LEAD SCORE CASE STUDY

Logistic Regression

Problem Statement

X Education is a company that offers online courses tailored for professionals in various industries. The courses offered by the company are promoted across various websites and search engines, including Google.

Despite generating a substantial number of leads, X Education struggles with a low lead conversion rate. X Education typically experiences a lead conversion rate of approximately 30% through the whole process of turning leads to customers by approaching them.

In order to enhance efficiency, the company aims to pinpoint the most promising prospects referred to as 'Hot Leads'. By successfully discerning this group of leads, the conversion rate is anticipated to increase. The implementation process for lead generation attributes lacks effectiveness in contributing to conversions.

Business Goal

The company requires to build a model.

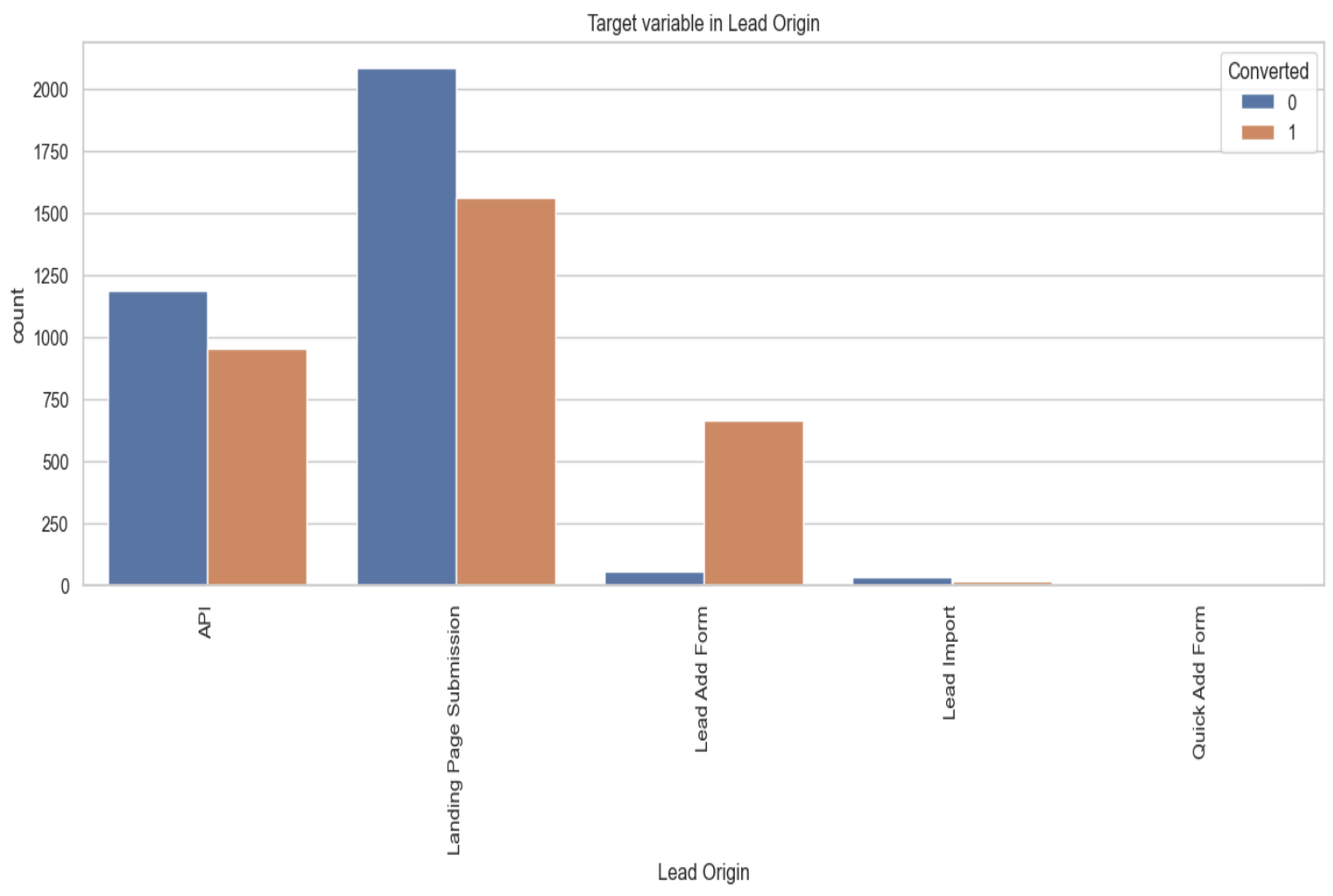
Allocate a lead score to each lead, ensuring that customers with higher scores are more likely to convert, while those with lower scores have a reduced likelihood of conversion.

Desired lead conversion rate is 80%.

