Lead Score Case Study Analysis

Submitted By:

Kishor Acharya

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

•Although X Education company receives a lot of leads but the lead conversion rate or we can say paying customers is only 30 %.

•In order to accelerate the rate of lead conversion, company wishes to identify their most potential leads or ‘HOT leads’ which can be done through nurturing them well i.e. by educating them about the product and constantly communicating.

Data

You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

- Leads.csv
- Leads Data Dictionary

And a file named Assignment Subjective Questions.doc has been given to answer a few questions about the analysis.

Goals and Expected Results

- To help the organization determine their most potential leads or paying customers.

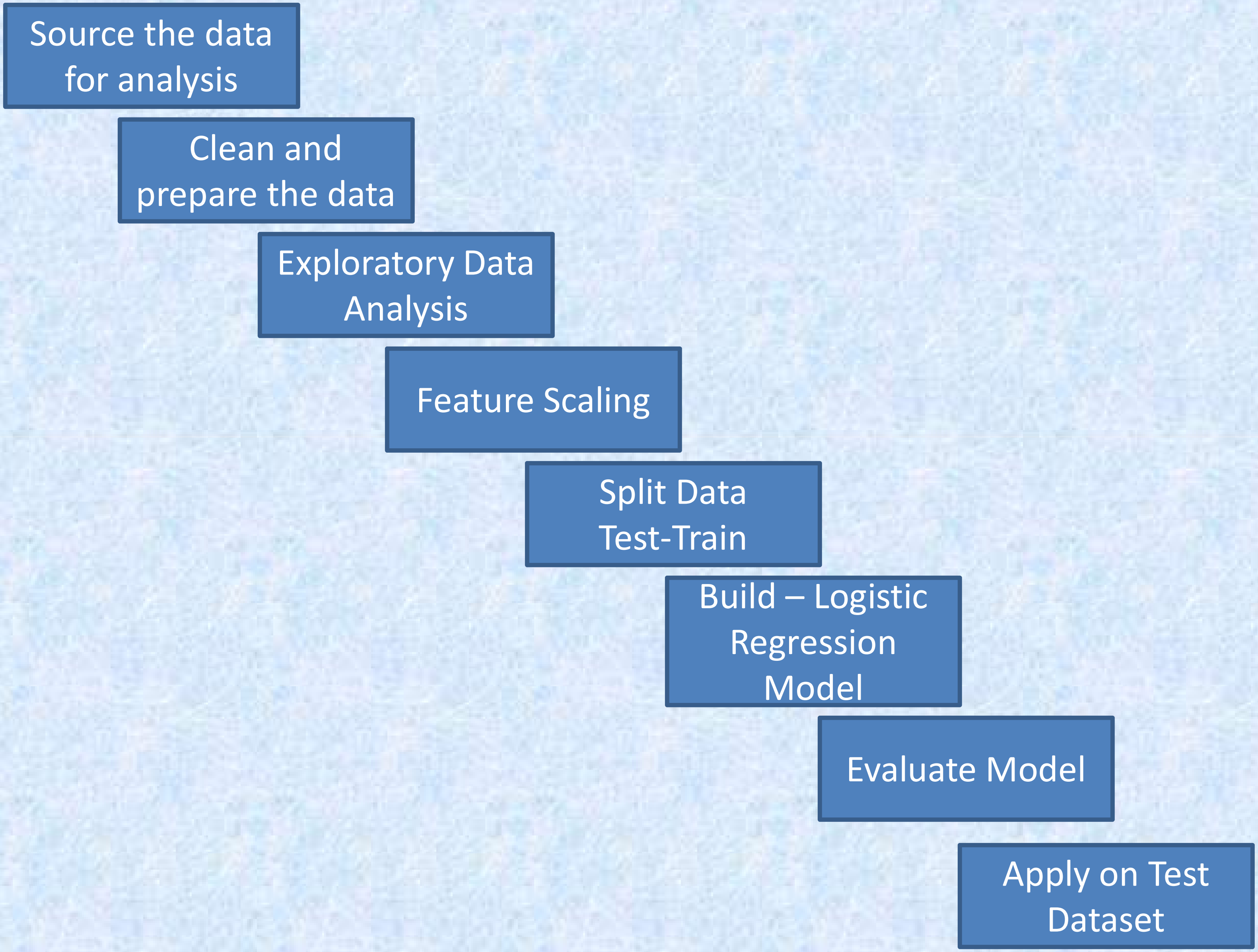
- Company requires a logistic regression model to be built that will assign a lead score to every lead bifurcating them as lead with higher lead score and a lower lead score as a lead having high score will have high conversion chance and a lead having low score will have low conversion chance.

- Company has a target of achieving the lead conversion rate of 80%.

Assumption

- **Leads.csv** is the dataset for past leads which will provide us all the information with Target Variable Target where 0 = Not Converted and 1=Converted
- Study of all variables provides the quantitative study of all variables given
- We have analyzed the data to understand which are converted leads and which are not converted leads.
- This study will help the CEO of X Education to understand the leads that are most likely to convert into paying customers.
- It could be a decisive factor for them if we loose a good customer which can be converted into a lead.

