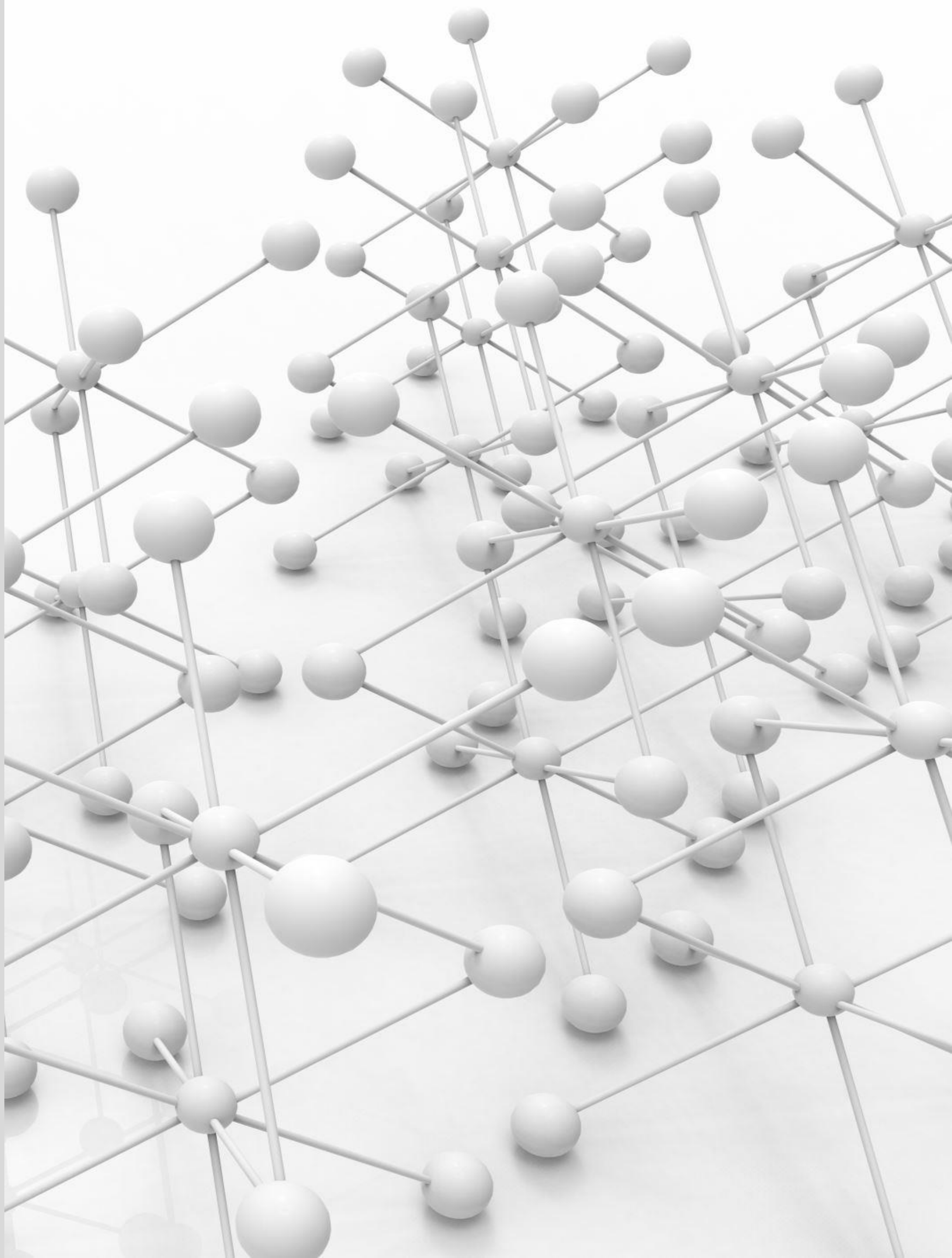
LEAD SCORE CASE STUDY

Submitted By :
Sonam Kulshrestha



Problem Statement

X Education Company sells online courses to industry professionals.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The **typical lead conversion rate** at X education is around **30%**.

X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as '**Hot Leads**'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.
- We should be able to identify and return the most prominent leads who will result in being a payable customer to the X Education Company

Assumptions

- The features having NULL/NAN values more than 40% are removed as most of the data having NULL cannot be imputed and will completely alter the analysis if done so.
- Features having majority of the data as null are also dropped.
- The features having the value as “Select” is treated and imputed as “Not Provided” due to the fact that these might be the default values and we should not make predictions on that.
- The dummy features created for “Not Provided” are also dropped.
- The numeric features are scaled in order to normalize the data and use it for further analysis

Solution Approach

1. **Data Cleaning :**
   - This stage is where we clean the data, removing the Null or Nan values. We removed / imputed based on the values and respective percentage of Null values.
   - The data having wrong values like “Select” are corrected.
2. **Checking Outliers :**
   - After correction, we analyzed the outliers and observed how to treat them wherever required.
3. **Univariate Analysis :**
   - Analyzed categorical columns and performed univariate analysis to analyze patterns in data.
   - Analyzed numeric columns and performed univariate analysis to analyze data patterns
4. **Bivariate Analysis :**
   - Analyzed multiple features and got the respected patterns and observations to understand and help in decision making.

Solution Approach (cont.)

5. Data Modelling :

- This stage is where we **Scale** the **numeric** features in order to keep the data normalized for our analysis.
- We also created **Dummy variables** for our categorical features to help our analysis.
- Building the Linear Regression model and used it for predictions.

