LEAD SCORE ASSIGNMENT CASE STUDY

SUBMITTED BY :

Kartik Galhotra
Mohith Kune
Janvi Iyer

PROBLEM STATEMENT

The organization X Education is an educational company who sells online courses for professionals. X Education needs to select the correct Leads from the given applicants.

Even while the company generates a lot of leads, not many of those leads end up becoming clients. These leads are coming from various platforms, like Google, email, advertisements, etc.

The average conversion rate for the business is currently 30%, but the CEO wants to
raise it to 80%.

For that need to build a model which help to choose the correct Lead and achieve the target.

GOAL

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file.

Please fill it based on the logistic regression model you got in the first step.

Also, make sure you include this in your final PPT where you'll make recommendations.

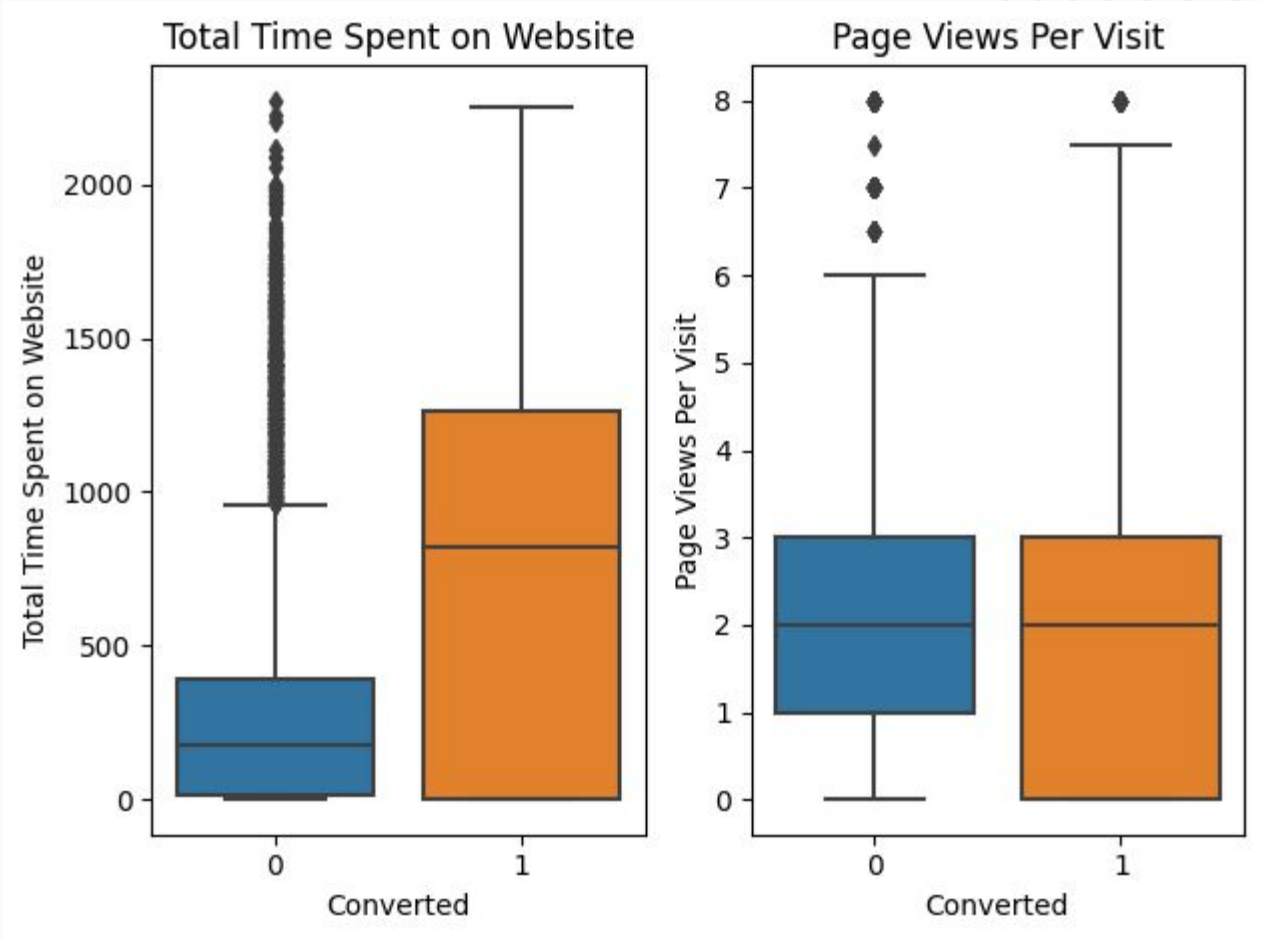
STEPS PERFORMED :

