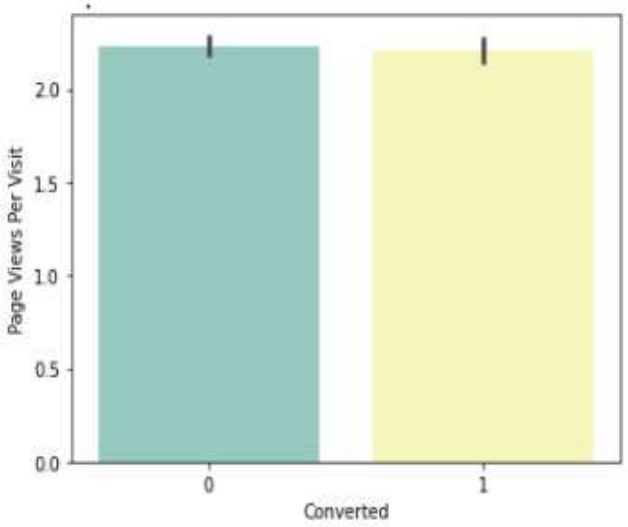
  - Feature Standardization

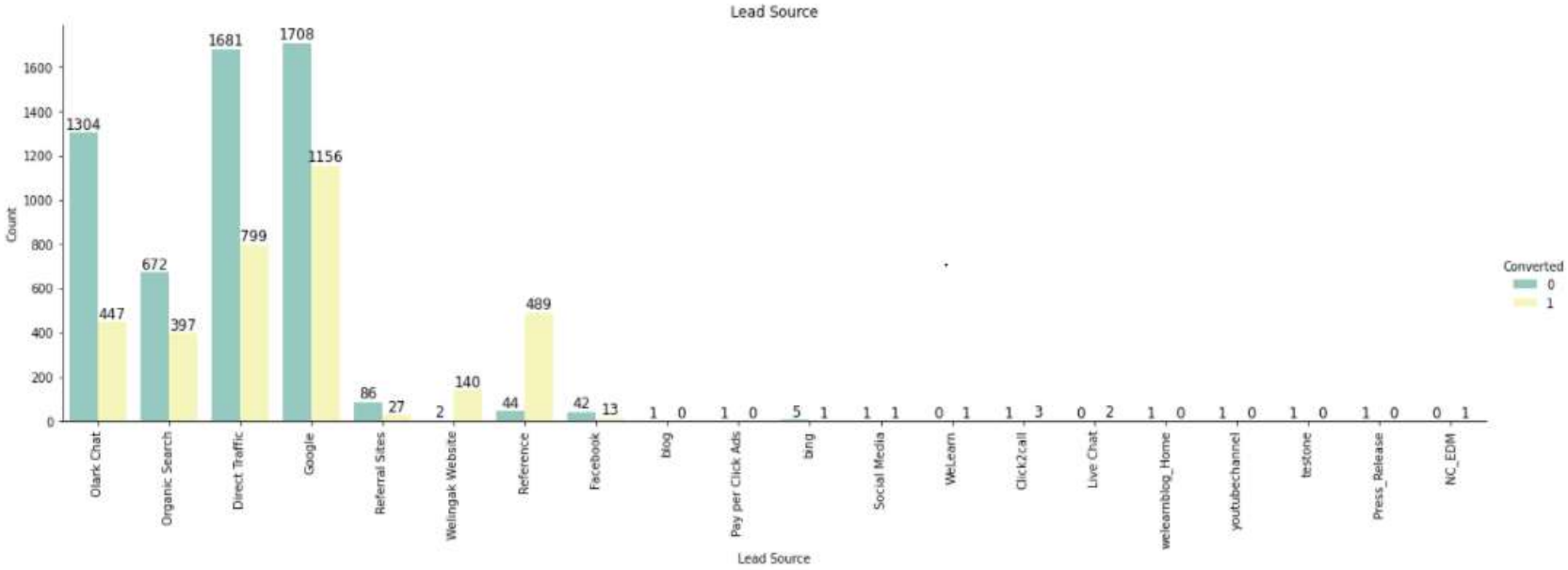
EDA Visuals:

The conversion rates were high for '**Total Visits**', '**Total Time Spent on Website**' and '**Page Views Per Visit**' columns.