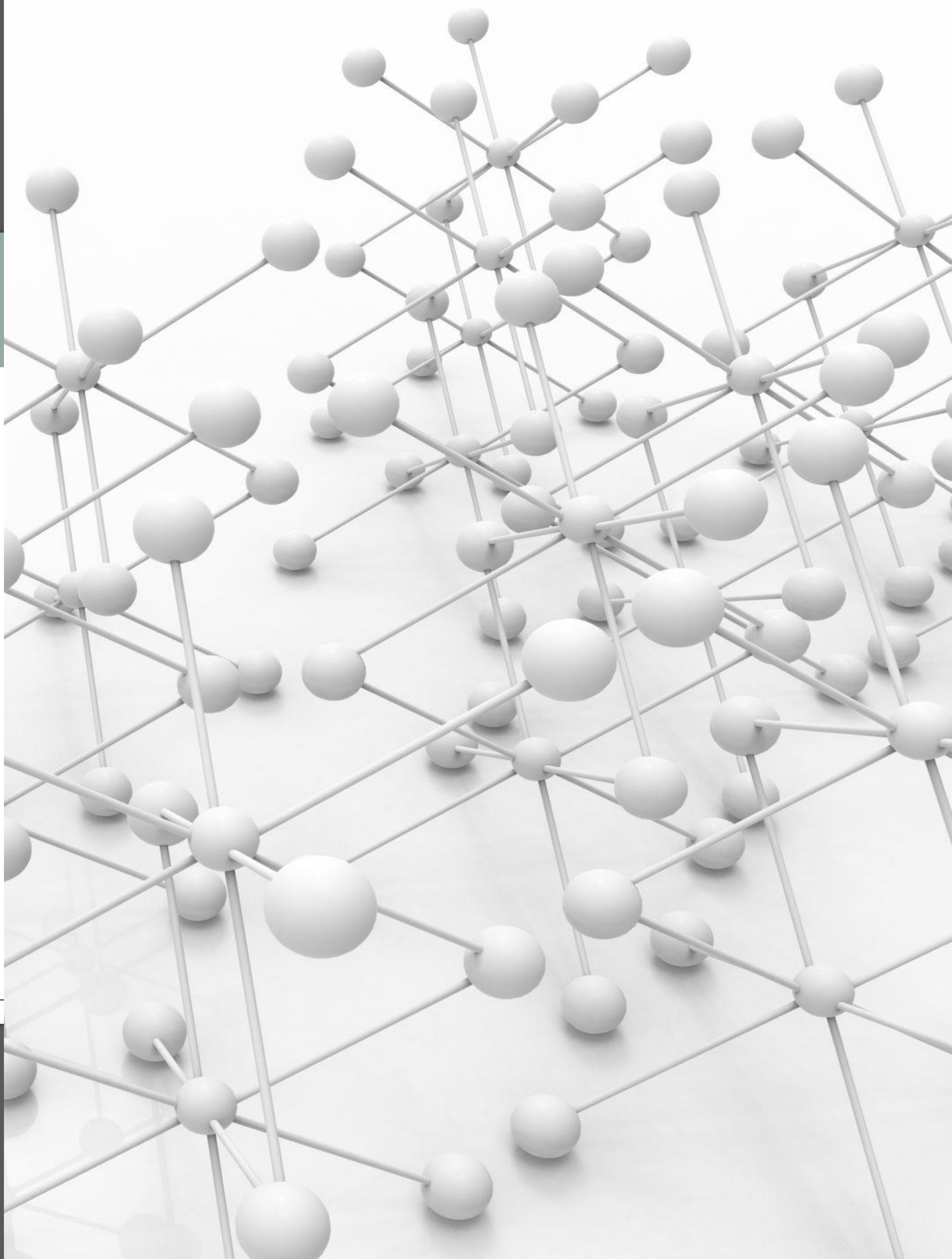
6. Model Evaluation :

- After we build the model, we evaluate it using different Evaluation metrics like **Sensitivity, Specificity, Prediction and Recall**.

7. Making Predictions :

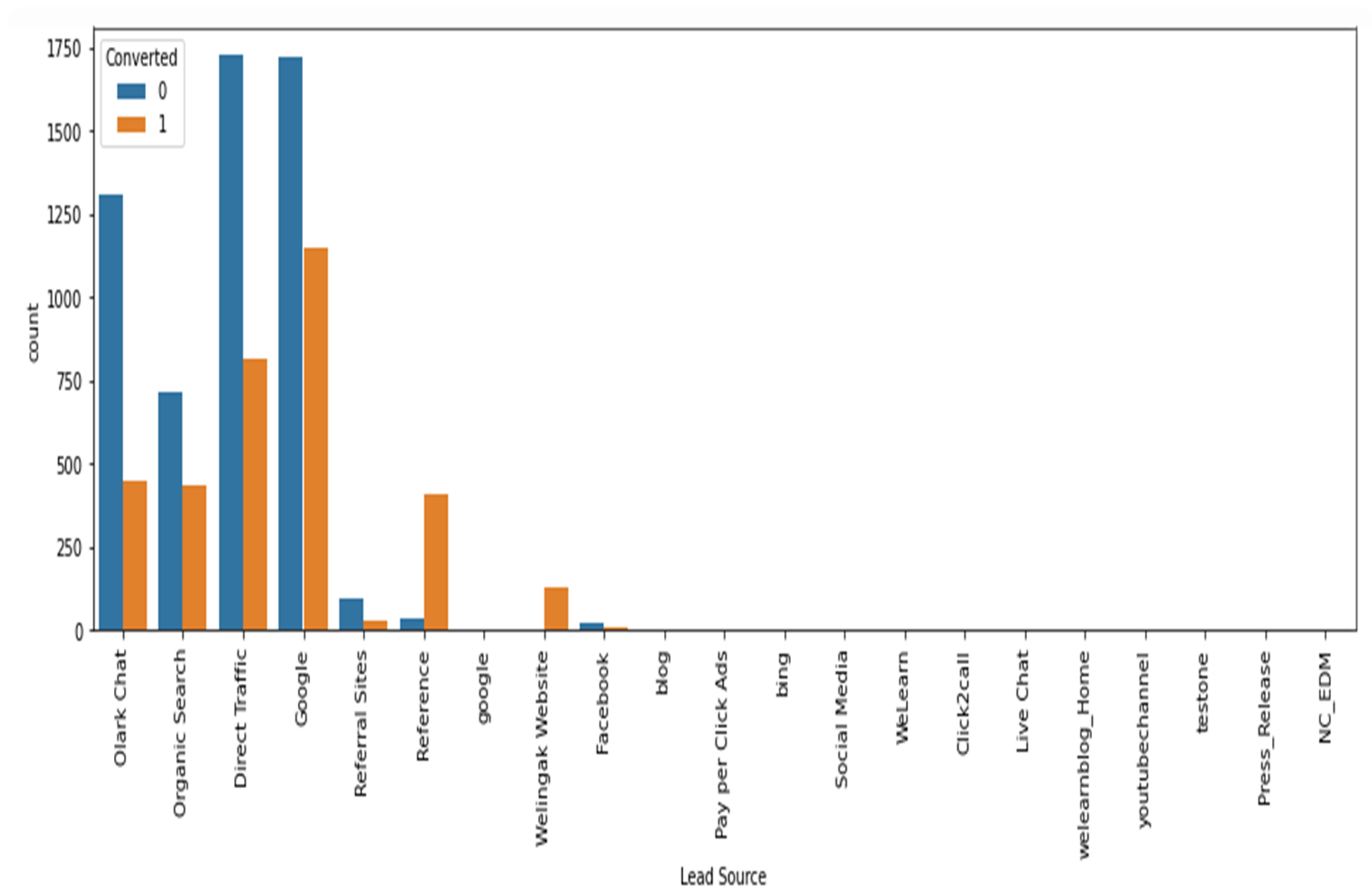
- After finalizing and getting the model, we use it for Test data and making predictions.