Strategy

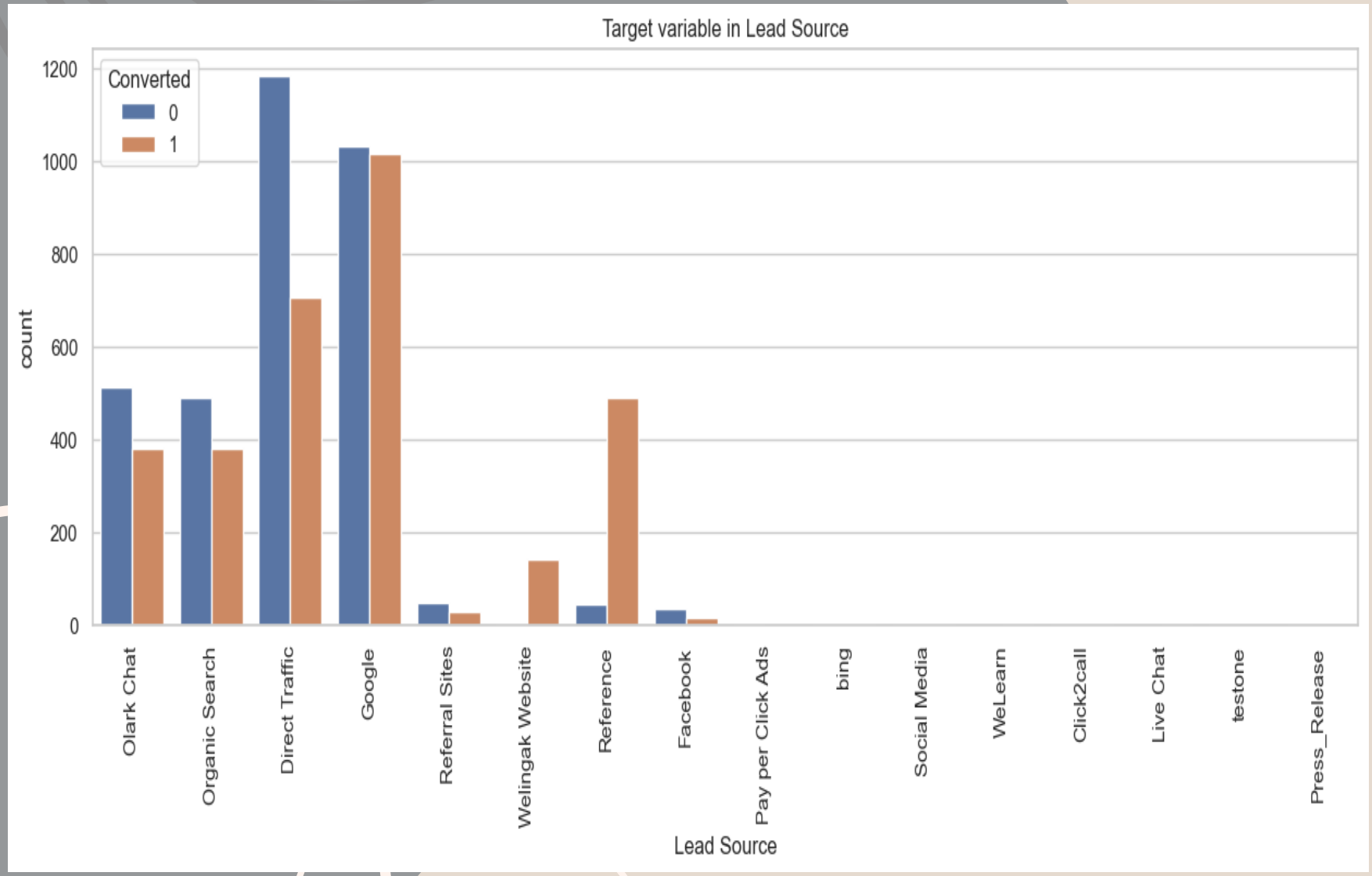
- Import Data
- Clean and prepare the data for further analysis
- Exploratory data analysis for selecting helpful attributes for conversion
- Selecting features
- Prepare the data for model building
- Build a logistic regression model
- Test the model on train set
- Evaluate model for different metrics
- Test the model on test set
- Measure the accuracy of model and other metrics for evaluation
- Assign a lead score for Hot leads