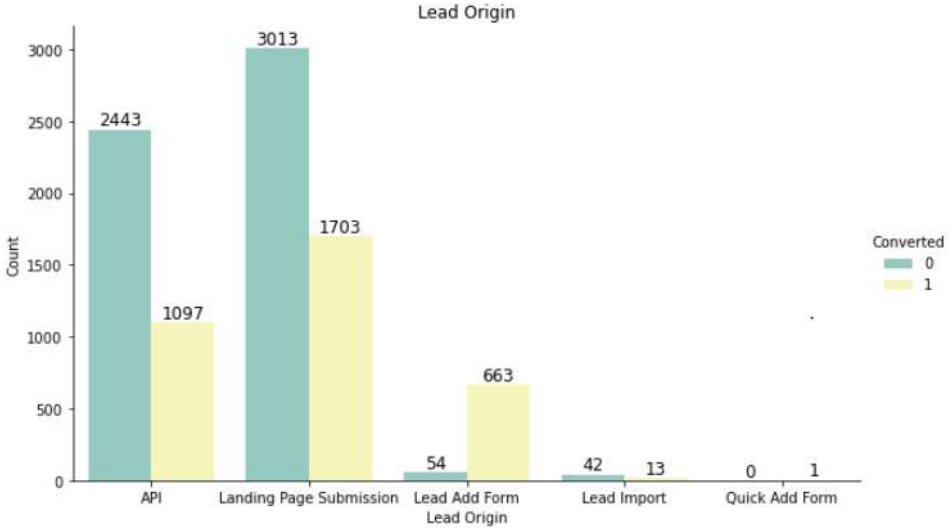






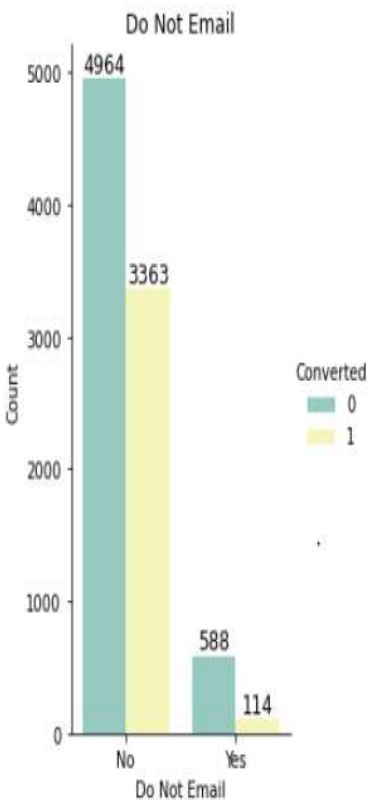
Major conversion in the '**Lead source**' is from Google

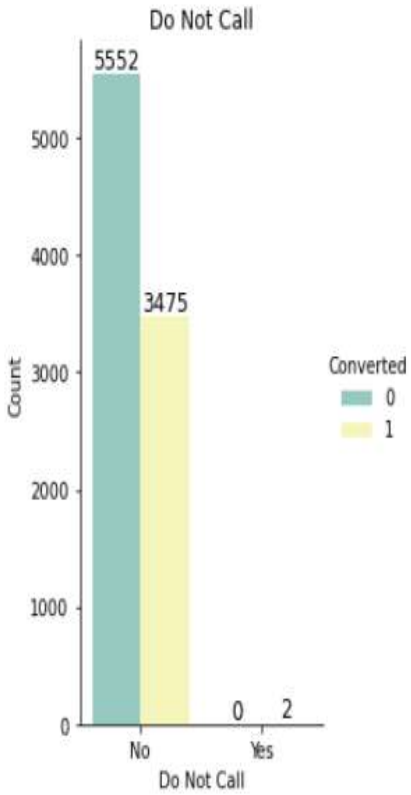
In '**Lead Origin**', maximum conversion happened from Lan Page Submission and API.