EXPLORATORY DATA ANALYSIS



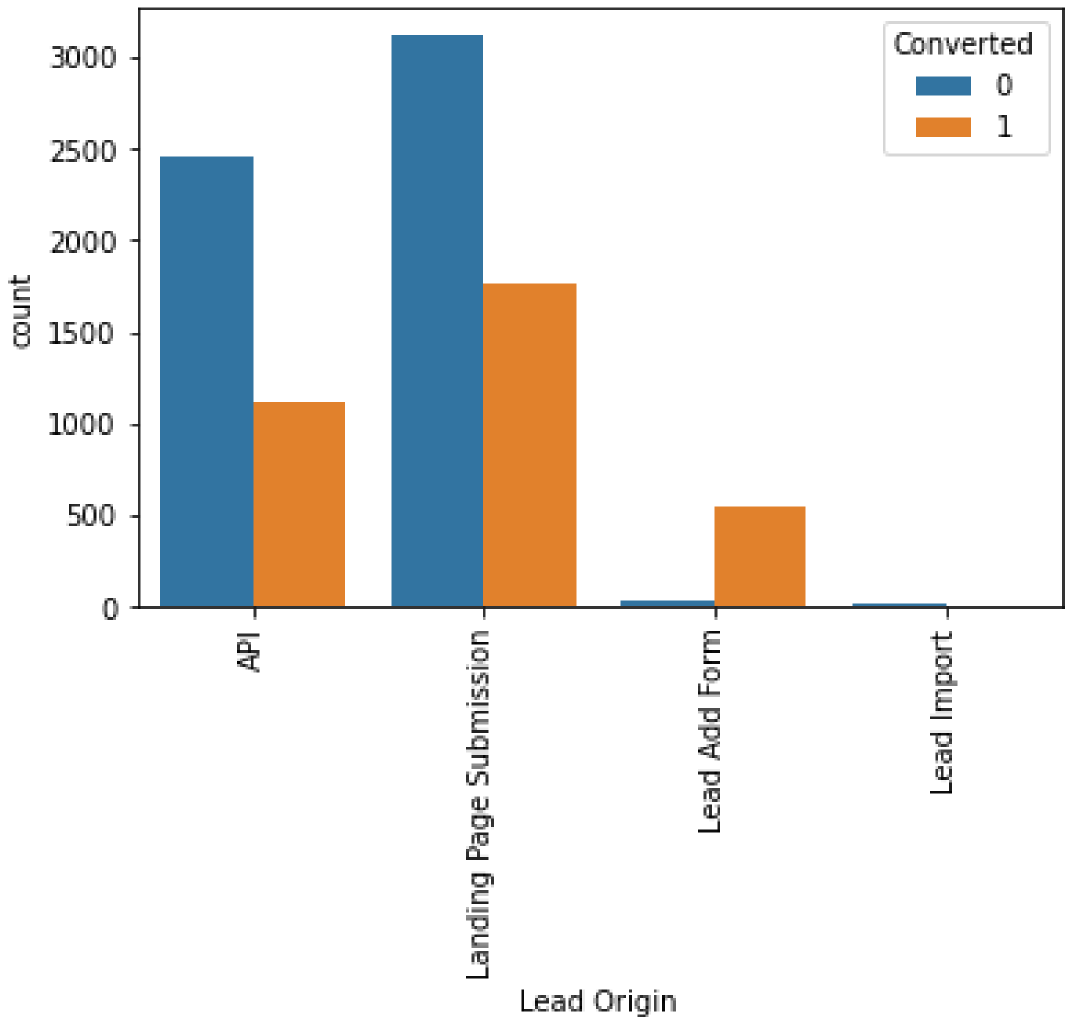
Categorical Univariate Analysis

- This plot is showcasing the Lead Source vs count
- The Lead Source as Google is having most prominent lead conversion.
- Lead Sources Direct Traffic, Organic Search, Olark Chat Conversation and Reference are also having probable positive Leads..



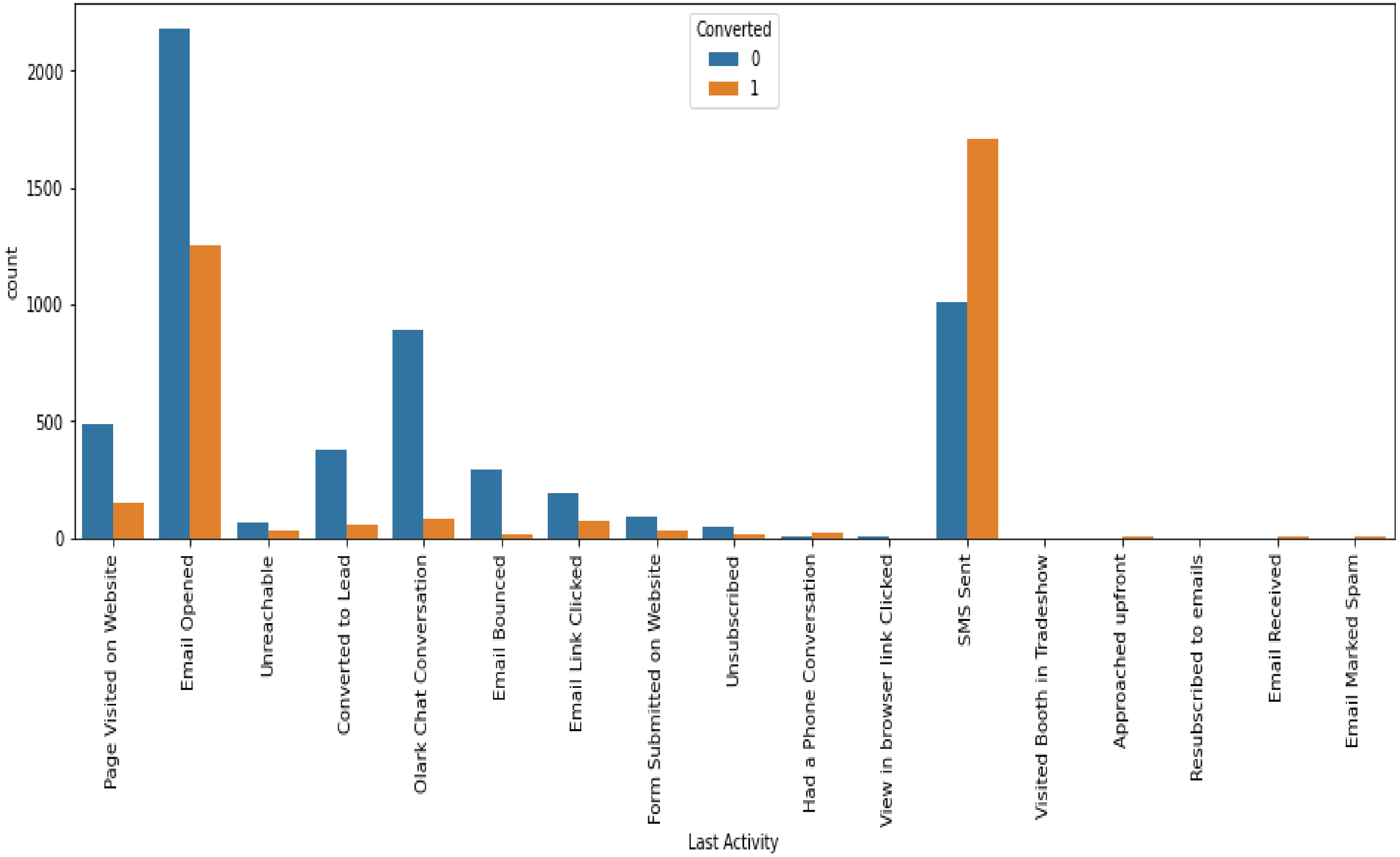
Categorical Univariate Analysis

- This plot is showcasing the Lead Origin
- Very clearly, Lead Add Forms are the most prominent Lead Converting factor.
- Lead Add Form followed by Landing Page Submissions and API are also considerate factors in positive Lead conversions.



