Lead Score Assignment Case Study

Submitted by:

Dishant Bahuguna

Pandiarajan Nammalwar

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 chose 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.

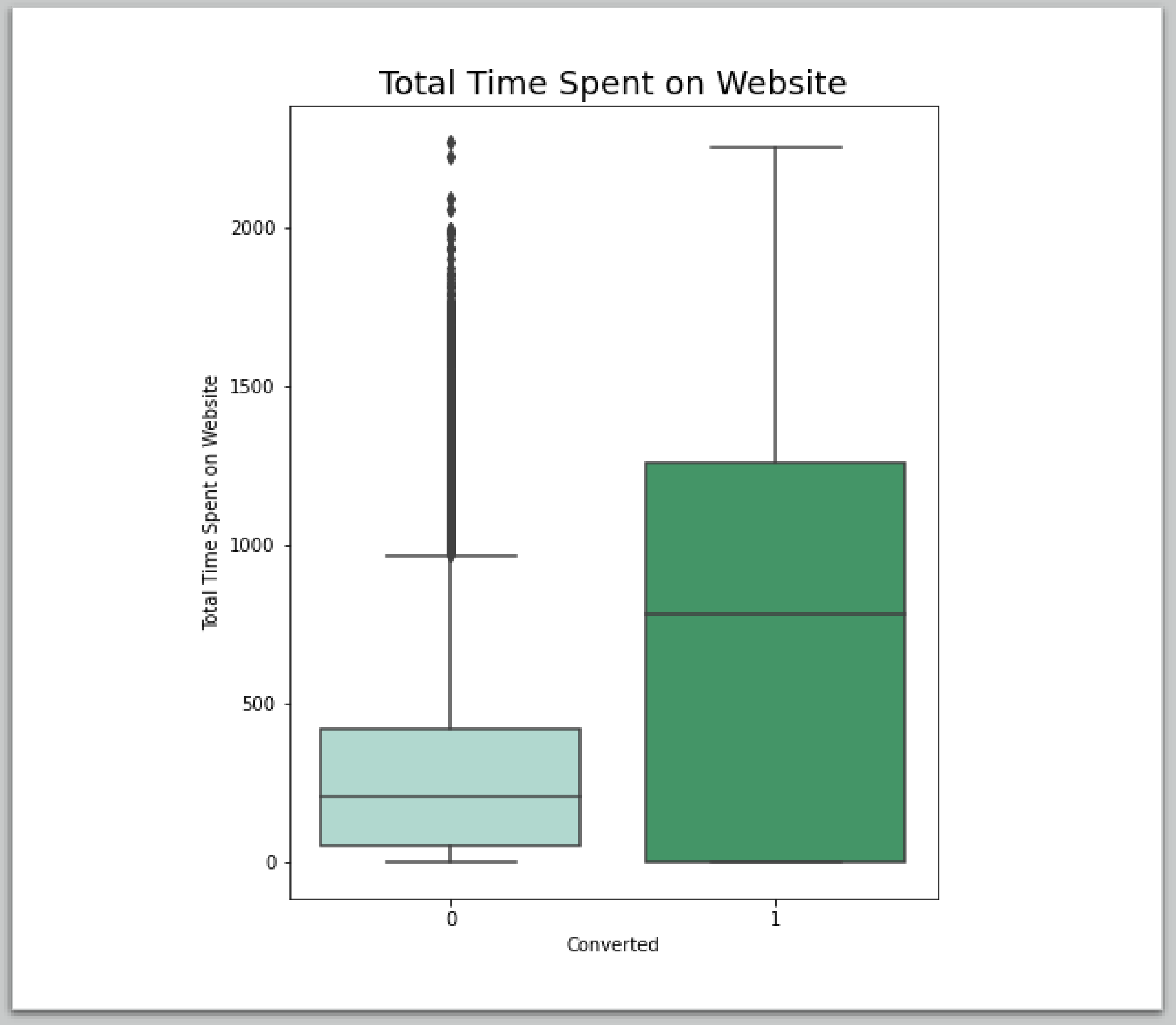
Steps Performed

- 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

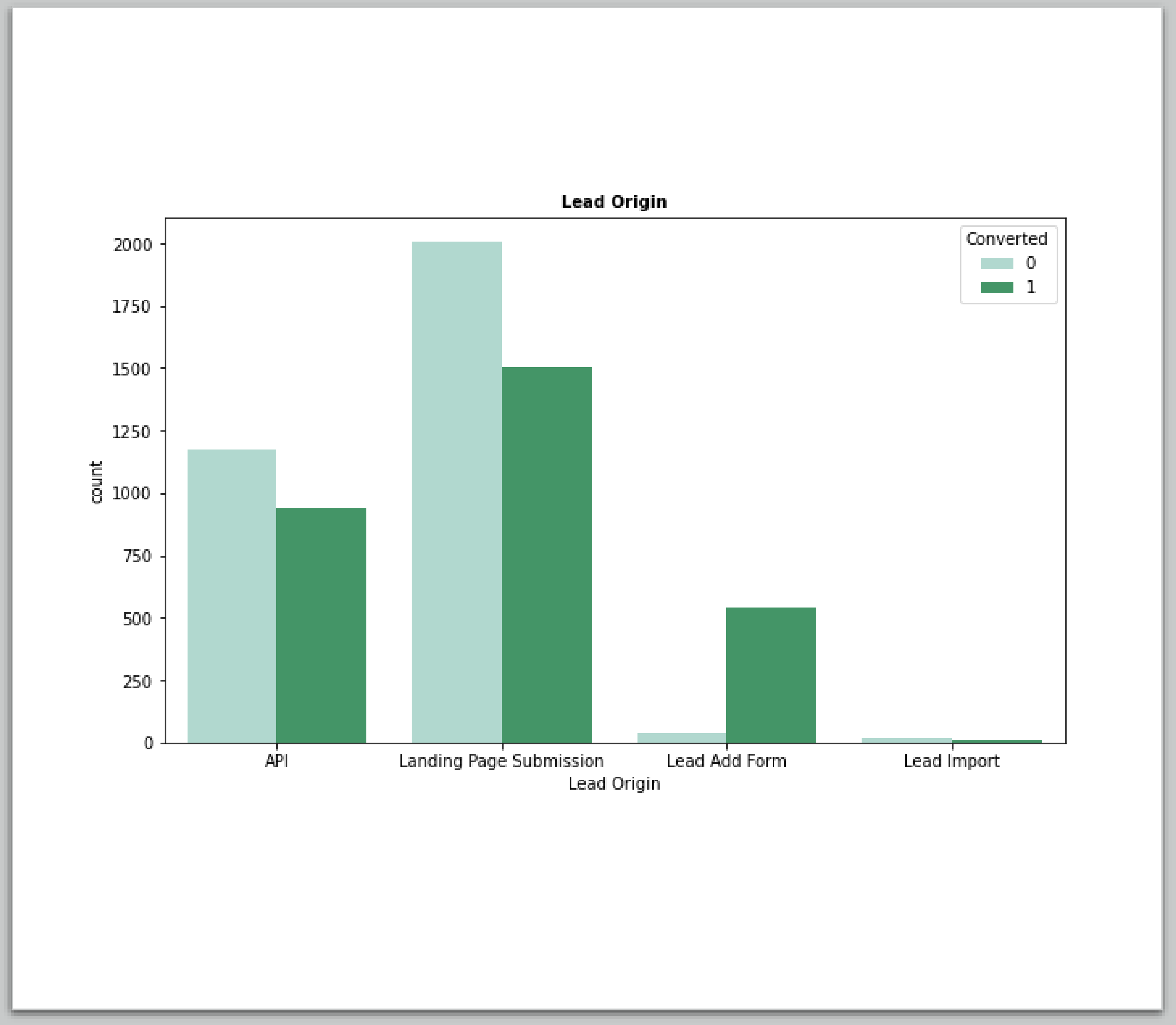
EDA

Total Time Spent on Website vs Converted:

