Lead Scoring Case Study

ADWAIT ATHAWALE

FARAZ AHMED

RADHIKA KUTE

Problem Statement

- X Education has a low lead conversion rate of 30%.
- The Company wishes to identify the most promising leads, also known as "hot leads," in order to increase the lead conversion rate to 80%.
- A model is to be build that assigns a lead score to each lead, such that the customers with higher lead scores have a higher conversion chance.
- The model will be trained on a historical dataset of leads. The model will then be used to score new leads.
- The sales team will focus their efforts on the leads with the highest lead scores. This will help to increase the lead conversion rate.

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ € ₹ $
Converted Leads

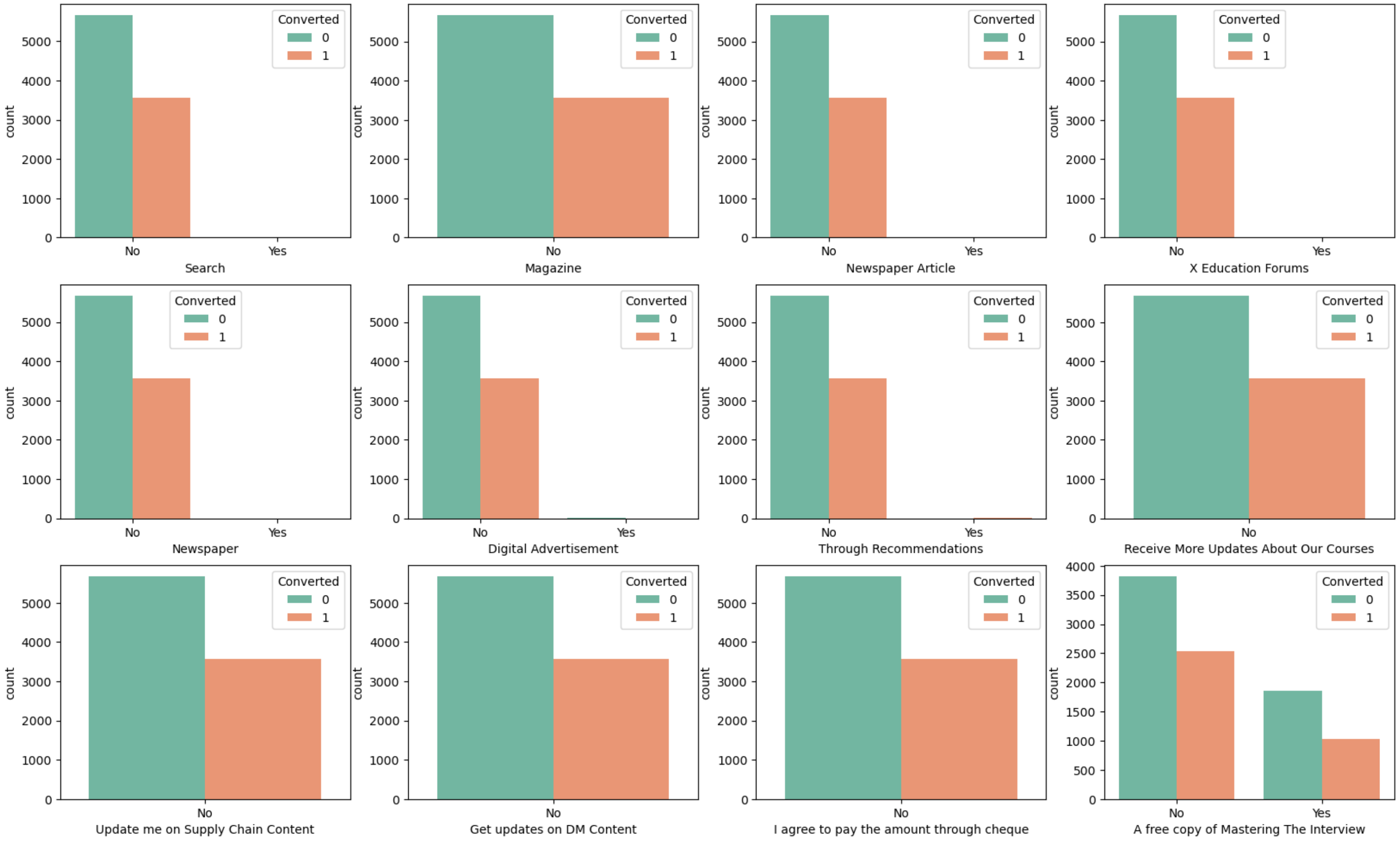
Approach

- Data Inspection
- Exploratory Data Analysis
  a. Data Cleaning
  b. Categorical Attributes Analysis
  c. Numerical Attributes Analysis
- Data Preparation
- Test-Train Split
- Feature Scaling
- Model Building
  a. Using Stats Model & RFE
  b. Using Decision Tree
- Conclusion