- Reading and Understanding the data
- Data Cleaning and Outlier Analysis
- Visualizing Data
- Creating dummy variable
- Splitting the Data into Training and Testing Sets
- Feature Scaling using Min/Max Scaling
- Looking at Correlations
- Feature Selection Using RFE
- Model Building-Assessing the model with StatsModels
- Creating Prediction
- Model Evaluation
- Plotting the ROC Curve ('Receiver Operating Characteristic’ Curve)
- Finding Optimal Cutoff Point
- Making predictions on the test set

EDA

We can observe that:

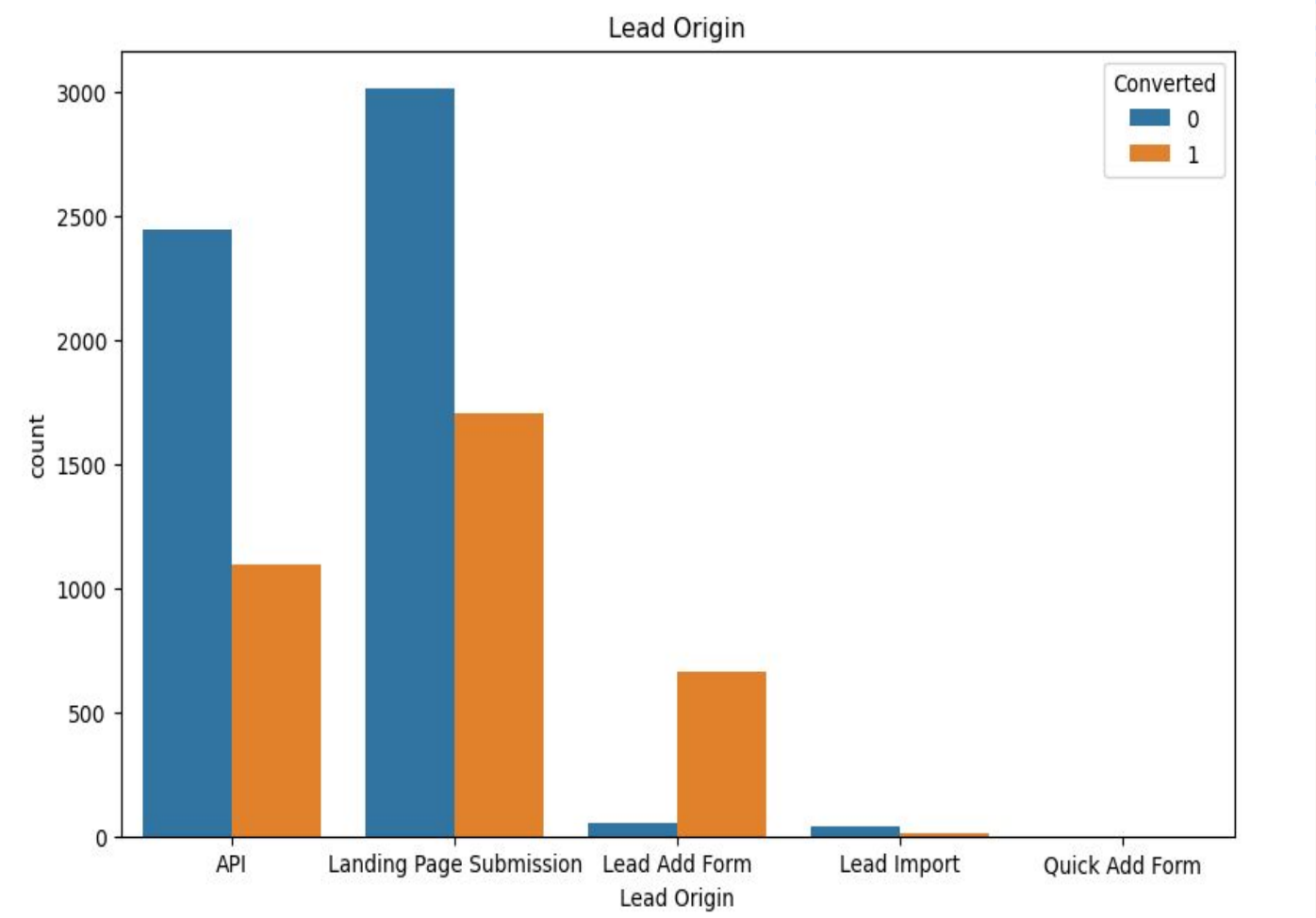
- People who Converted have spent larger time on the website than others.
- The medians and 75 quantiles for ‘Page Views per visit’ are same for both converted and non-converted.



