Lead Score Case Study

Submitted by :

Avinash Jijynasu

Vinay Kumar Mishra

Latheef D

Public

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%.

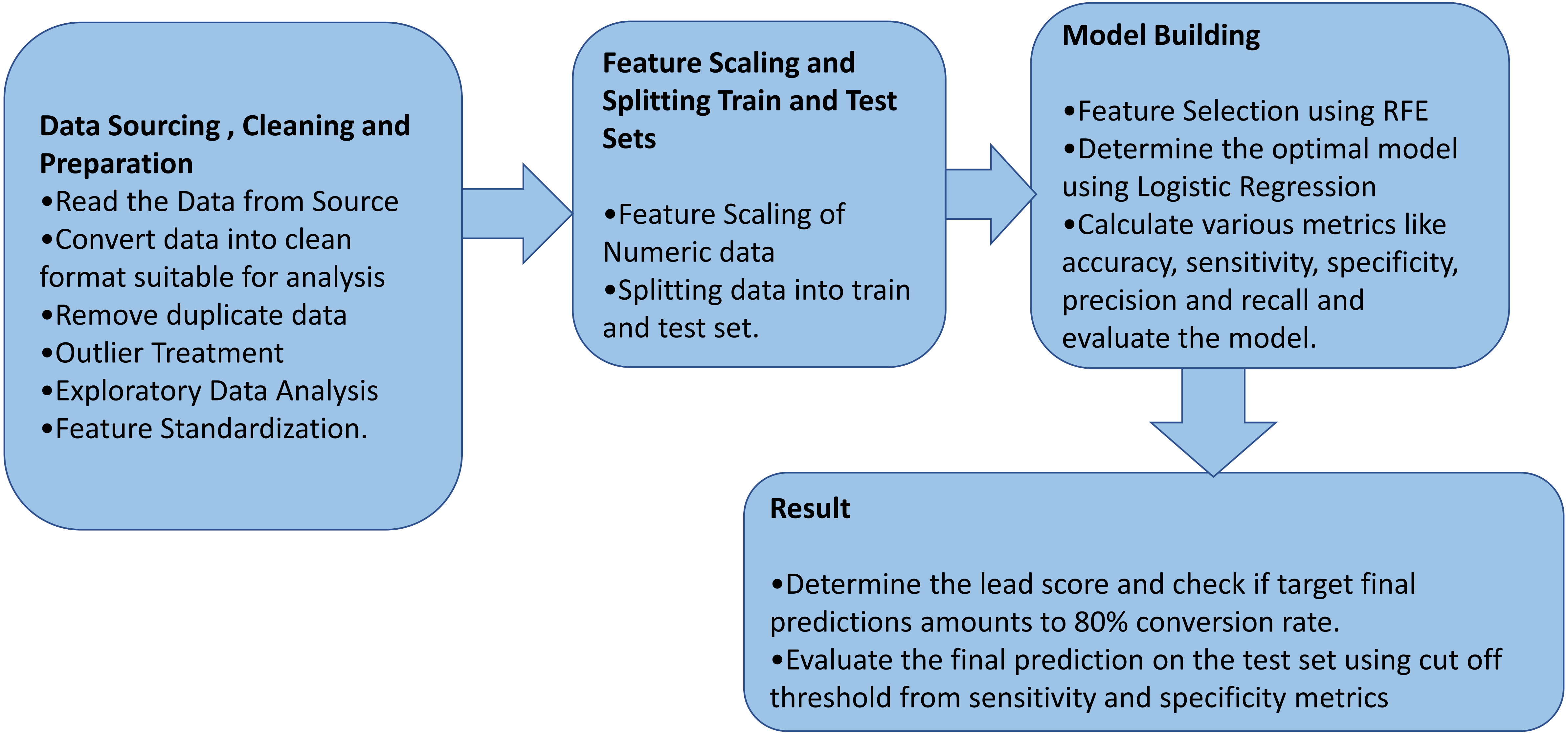
Public

Strategy:

- ✓ Source the data for analysis
- ✓ Clean and prepare 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 -Specificity and Sensitivity or Precision and Recall.
- ✓ Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Public

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 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

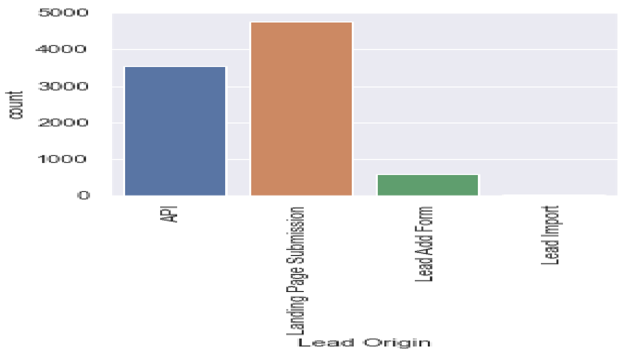
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

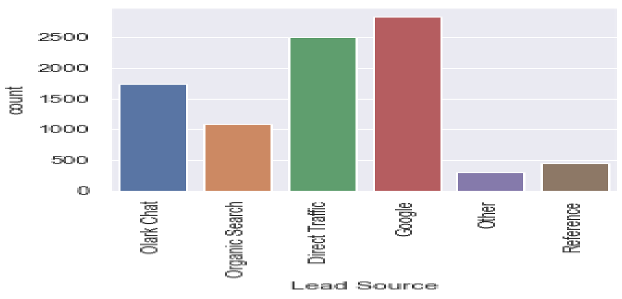
Public

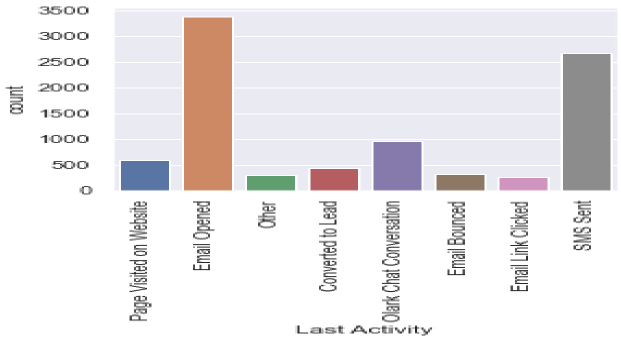
Exploratory Data Analysis

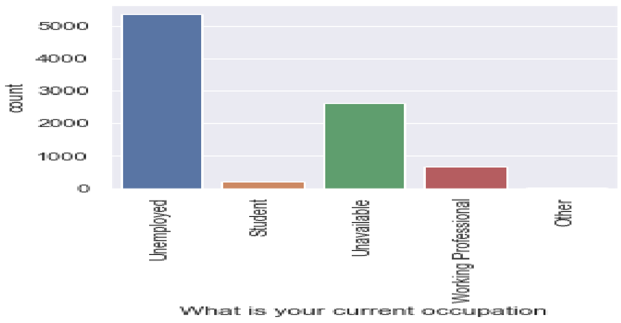
Univariate Analysis (Categorical)

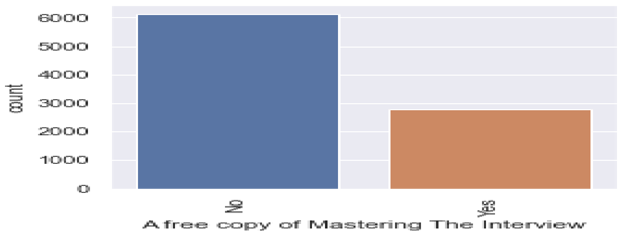
- ✓ API & Landing Page Submission are two major contributor of Lead Origin.
- ✓ Direct Traffic and Google are the two main source of Leads.
- ✓ Email Opened and SMS Sent are the major Last Activity.
- ✓ Most of the lead generated by Unemployed.
- ✓ Majority don't want a free copy of Mastering The Interview.