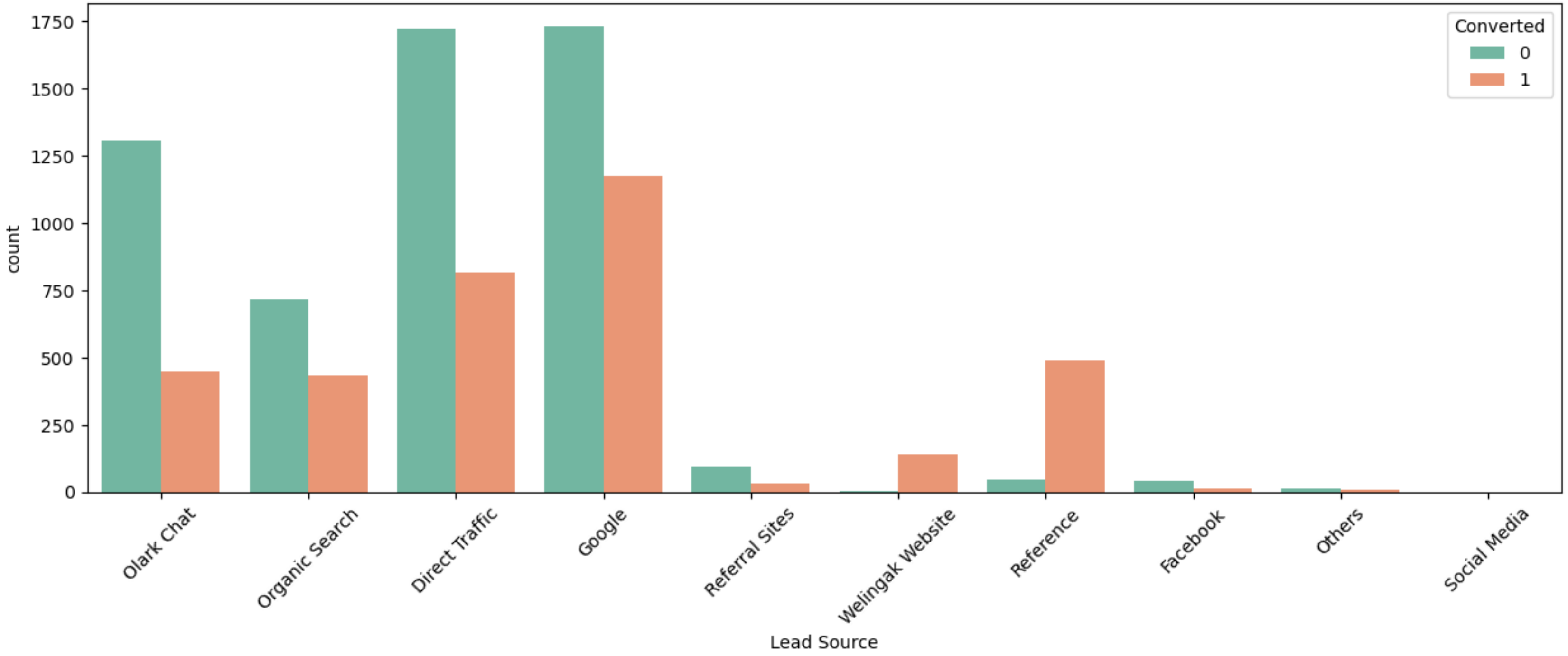
Exploratory Data Analysis

Categorical Attributes Analysis – *Imbalanced Variables*

- For all these columns except 'A free copy of Mastering The Interview' data is highly imbalanced, thus we will drop them
- "A free copy of Mastering The Interview" is a redundant variable so we will include this also in list of dropping columns.

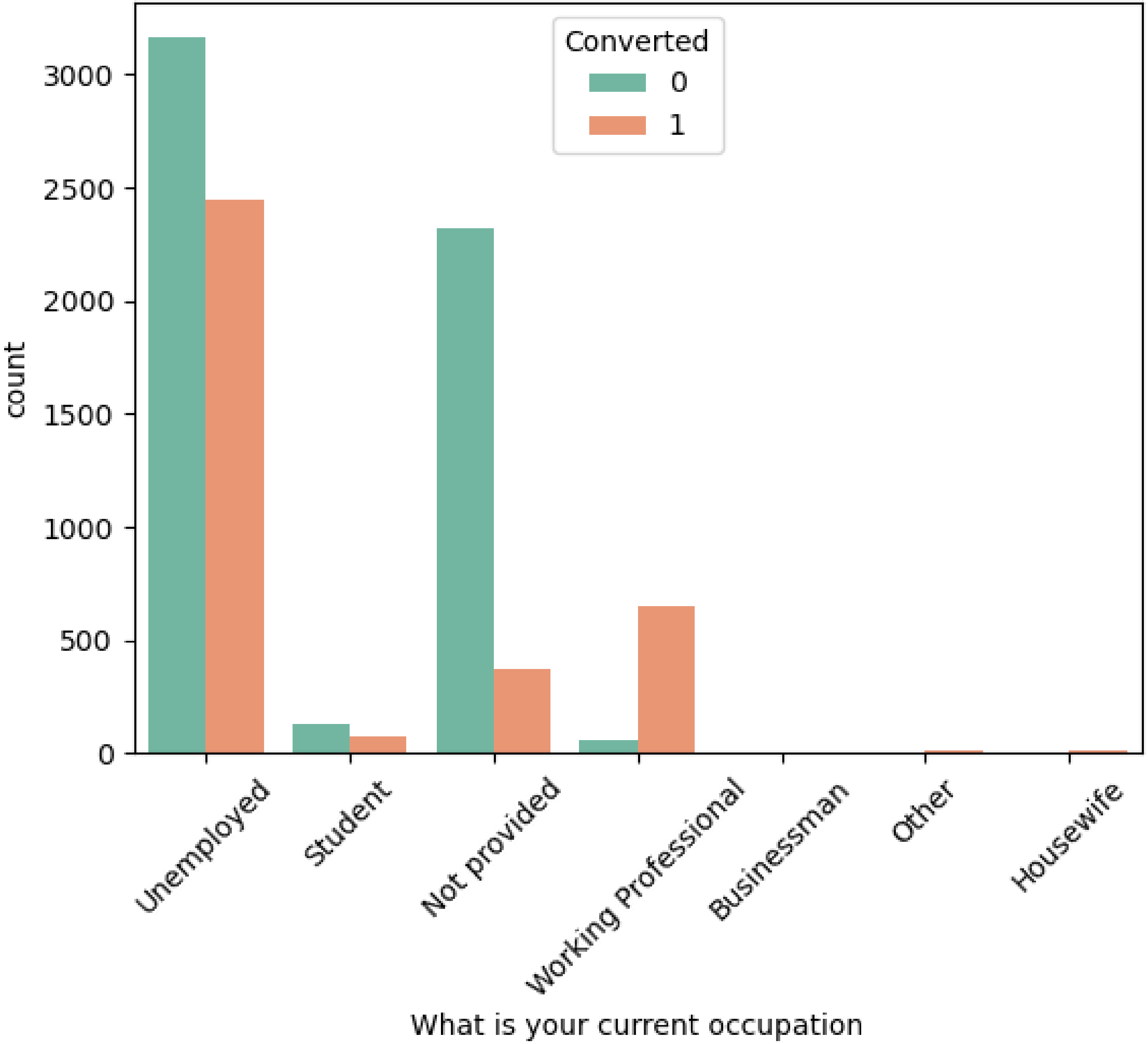


Categorical Attributes Analysis – *Lead Source*



- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.