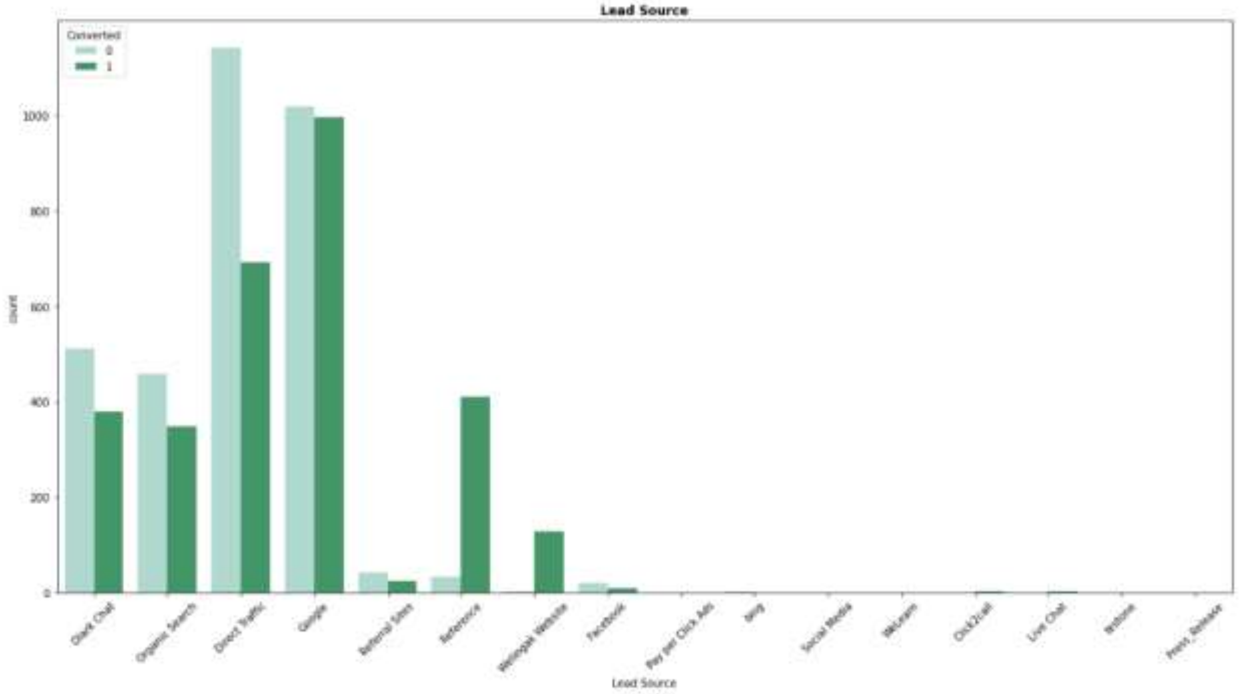
People spending more time are promising Leads



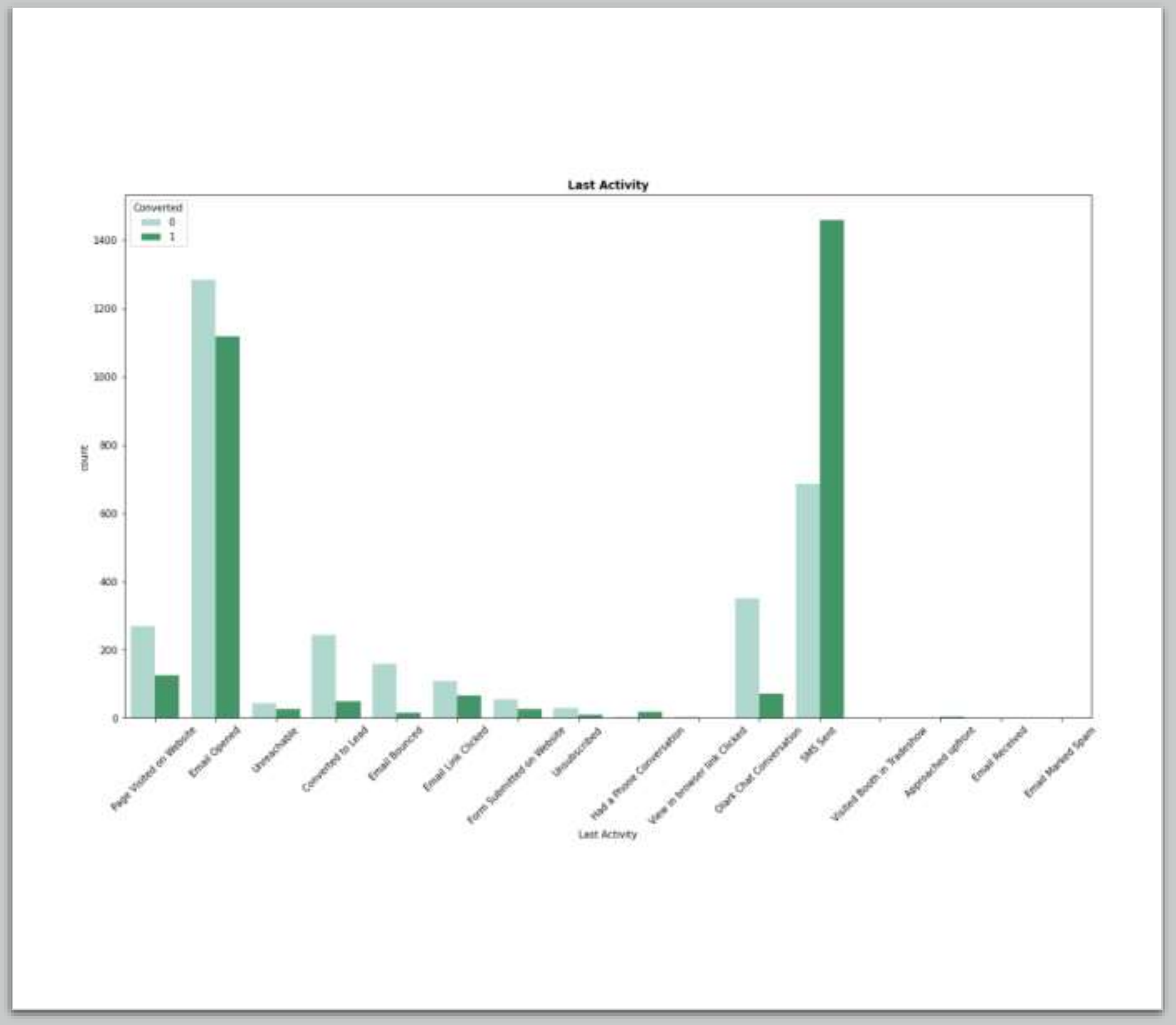
The Lead Origin- Landing Page Submission has the highest conversion rate among others.