Exploratory Data Analysis



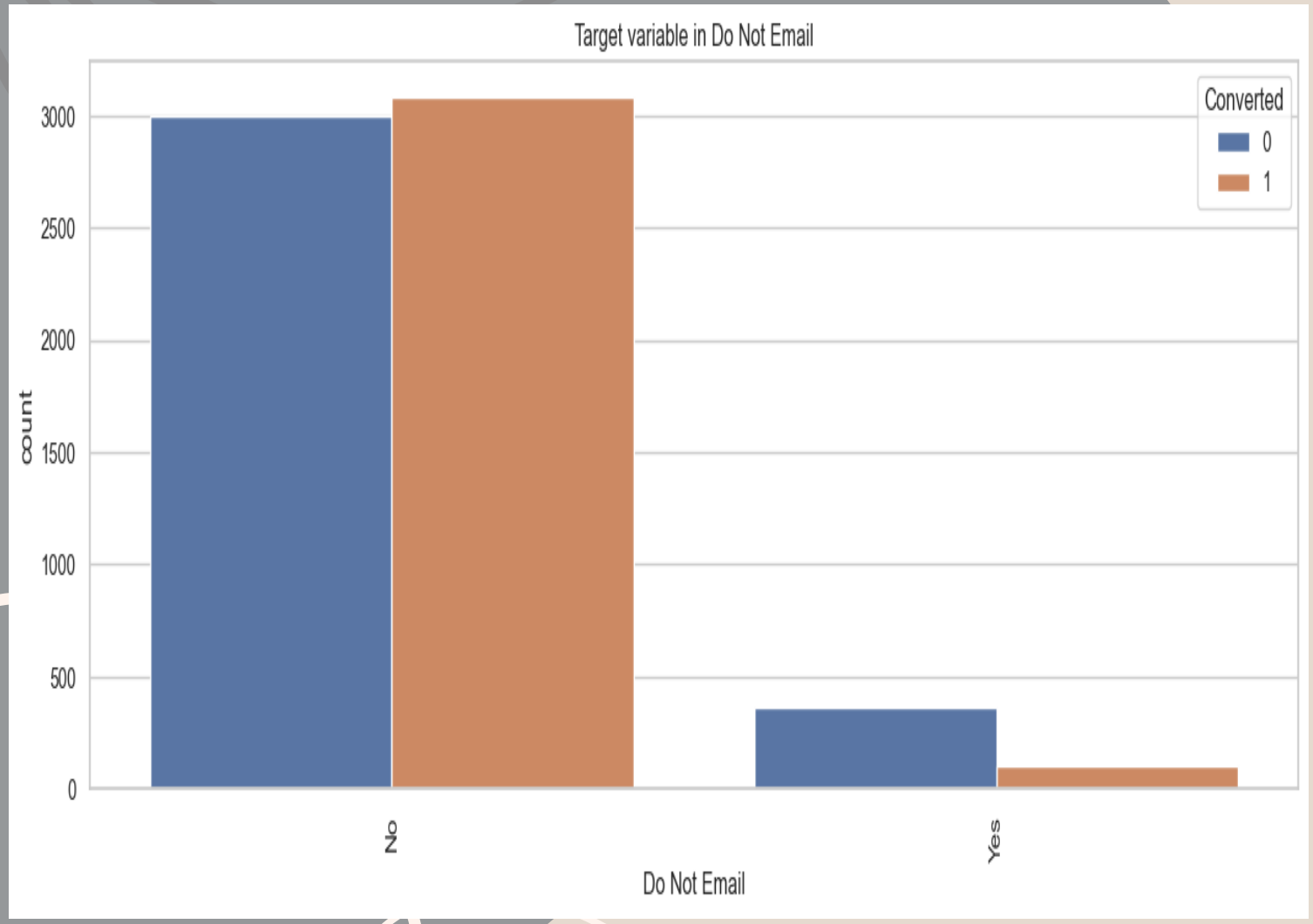
Lead Origin

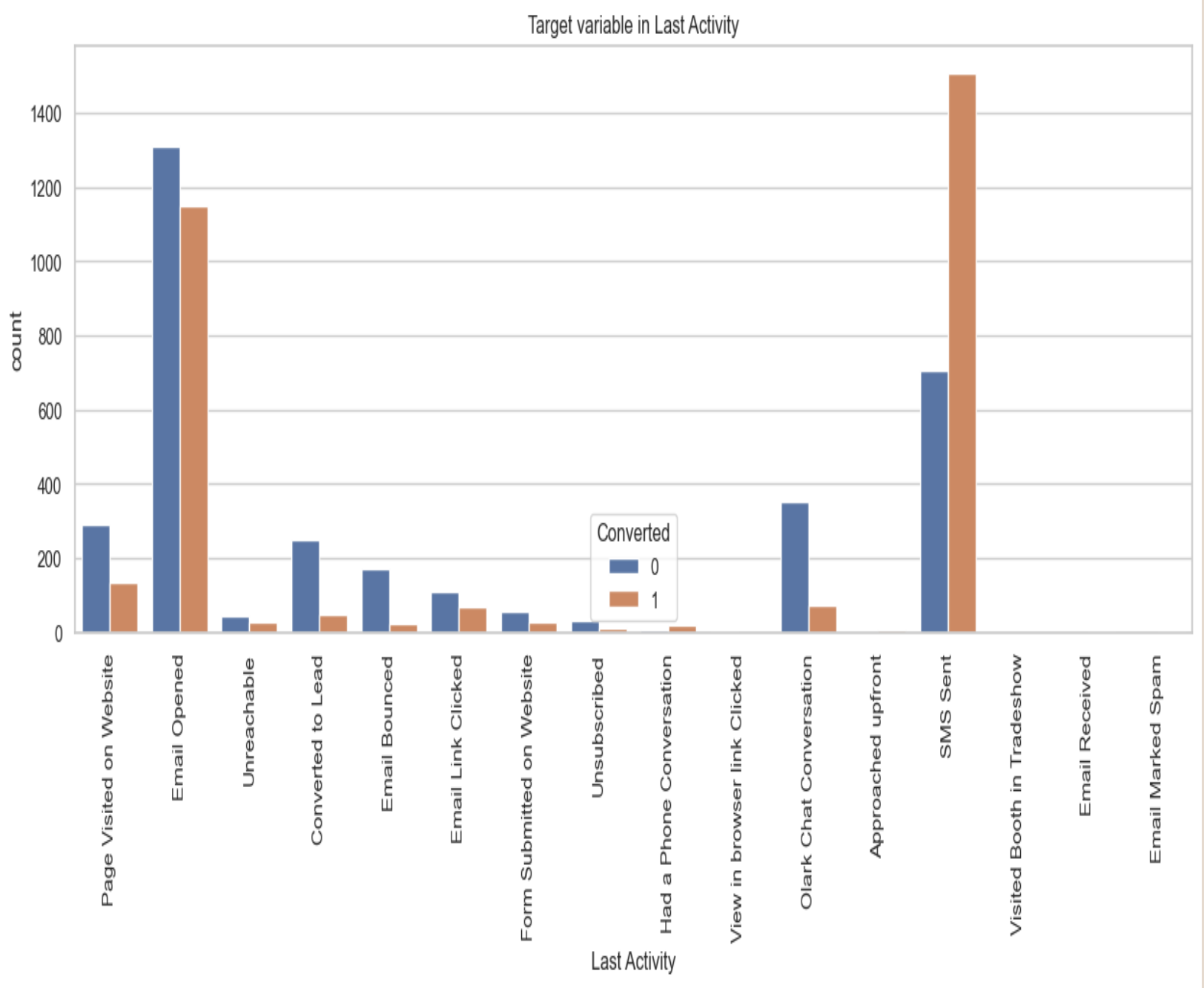
'Landing Page' submission has greatest source of audience, but conversion rate is higher in 'Lead and Form'



