Lead Score - Case Study

UpGrad

Siddhartha Pramanik

Approach of the Analysis

Problem Statement

- To find out the ‘Hot Leads’, most potential leads, based on the various aspects provided for a particular lead
- Lead reach the website using various methods and are tagged accordingly, we need to identify using these tags and other parameters like the amount of time spent on the website, etc, if the lead can convert into an actual student for the website.

Approach of the Analysis

Overview of Steps Followed

- After importing the data, we check for the missing values and columns with huge chunk of data missing are plotted to check if they affect our target variable i.e. Converted, significantly
- The columns which does not show much influence on the target variables are dropped, rest of the columns are imputed either using median or mode
- Identifier columns and columns with redundant informations are dropped
- We check for significant class imbalance in the data and drop the columns accordingly

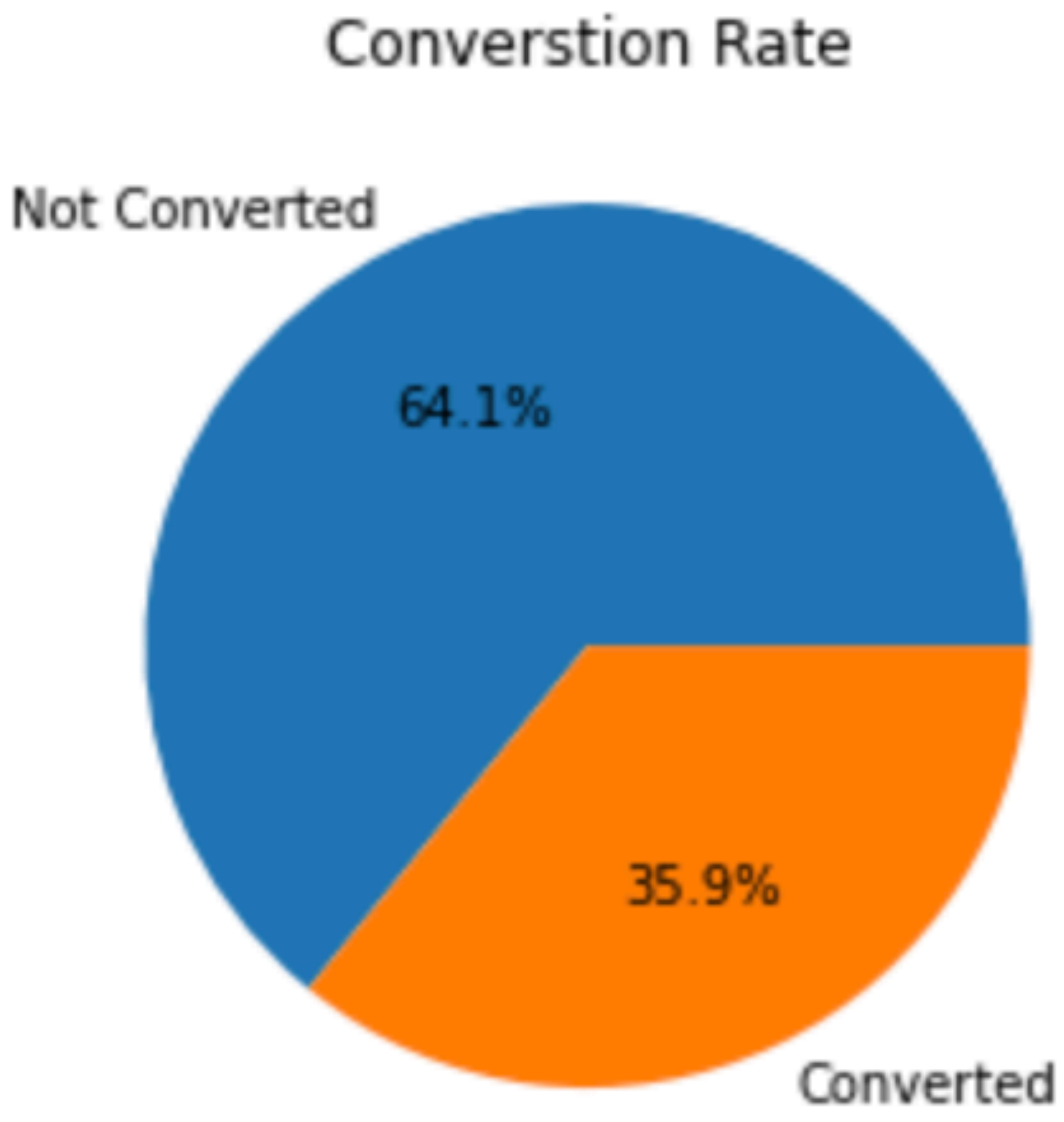
Approach of the Analysis

Overview of Steps Followed

- Perform further EDA on the columns left
- Pre-process the data, creating dummy variables, converting data types, scaling etc
- Build a basic logistic model with all the features present
- Evaluate the features based on the p-value and VIF
- Re-build a model and re-check the p-value and VIF until the numbers are in the acceptable range
- Calculating various evolution matrices and ROC curve, further finding the optimal cut off/ threshold value
- Evaluating the model on unseen data