Strategy



Source the data for analysis

Clean and prepare the data

Exploratory Data Analysis

Feature Scaling

Split Data Test-Train

Build – Logistic Regression Model

Evaluate Model

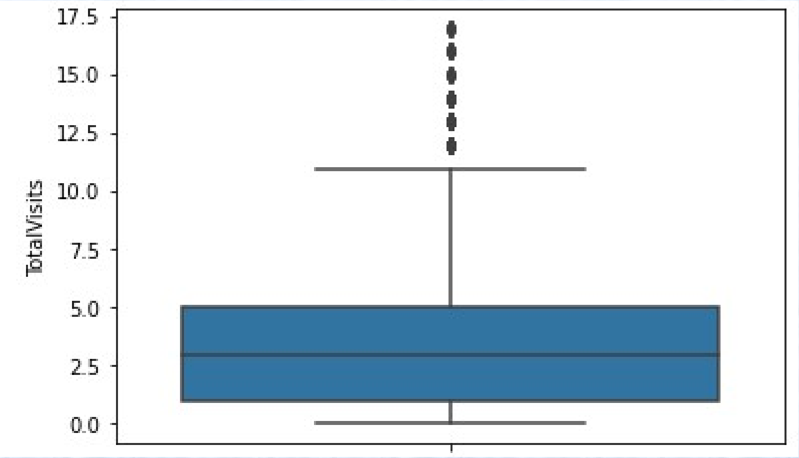
Apply on Test Dataset

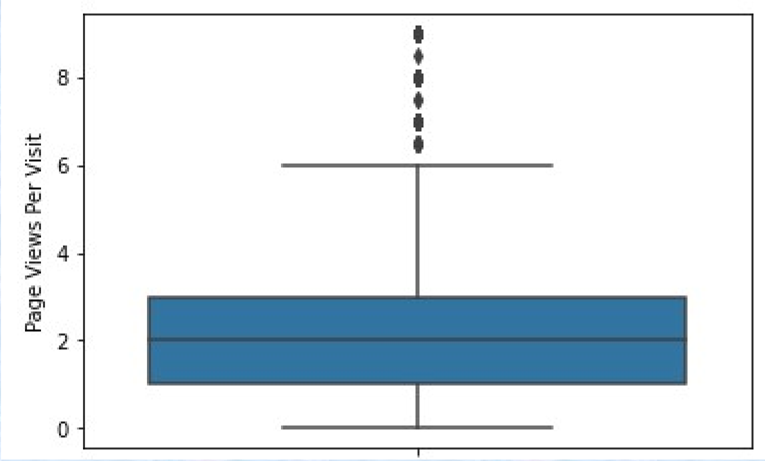
Approach

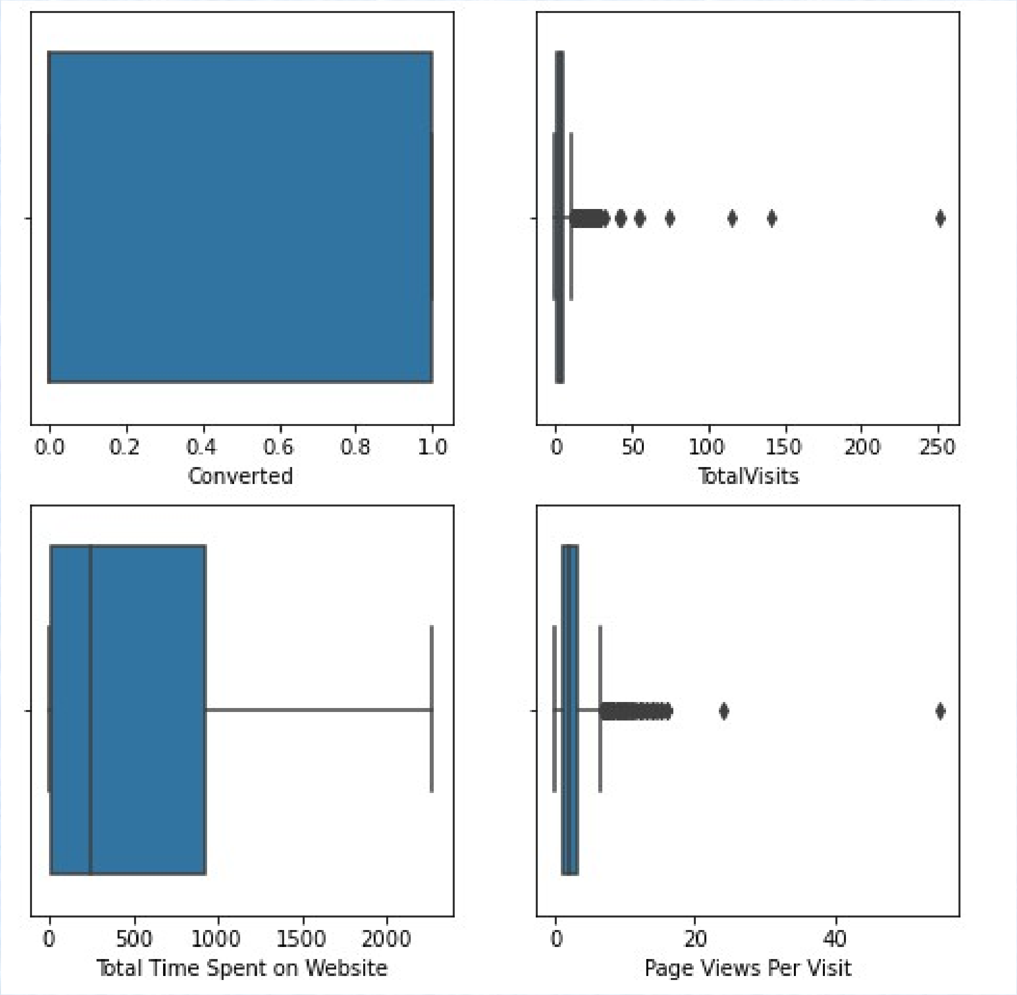
- First of all we have loaded data of current dataset
- Then we have cleaned it and categorized it
- Then we have prepared the data for column-wise analysis
- Impute the values which are NaN depending upon the requirement
- Dropping un-useful columns and Select value areas
- Treating outliers with IQR i.e. Inter Quartile Range Method
- Visualizing the data to understand reasons behind Converted Leads
- Dropping the Categorical columns which were skewing the data and not deriving anything meaningful
- Identify the correlation
- Create dummy variables and standardize the dataset
- Perform training on the training set and find the best model using evaluation metrics like Specificity and Sensitivity or Precision and Recall.
- Test and evaluate the same of Test Data set to prove the solution