We can observe that:

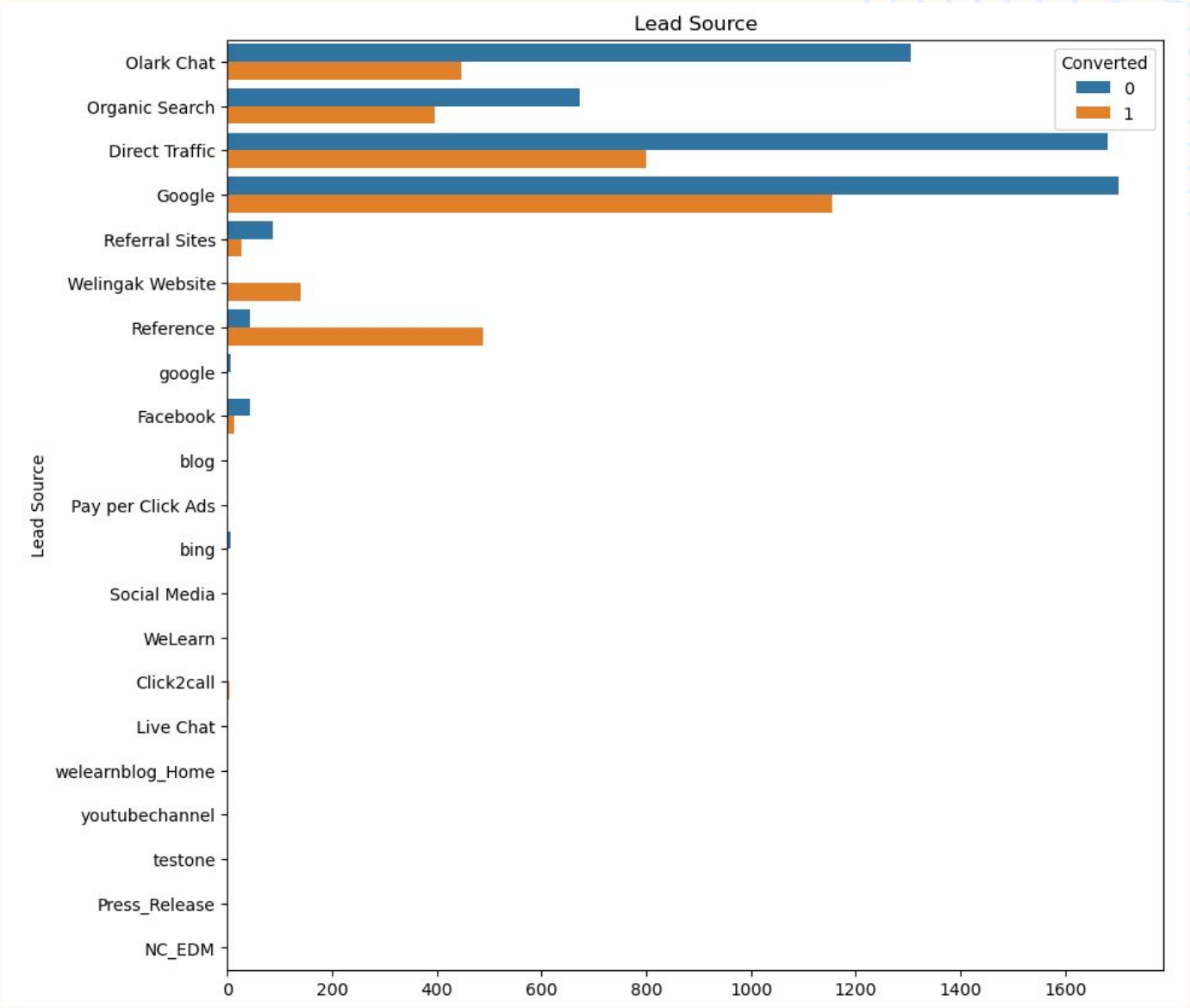
- Most of the Lead Origins have appeared with "Landing Page Submission" followed by "API" identifiers.