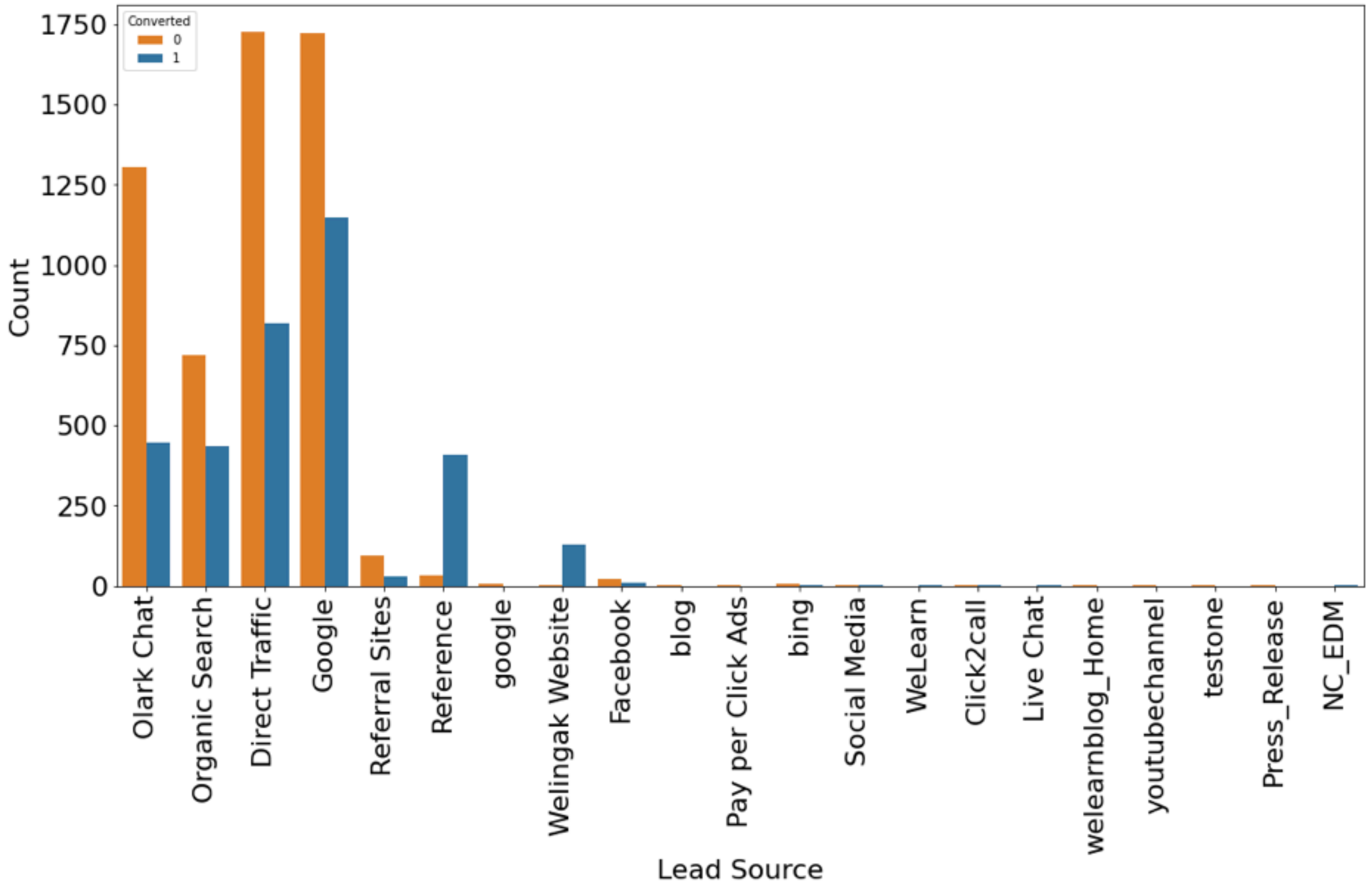
Current Conversion Rate

- The conversion rate of leads is ~36%



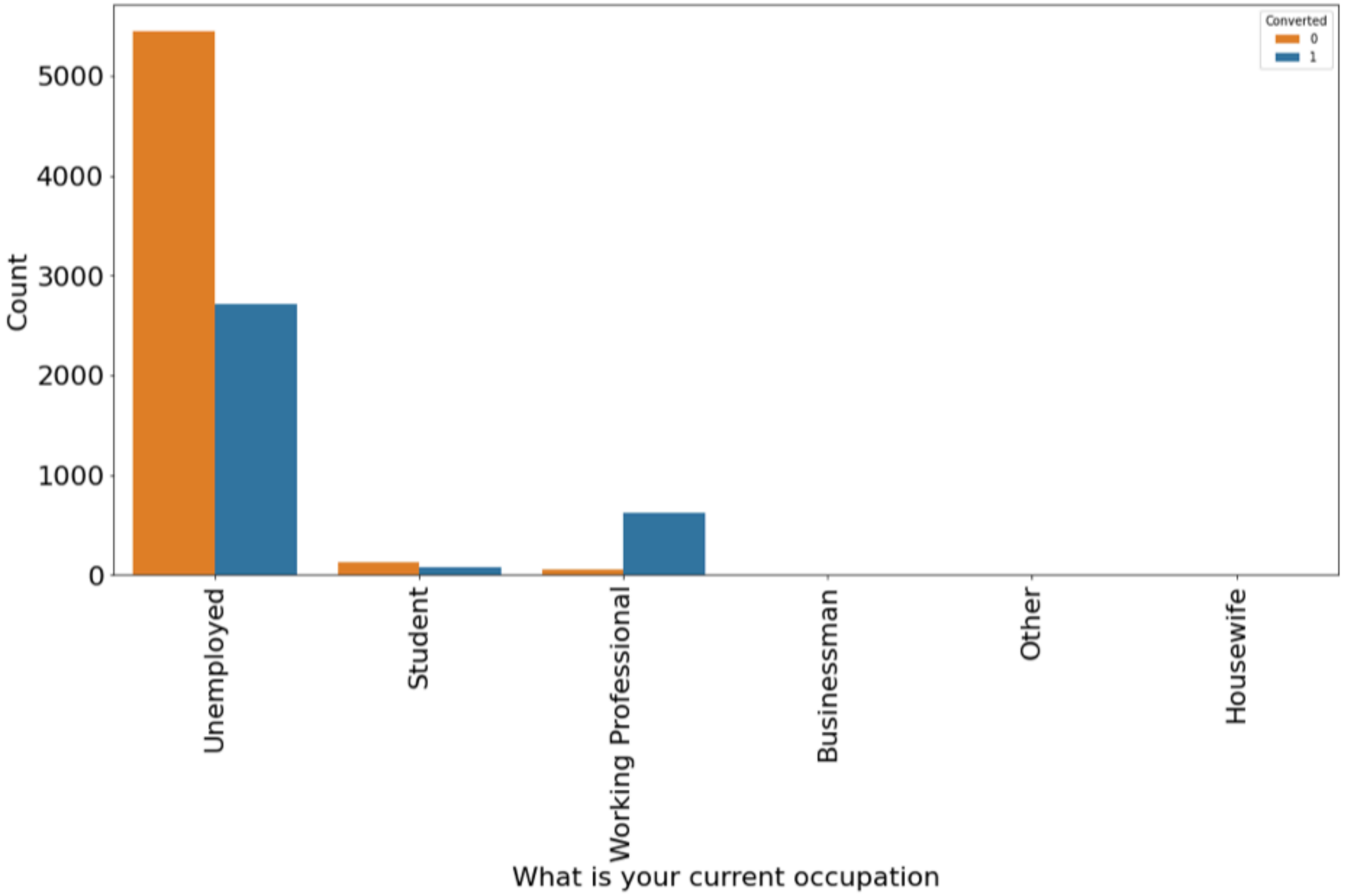
Lead Source Analysis

- We can invest more on the google traffic building activities as not only the generate the highest number of lead they also have great conversion rate