Exploratory Data Analysis

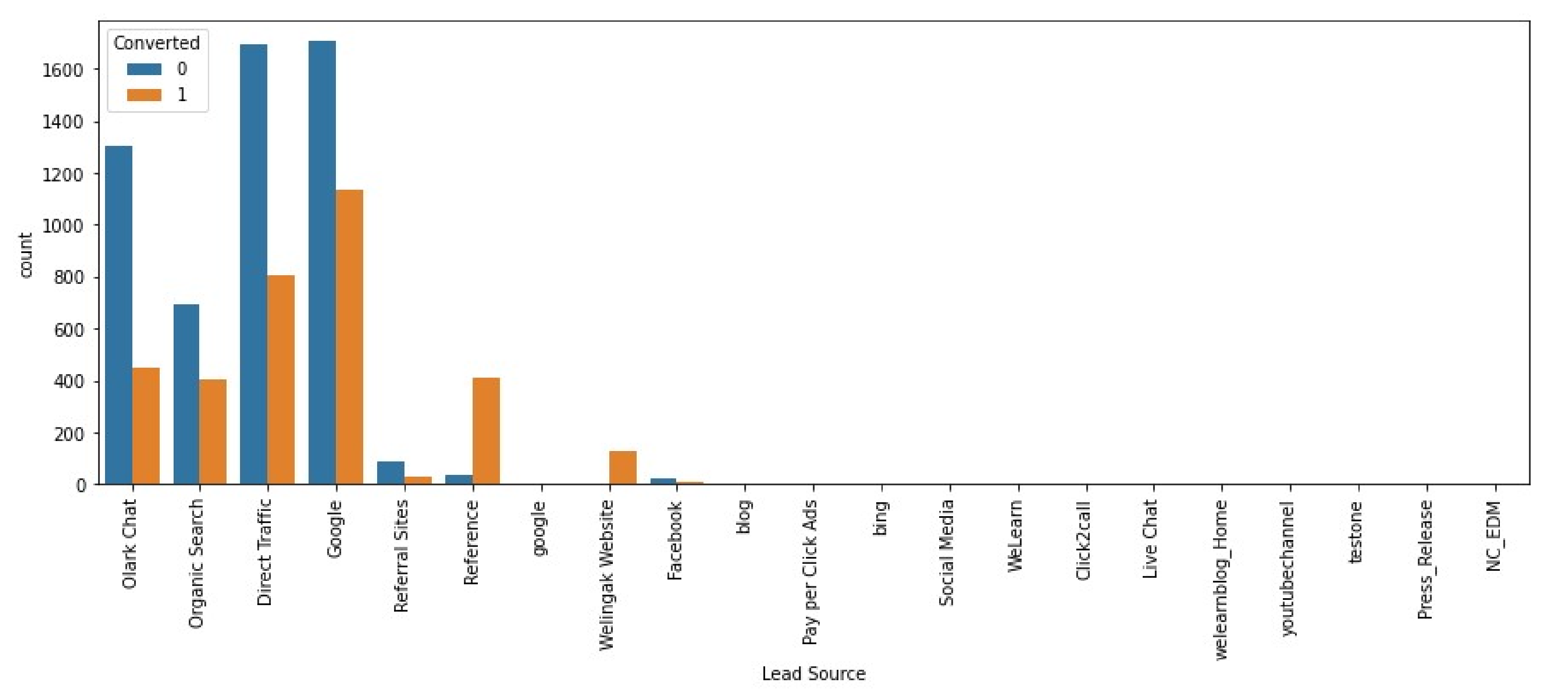
- Total visits have too many outliers in the data
- Page Views per visit also have outliers in the data
- This has been fixed by the IQR method