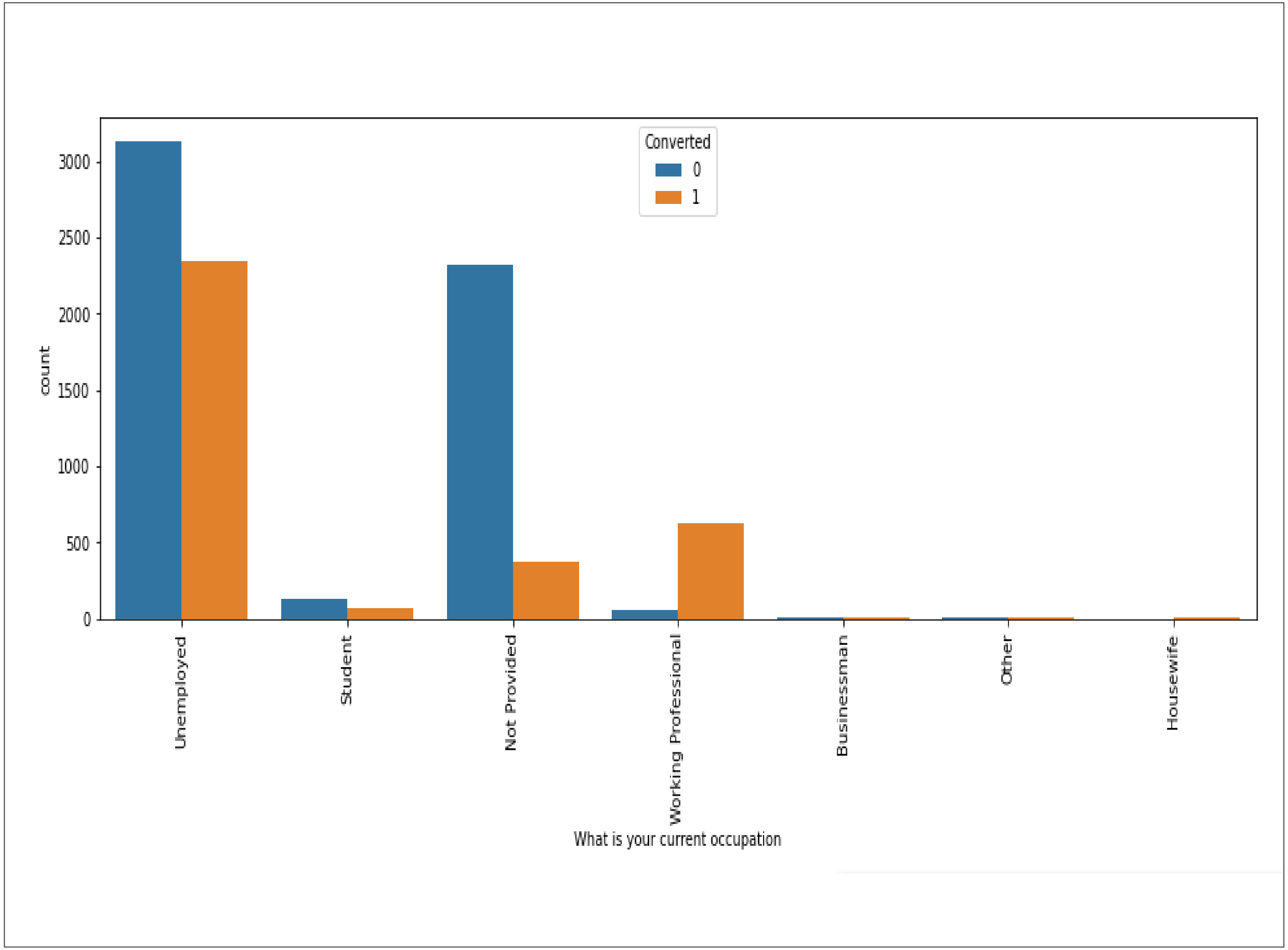
Categorical Univariate Analysis

- This plot is showcasing the Last Activity for both Convert values.
- The people having their Last activity as **SMS Sent** followed by **Email Opened** must be considered as they can be most prominent features for positive Lead Conversion.



