Lead Score Case Study

Submitted by :
Ashutosh singh
G Venkatesh

Lead Score Case Study for X Education

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%.
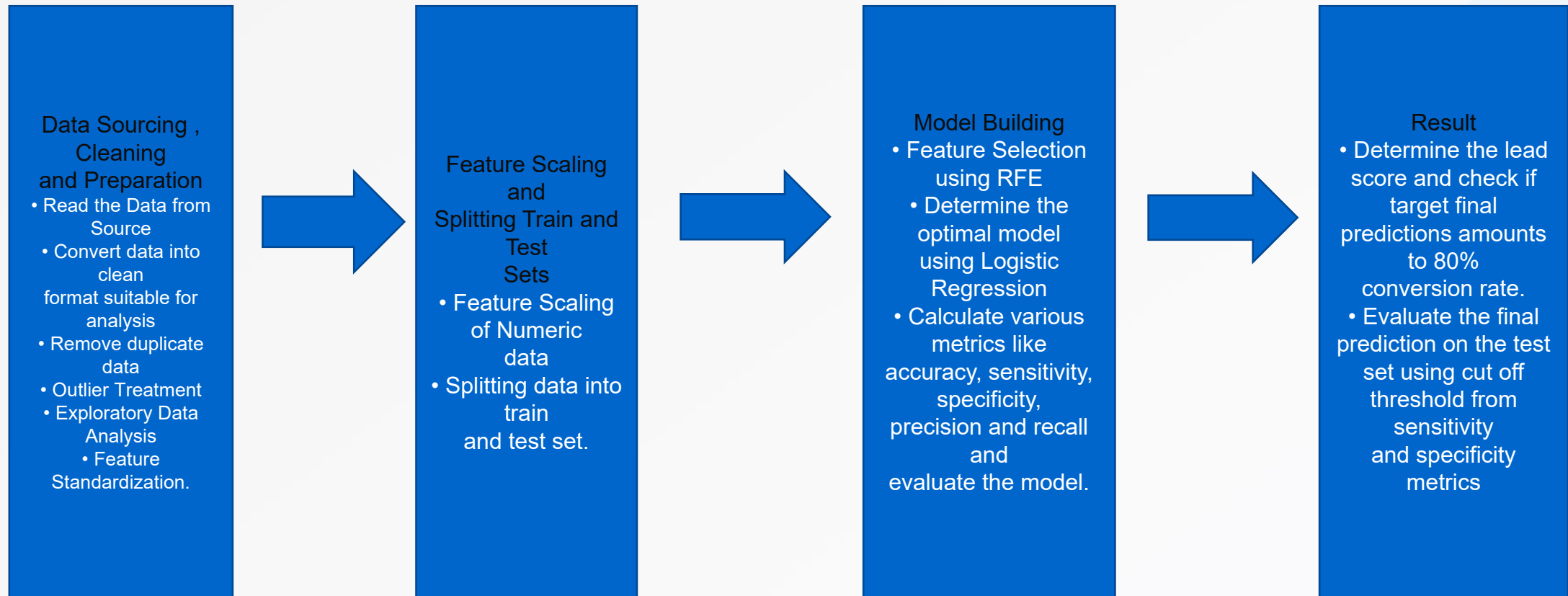
Business Goal:

- 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%.
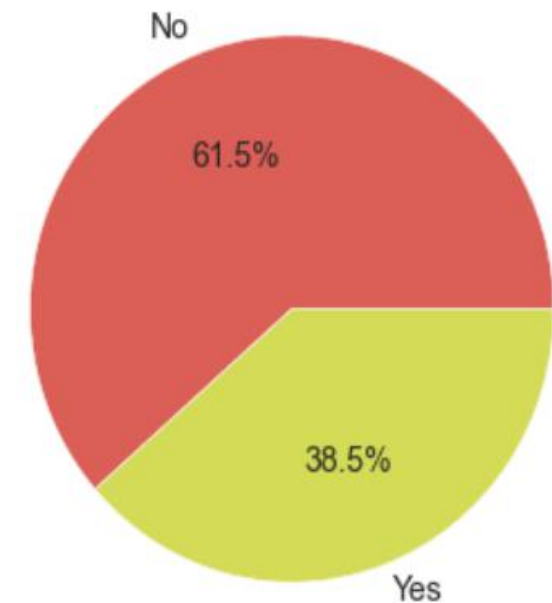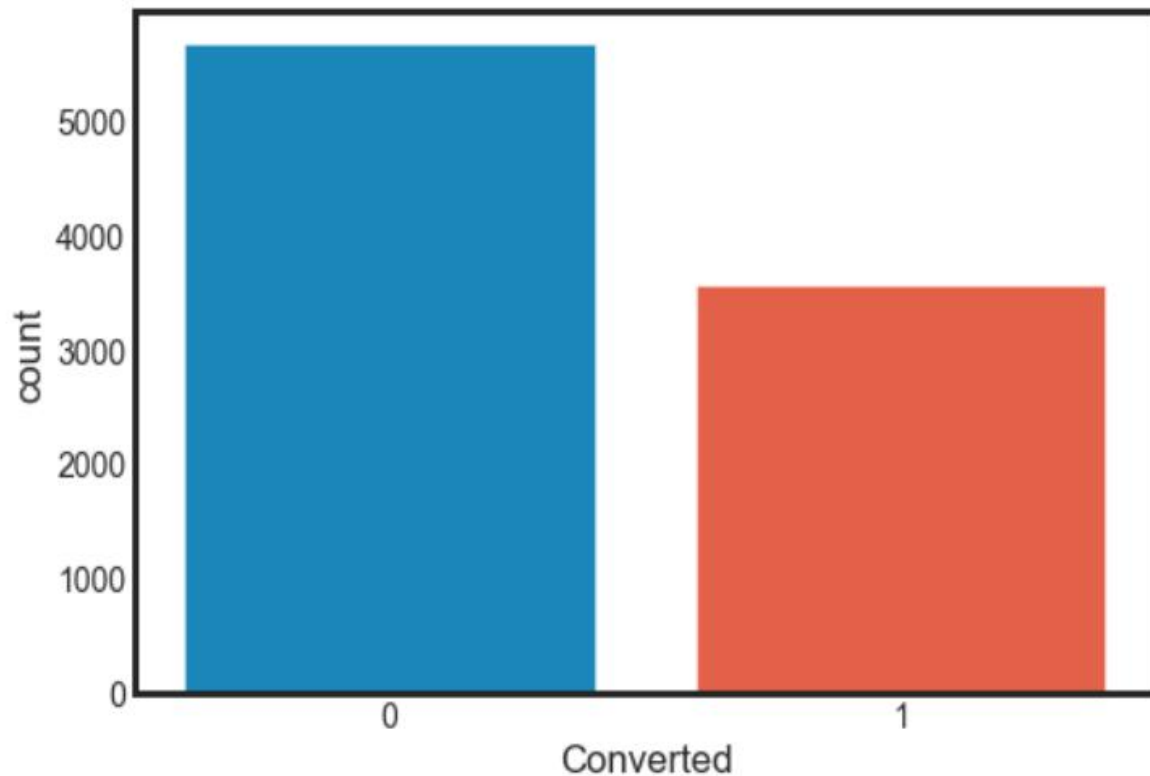
Strategy

➢ 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.