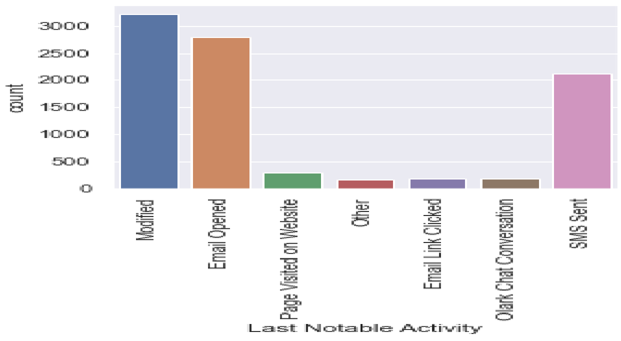










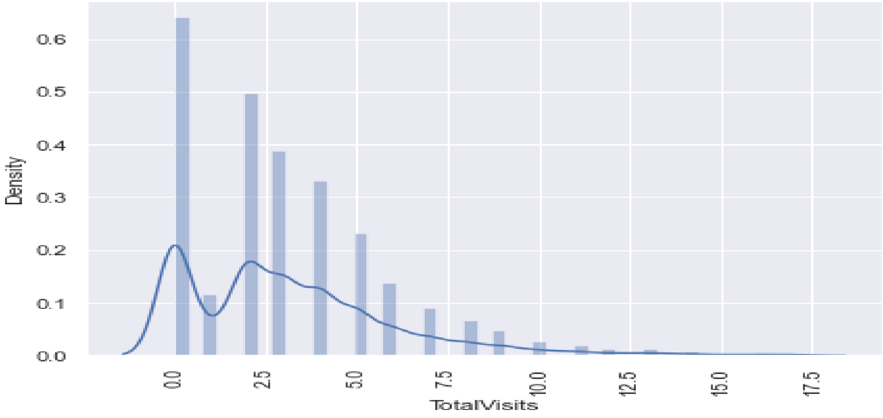


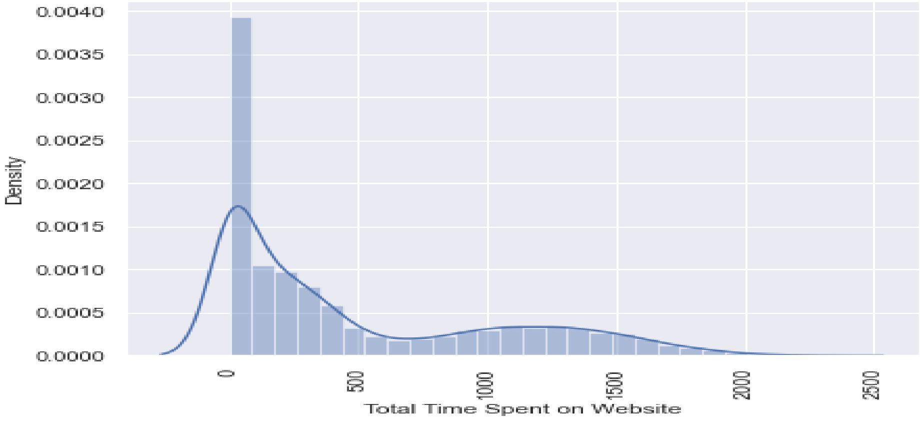
Public

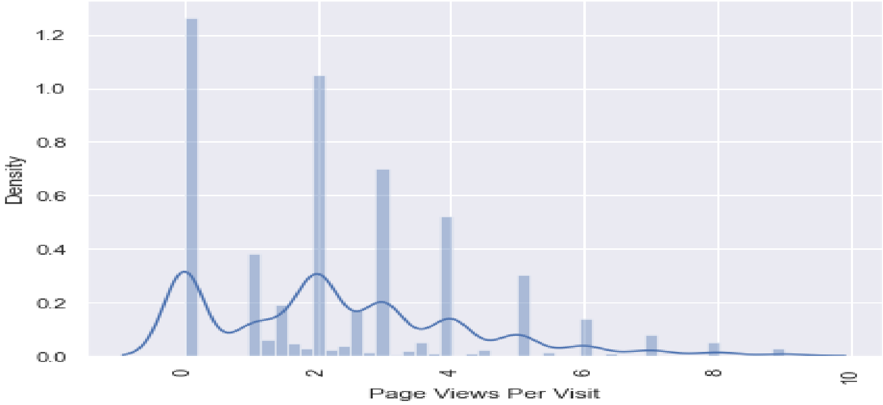
Exploratory Data Analysis

Univariate Analysis(Continuous)

- ✓ None of the continuous variables are normally distributed.
- ✓ Outliers' presence are not there.
- ✓ Totalvisits values are between 0-17, Total Time Spent on Website values are between 0-2500 and Page Views Per Visits values are between 0-10