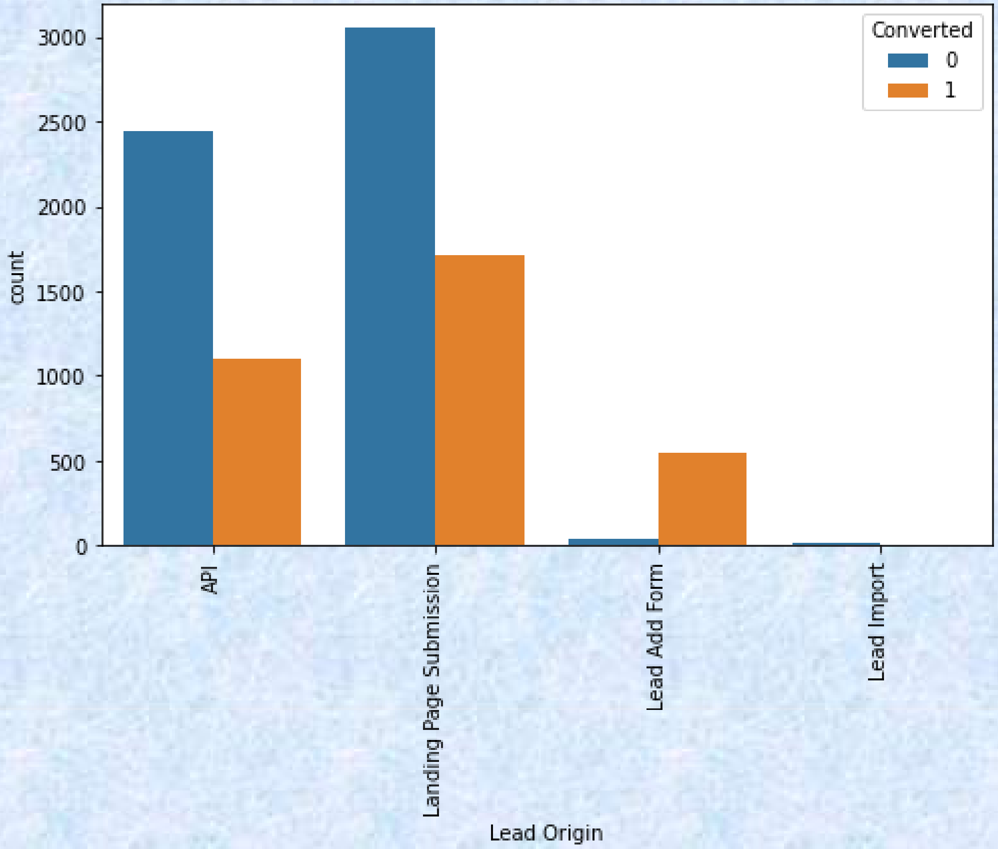


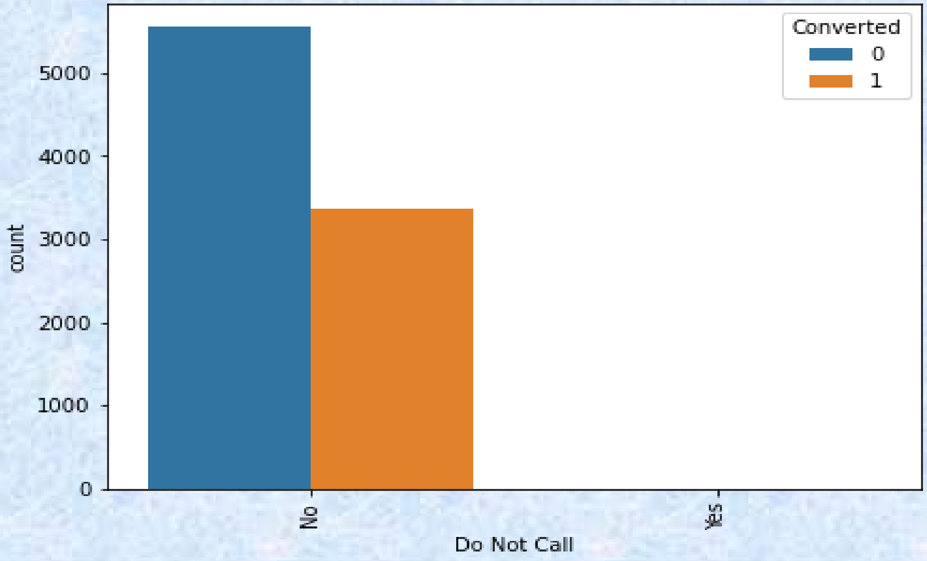


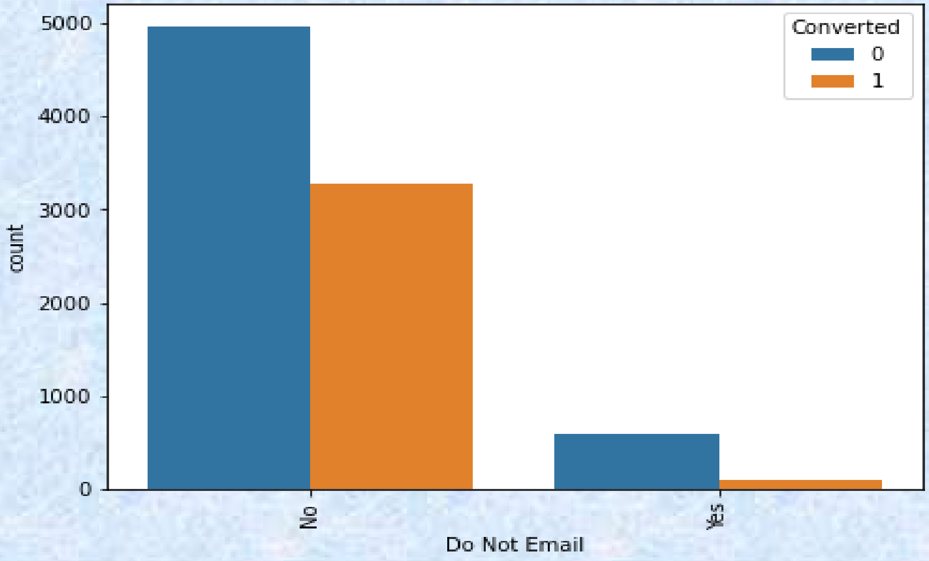
- Below figure explains the Distribution of Leads source with Converted
- Majority of the Leads were converted through Google, Olark Chat, Organic Search and Direct Traffic