- The Conversion rate is higher with the identifier "Lead Add Form Lead Orign"



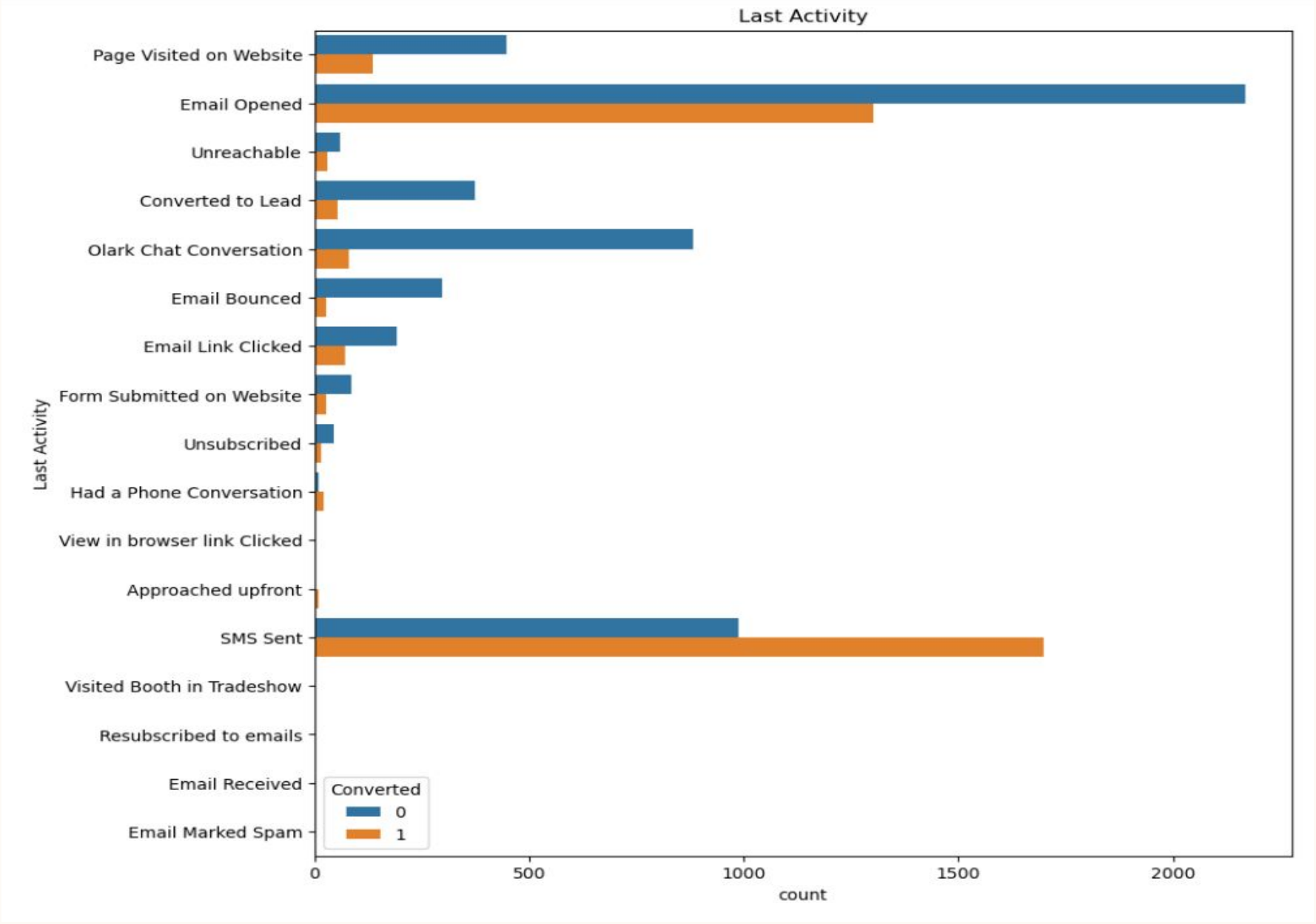
We can observe that:

- Most Lead Source are identified from "Google" followed by "Direct Traffic", "Olark Chart"
- "Google" has highest conversion rate, followed by "Direct Traffic", "Olark Chart"
- Most difference in Converted and non-Converted ones can be seen in "Reference" Lead source.