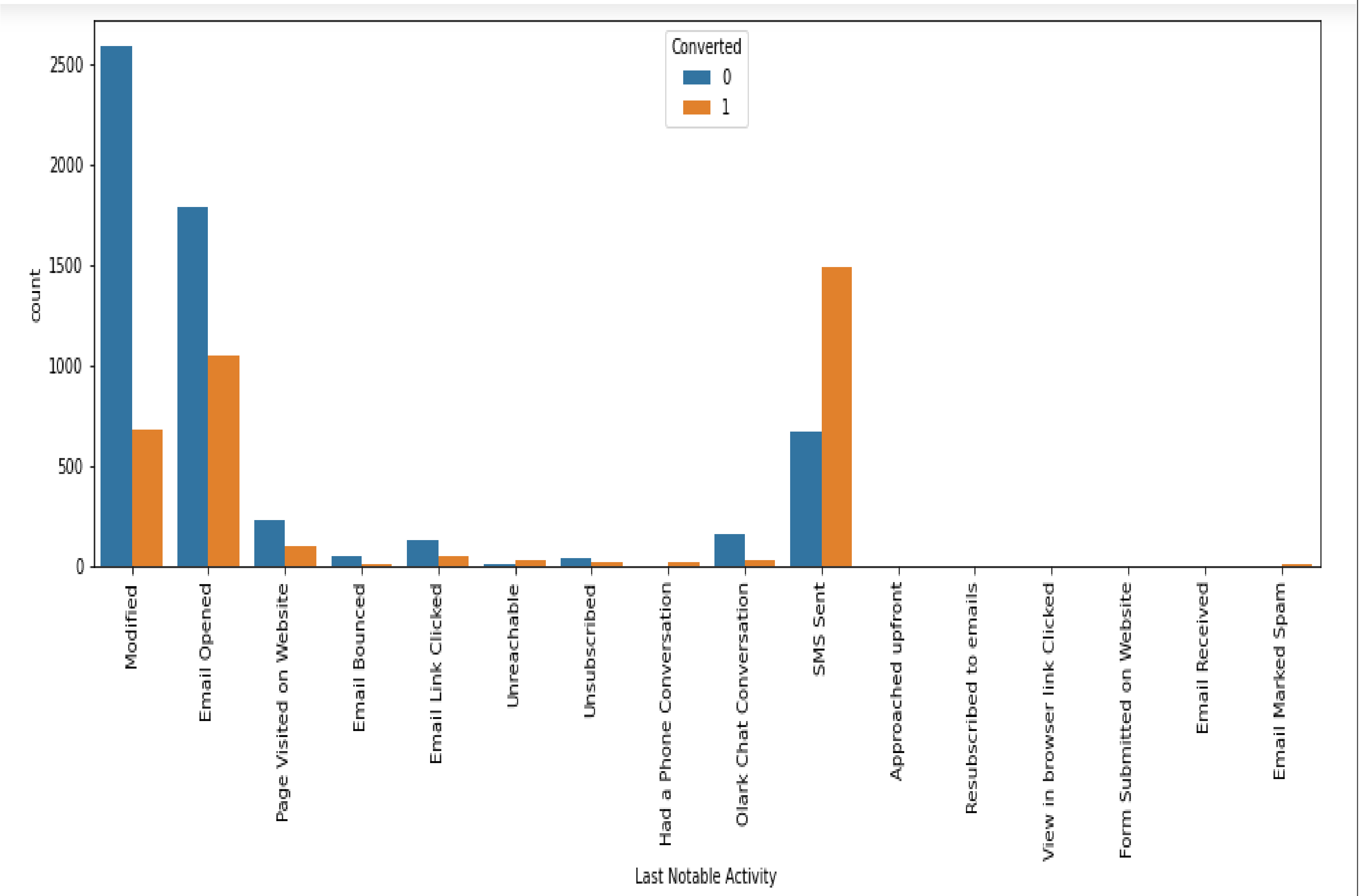
Categorical Univariate Analysis

- This plot is showcasing the Current Occupation.
- The people who are Unemployed are possibly highest Positive or more prominent Lead converters.
- Working Professionals can also be targeted in order to get more prominent Leads.



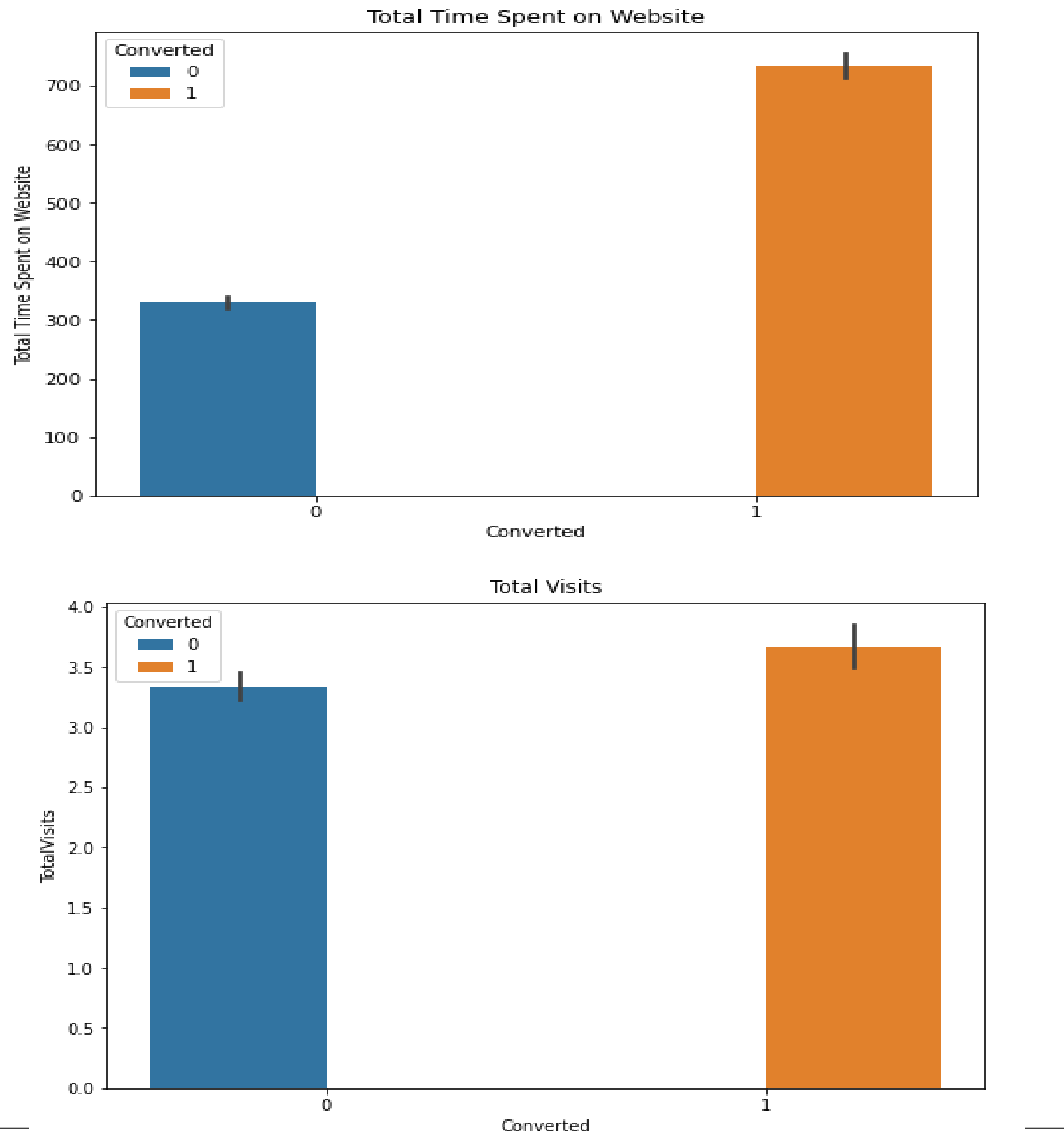
Categorical Univariate Analysis

- This plot is showcasing the Last Notable Activity.
- The leads whose Last notable activity is SMs Sent are most prominent Leads which can be converted to payable customers to the company.



Numerical Univariate Analysis

- The leads with higher number of Total Visits are clearly prominent Leads. They might be visiting frequently in order to get more details.
- The plot for Total Time Spent on Website also shows that the more is the Time spent, more were the prominent leads. They might be gathering information from their website.

