Lead Scoring Case Study – Master of Data Science – upGrad

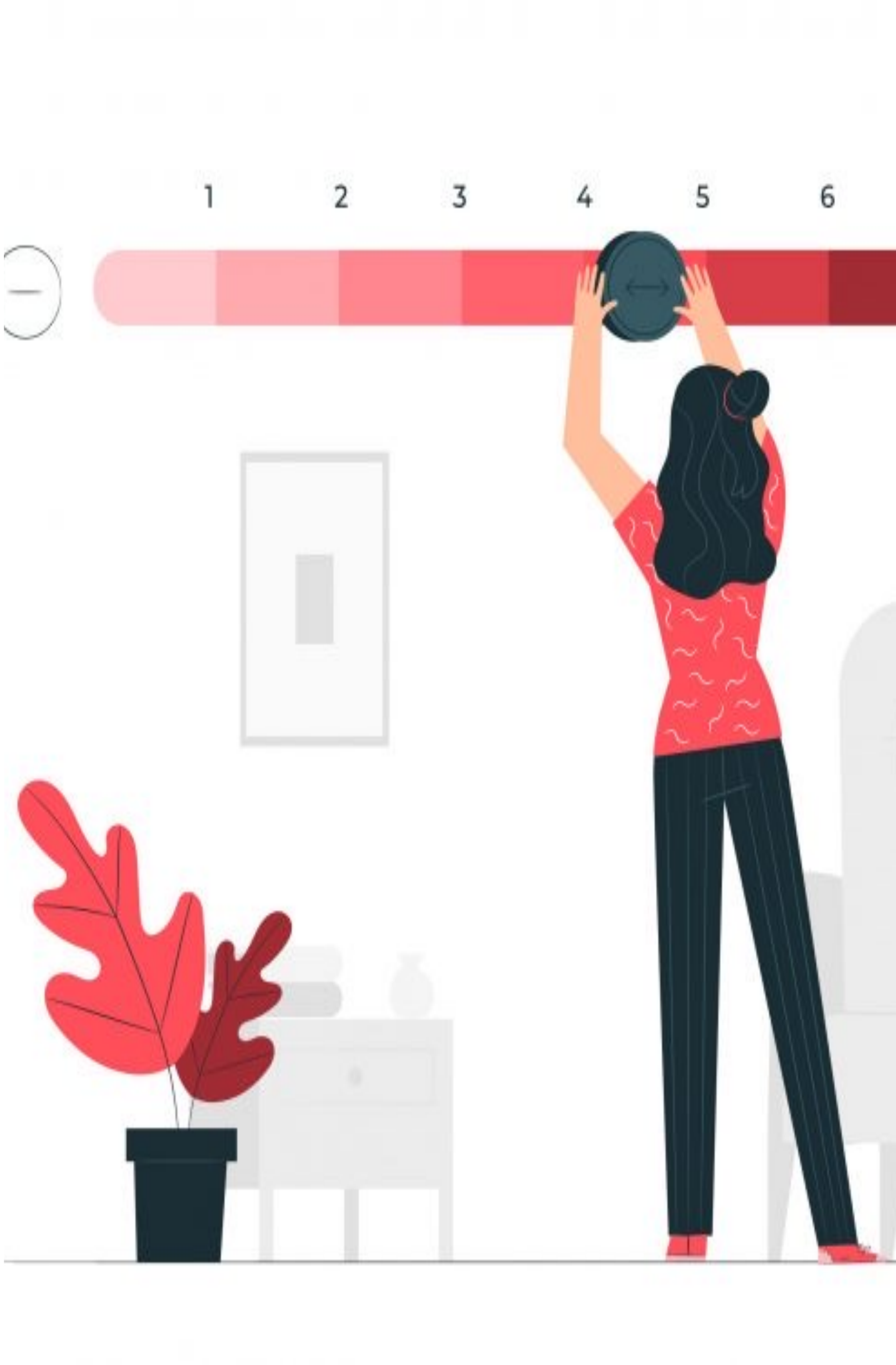
By: Anh Quoc Duy Ngo & Truong Hai Vu

Course 2: Machine Learning - I – Lead Scoring case Study

February 27th, 2023

Table of Content

- I. Introduction
- II. Data Cleaning
- III. Exploratory Data Analysis
- 3.1. Univariate Analysis and Bivariate Analysis
- 3.2. Data Preparation
- IV. Building Model
- V. Making Predictions
- VI. Model Evaluation
- VII. Conclusion