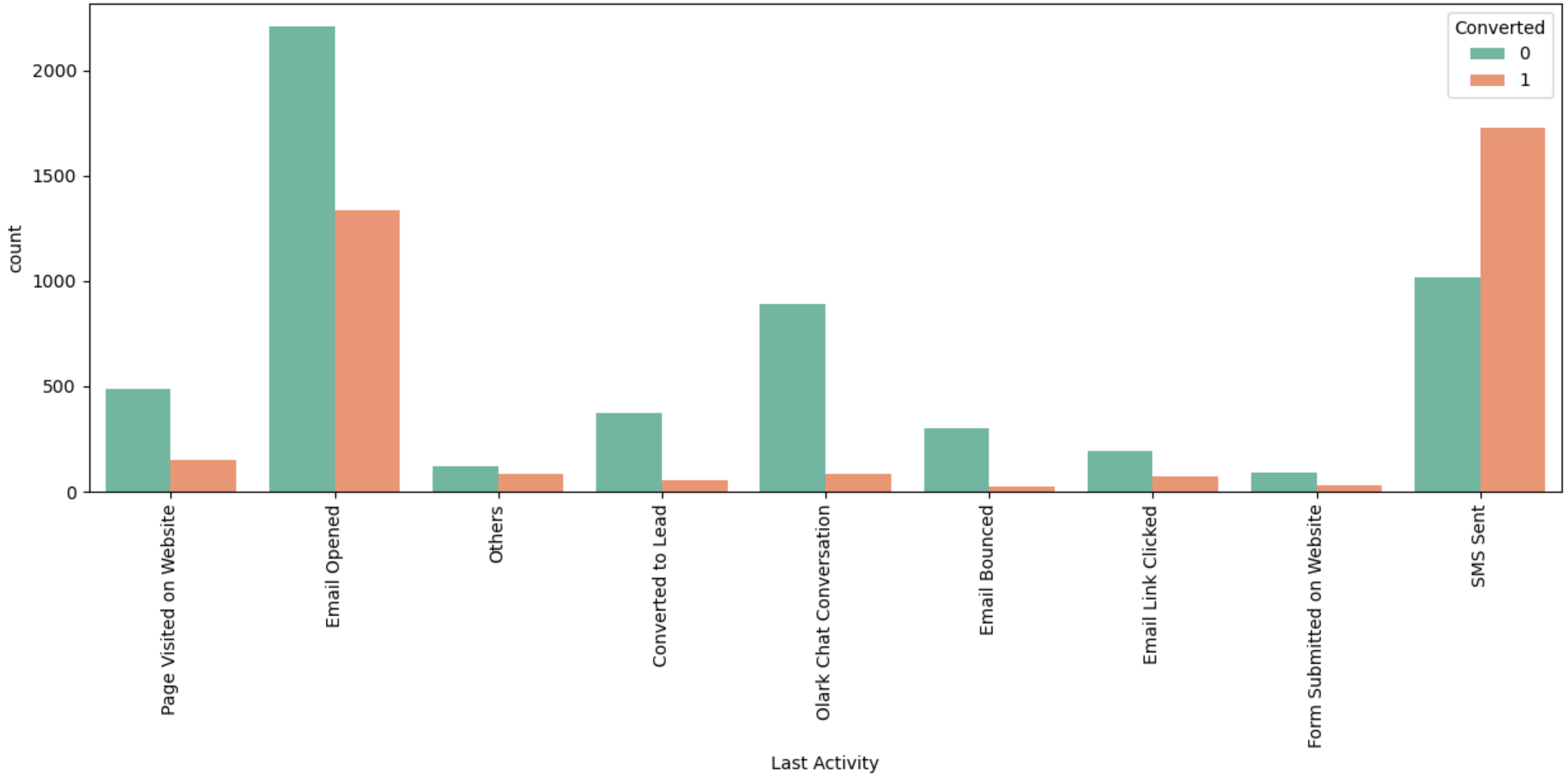
Categorical Attributes Analysis – *What is your current occupation*



- Maximum leads generated are unemployed and their conversion rate is more than 50%.
- Conversion rate of working professionals is very high.

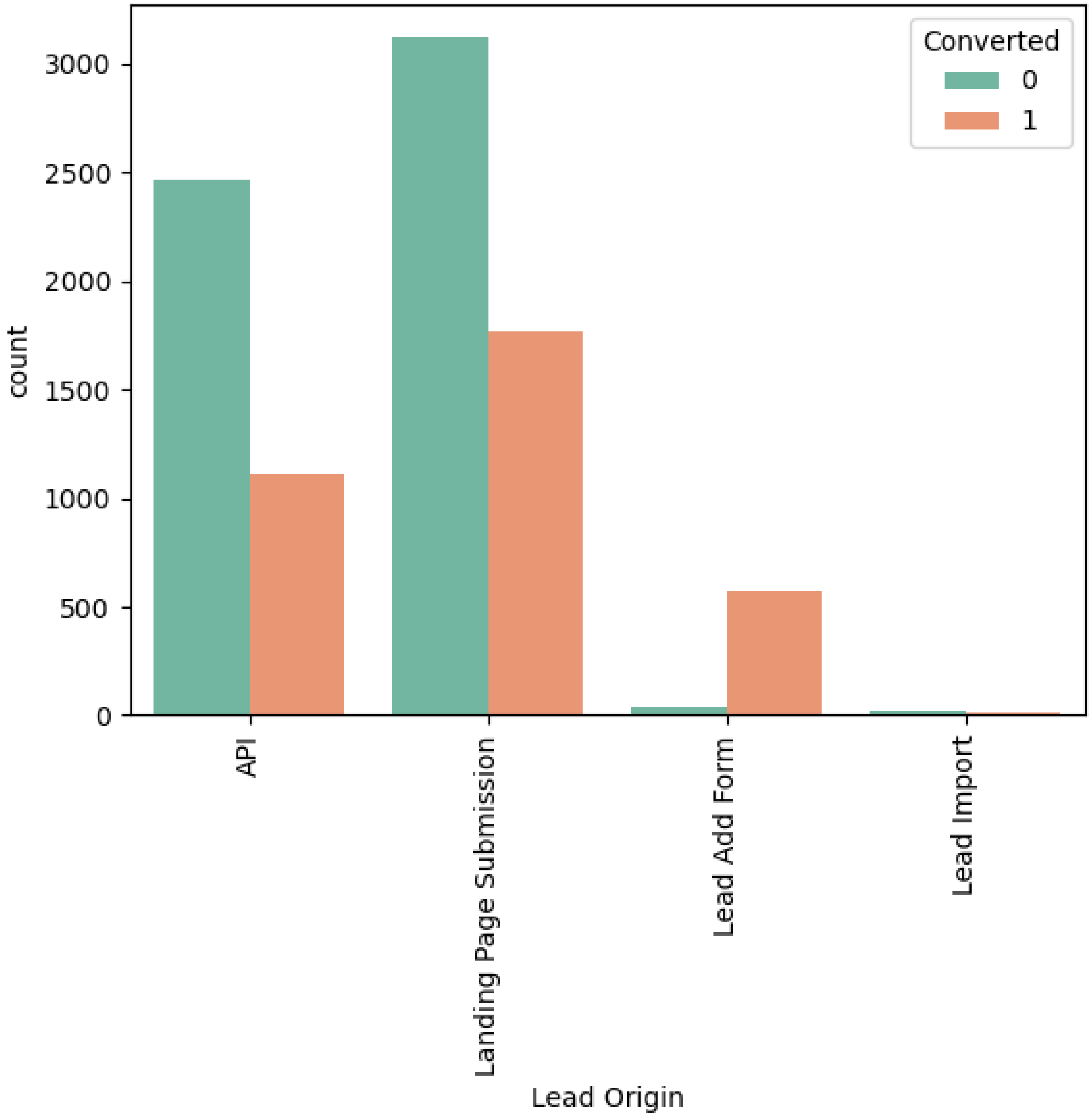
Categorical Attributes Analysis – *Last Activity*

- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last activity has high conversion rate.