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

Exploratory Data Analysis

We have around 39% Conversion rate in Total
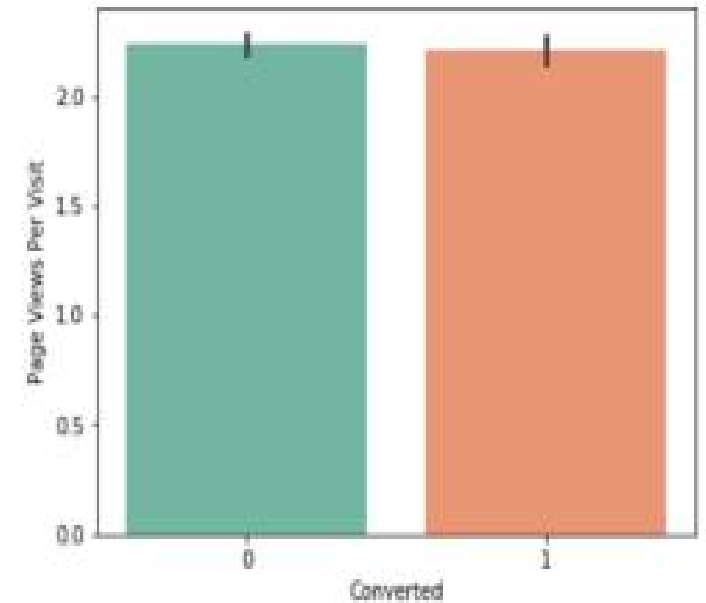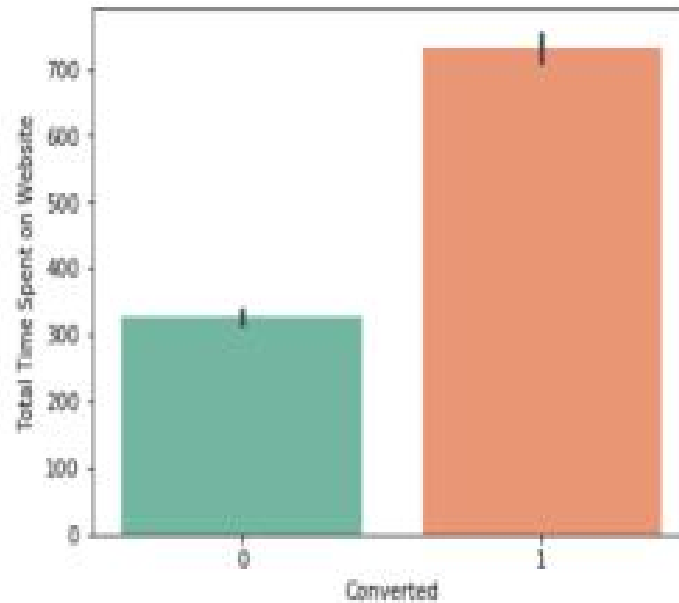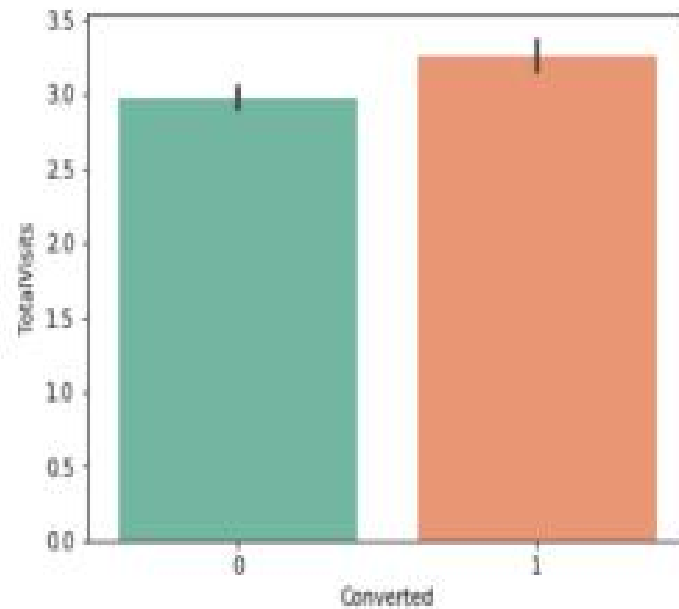
# In Lead Origin, maximum conversion happened from Landing Page Submission
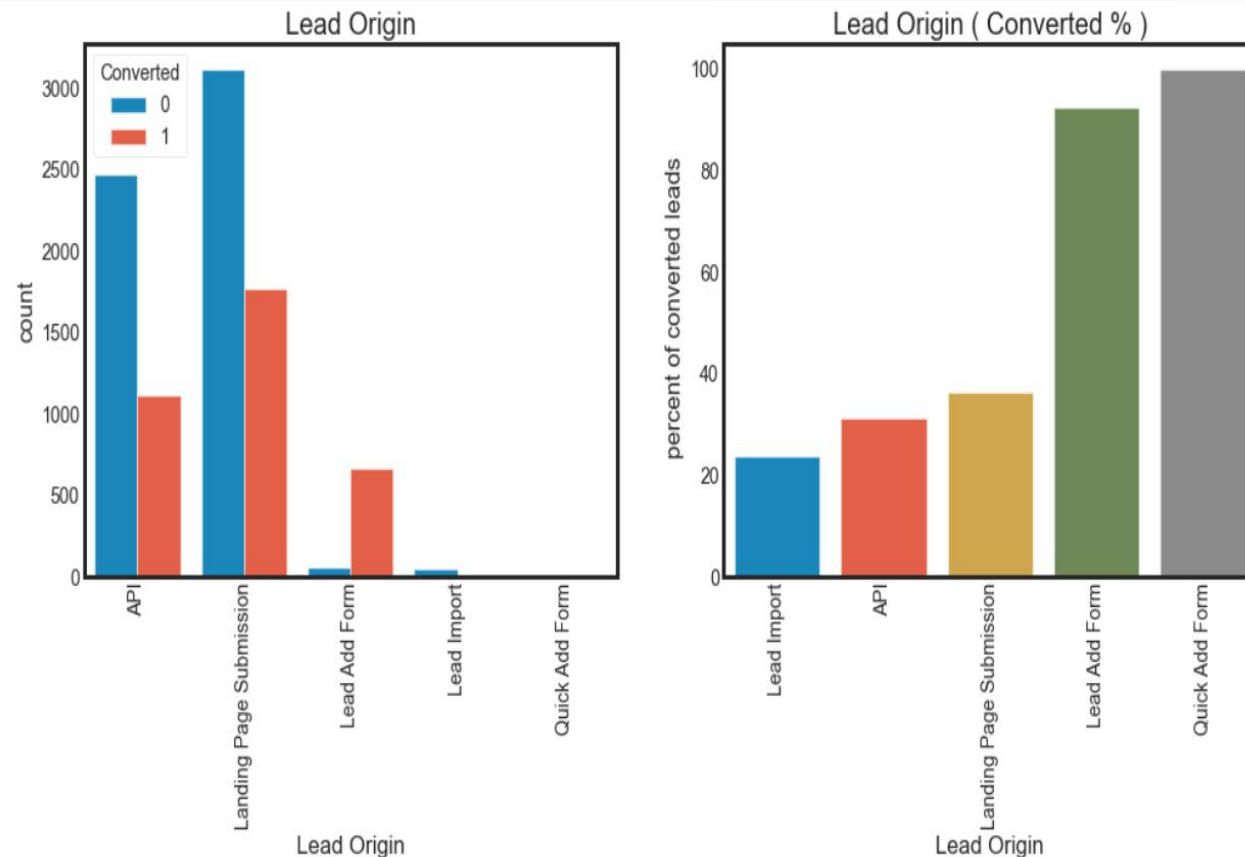
- Most Leads originated from submissions on the landing page and around 38% of those are converted followed by API, where around 30% are converted.