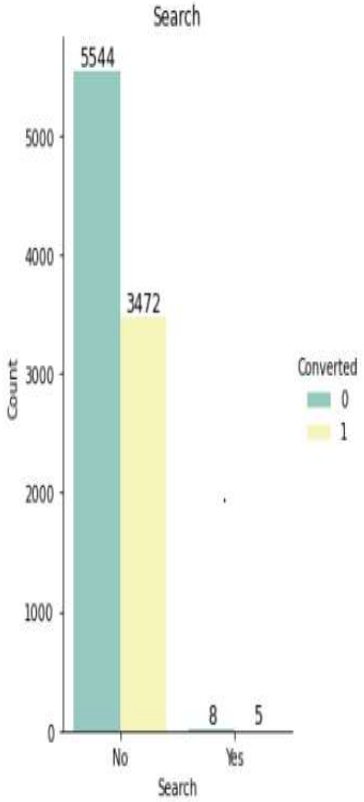


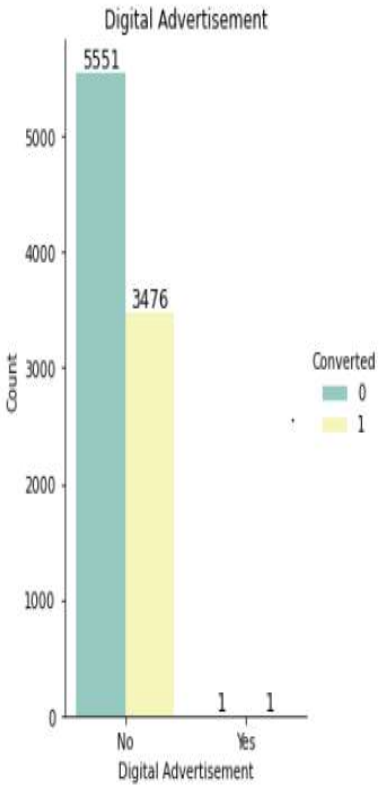
Major conversion has happened from **Emails** sent and **Calls** made

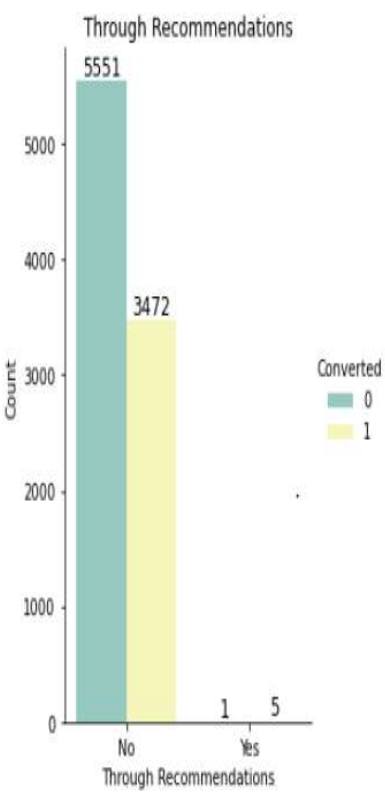
We do not see much impact on conversion rates through **Search**, **digital advertisements** and **through recommendations**