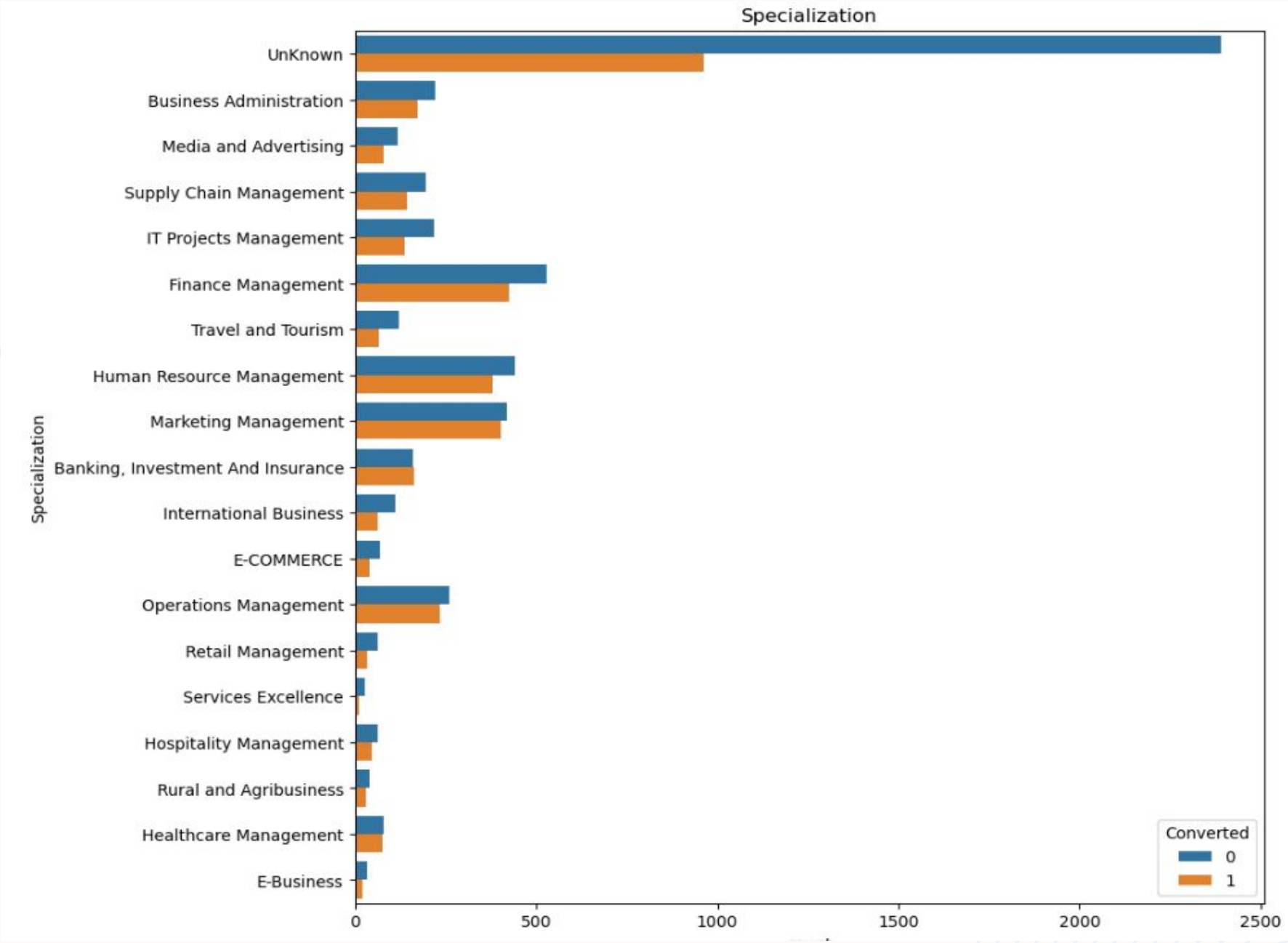
We can observe that:

- "Email Opened" and "SMS Sent" are the Most Last Activity of customers.
- Those whose Last Activity was SMS sent had the best conversion rate.