- Even though Lead Origins from Quick Add Form are 100% Converted, there was just 1 lead from that category. Leads from the Lead Add Form are the next highest conversions in this category at around 90% of 718 leads.
- Lead Import are very less in count and conversion rate is also the lowest
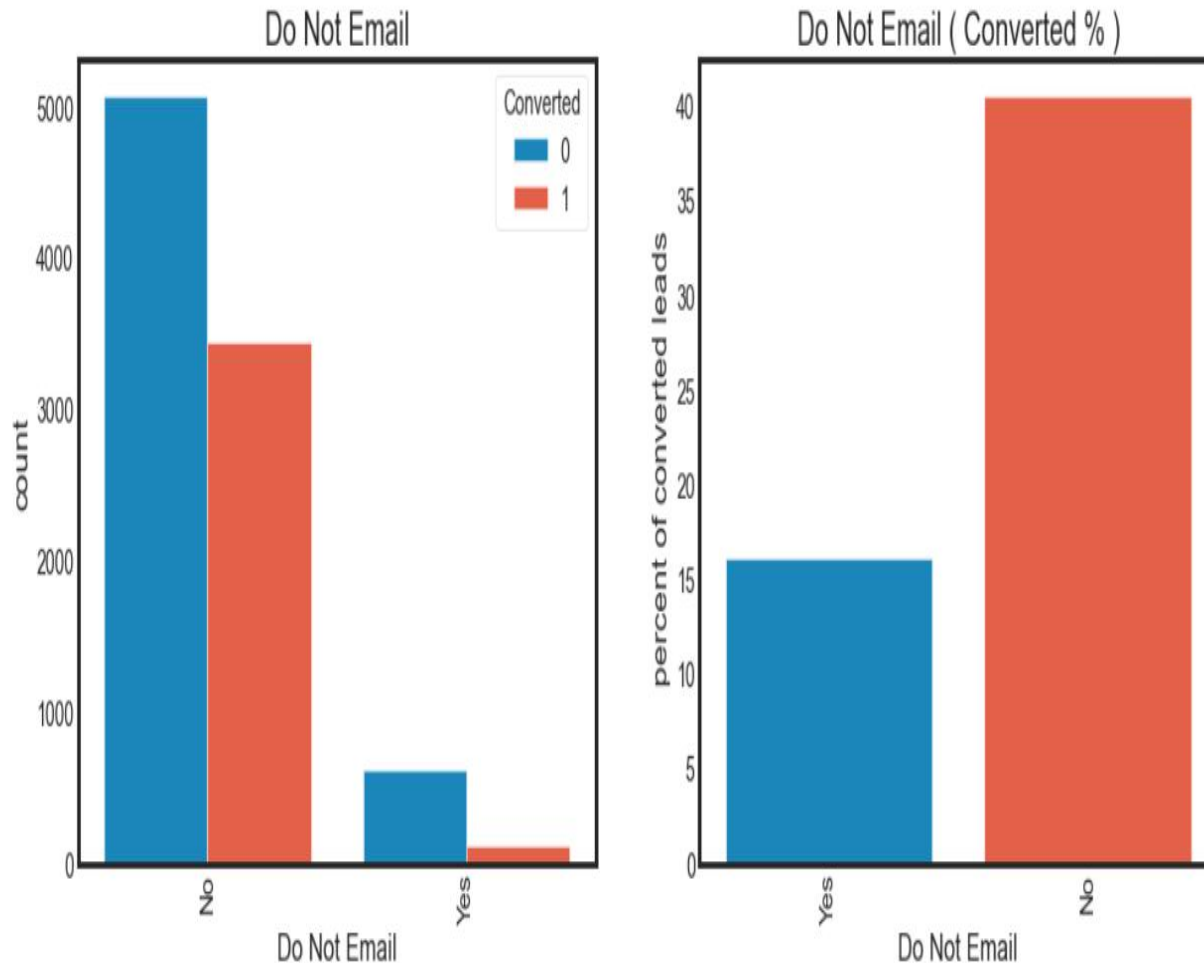
- Majority of the people are ok with receiving email (~92%)
- People who are ok with email has conversion rate of 40%
- People who have opted out of receive email has lower rate of conversion (only 15%)