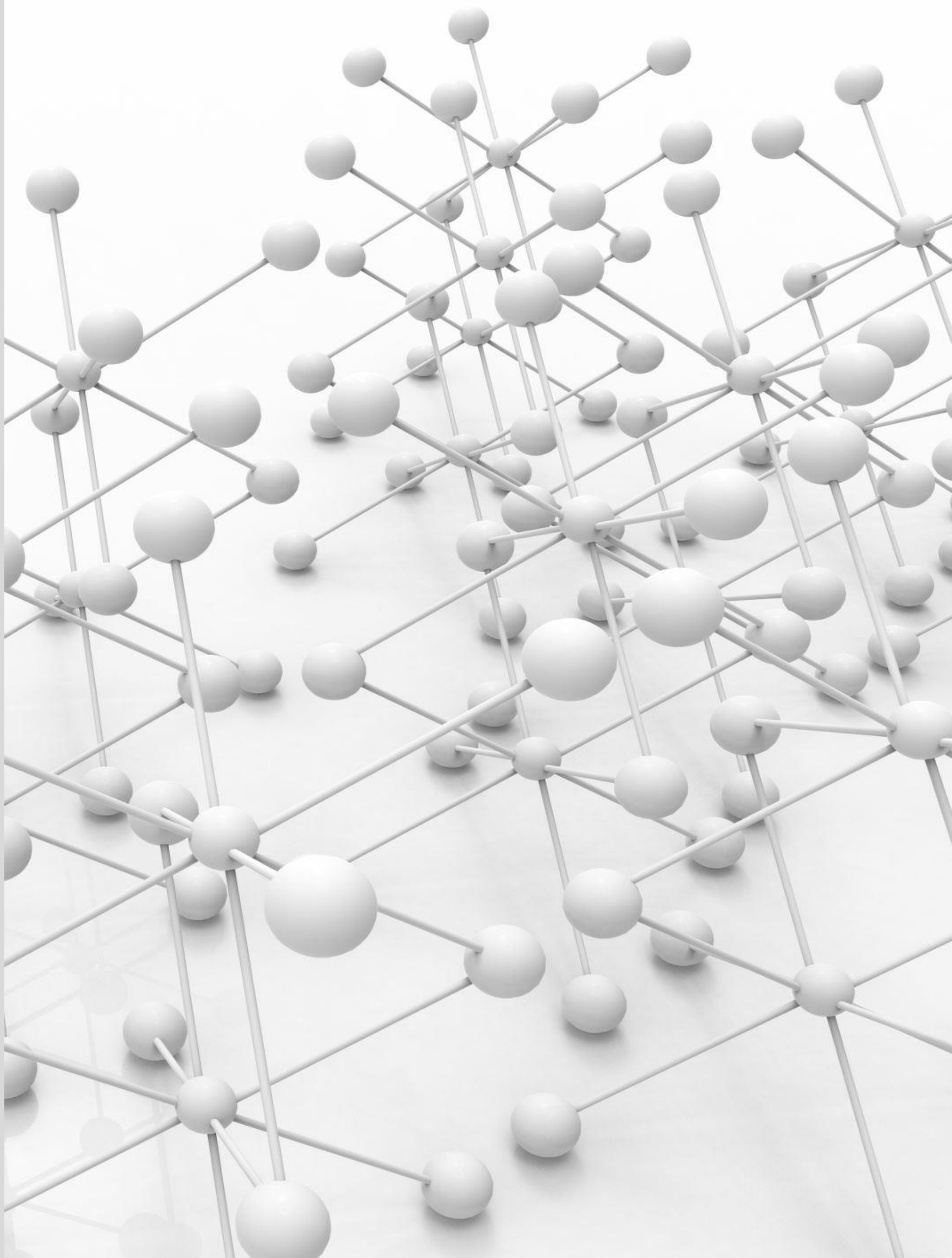


Impacting Variable for our Conversion Rate

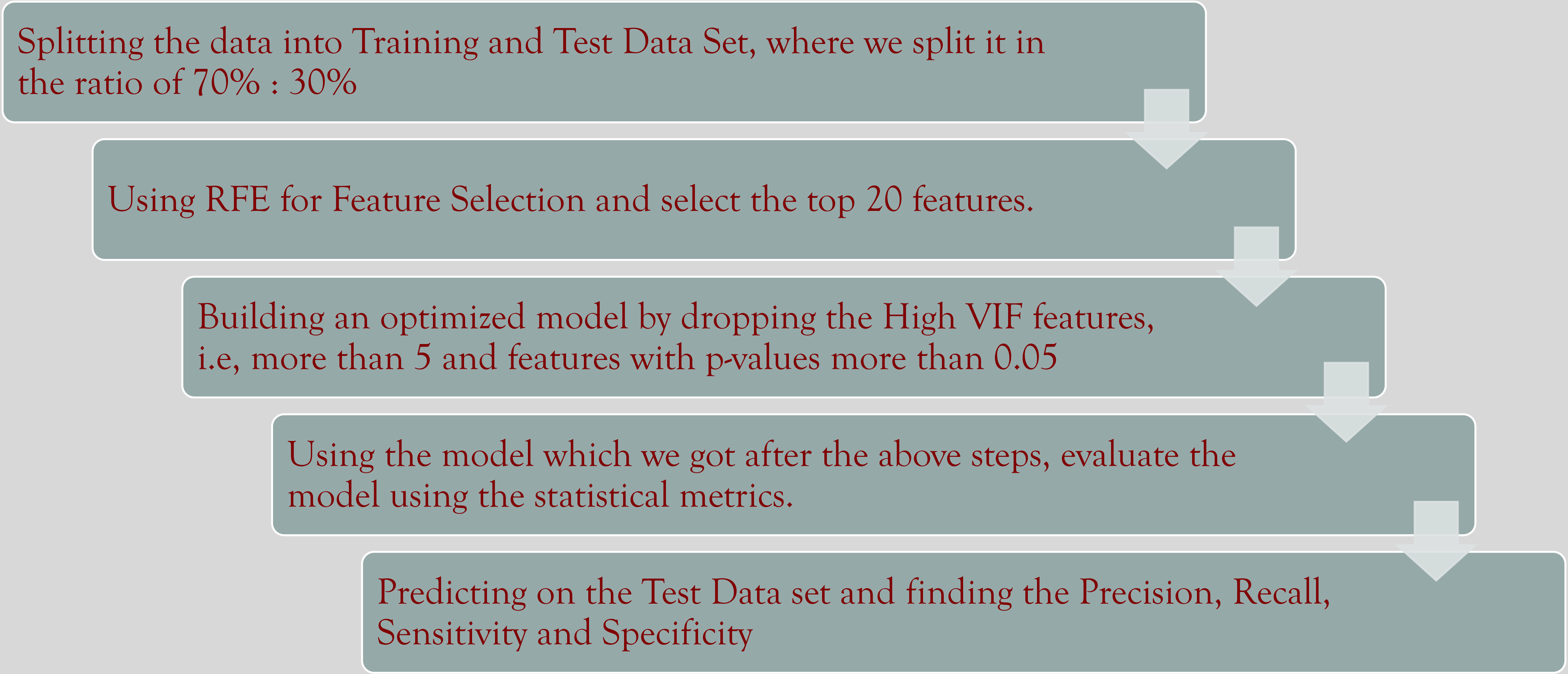
From the above visualizations, we can observe that a few features are prominently impacting in order to increase the positive leads. The company should work on these factors and use them in order to increase the prominent leads, following are the features :