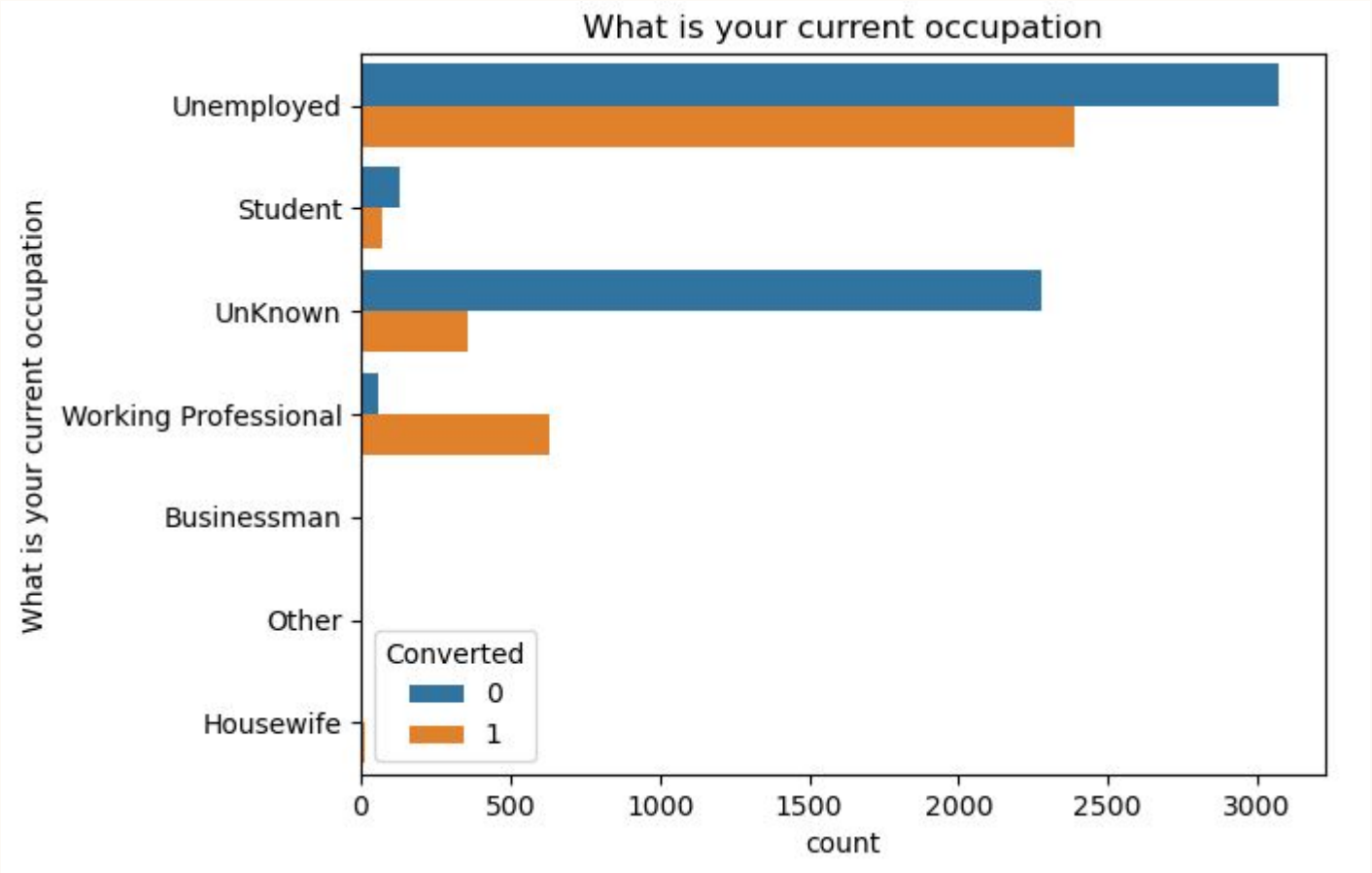
We can observe that:

- Whose Specialization is unknown has the highest rate of conversion.
- Other than "Unknown"s high conversion is seen in "Finance Management", "human resource management", "Marketing management"