Lead Source

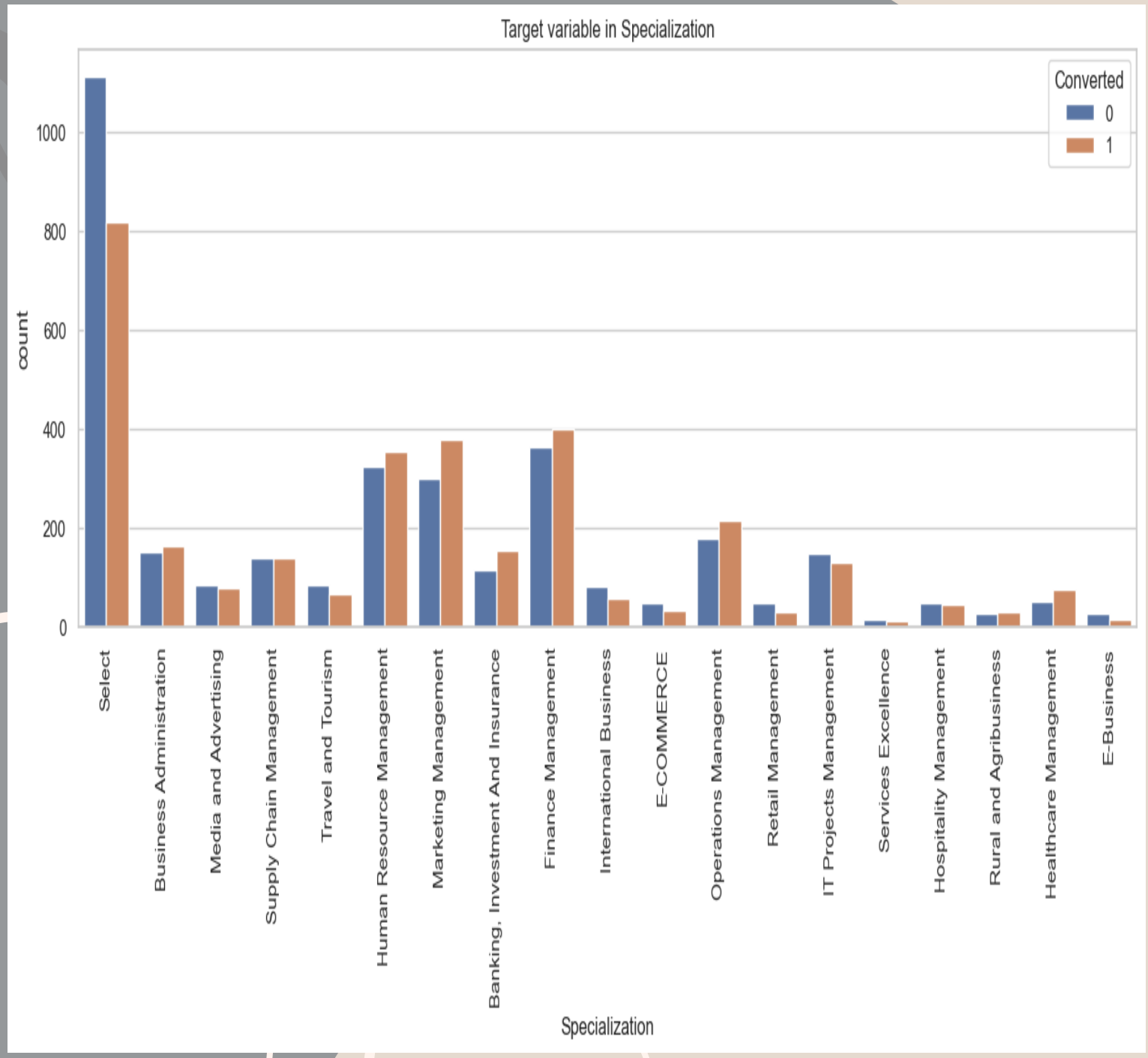
'Google' attracted more audience, but the 'Reference' And 'Welingak' website has higher conversion rate.


Target variable in Do Not Email
Converted
0
1
3000
2500
2000
1500
1000
500
0
count
No
Yes
Do Not Email