Major conversion in the lead source is from Google
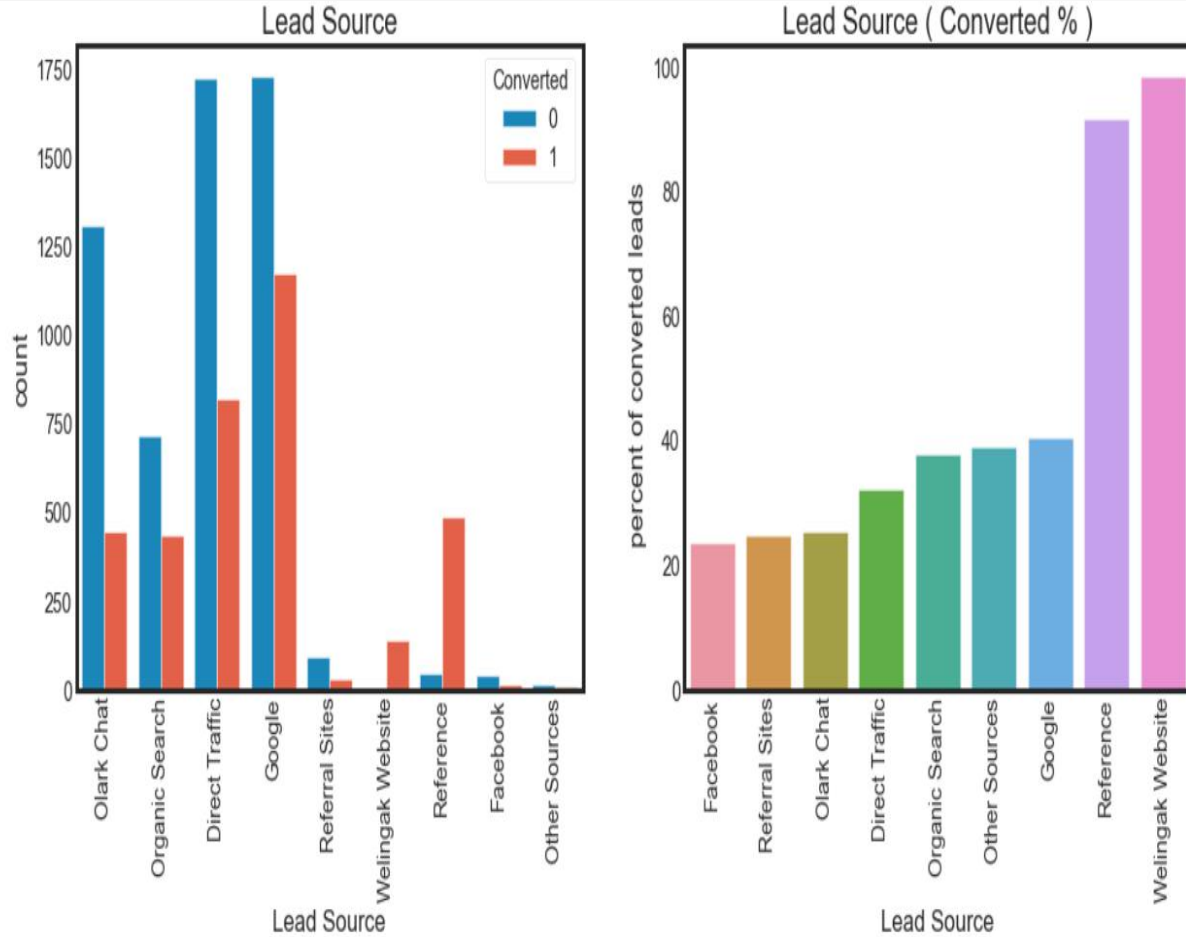
- The source of most leads was Google, and 40% of the leads converted, followed by Direct Traffic,Organic search and Olark chat where around 35%, 38% and 30% converted respectively.