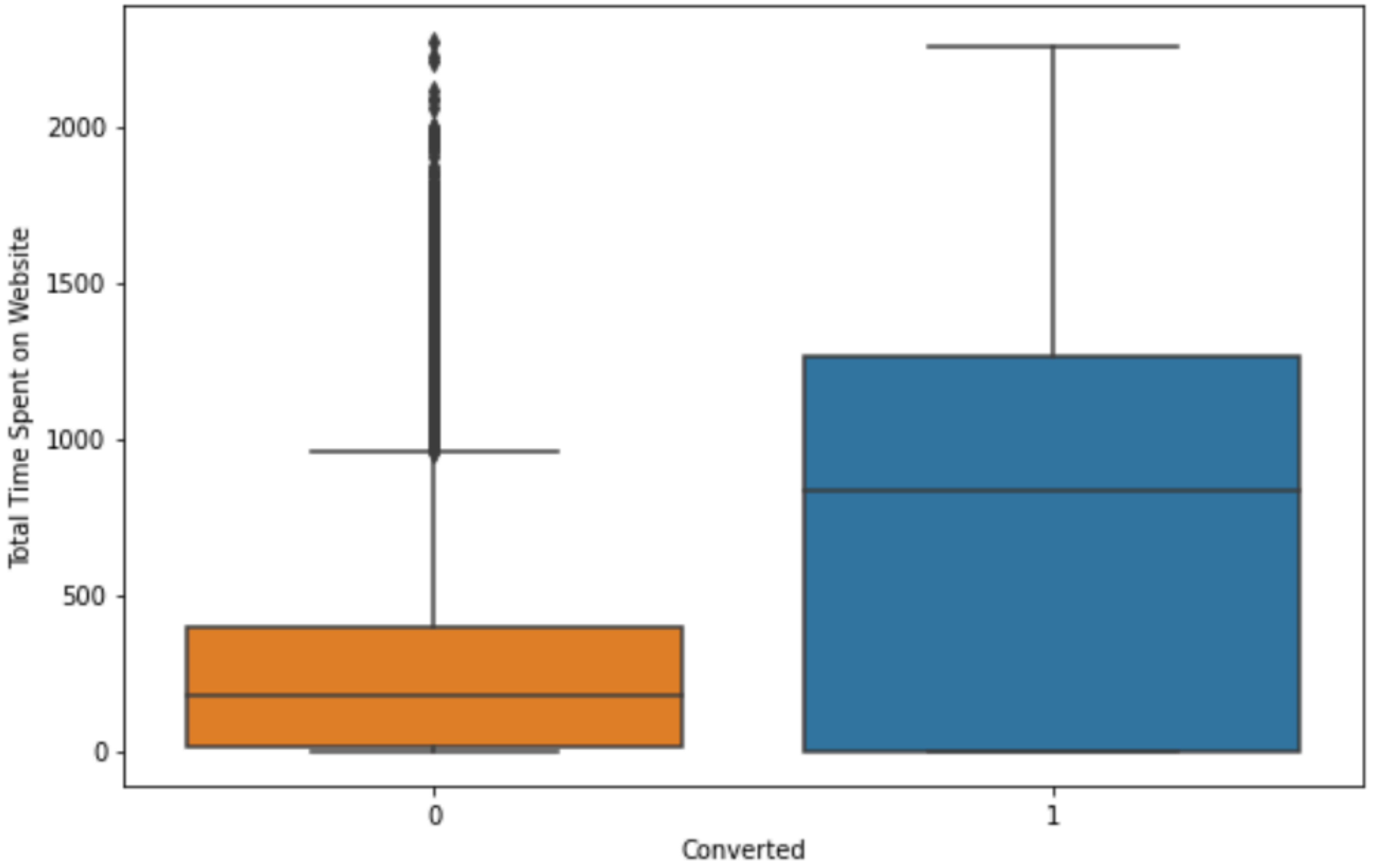
Occupation Type Analysis

- Employed leads are most likely to convert



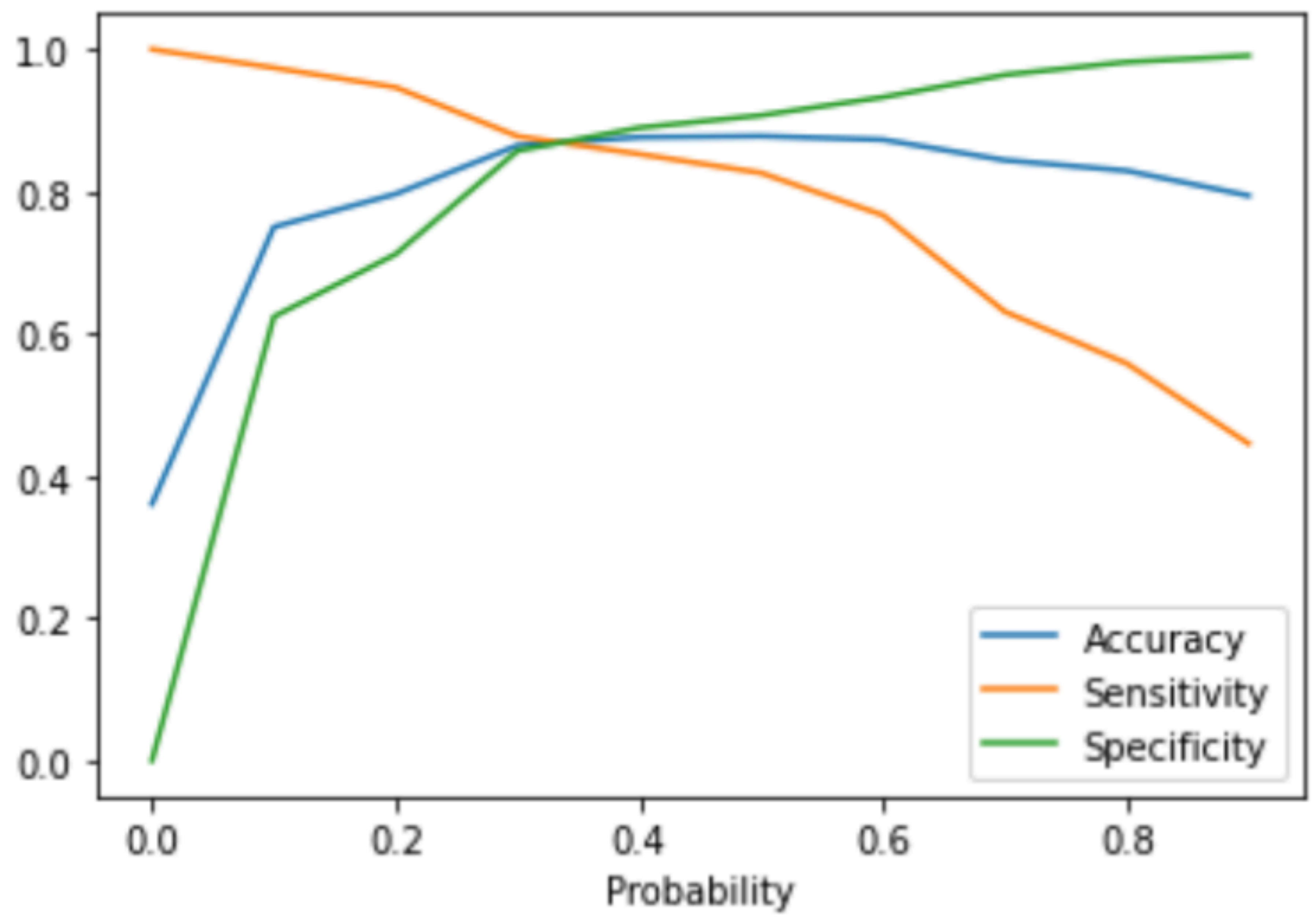
Time Spent on Website

- We can invest more on building on our website even more informative and engaging as people spending more time online are converting better.



Optimal Cutoff/ Threshold

- From the graph below we can tell that the value closer to 0.4 has the perfect mix of all three major matrices



Some Key Observations

On data

- Students/Leads coming through Google are most likely to be converted into successful leads
- Thought the most command last activity is opening email, the leads that sent SMS tend to convert more
- Unemployed sectors generate the highest numbers of lead but the professional class turns out to be more successful which reside with the fact that they can pay the fee easily
- People who revert back on once the email is read tend to have high success rate for converting, which resides with the ideology that we usually intereact more when we are actually interested
- Leads who didn't take a free copy of Mastering the interview appears to be more successful but we need to consider the fact that the ratio for yes:no and the conversion is almost same

Some Key Observations

for data points with missing values

- Column `How did you hear about X Education` seems to have no exact effect on the target variable as the plot every label in the column have almost equal numbers for 0 and 1
- For column `Lead Profile` only the label `Potential Lead` has some subtential amount of data and hence it shows the conversion the highest
- As columns `How did you hear about X Education` & `Lead Profile` has almost all rows empty, imputing them will result in bias or false weightage to these columns during model building. So we can drop them without any doubt.
- Column `Lead Quality` does have some more data compared to the other columns in this category, as well as it shows some sort of correlation for target variable, but as this variable is defined on the intiution of human it may vary. Also the mode for the data is 'Might be' which will further increase the percentage of the label if imputed, hence creating a class imbalance.
- We decide to drop all 3 of these columns