- Rest of the population is of very low frequency



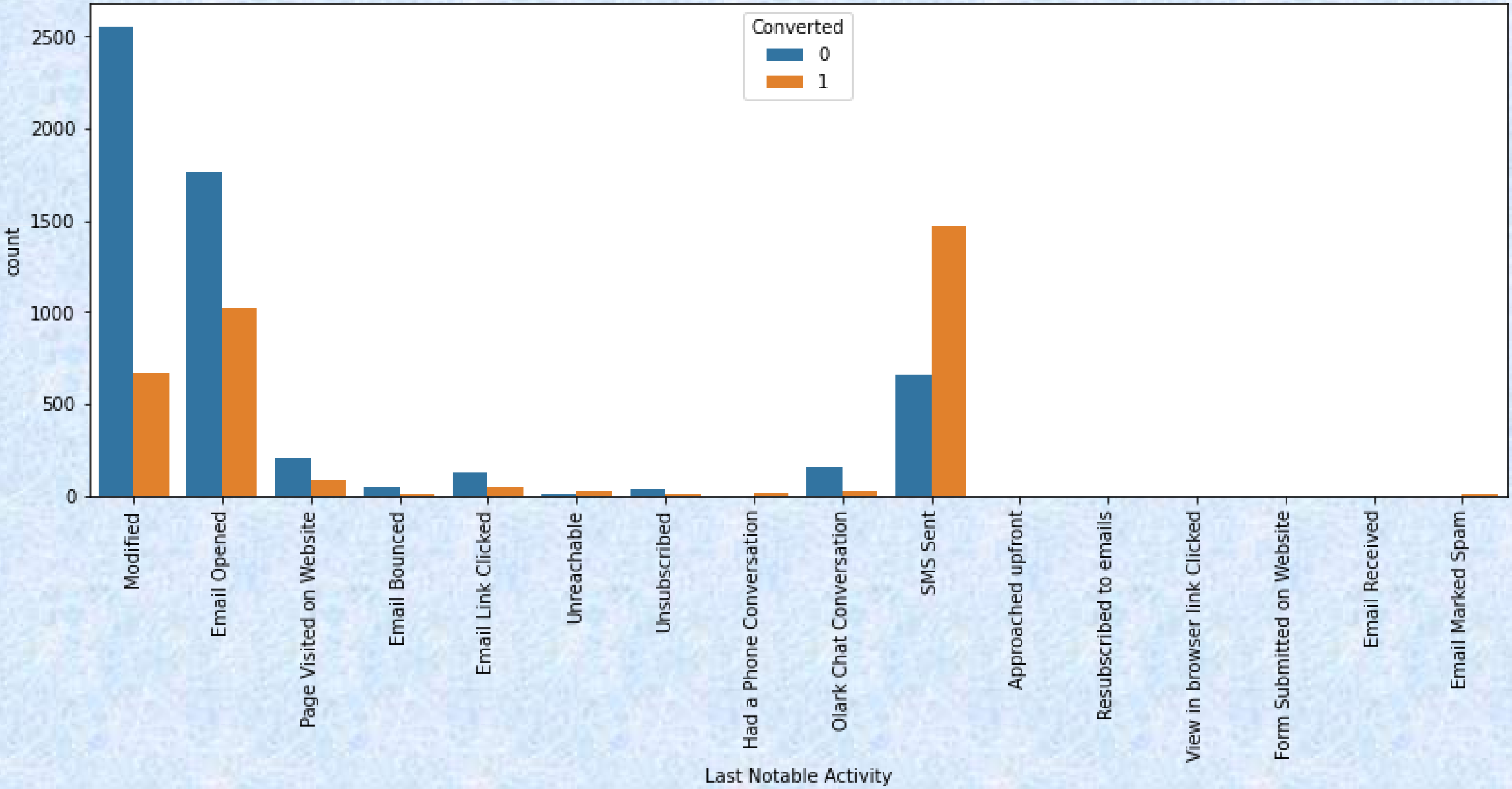
- The conversion rate for Landing page submission is higher than others
- Conversion rate is similar for both Do not call / email → NO Leads