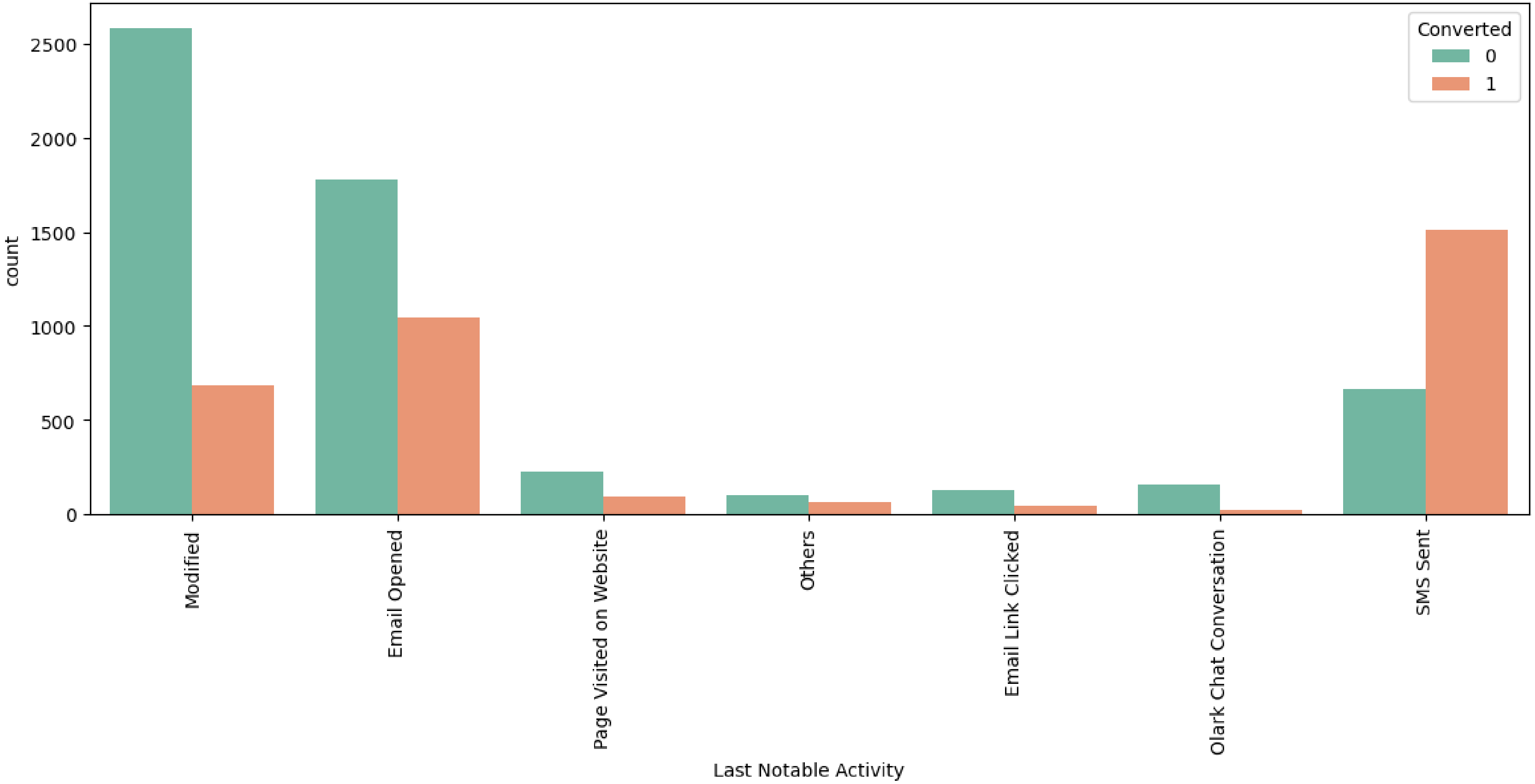
Categorical Attributes Analysis – *Lead Origin*

- Maximum leads are generated having lead origin as Landing Page Submission but conversion rate is low.
- Lead Add Form lead origin has high conversion rate.



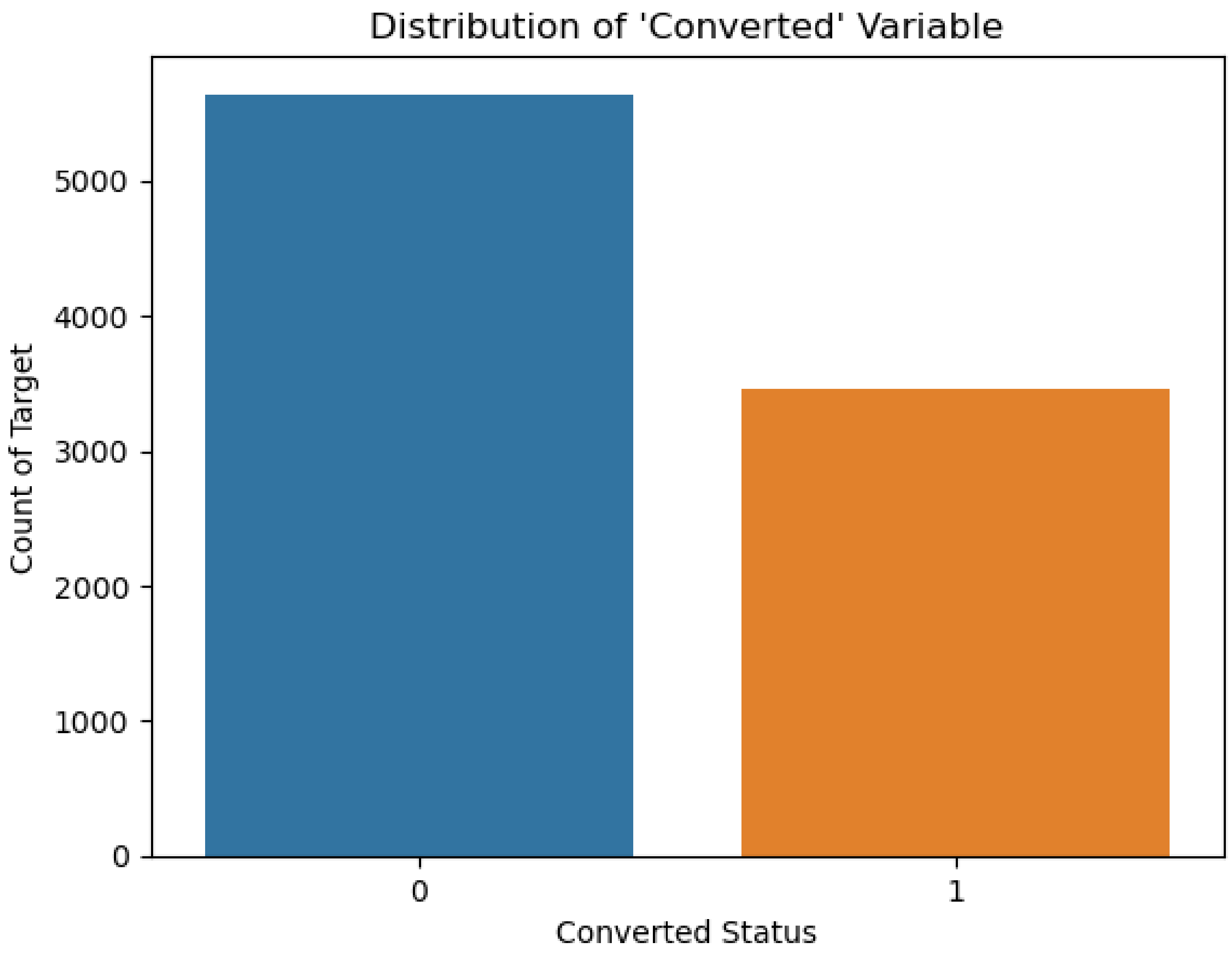
Categorical Attributes Analysis – *Last Notable Activity*

- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last activity has high conversion rate.



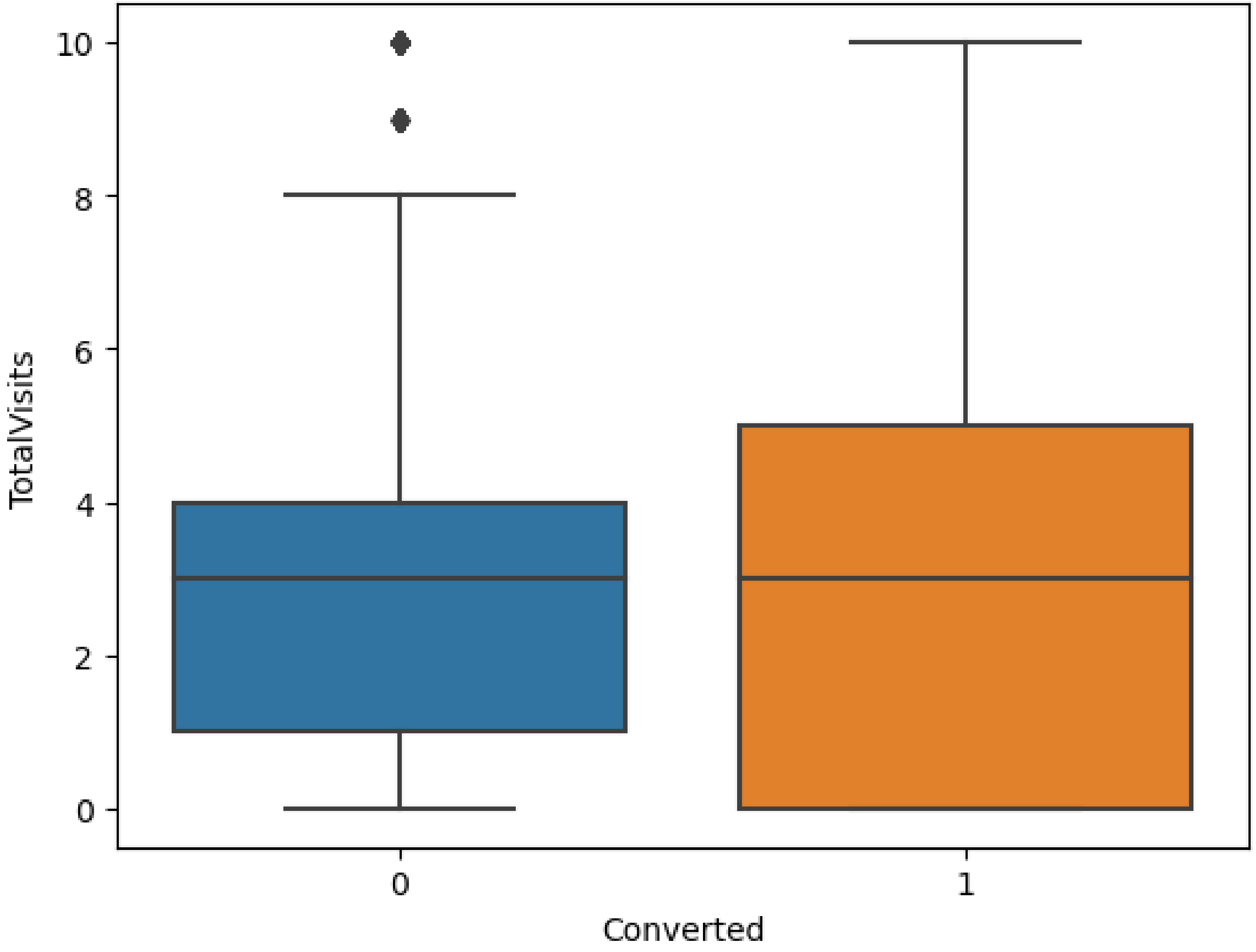
Numerical Attributes Analysis - *Converted*