- A lead that came from a reference has over 90% conversion from the total of 534.

- Welingak Website has almost 100% lead conversion rate. This option should be explored more to increase lead conversion

- To increase lead count, initiatives should be taken so already exitsing members increase their referrals

# Not much impact on conversion rates through Search
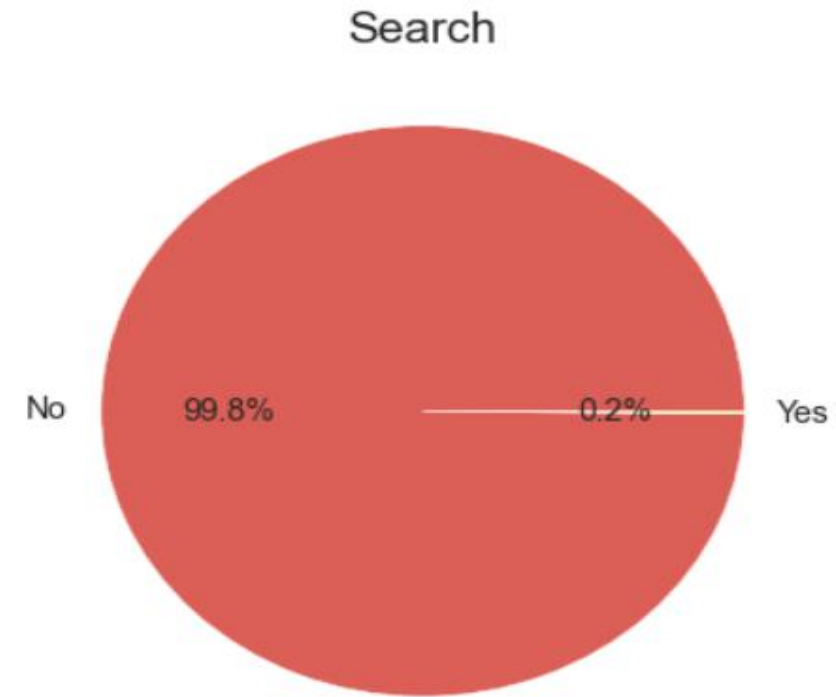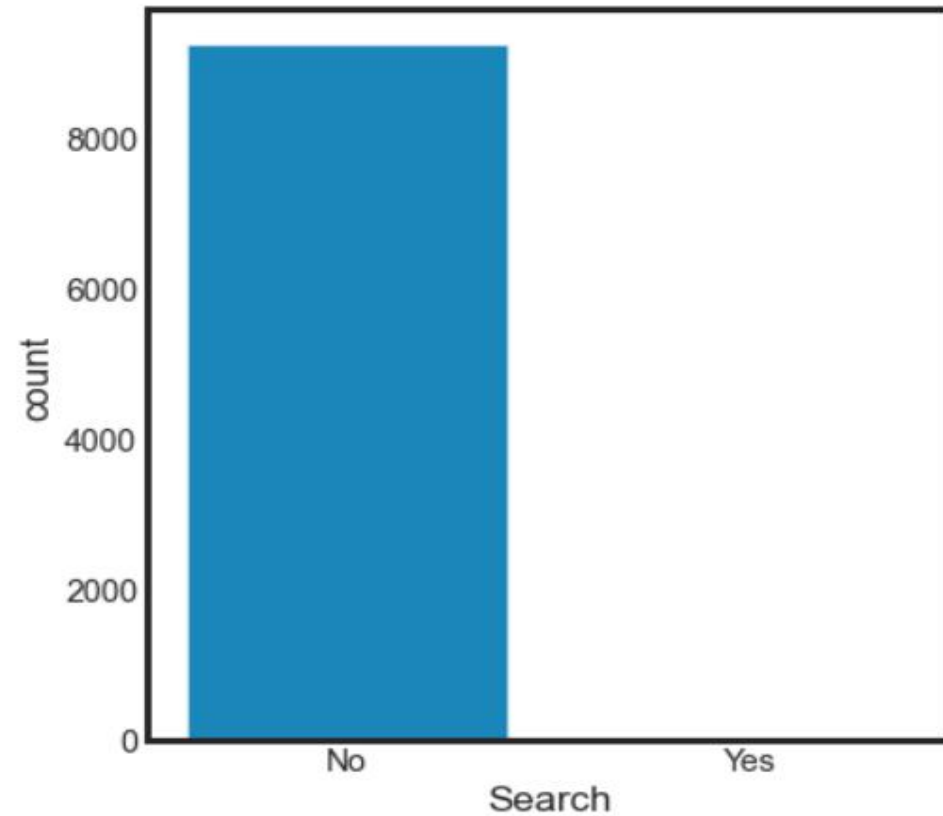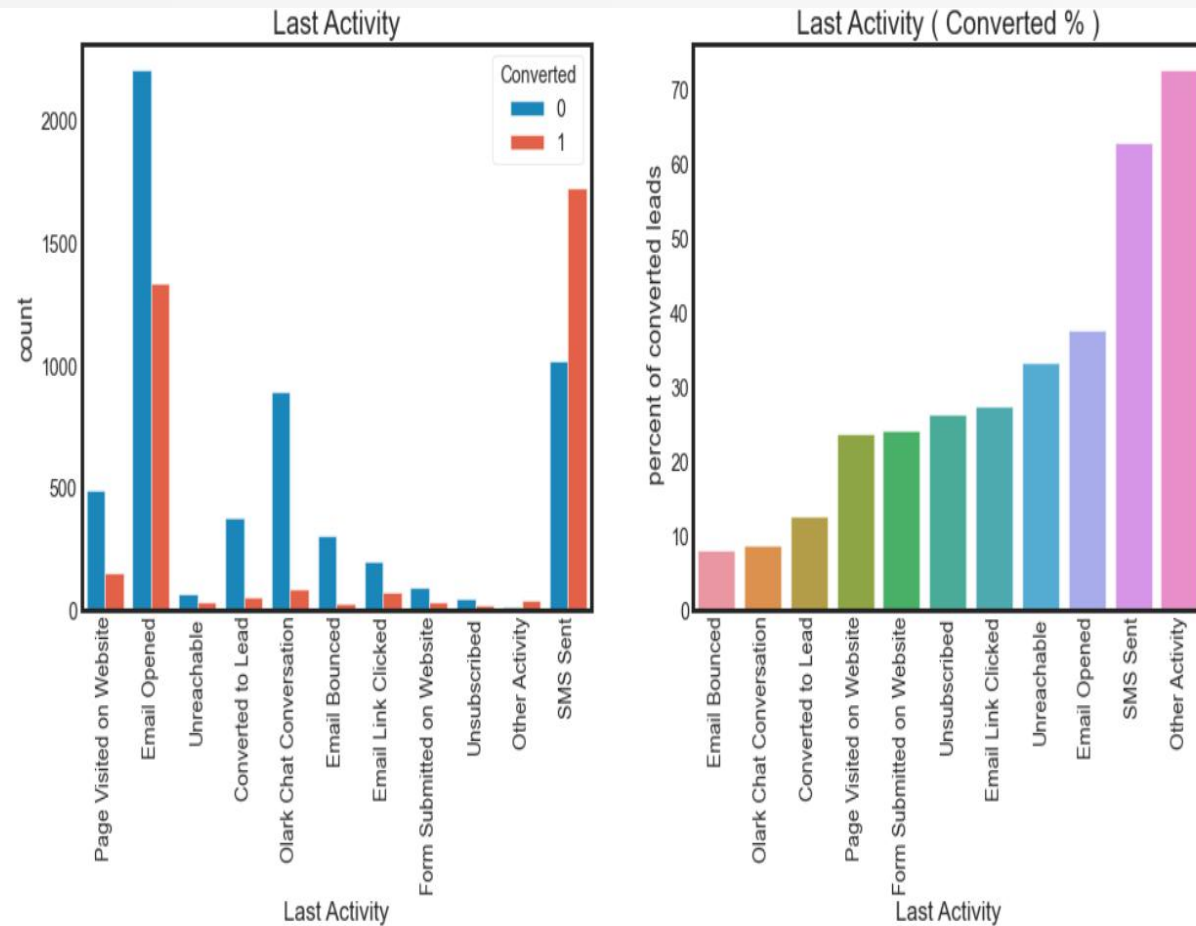
- Most of the lead have their Email opened as their last activity
- Conversion rate for leads with last activity as SMS Sent is almost 60%