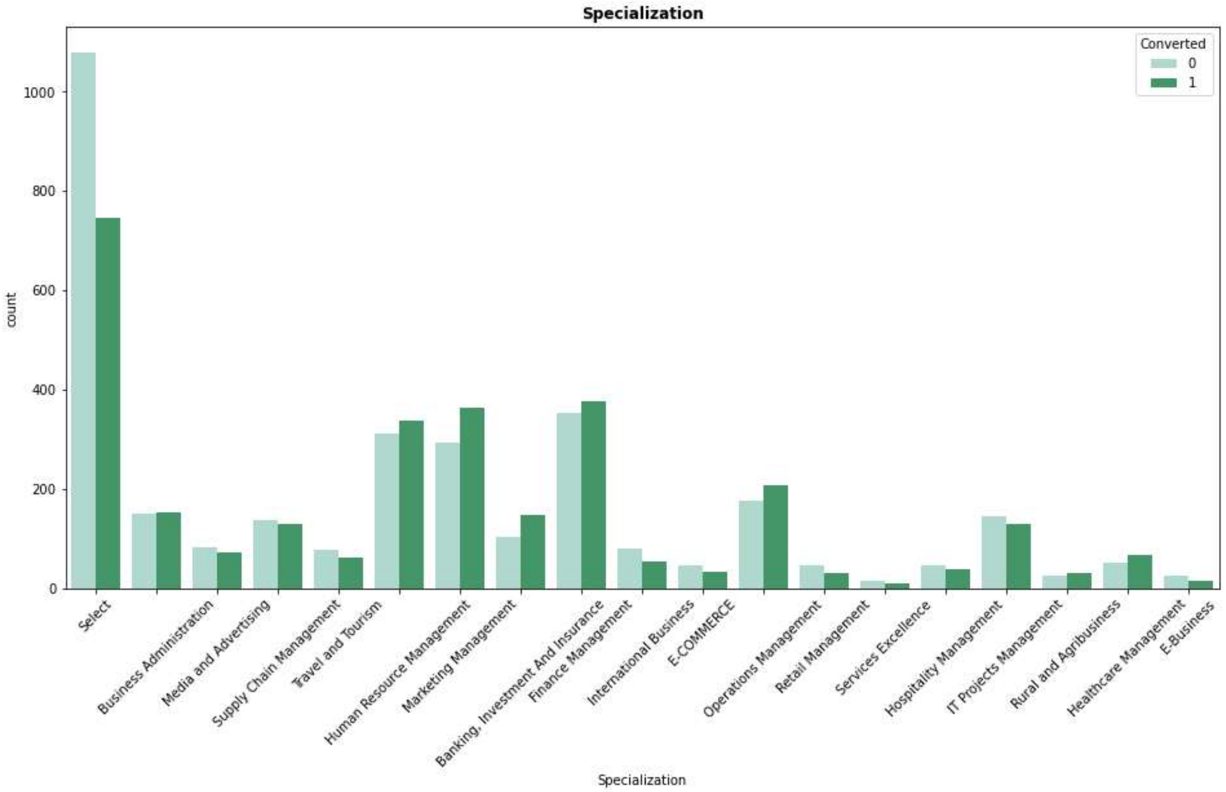
Google has the highest conversion rate.