- Total Time Spent on Website
- Total Visits
- Current Occupation – Not Provided followed by the Working Professionals
- Following Lead Sources –
  - Direct Traffic
  - Google
  - Organic Search
- Last Notable Activity as SMS Sent or Email Opened
- Lead Origin as Lead Add Form

MODEL BUILDING



Steps For Model Building



Splitting the data into Training and Test Data Set, where we split it in the ratio of 70% : 30%

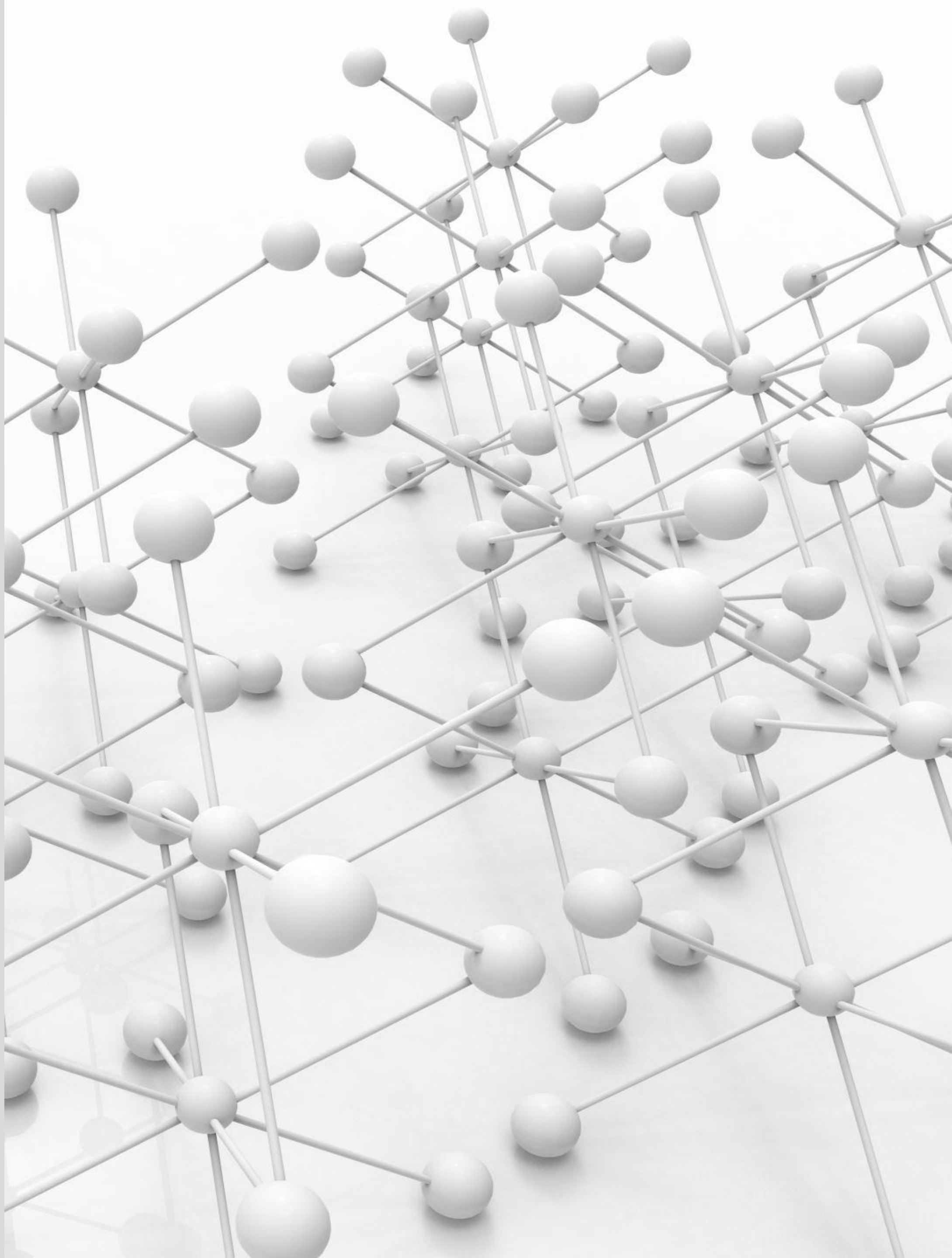
Using RFE for Feature Selection and select the top 20 features.

Building an optimized model by dropping the High VIF features, i.e, more than 5 and features with p-values more than 0.05

Using the model which we got after the above steps, evaluate the model using the statistical metrics.

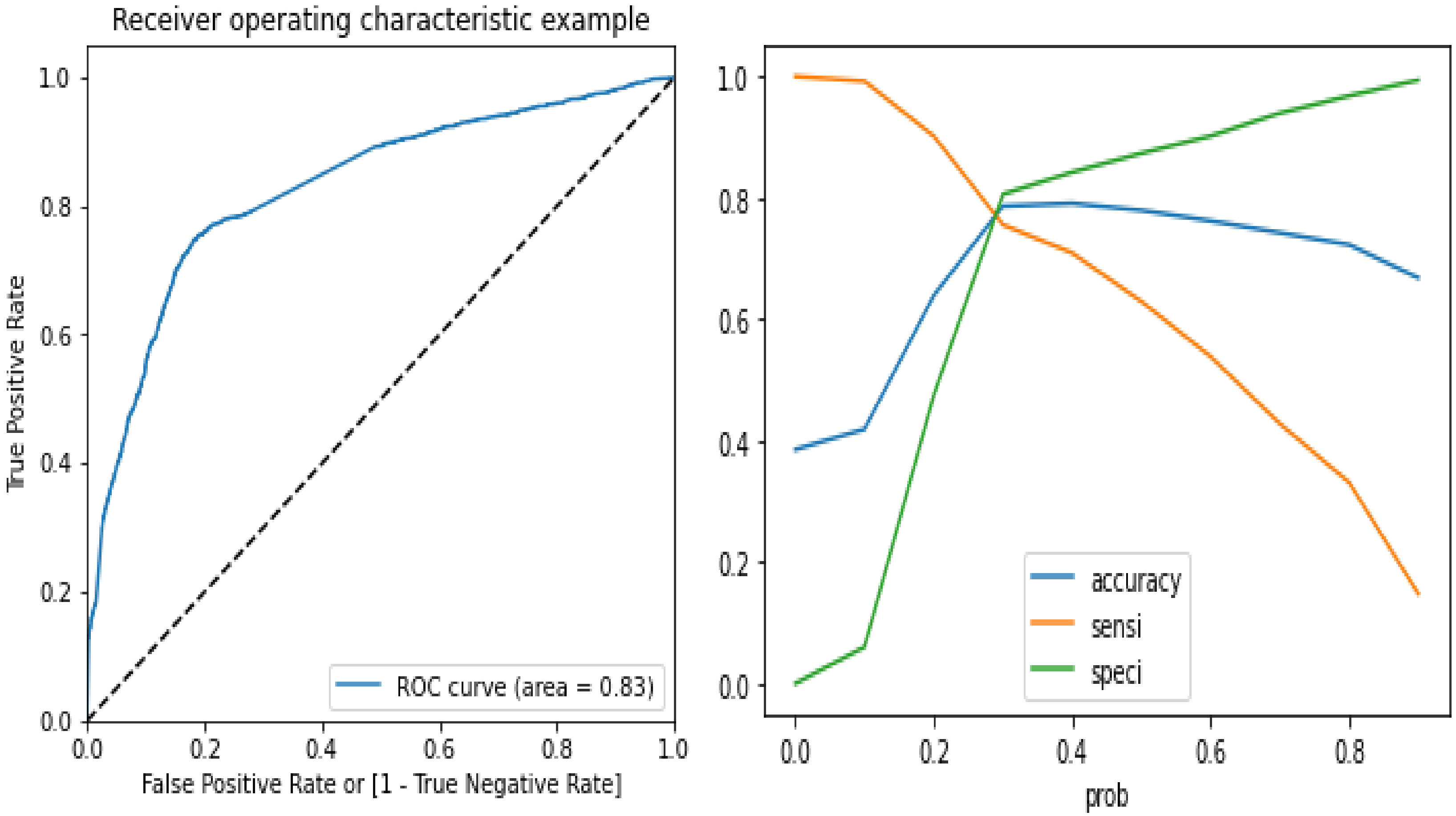
Predicting on the Test Data set and finding the Precision, Recall, Sensitivity and Specificity