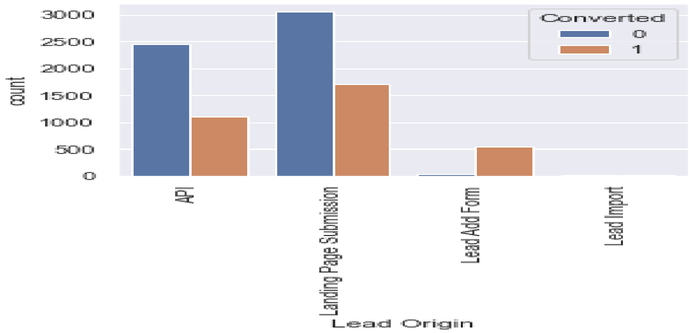


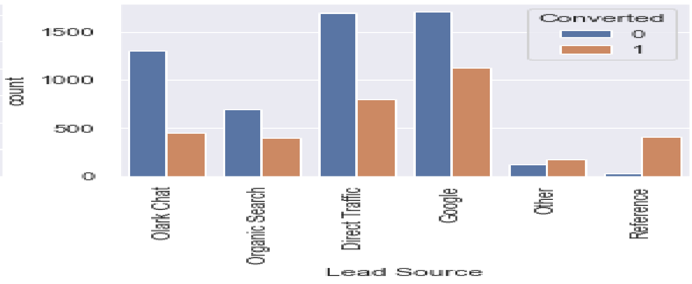
Public

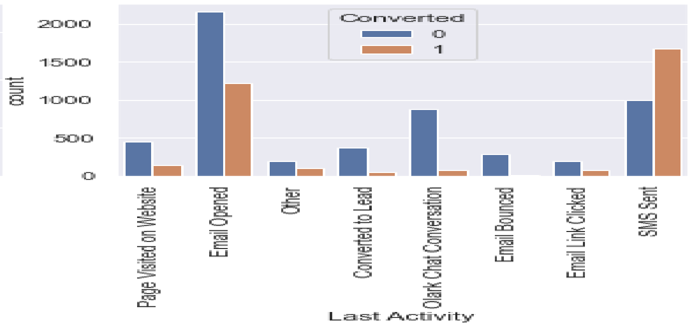
Exploratory Data Analysis

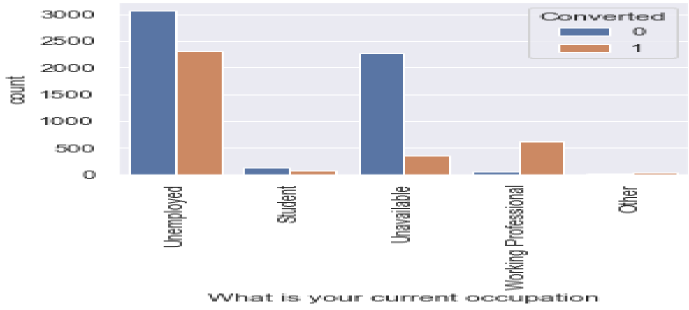
Bivariate Analysis

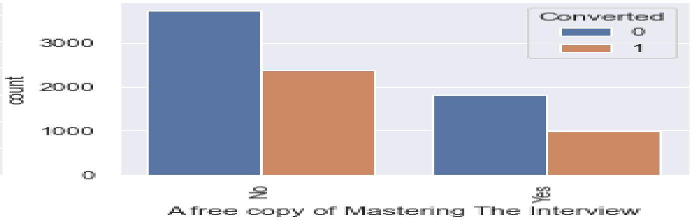
- ✓ Lead Origin : Hot leads are more in Landing Page Submission, API and Lead Add Form.
- ✓ Lead Source: Hot leads are higher in Direct Traffic and Google.
- ✓ Last Activity: Hot leads are higher in SMS Sent and EMAIL Opened.
- ✓ What is your current occupation: Hot leads are mostly generated by Unemployed and Working Professional.
- ✓ A free copy of Mastering The Interview: Hot leads are more with answer No.
- ✓ Last Notable Activity: Similar to Last Activity.