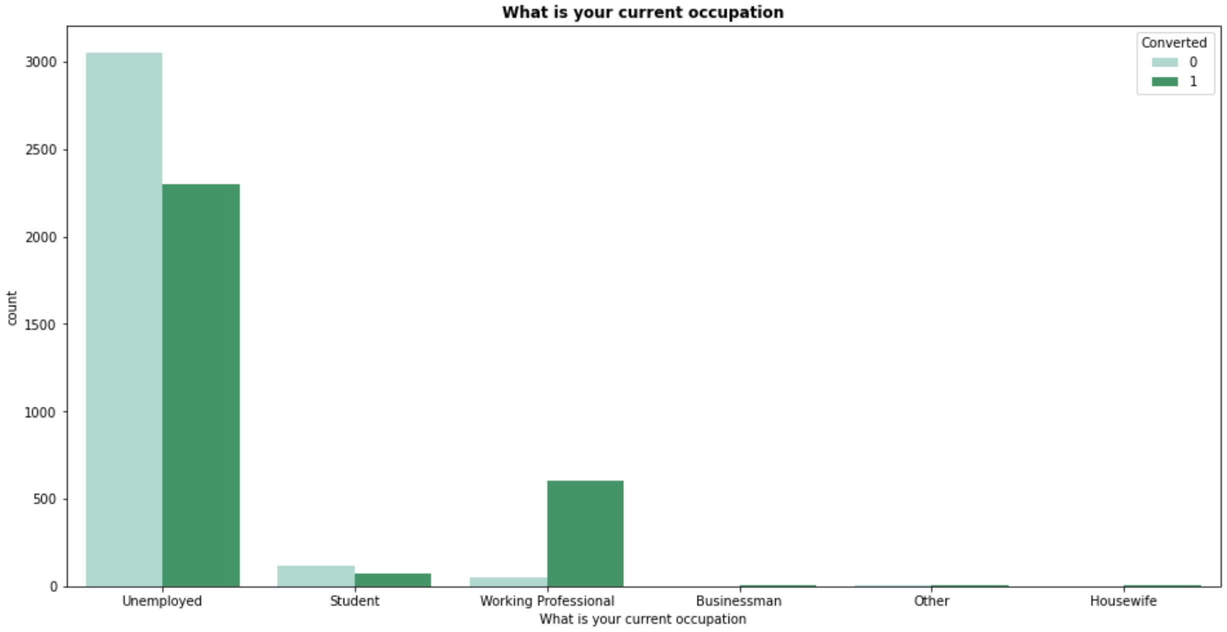
Leads whose Last Activity was SMS sent had the best conversion rate.



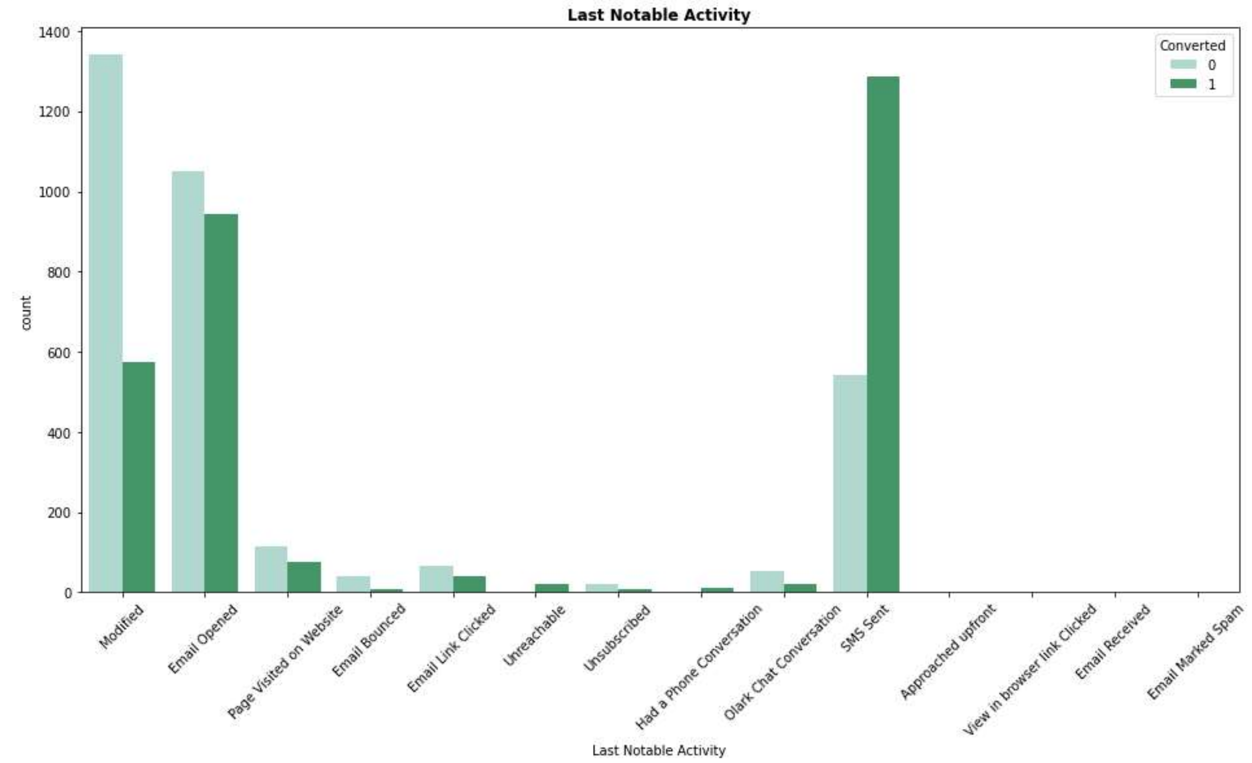
Lead from Specialization who are unknown/Select columns has the highest rate of conversion.