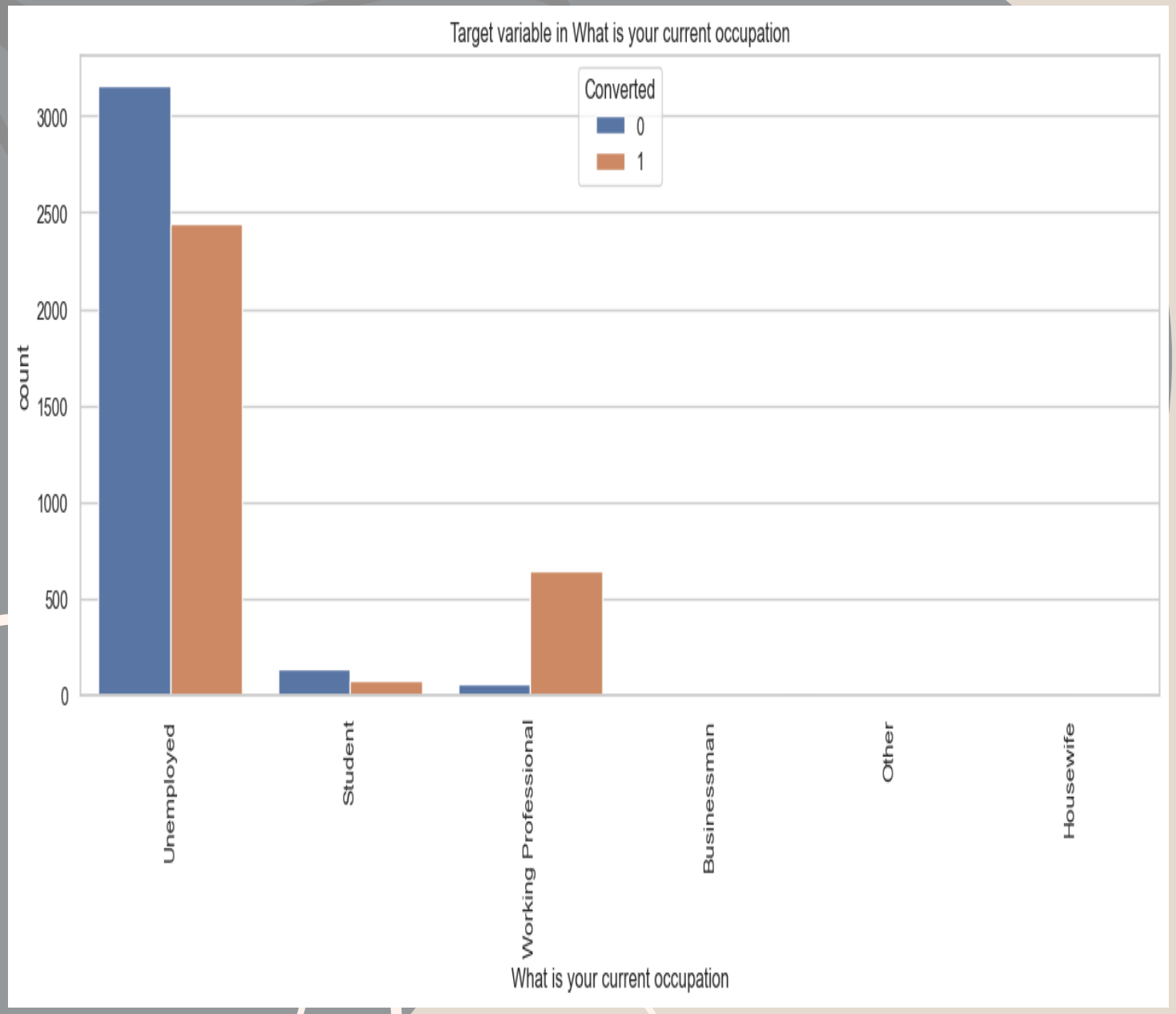
Last Activity

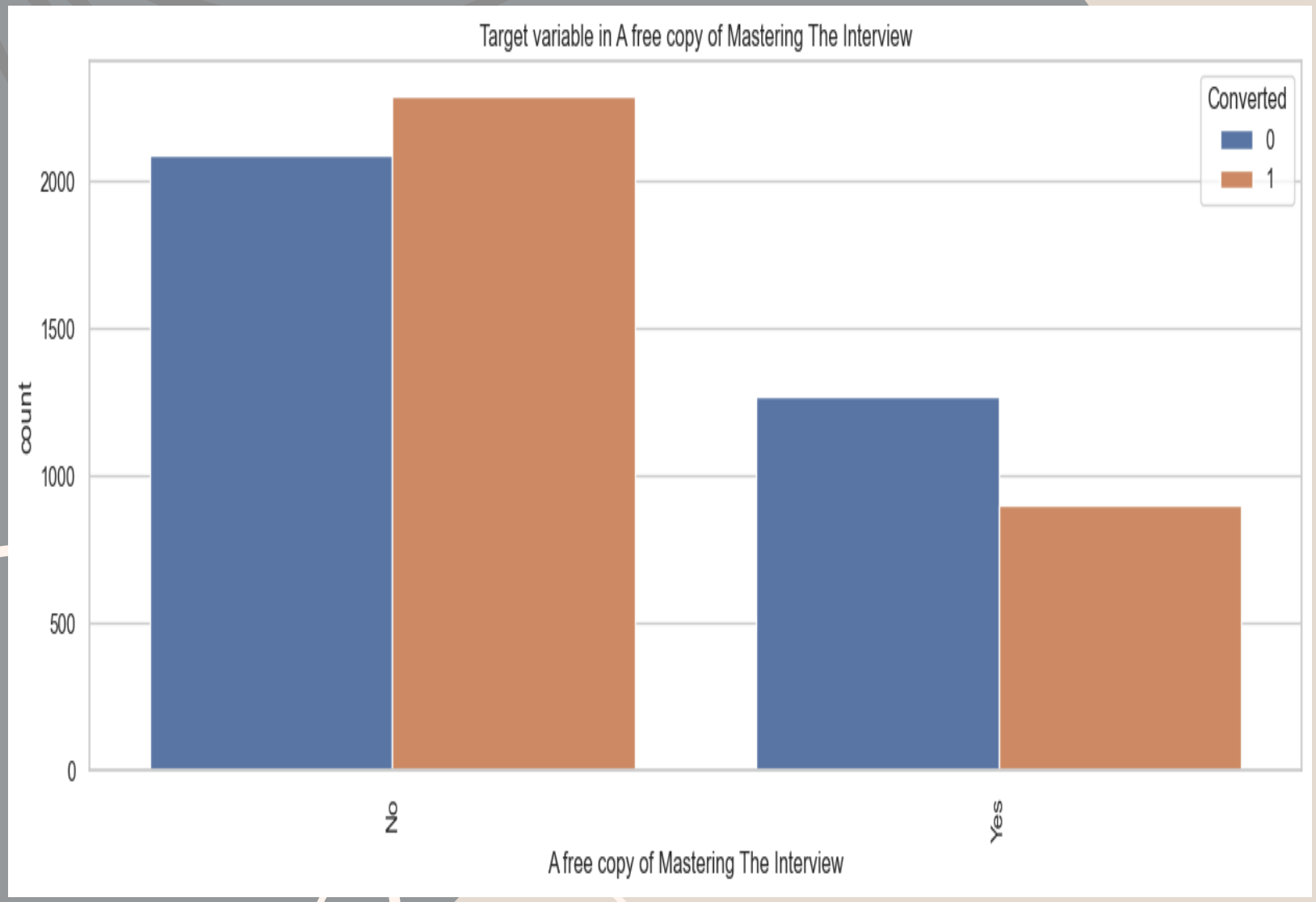
'Email Opened' has the highest range of customers, but conversion rat is high in 'SMS sent'.