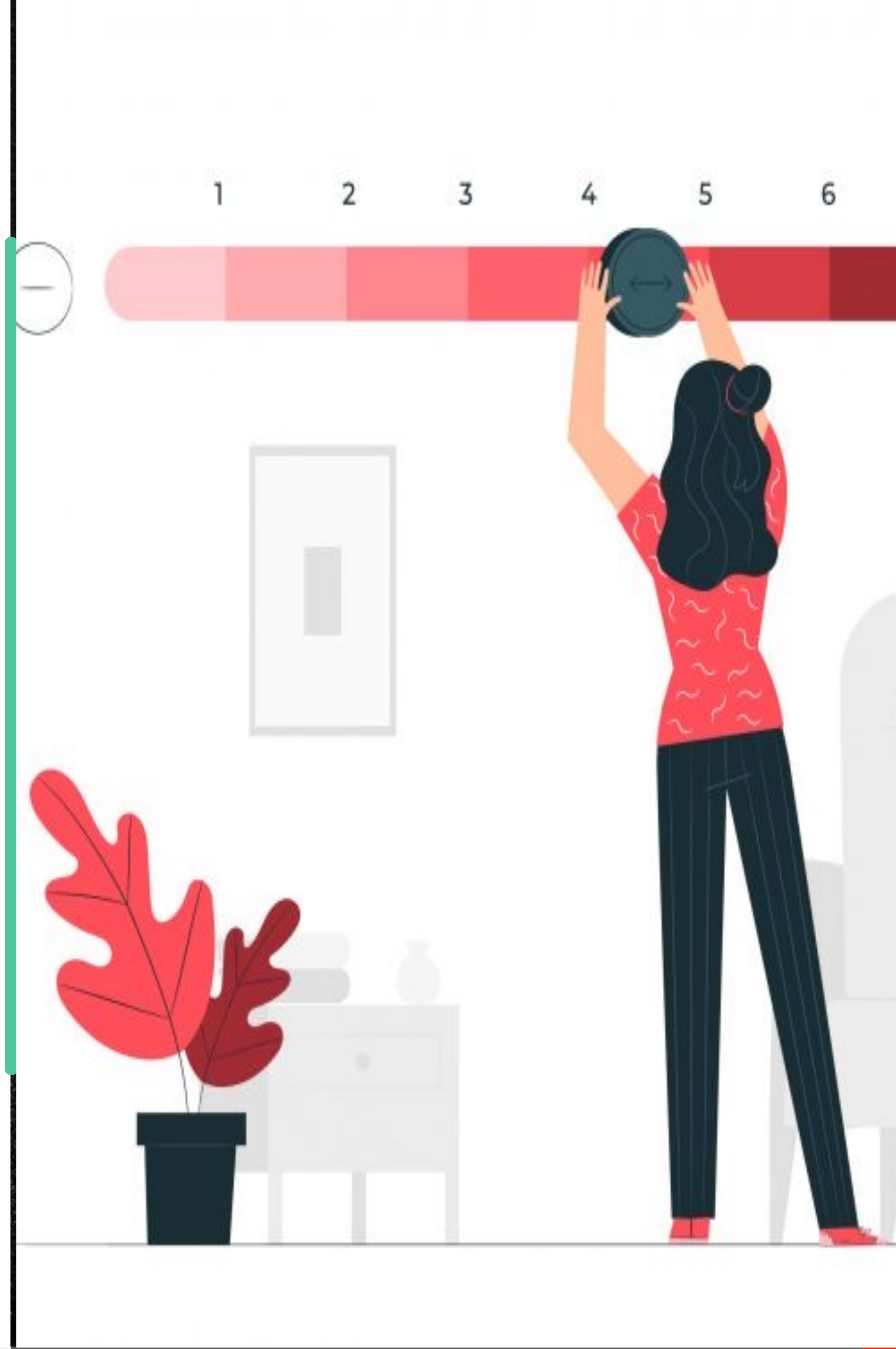


I. Introduction

- X Education sells online courses to industry professionals on several websites and search engines.
- The typical lead conversion rate at X education is around 30%.
- The company wants to identify the most potential leads (Hot Leads) and improve lead conversion rate by building a model assigning a lead score to each of the leads.
- The CEO has given a ballpark of the target lead conversion rate around 80%.