Person who are unemployed has the highest conversion rate comparatively to working professional.



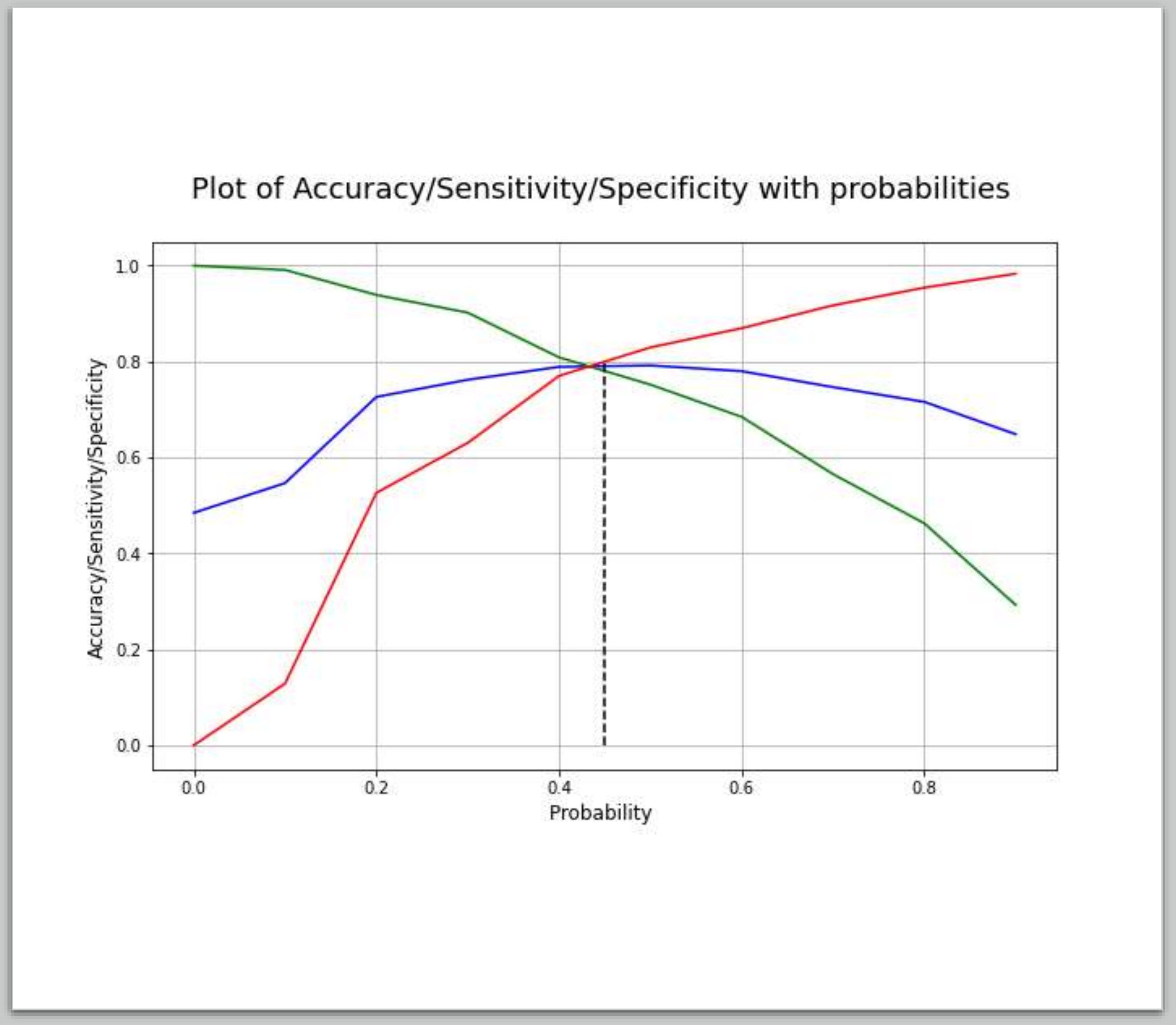
Students whose Last Notable Activity was found to be SMS Sent had the best conversion rate.