- This is the target Variable
- Currently, the lead Conversion rate is 38% only



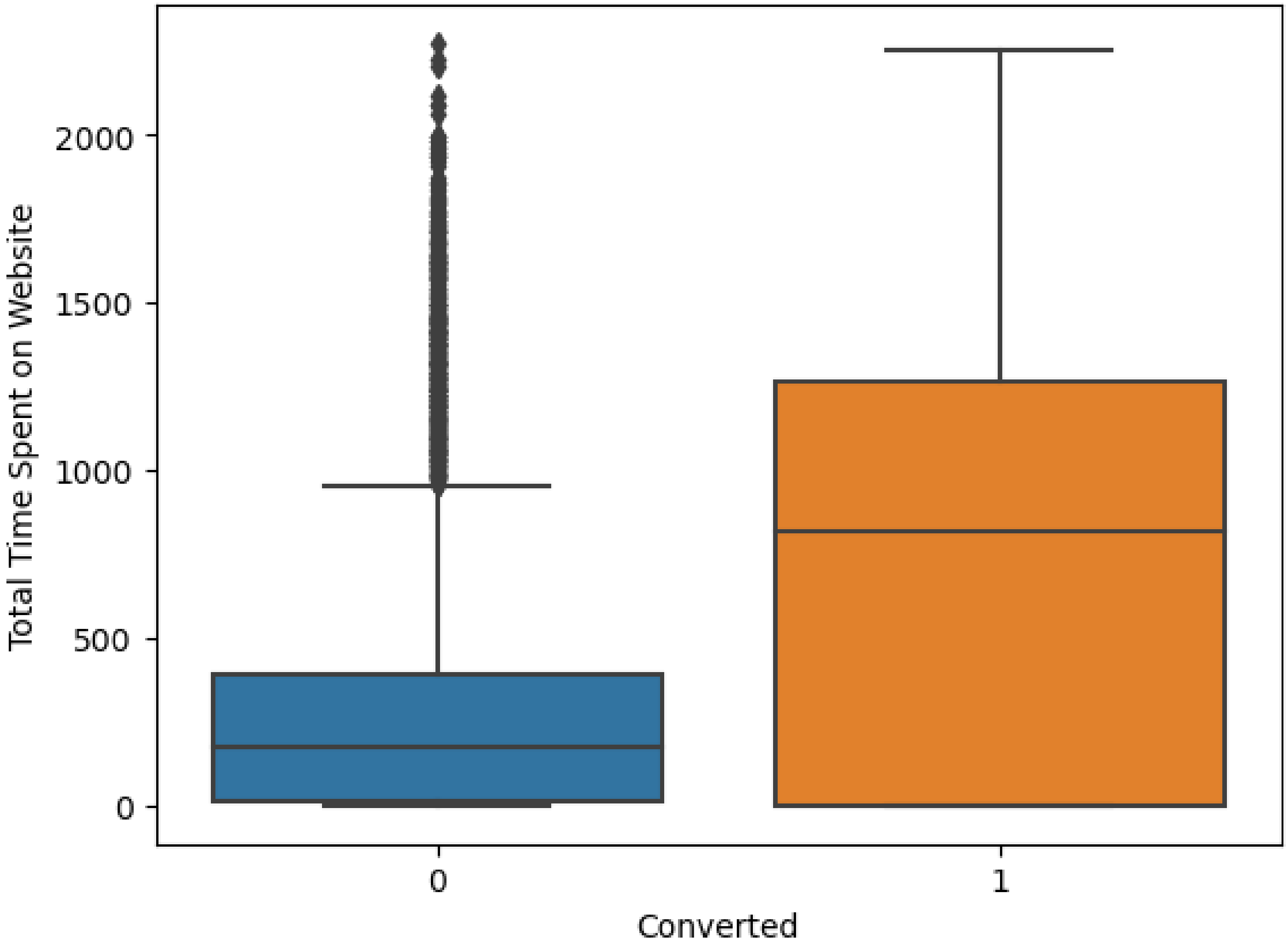
Numerical Attributes Analysis *– Total Visits w.r.t Target Variable*

- As the median for both converted and non-converted leads are same , nothing conclusive can be said on the basis of variable Total Visits



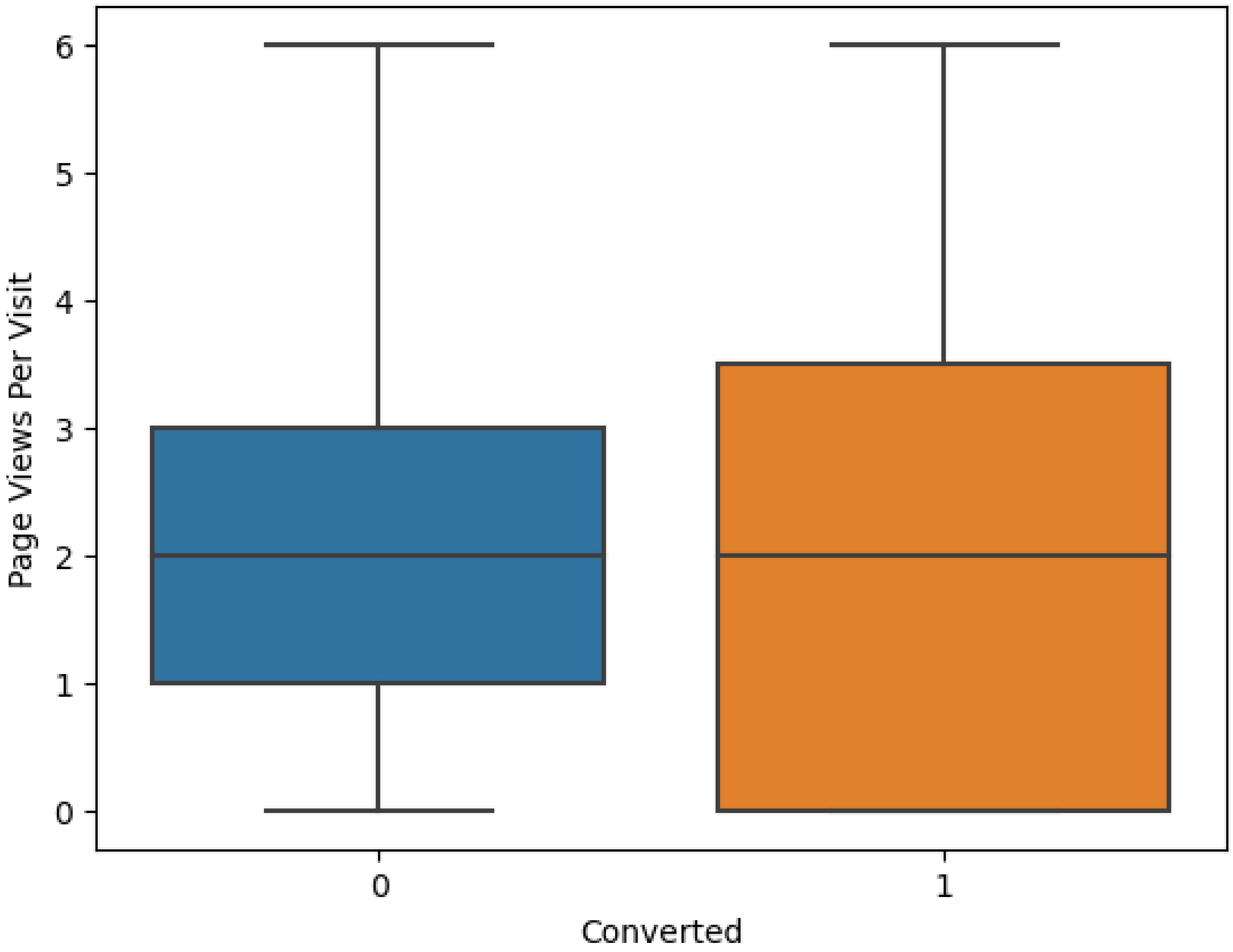
Numerical Attributes Analysis *– Total Time Spent on Website w.r.t Target Variable*