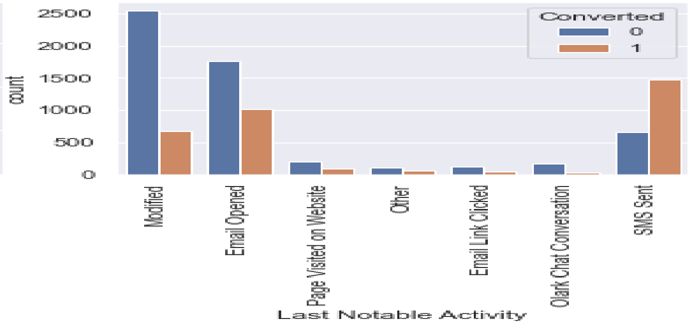












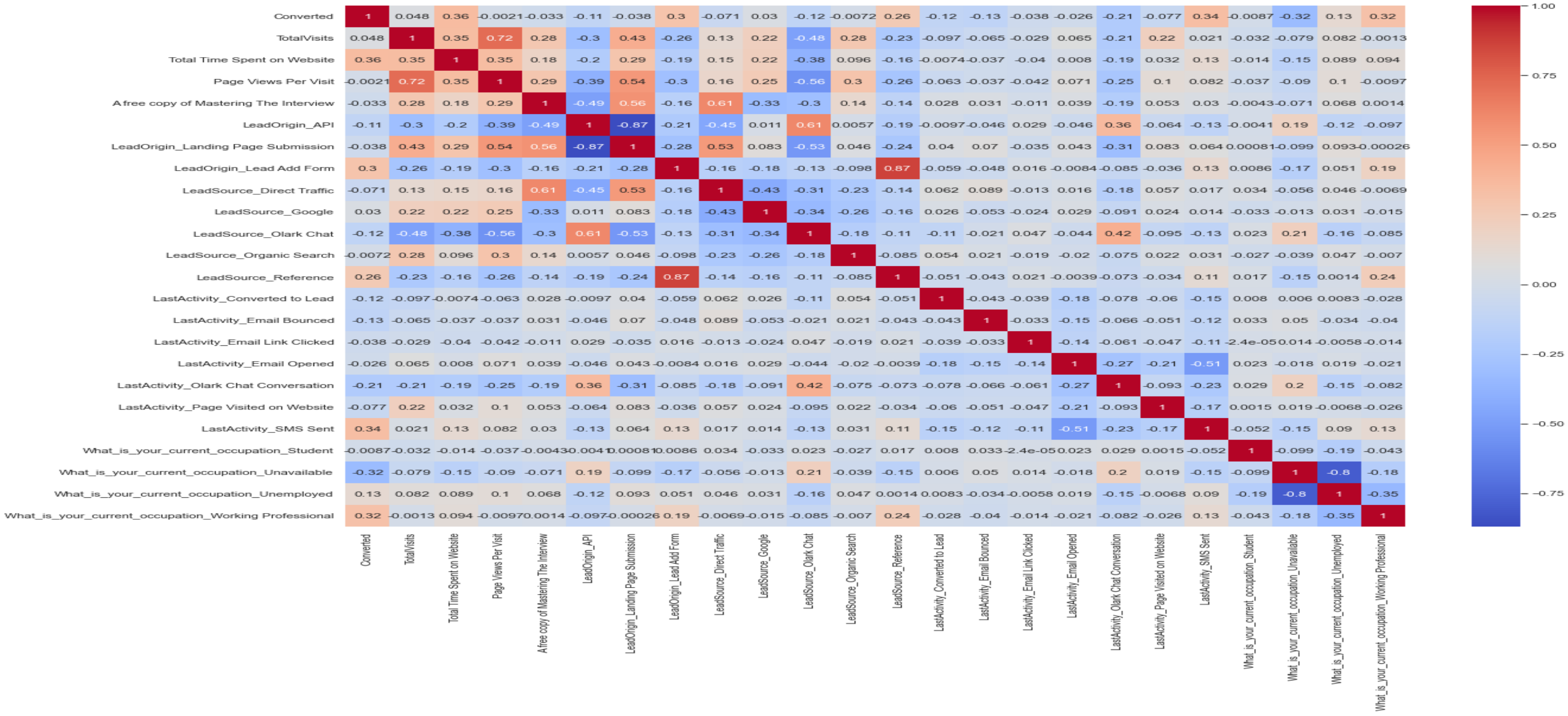
Public

Exploratory Data Analysis

Correlations

✓ From the heatmap it is clear that some variables are highly correlated.

✓ We have decided based on RFE to drop the correlated column



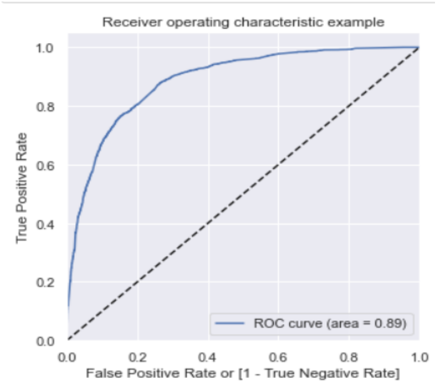
Model Evaluation

Plotting the An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

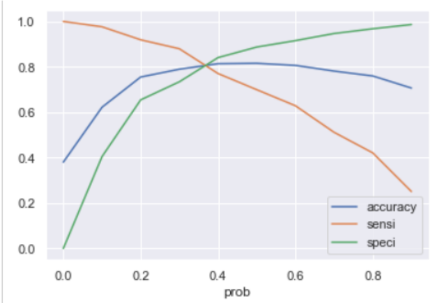


Inferences:

•From graph, ROC Curve area is 0.89, which indicates that the model is good and not overfitting.

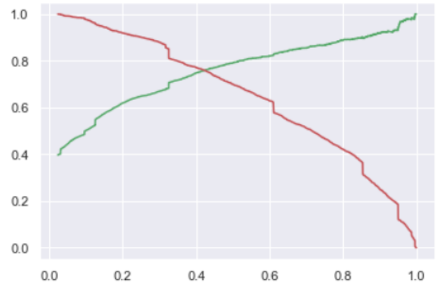
Finding Optimal Cutoff Point

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity



- From the curve above we can take 0.32 as optimum final cutoff.

Precision and recall tradeoff



- From the above curve we can see that precision & recall intersects at 0.41
- From the graph it is also clear that for having Recall >= 80% we have to keep cutoff <=0.32

Public

Conclusion :-

After analyzing the dataset and model building using logistic regression, company should focus on following features to increase the lead conversion

- Total Time Spent on Website.
- When 'Lead Origin’ is Lead Add Form
- When 'What is your current occupation’ is Working Professional
- When 'Lead Source’ is Olark Chat
- Total Visits.
- When 'Last Activity’ is SMS Sent

Public

Thank You

Public