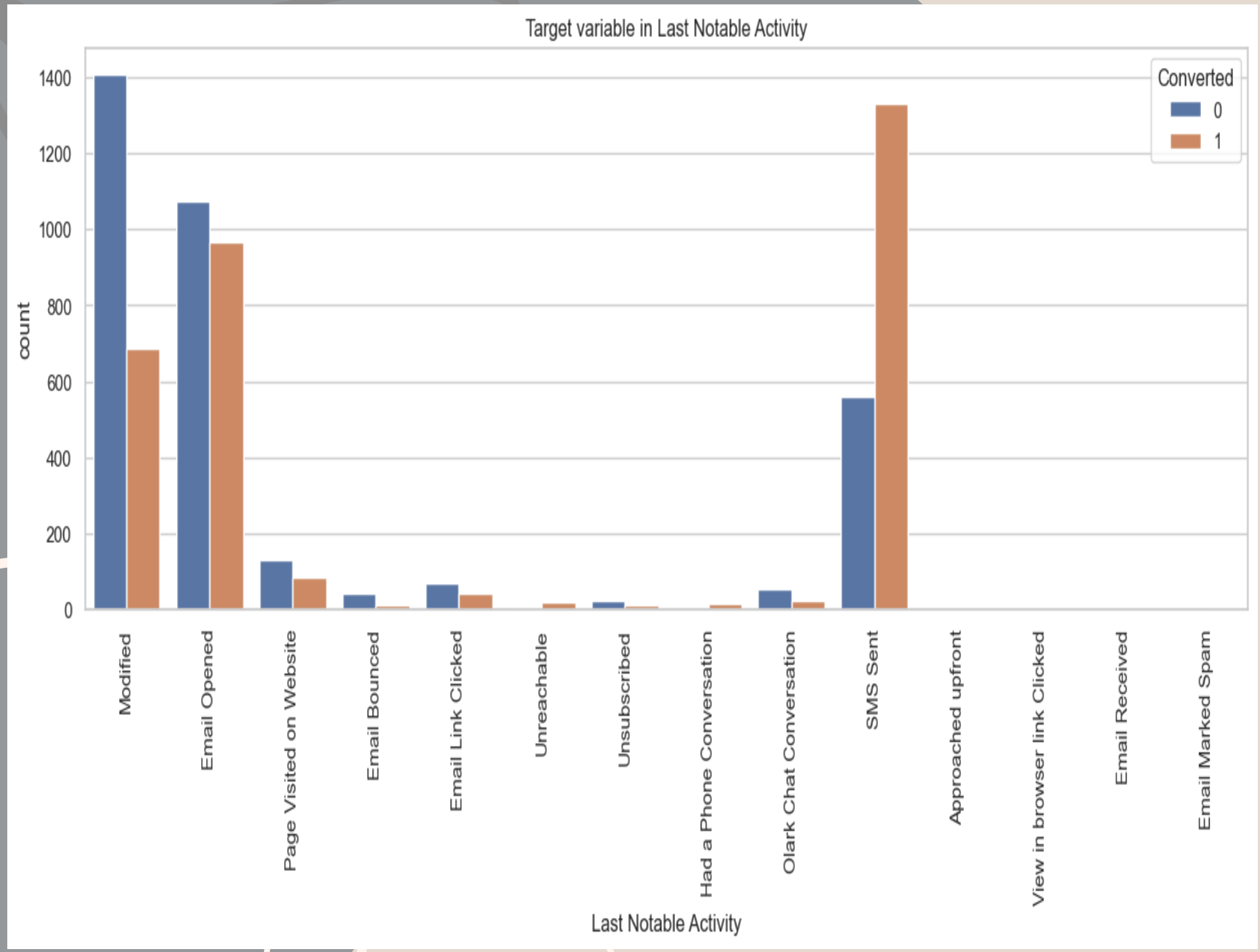


Specialization

Majority of people didn't select any specialization.so, from the remaining, we can see 'Marketing Management' has the highest conversion rate.


Target variable in What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Target variable in A free copy of Mastering The Interview
Converted
0
1
count
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview



Last Notable Activity

Highest is seen in 'Email Opened', but the conversion rate is higher in 'SMS Sent'