MODEL EVALUATION



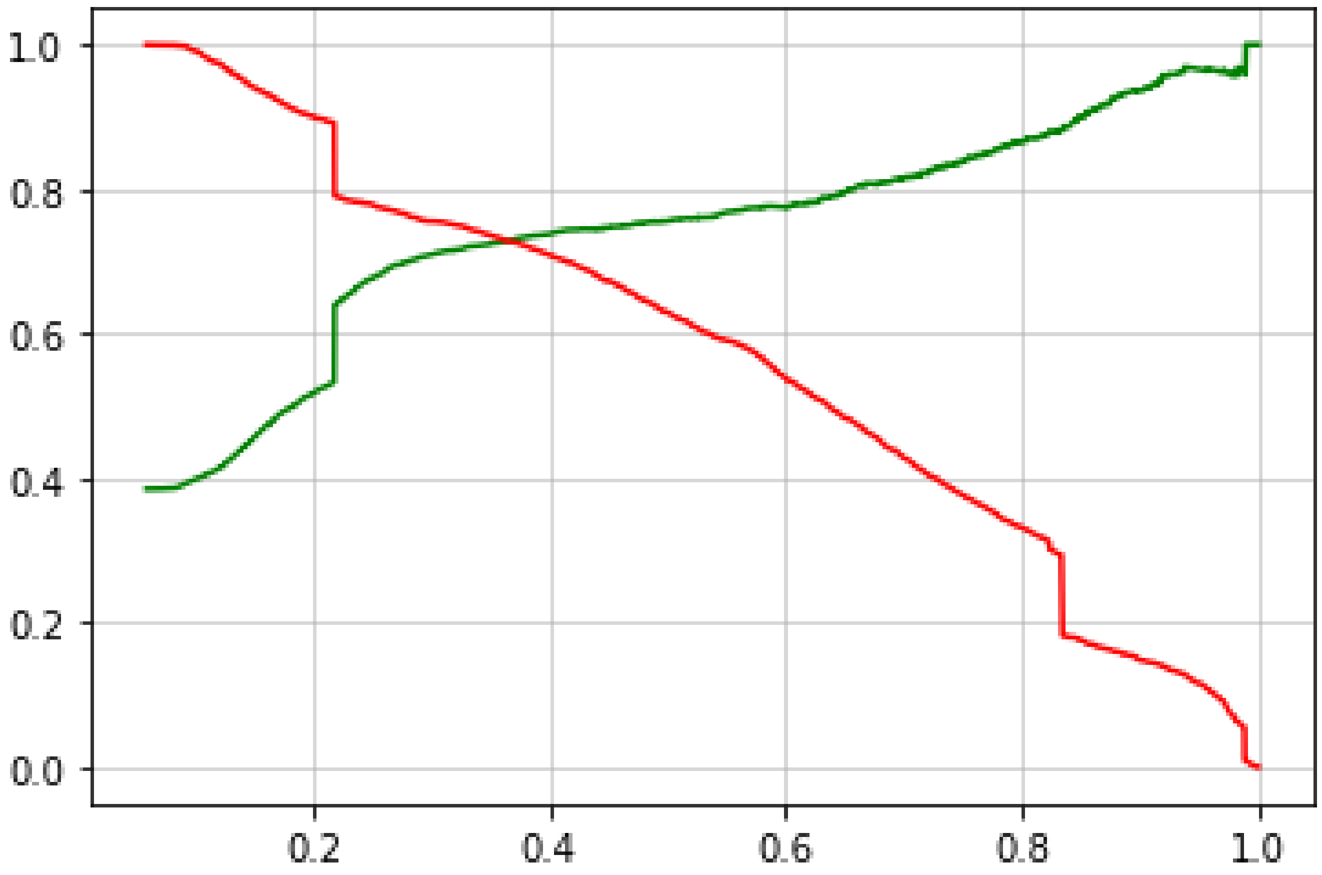
ROC Curve & Specificity, Sensitivity on Train Data Set

- The optimal cutoff we have here is 0.32.
- The area under ROC curve is 0.83 which is quite good value which means we have got a good model.
- Sensitivity – 63%
- Specificity – 87%



Precision & Recall on Train Data Set

- The optimal cutoff we have here is 0.38.
- Precision – 73%
- Recall – 72%



Conclusion

The statistical evaluation on both Train and Test sets seems to be similar and fine as well. We were able to achieve 79% accuracy on test data and 78% on the train data. The most prominent features which can result with increase the X company's positive leads are as follows :

- Total Time Spent on Website
- Total Visits
- Current Occupation – Not Provided followed by the Working Professionals
- Following Lead Sources –
  - Direct Traffic
  - Google
  - Organic Search
- Last Notable Activity as SMS Sent or Email Opened
- Lead Origin as Lead Add Form

Using these details, the X company can enhance their probability of getting the prominent buyers change their mind and get converted into the payable customers who might end up taking the courses.

THANK YOU