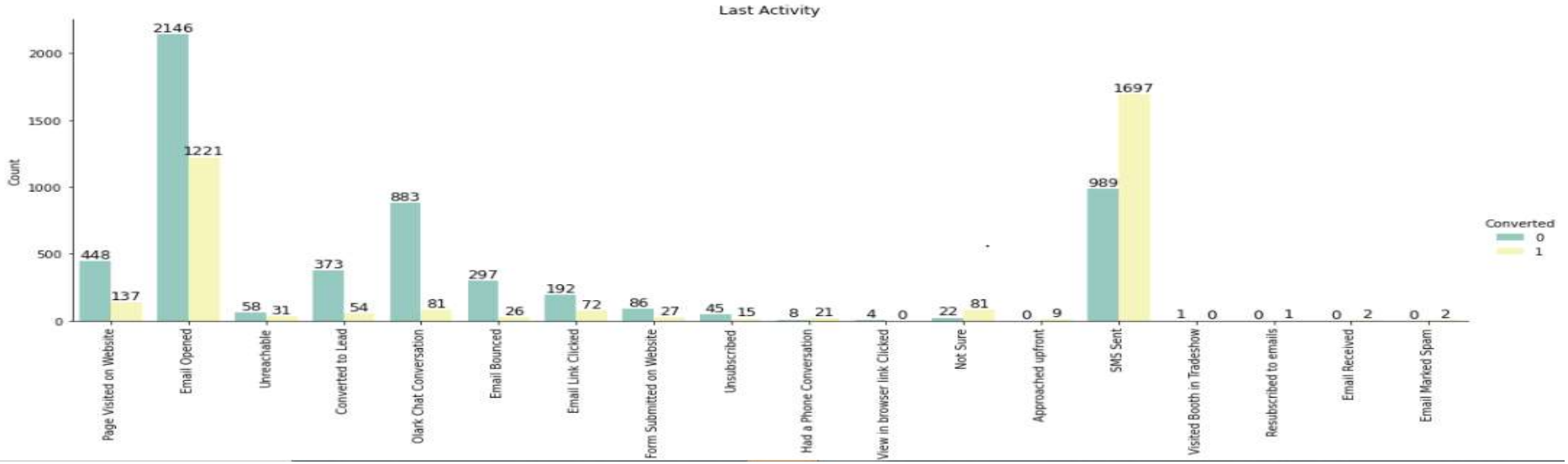




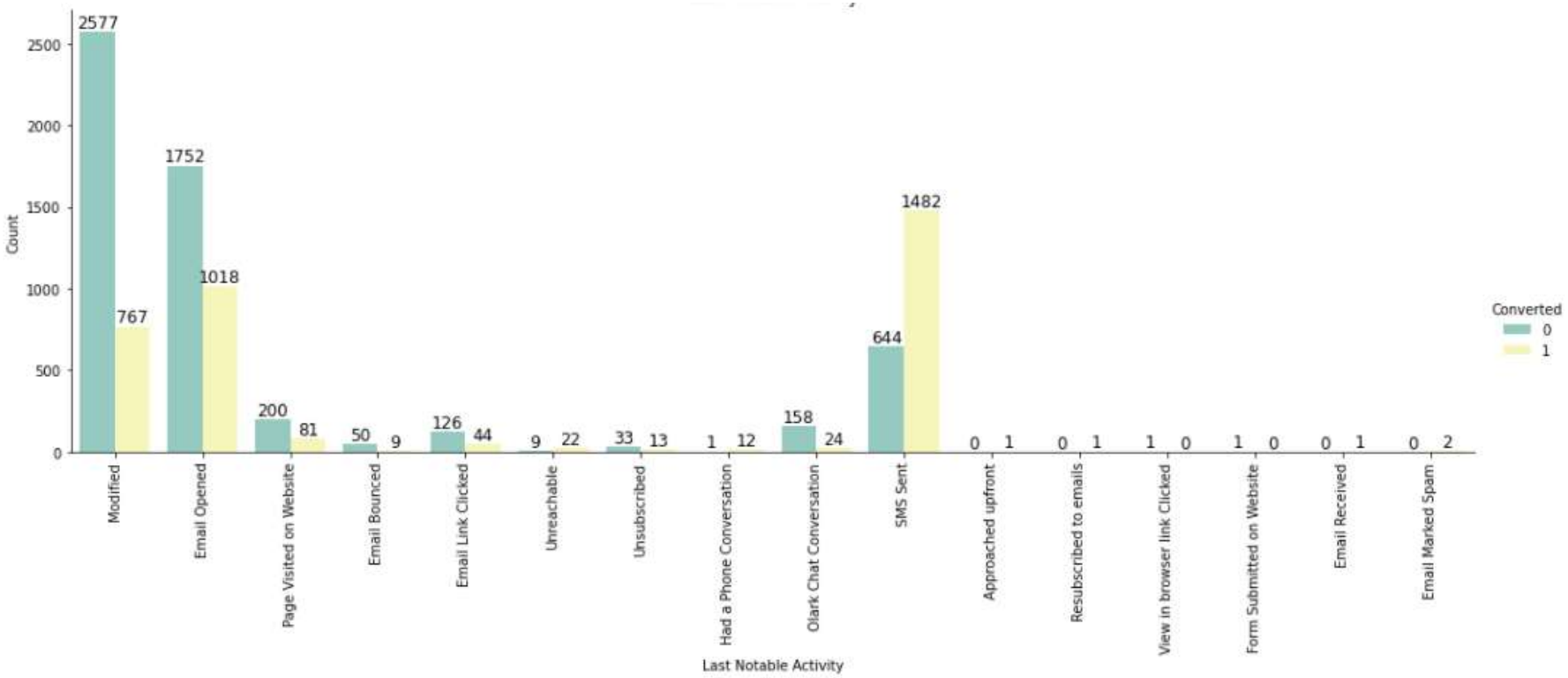


More conversion happened with people who are unemployed. It can also be noticed that most working professional leads converted. - **Occupation**

As per the below graph, **last activity** value of 'SMS sent' had the most conversion rate.



It can be noticed that the conversion rate is high for "SMS Sent" and also a high number of people who ‘opened the emails’ got converted in the **Last Notable Activity** column



- **Splitting the Dataset**