- **Target variable**:
  - 1 – lead was converted
  - 0 - lead was not converted

- **Goals**:
  - Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads
  - Suggest solutions to related problems to cope with changes in the future.



II. Data Cleaning

- Since there are many 'Select' values in our dataset as some customers did not select any option from the list, those are as good as NULL. Hence, we will convert them to null values.

- We will drop the columns with missing values > 40%.

- Columns with moderate missing values:

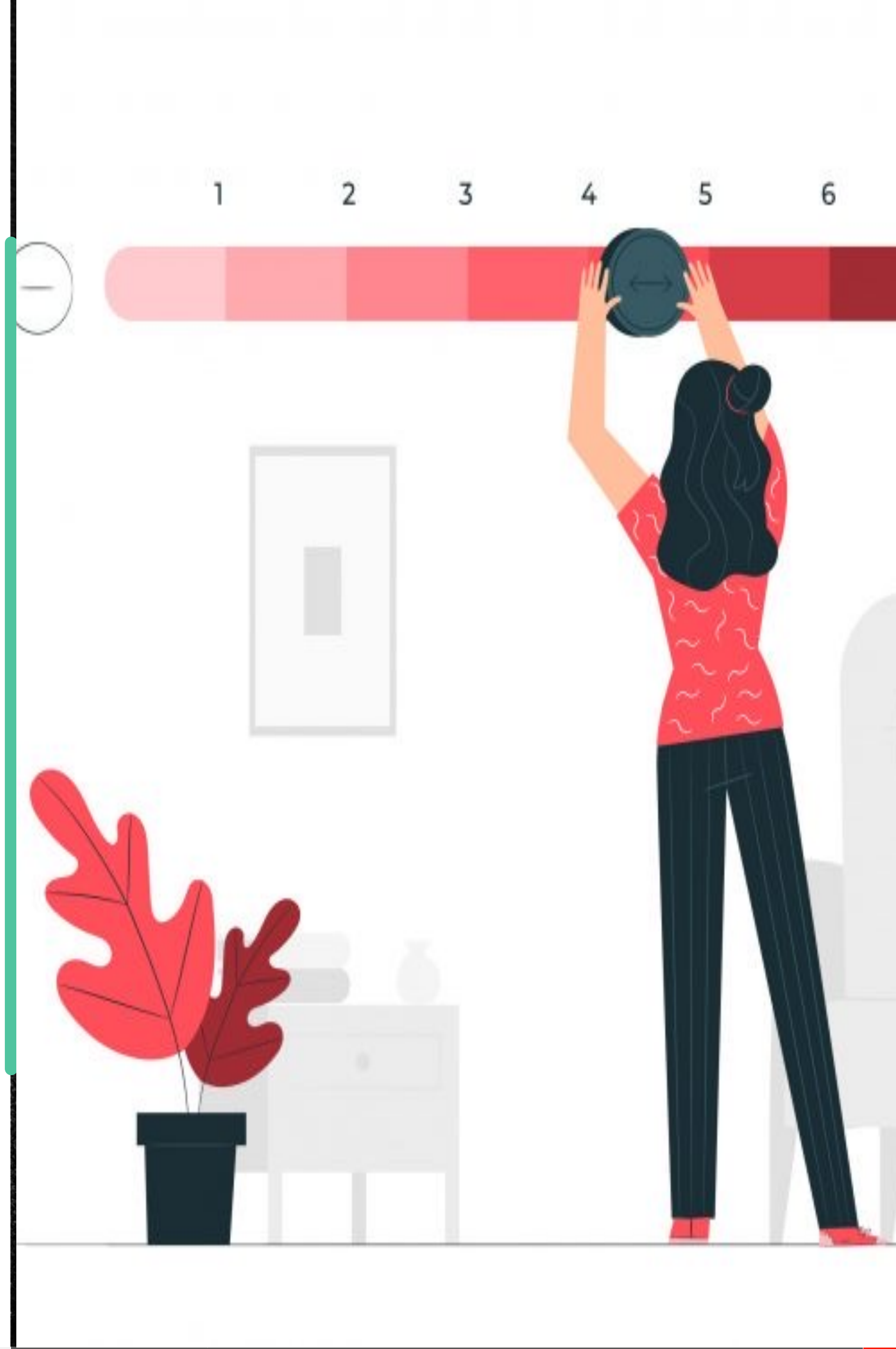
✓ Impute missing values with mode

✓ Create ‘Others’ category for insignificant values

✓ Should values fall into only one unique value, we drop other columns.

⇒ At this stage, we have retained 98% of the rows after cleaning the data.

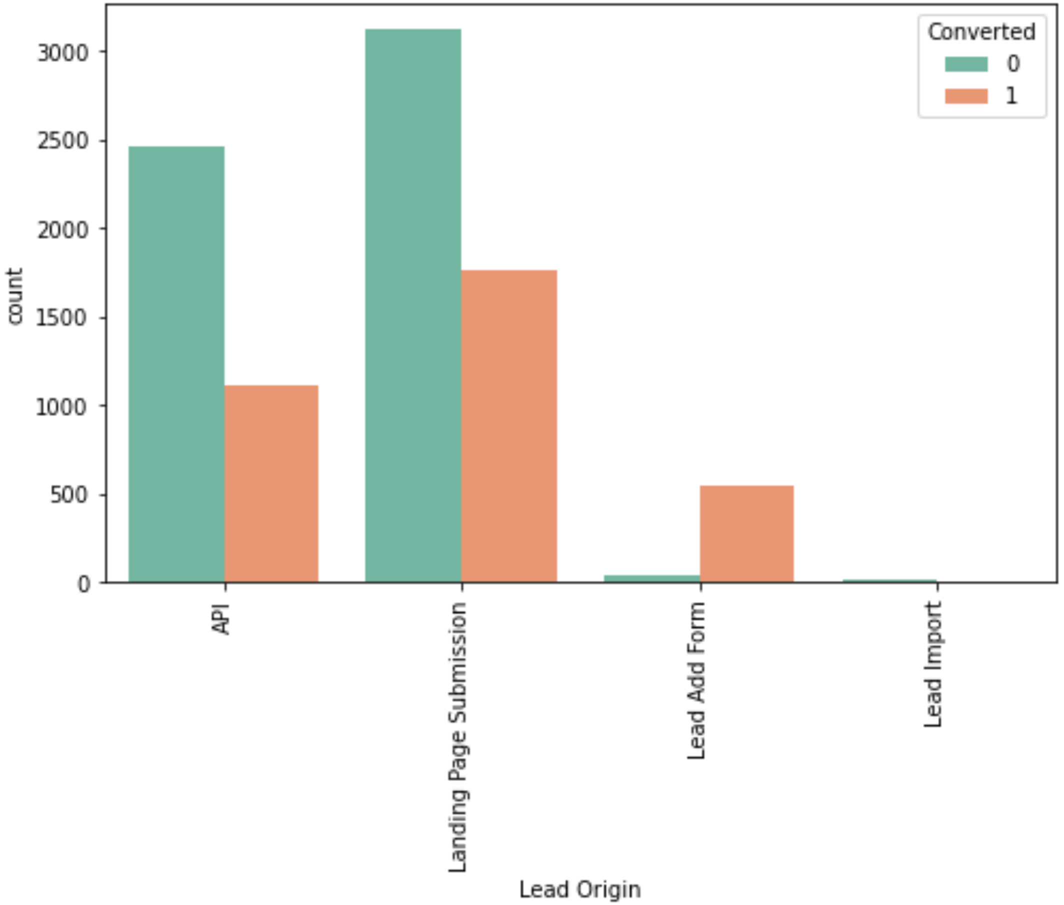
⇒ Null percentages < 1%



III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Lead Origin



Findings:

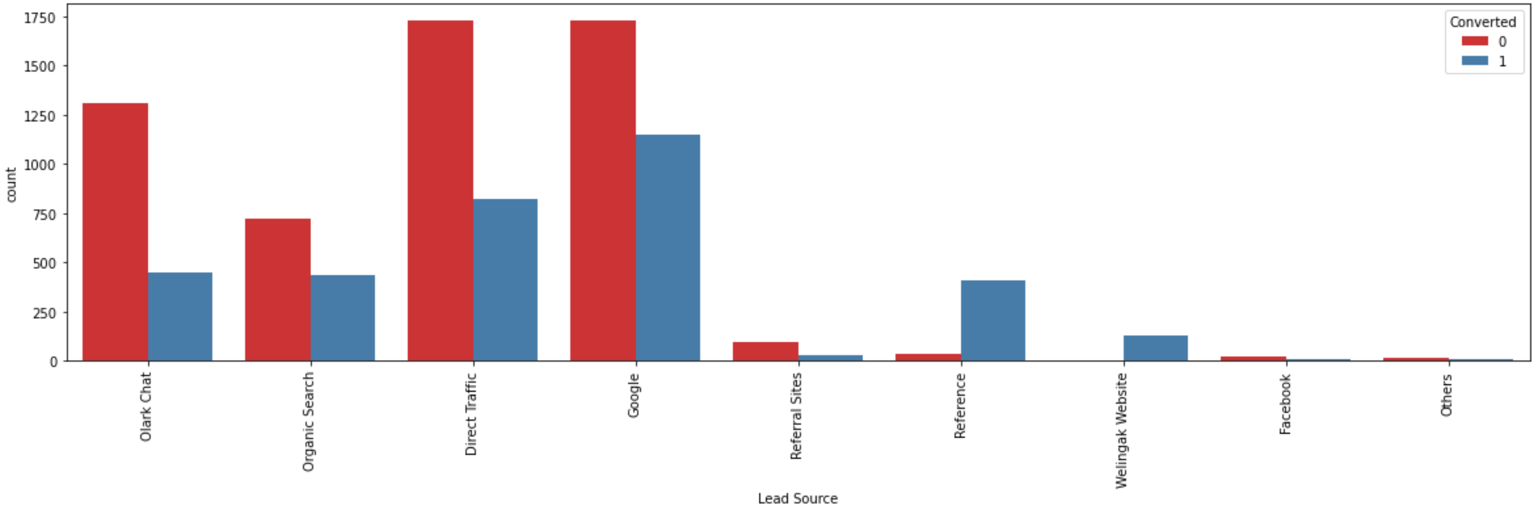
- API and Landing Page Submission have around 30-35% conversion rate while the number of lead originated from them are considerable.
- Lead Add Form has > 90% conversion rate but the number of lead are not significant.
- Lead Import are also not noticeable.

=> To improve the lead conversion rate, X Education should focus on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Lead Source



Findings:

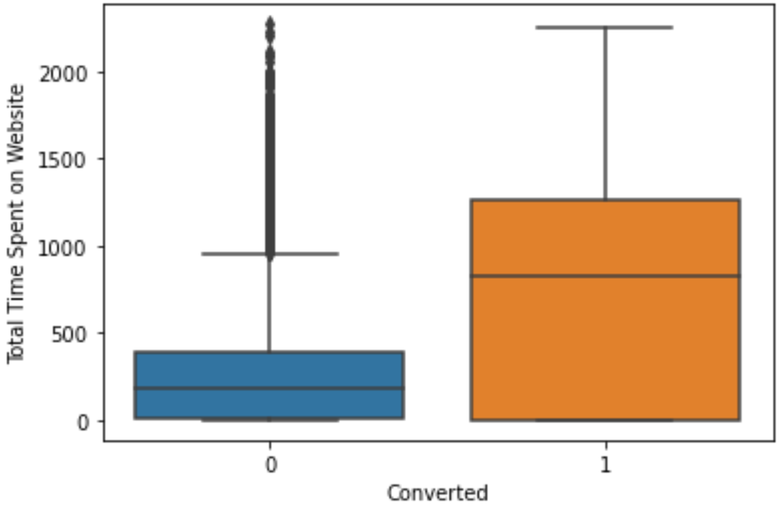
- Google and Direct traffic generates most of leads.
- Conversion Rate of Reference leads and Welingak Website is high.