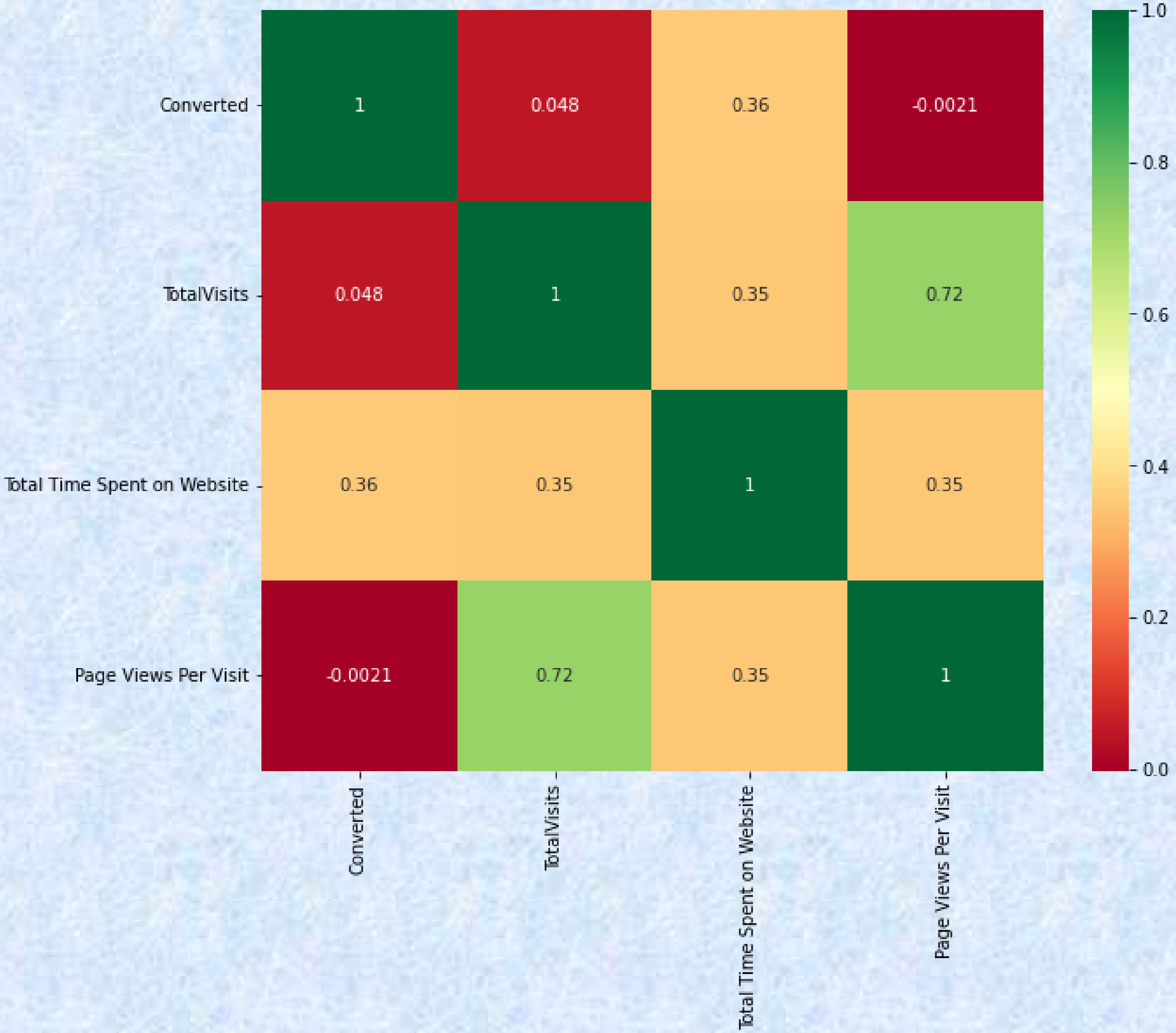




- Emailed opened and SMS sent Leads were more of seems to be converted here.