Variable Impacting The Conversion

- Total Time Spent on Website
- TotalVisits
- LastActivity_SMS Sent
- LeadOrigin_Lead Add Form
- LeadSource_Welingak Website
- LeadSource_Olark Chat
- CurrentOccupation_Working Professional
- LastActivity_Olark Chat Conversation
- Do Not Email
- LastActivity_Had a Phone Conversation
- LastNotableActivity_Unreachable

MODEL EVALUATION TRAIN AND TEST SET



TRAIN SET:

Accuracy = 0.79072

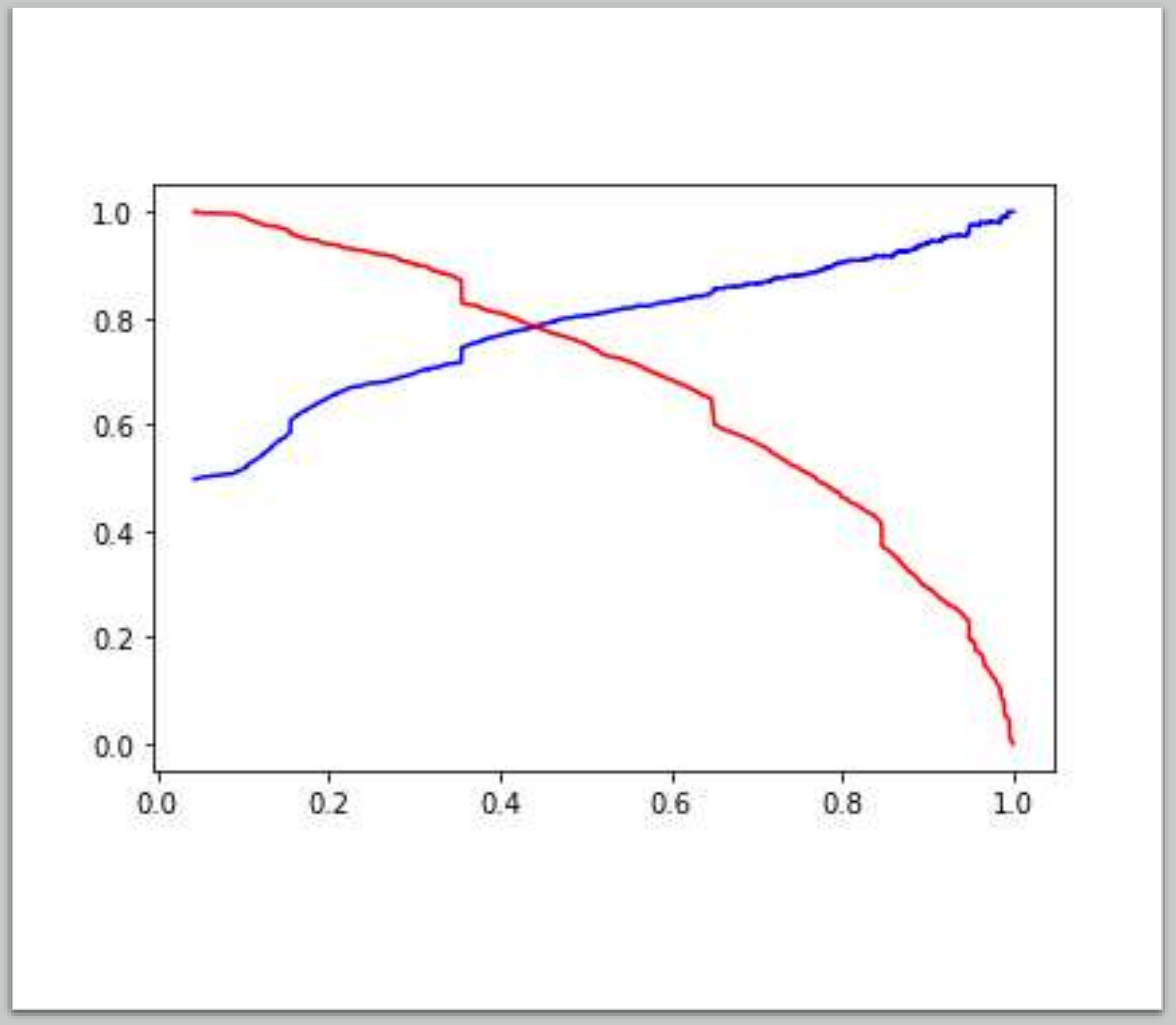
Sensitivity = 0.79071

Specificity = 0.79073

Precision = 78%

Recall =79%

The cut-off value is approximately 0.43