  - Feature Scaling of Numeric data

  - Splitting data into train and test data, by choosing the 70 – 30 ratio

- **Model Building**

  - Feature Selection using RFE, with 15 variables as output

  - Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5

  - Determine the optimal model using Logistic Regression

  - Predict on Test dataset

  - Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

- **Final Evaluation**

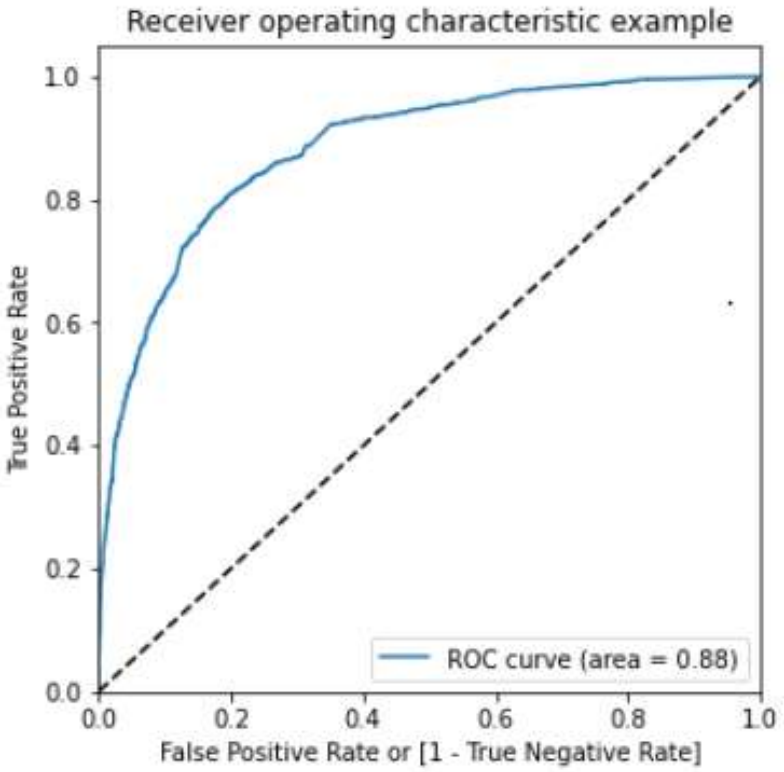
  - Determine the lead score and check if target final predictions amounts to 80% conversion rate.

  - Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Variables Impacting the Conversion rate

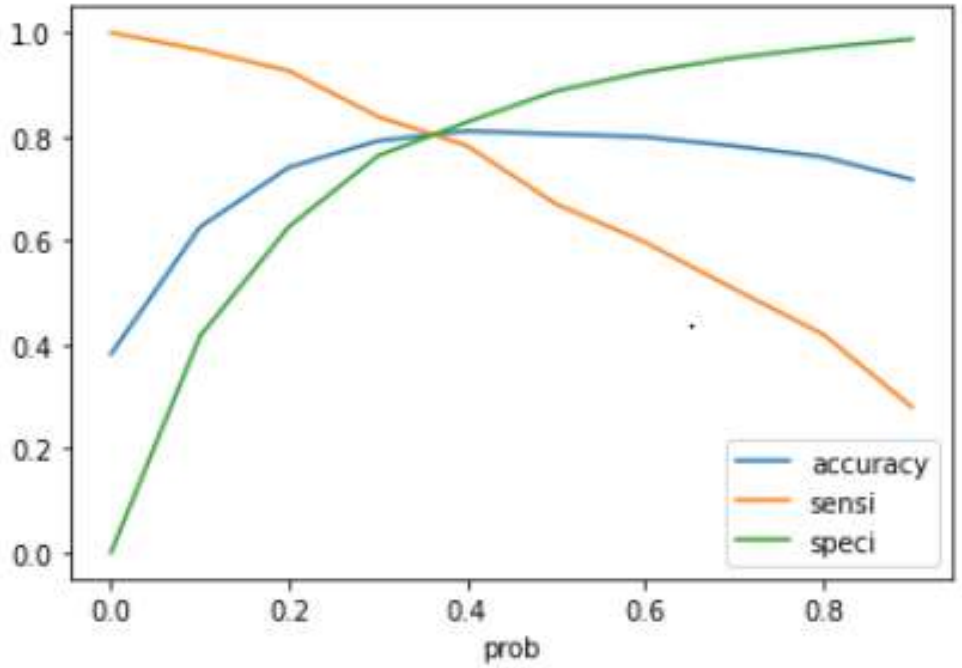
- LeadOrigin_Lead Add Form
- LeadSource_Welingak Website
- CurrentOccupation_Working Professional
- LastActivity_SMS Sent
- Total Time Spent on Website
- LastNotableActivity_Modified
- LeadSource_Direct Traffic
- LeadSource_Organic Search
- CurrentOccupation_Info not available
- LastActivity_Email Bounced

ROC Curve



Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.35 based on Accuracy, Sensitivity and Specificity



The calculated Metrics of Train Data set are as:

Accuracy : 80.2%
Sensitivity : 81.0%
Specificity : 79.76%
Precision: 78.63%
Recall: 67.02%

Model Evaluation - Sensitivity and Specificity on Test Data Set

The calculated Metrics of Test Data set are as:
Accuracy : 79.4%
Sensitivity : 81.4%
Specificity : 78.2%
Precision: 70.79%
Recall: 81.42%

Conclusion

The lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around **81%,** as expected by the CEO. Hence we conclude that the model is good

The top 3 variables that contribute for lead getting converted in the model are

1. Lead Add form from Lead Origin
2. Working Professionals in Current Occupation
3. Wellingak Website in Lead Source

We would advise 'X Education' to focus on the leads through the above three sources (along with the other sources identified in slide 10), to increase their conversion rate.