- Below is the correlation matrix which resembles highly correlated with Total Time spent on the website.



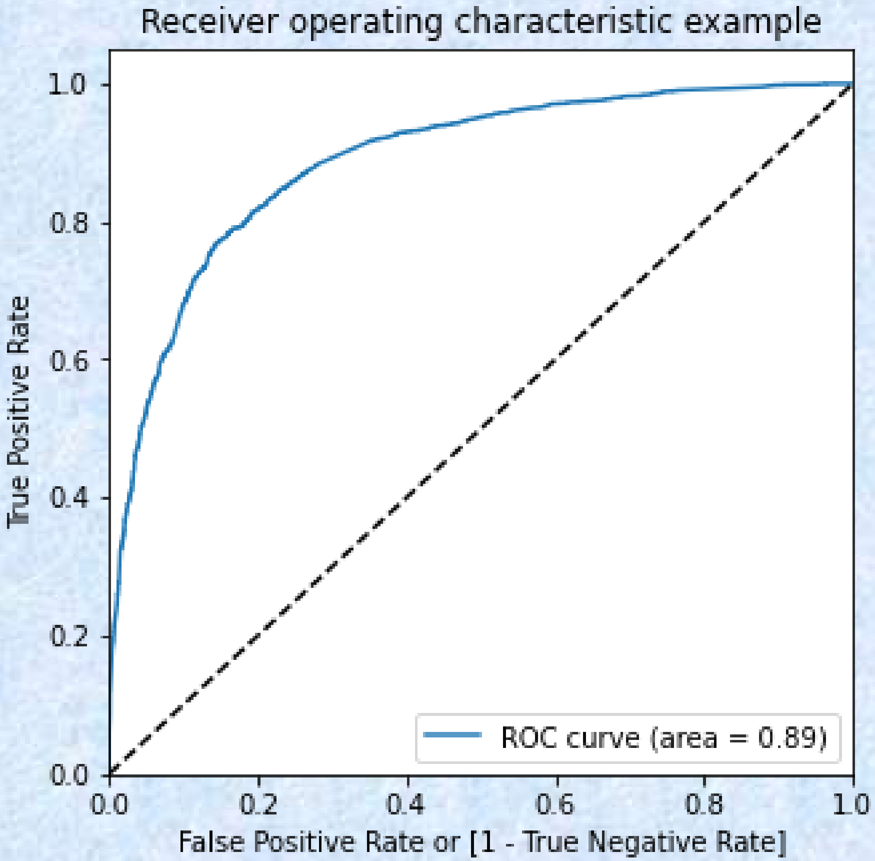
Variables Impacting Conversion Rate

- Total number of visits
- Time spent on website
- Lead Source
  - Google
  - Direct Traffic
  - Organic Search
  - Welingak Website
  - Reference
- Lead Origin
  - SMS
  - Olark Chat

- Variables after model finalisation
  - Last Notable Activity_Had a Phone Conversation
  - Lead Source_Welingak Website
  - Total Time Spent on Website
  - Lead Source_Reference

ROC Curve

- 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.
- Area under ROC curve is 0.89 out of 1 which indicates a good predictive model.
-