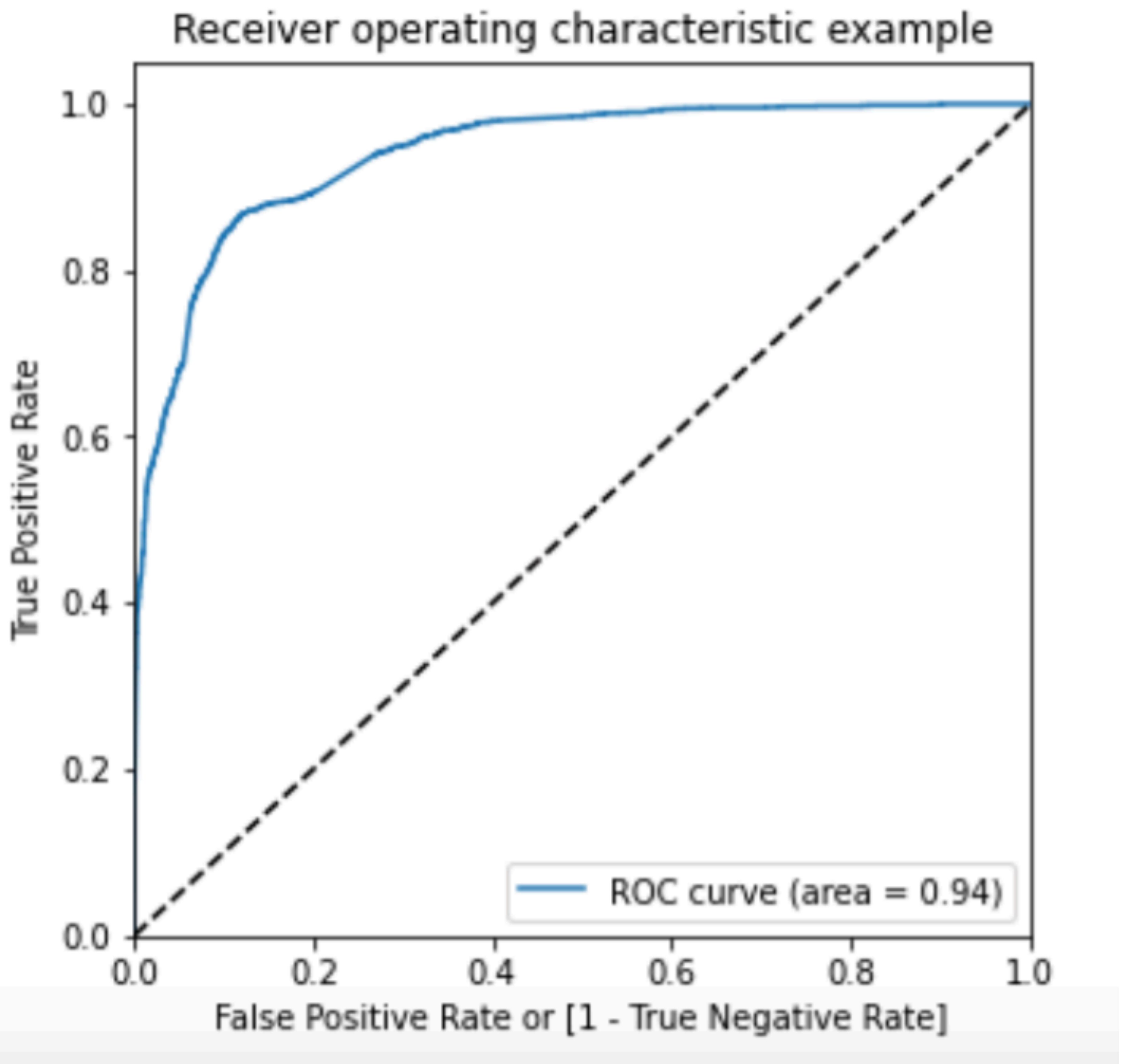
Some Key Observations

On Model

- Accuracy of the model is 88%
- Sensitivity for the model is 85%
- Specificity for the model is 88%
- Recall of the model is 85%
- Precision of the model is 81%
- Leads having a score of 40 or more have a high probability of converting

Some Key Observations

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.94)