Heat Map

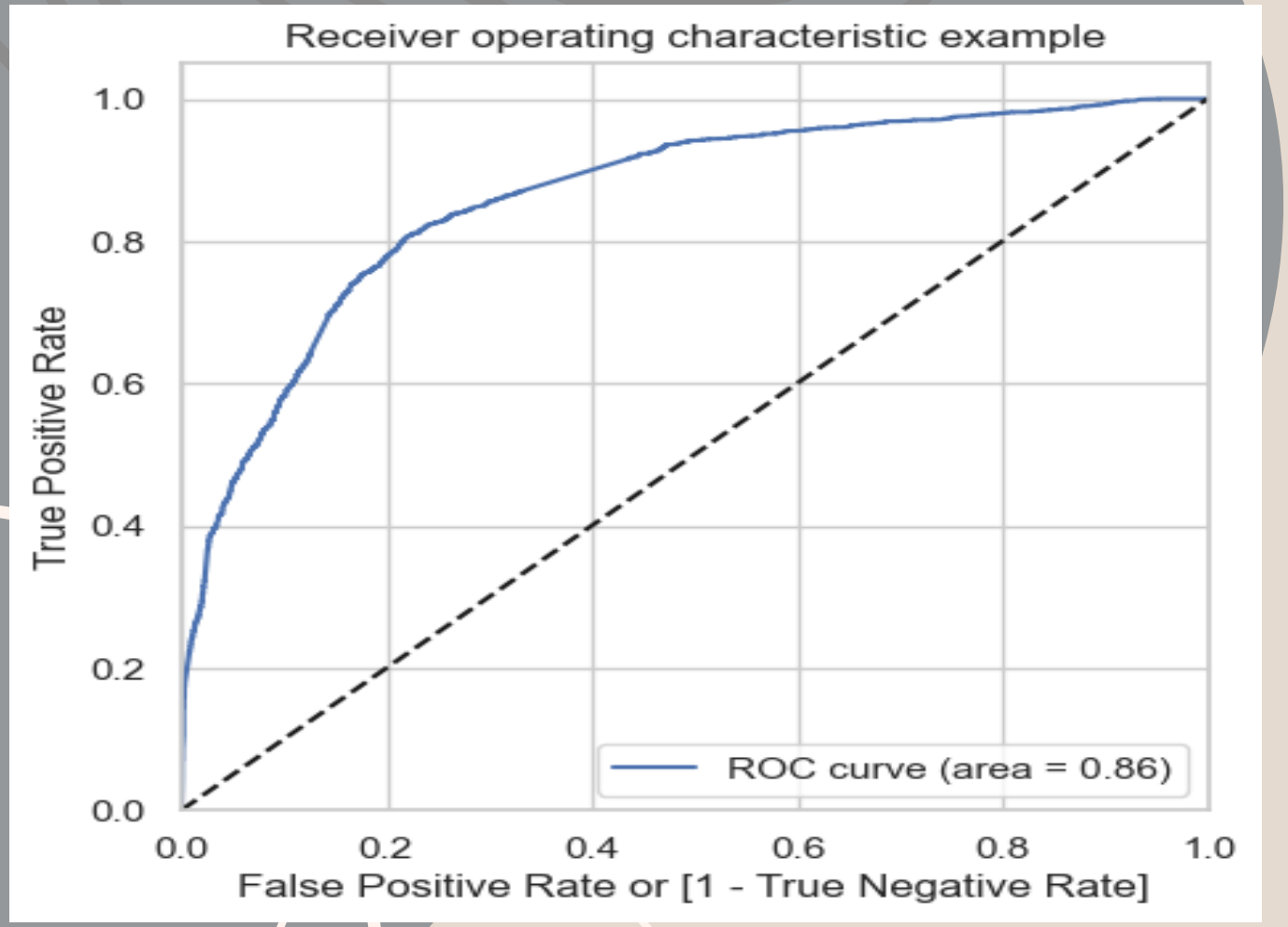
➢ `Page views Per Visit` and `Converted` has highest negative correlation of `-0.063`

➢ `Total Visits` and `Page Views Per Visit` shows the highest correlation of `0.49`

➢ `Total Time Spent on Website` and `Converted` shows` 0.31` correlation.

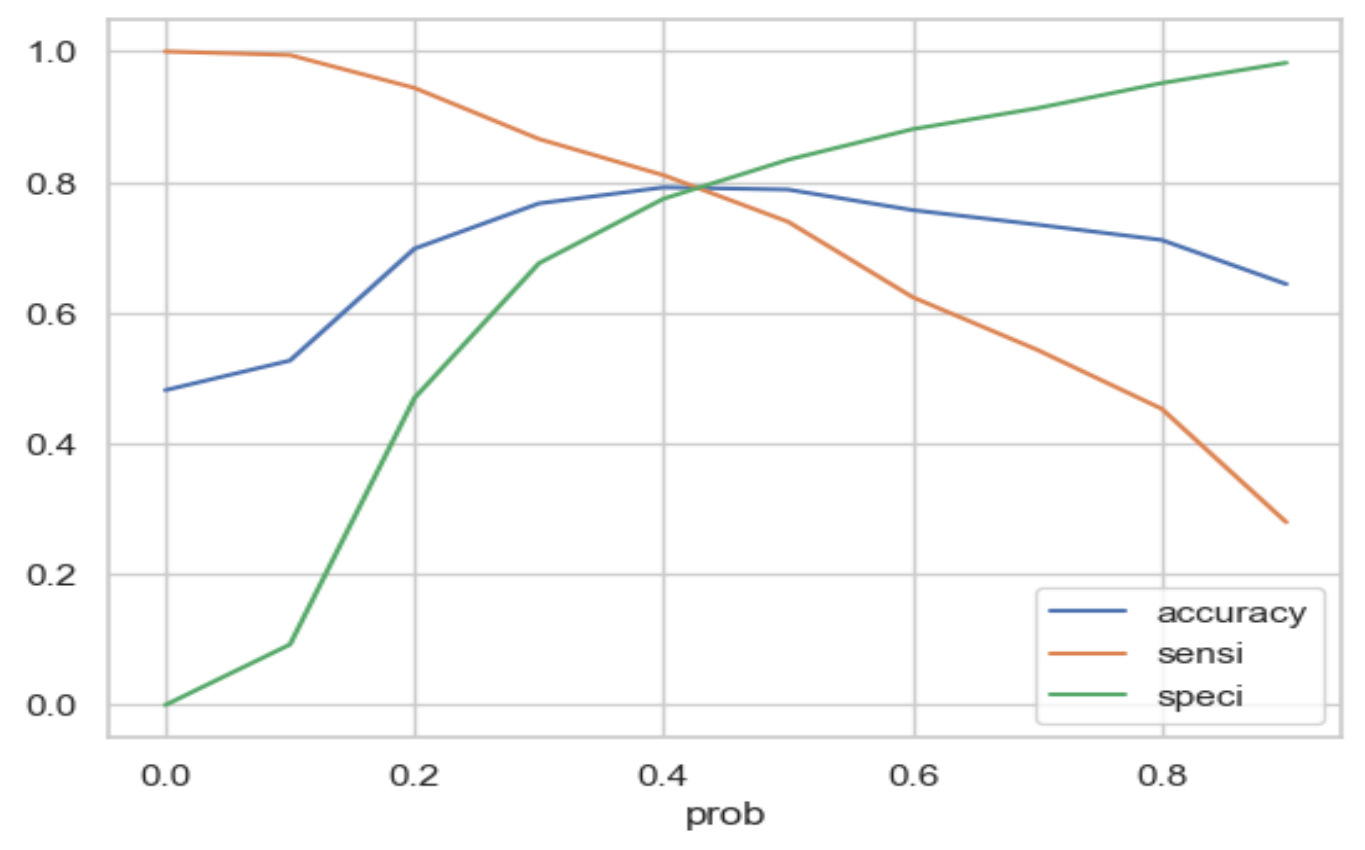
Model Building

- Splitting the data into Train and test sets
- Scaling the variables
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on p-values
- Check VIF value for all the existing columns
- Predict the train set
- Evaluate accuracy and other metrics
- Predict test set
- Precision and recall analysis



The ROC curve illustrates the trade-off between a binary classifier's true positive rate and false positive rate across varying classification thresholds, aiding performance evaluation.
From the ROC curve, the area under curve is 0.86 therefore our model is good.