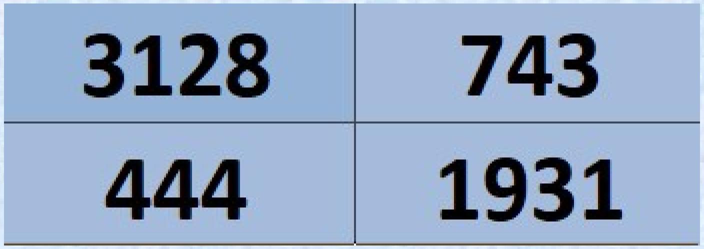


Model Evaluation - Sensitivity and Specificity on Train

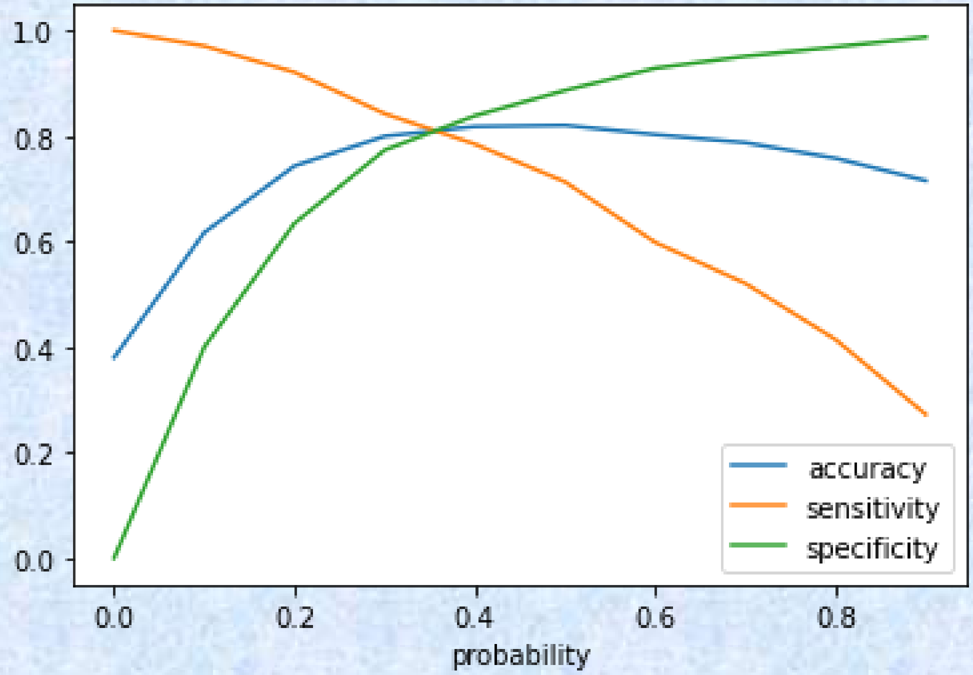
- The graph depicts an optimal cut-off of 0.35 for which we have calculated the below values on our final model which appears to be going with what's required.

- Accuracy = 80.99%
- Sensitivity = 81.30%
- Specificity = 80.80%
- Precision = 72.21%
- Recall = 81.30%

Confusion Matrix



3128	743
444	1931

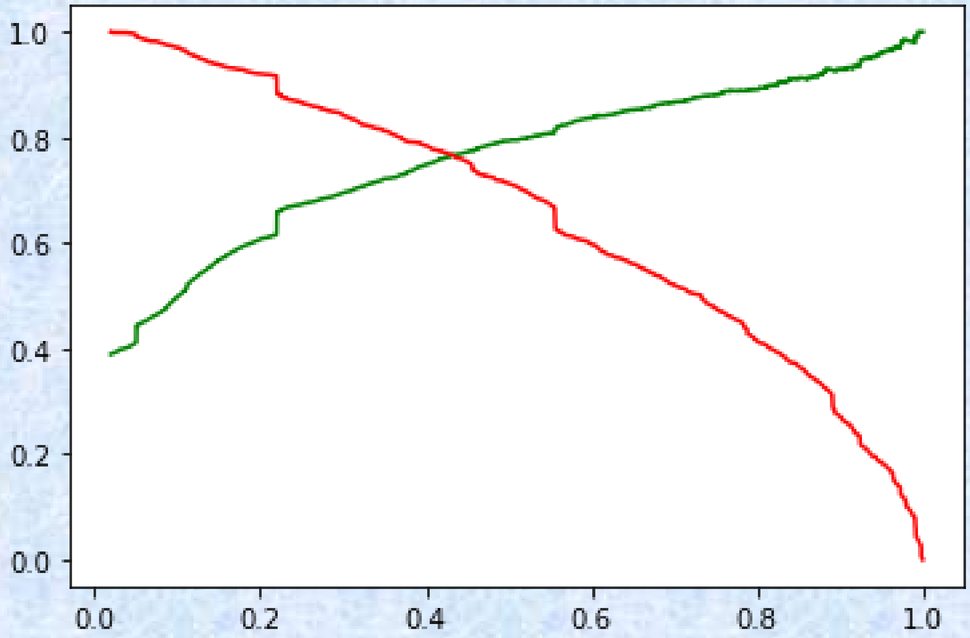


Model Evaluation - Sensitivity and Specificity on Test

- The graph depicts an optimal cut off of 0.42 based on Precision and Confusion Matrix Recall
- Accuracy = 80.02%
- Sensitivity = 80.18%
- Specificity = 79.92%
- Precision = 70.22%
- Recall = 80.18%

Confusion Matrix

1346	338
197	797



Conclusion & Recommendation

- Model is doing good as metrics are close to each other.
- ➡ Train Data Set
- 👈Accuracy : 81% 👈Sensitivity: 81.31% 👈Specificity : 80.81%
- ➡ Test Data Set
- 👈Accuracy : 80.02% 👈Sensitivity: 80.18% 👈Specificity : 79.93%

- We have achieved the goal of meeting 80% success rate.

- The top three variables which are important for getting a lead are Total Time on website, Lead Source and Total Visits and according to below the variables with negative coefficient can be worked upon for better results.
- All these are positive columns Last Notable Activity_Had a Phone Conversation, Lead Source_Welingak Website, Total Time Spent on Website, Lead Source_Reference, What is your current occupation_Working Professional, Last Notable Activity_Unreachable, Last Activity_Had a Phone Conversation, TotalVisits, Last Notable Activity_SMS Sent, Lead Source_Olark Chat, Last Activity_SMS Sent, Last Activity_Email Opened
- All negatives should be worked upon : Last Activity_Olark Chat Conversation , Lead Origin_Landing Page Submission , Specialization_Others , Last Activity_Email Bounced , Page Views Per Visit.