Variables Impacting the Conversion

- 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 – Olark Chat Conversation
- 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 - Sensitivity and Specificity on Train Data Set
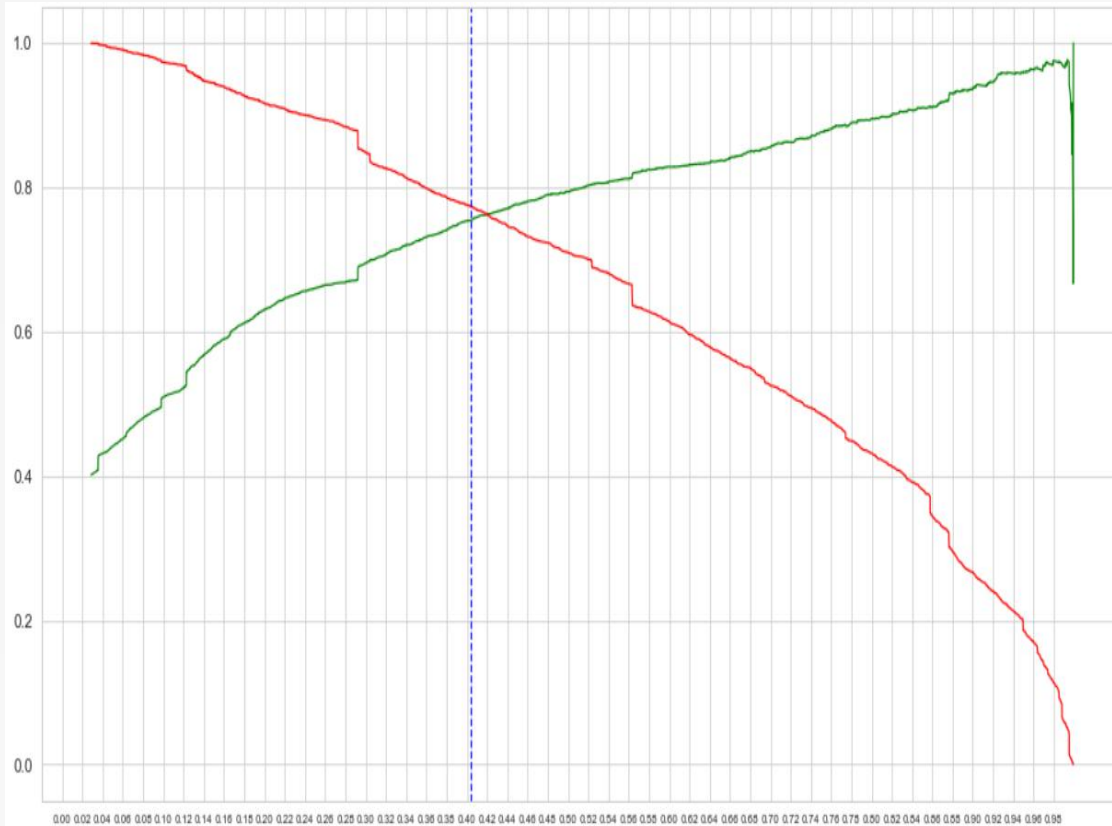
- ## Confusion Matrix
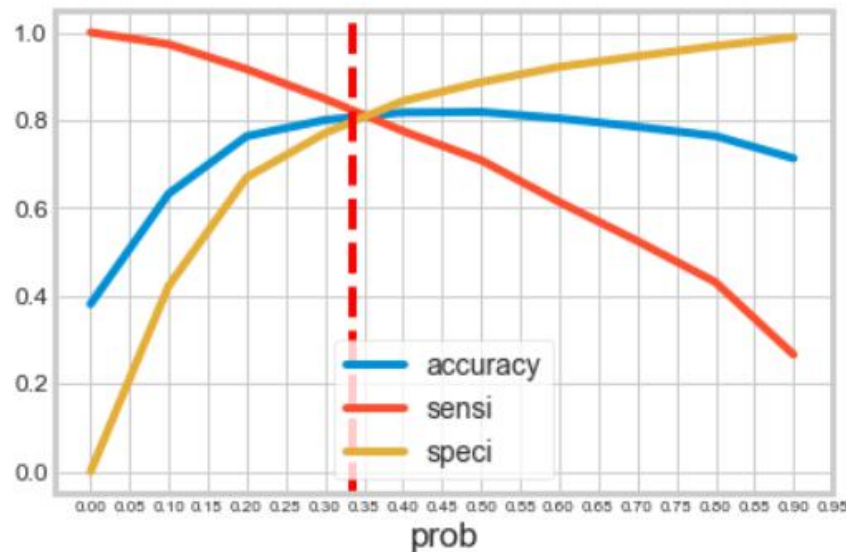
- From the above graph, 0.335 seems to be ideal cut-off points

3203	799
449	2017

Model Accuracy value is   : 80.71 %
Model Sensitivity value is   : 81.79 %
Model Specificity value is   : 80.03 %
Model True Positive Rate (TPR)   : 81.79 %
Model False Positive Rate (FPR)   : 19.97 %
Model Poitive Prediction Value is   : 71.63 %
Model Negative Prediction value is   : 87.71 %

Model Evaluation- Precision and Recall on Train Dataset

- Confusion Matrix

3377	625
554	1912

Model Precision value is : 75.36 %
Model Recall value is : 77.53 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Confusion Matrix

1341	336
199	896

- Model Accuracy value is   : 80.7 %
- Model Sensitivity value is   : 81.83 %
- Model Specificity value is   : 79.96 %
- Model Precision value is   : 72.73 %
- Model Recall value is   : 81.83 %
- Model True Positive Rate (TPR)   : 81.83 %
- Model False Positive Rate (FPR)   : 20.04 %
- Model Poitive Prediction Value is   : 72.73 %
- Model Negative Prediction value is   : 87.08 %

Conclusion

➢ While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on

➢ Sensitivity and Specificity for calculating the final prediction. –

➢ Accuracy, Sensitivity and Specificity values of test set are around 81%, 80% and 82% which are approximately closer to the respective values

➢ calculated using trained set.

➢ Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 80% in test set

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

➢ Lead Source_Welingak Website

➢ Lead Origin_Lead Add Form

➢ Occupation_Working Professional

➢ Hence overall this model seems to be good.