- As can be seen, leads spending more time on website are more likely to convert , thus website should be made more engaging to increase conversion rate

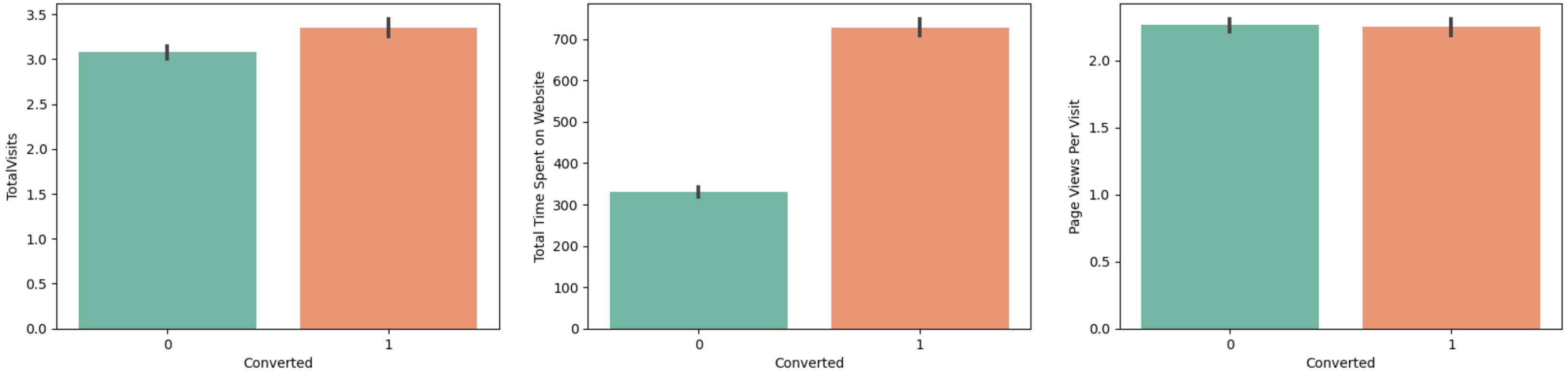


Numerical Attributes Analysis *– Page Views Per Visit w.r.t Target Variable*

- Median for converted and not converted leads is almost same.
- Nothing conclusive can be said on the basis of Page Views Per Visit.



Numerical Attributes Analysis – *All Variables*



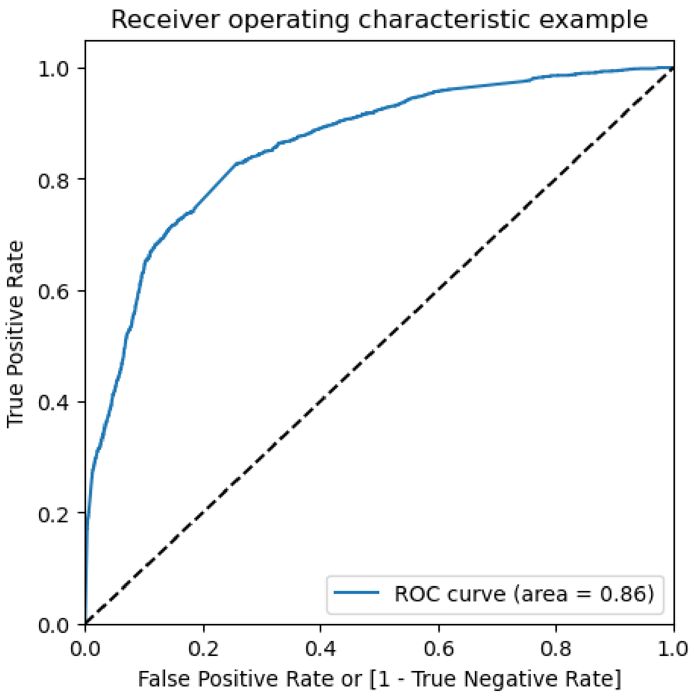
- The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit

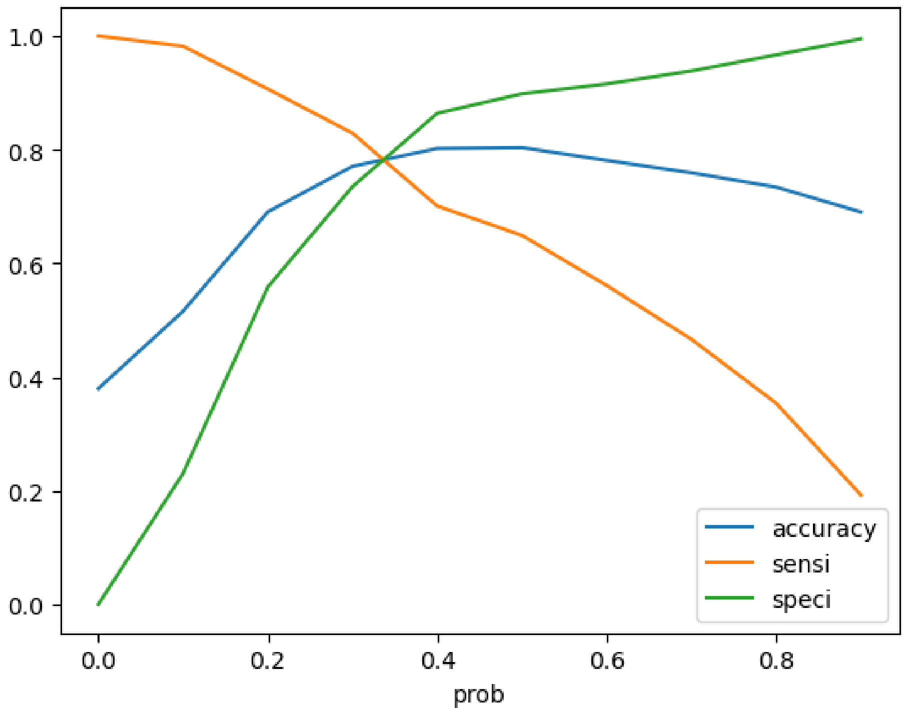
Model Building using Stats Model & RFE

- Feature Selection using RFE
- Running RFE with 15 variables as output
- Determined Optimal Model Using Logistic Regression
- Checked for high p-value & VIF value
- Calculated Metrics – accuracy, sensitivity, specificity, precision & recall to evaluate the model