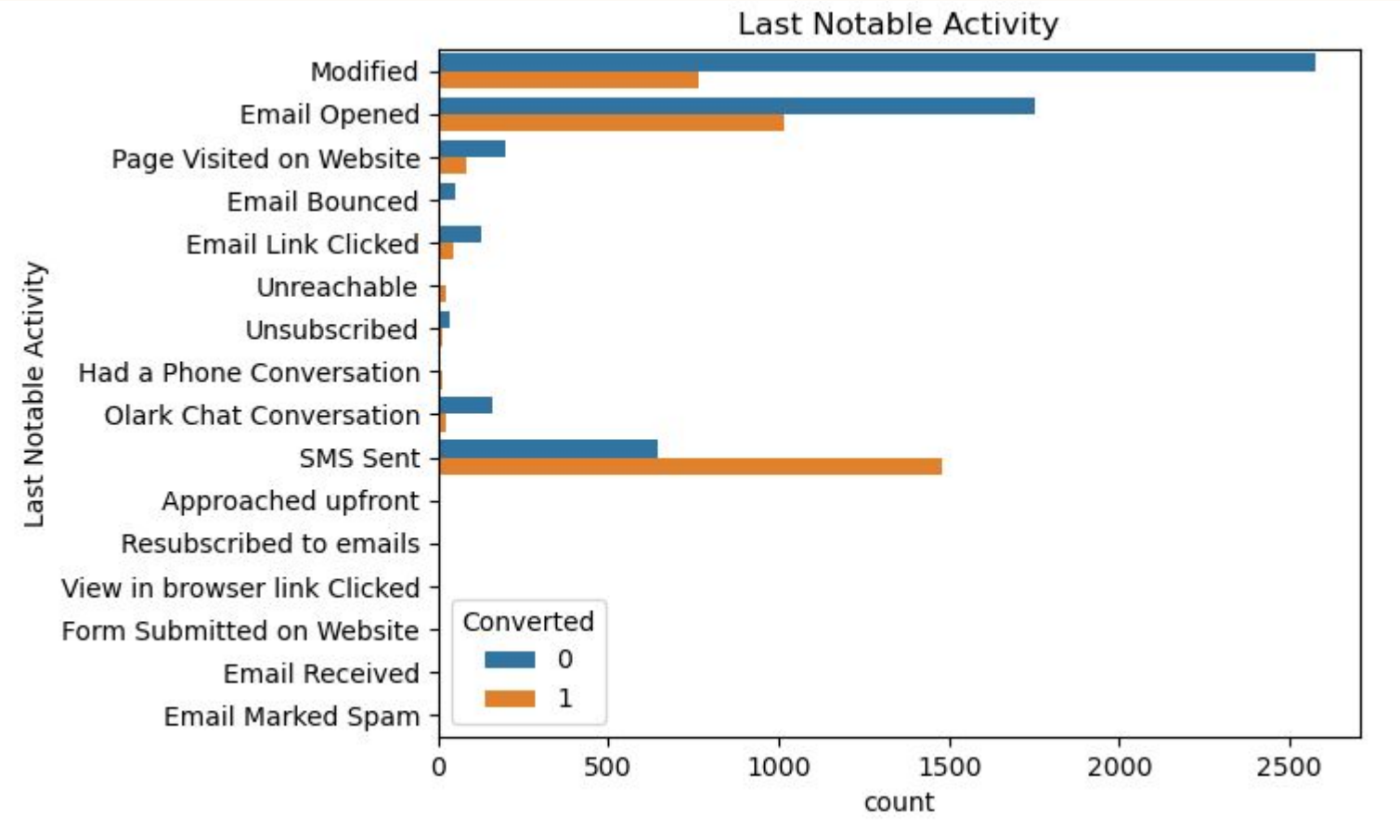
We can observe that:

- Person who are unemployed has the highest conversion rate comparatively to working professional.
- Those whose occupation in "Unknown" has high number of non-conversions.
- Among "Working Professionals" very high conversion rate is seen.



We can observe that:

- Those whose Last Notable Activity was found to be "SMS Sent" have the best conversion rate, followed by "Modifier", "Email Opened".