Model Evaluation-Train set

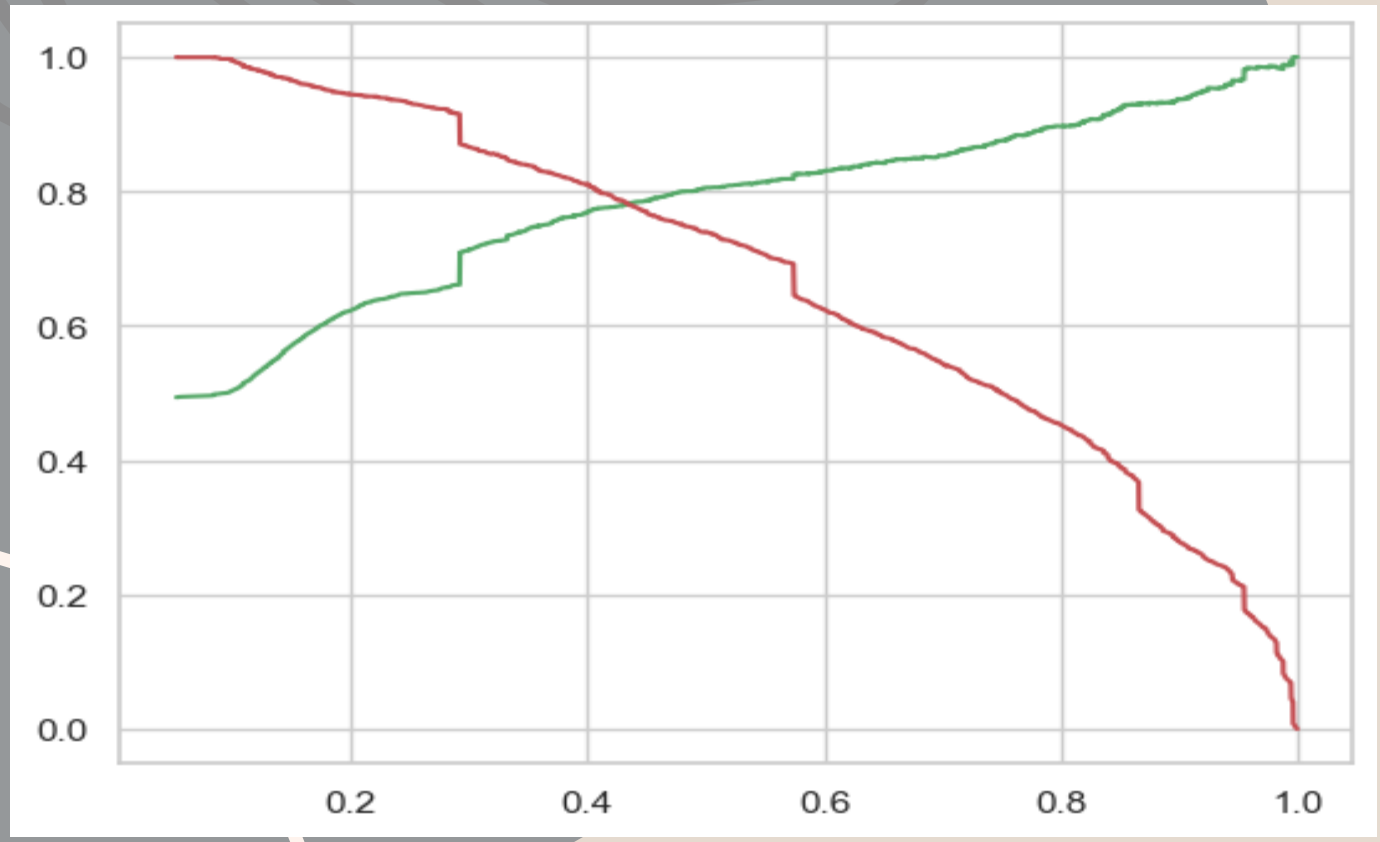


Confusion Matrix
1852 460
479 1670

Precision=78.4
Recall=77.71

Model Evaluation –Test Set

Precision-Recall Trade –off curve



Confusion Matrix

801 195

213 703

Precision=78.28

Recall=76.74

Summary

1. Lead scoring case study has been done using logistic regression model to meet the constraints as per business requirements.
2. There are a lot of leads in the initial stage but only a few of them are converted into paying customers. The most numbers of leads are from INDIA and in terms of city highest number are from Mumbai.
3. There are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. To get some useful data we have to make compulsory selection. Likewise, Customer occupation, Specialization, etc.
4. The high number of total visits & Total time spent on platform may increasing the chances of lead to be converted.
5. The leads are joined course for Better Career Prospects, most of having Specialization from Finance Management. Leads from HR, Finance & marketing management specializations are high probability to convert.
6. Talking to last notable Activity, making improvement in customer engagement through email & calls will help to convert leads. As the leads which are opening email have high probability to convert, Same as Sending SMS will also benefit.
7. Most of leads current occupation is Unemployed, which means gave more focus on unemployed leads.