TEST SET:

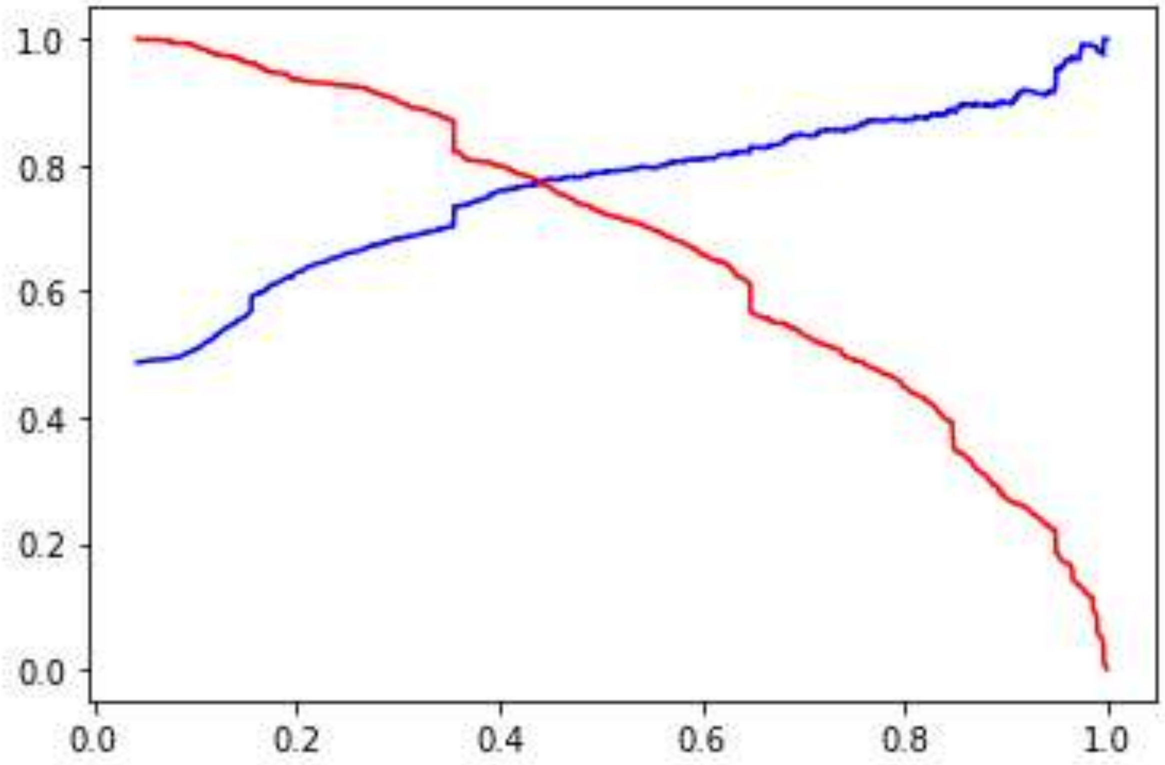
Accuracy = 0.785

sensitivity = 0.799

Specificity = 0.773

Precision = 77%

Recall = 78%



Conclusion

After assessing our model, we can see that the accuracy, sensitivity, and specificity values for both the train and test data have been around 79%.

The prediction was done on an optimal cut off of 0.43.

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

1. Current Occupation – Unemployed
2. Total time spent on website
3. LeadOrigin_Lead Add Form
4. Last Activity as SMS sent
5. Lead Source as Olark chat
6. TotalVisits