VARIABLE IMPACTING THE CONVERSION

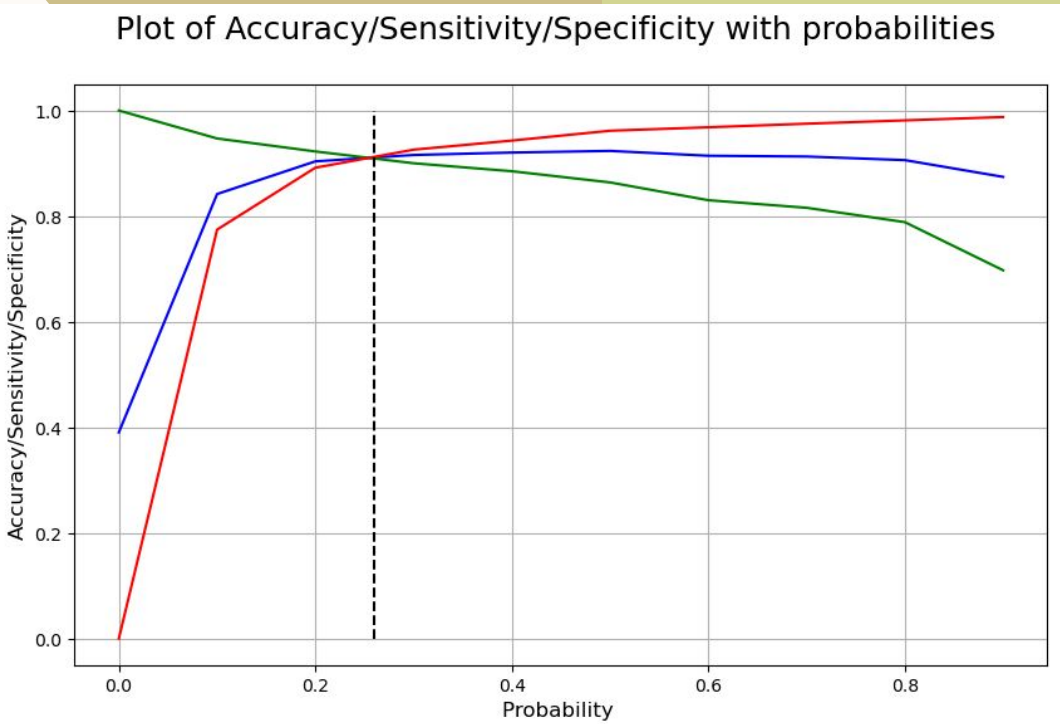
	coef
const	-2.5206
Total Time Spent on Website	3.0722
Lead Source_Welingak Website	5.4522
Last Activity_Email Bounced	-1.6763
What matters most to you in choosing a course_UnKnown	-0.6779
Tags_Busy	0.7109
Tags_Closed by Horizzon	6.7644
Tags_Lost to EINS	5.2938
Tags_Ringing	-3.4247
Tags_Will revert after reading the email	4.6936
Tags_switched off	-4.6942
Last Notable Activity_SMS Sent	2.8158

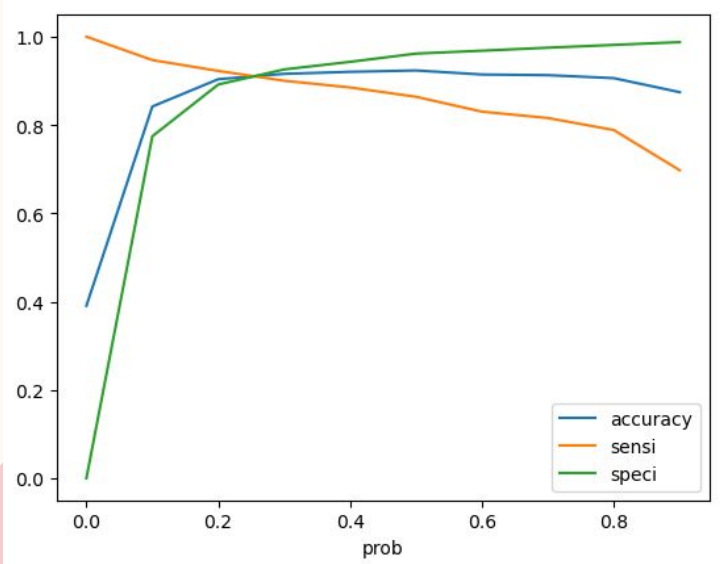
MODEL EVALUATION TRAIN AND TEST SET

- The probability cut off threshold value is approximately 0.26

- Metrics on Train Data:
  - Accuracy = 0.91
  - Precision = 0.87
  - Sensitivity = 0.91
  - Specificity = 0.92

- Metrics on Test Data:
  - Accuracy = 0.91
  - Precision = 0.85
  - Recall/sensitivity = 0.92
  - Specificity = 0.91





CONCLUSION

After reviewing our model, we can see that the accuracy, sensitivity, and specificity values for both the train and test data have been around 91%.

The prediction was done on an optimal cut off of 0.26.

We found that the following columns matter the most for our evaluation:

- Total Time Spent on Website
- Tags: Whether the Lead is tagged by
  - "Ringing"
  - "Closed by Horizzon"
  - "Lost to EINS"
  - "Will revert after reading the email"
  - "Switched off"
- Lead Source_Welingak Website
- Last Notable Activity_SMS Sent
- Last Activity_Email Bounced