Model Evaluation on Train data – *ROC Curve & Optimal Cutoff Point*

- The ROC curve has a value of 0.86, which is very good. We have the following values for the Train Data:
  - Accuracy : 77.05%
  - Sensitivity :82.89%
  - Specificity : 73.49%
- From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

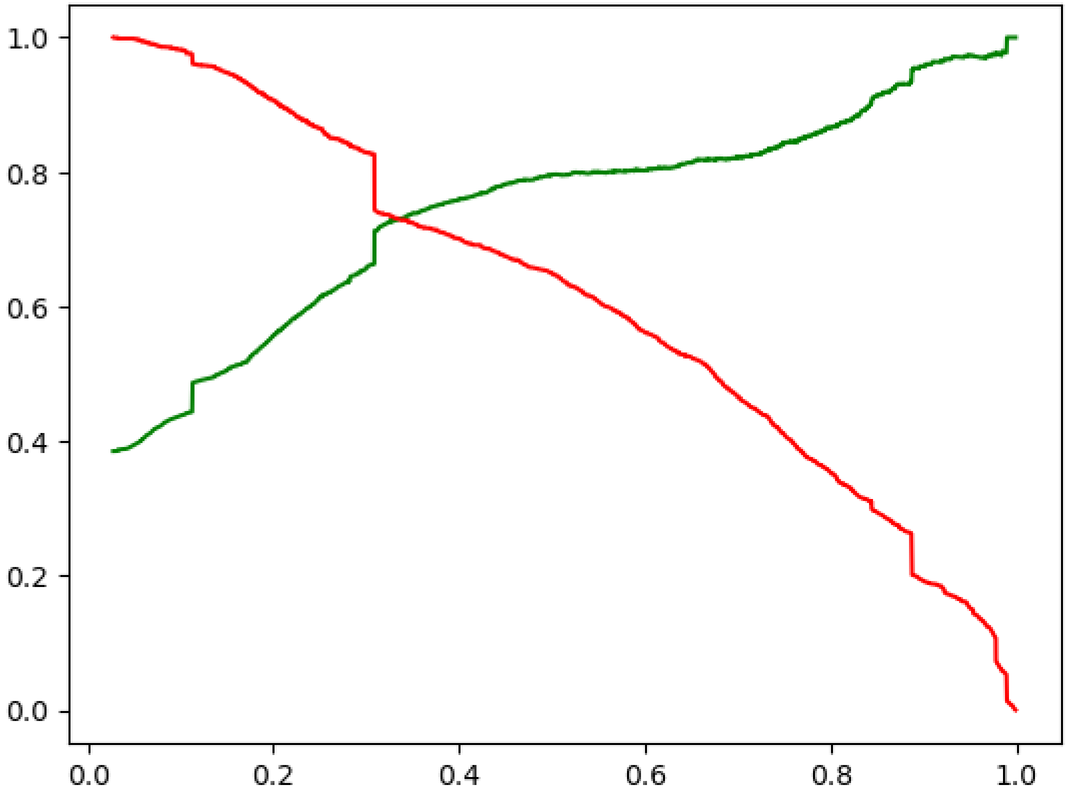




Model Evaluation on Train data – *Precision & Recall*

Based on below Precision & Recall values we are getting the given curve

- Precision : 65.67%
- Recall : 82.88%



Model Evaluation on Test data – Overall Metrics

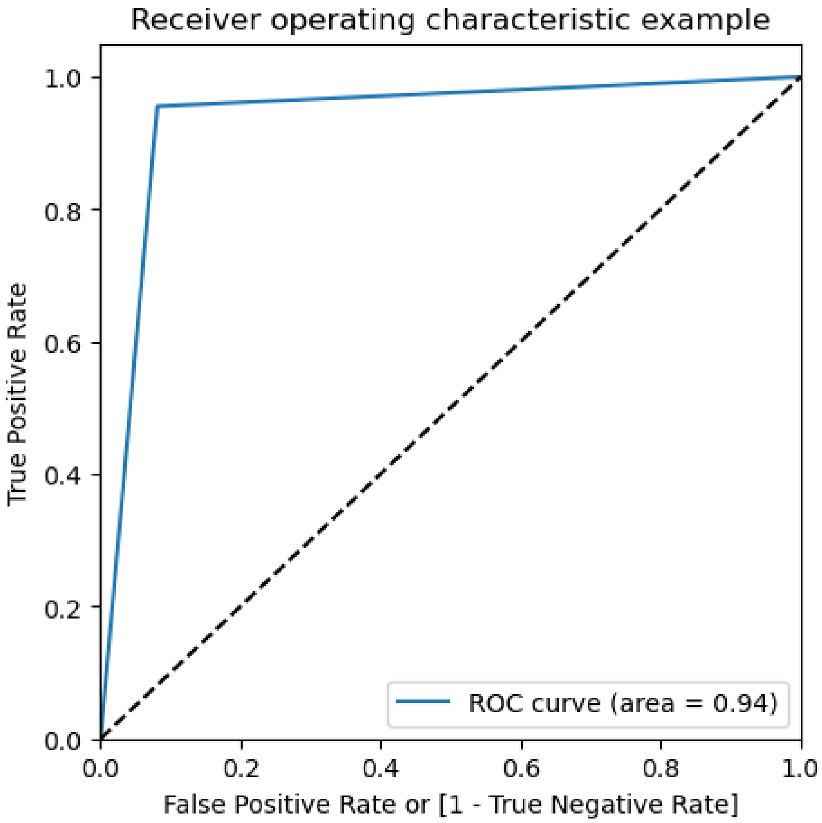
After running the model on the Test Data these are the figures we obtain:

- Accuracy : 77.52%
- Sensitivity : 83.01%
- Specificity : 74.13%
- Precision : 66.43%
- Recall : 83.01%

Decision Tree Model Evaluation

Values obtained for Train & Test dataset:

- Train Data Accuracy : 93.27%
- Train Data Sensitivity : 95.54%
- Train Data Specificity : 91.88%
- Test Data Accuracy : 72.68%
- Test Data Sensitivity : 95.54%
- Test Data Specificity : 91.88%



Conclusion:

- The logistic regression model is performing better than the decision tree model.
- The decision tree model has a high specificity on the test set, but this is at the expense of accuracy. The decision tree model is also overfitting the training set, as evident by the fact that the accuracy on the training set is 93% which is much higher than the 73% accuracy on the test set.
- 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 on the logistic regression model.
- For the logistic regression model, the Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Important Features

Features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Total Time Spent on Website