=> To improve lead conversion rate, X Education should focus on improving lead conversion of Olark Chat, Organic Search, Direct Traffic, and Google leads and generate more leads from Reference and Welingak Website.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Total Time Spent on Website



Findings:

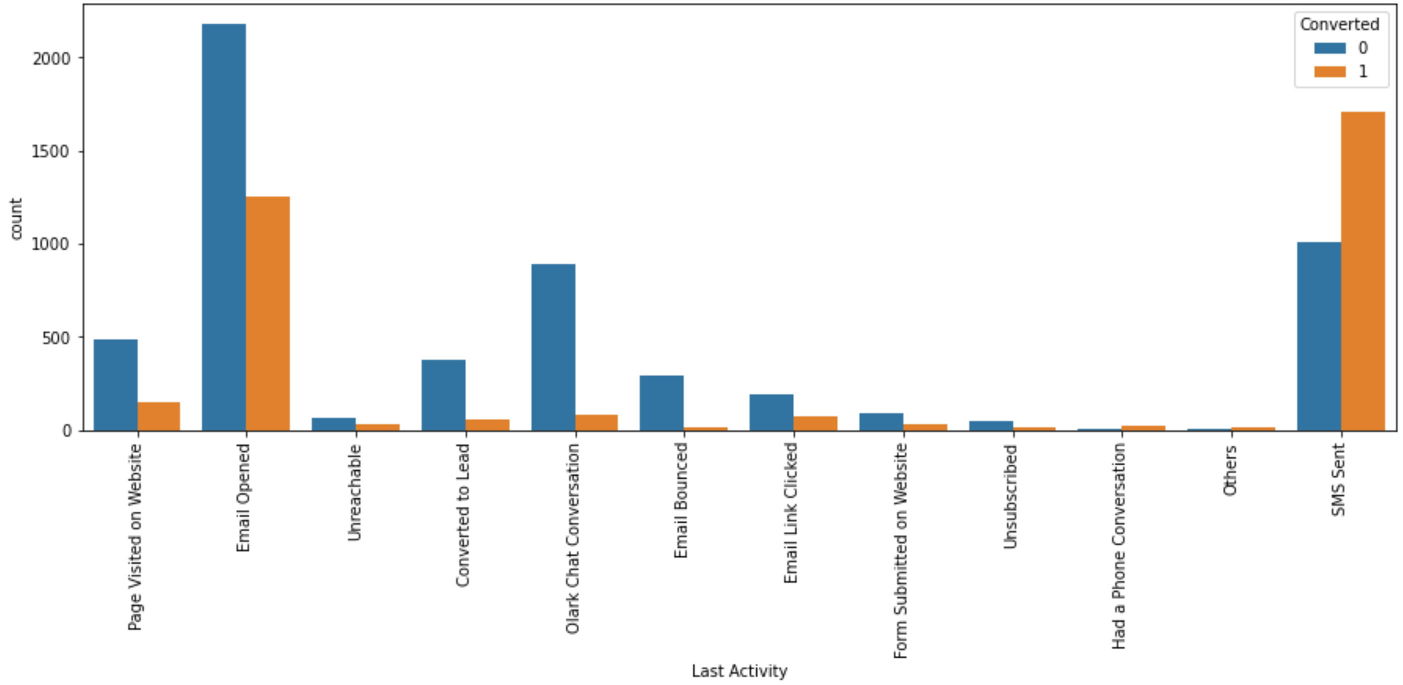
- Leads spending more time on the website are more likely to be converted.

=> Website should be made more attractive and engaging to make leads spend more time on the website.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Last Activity



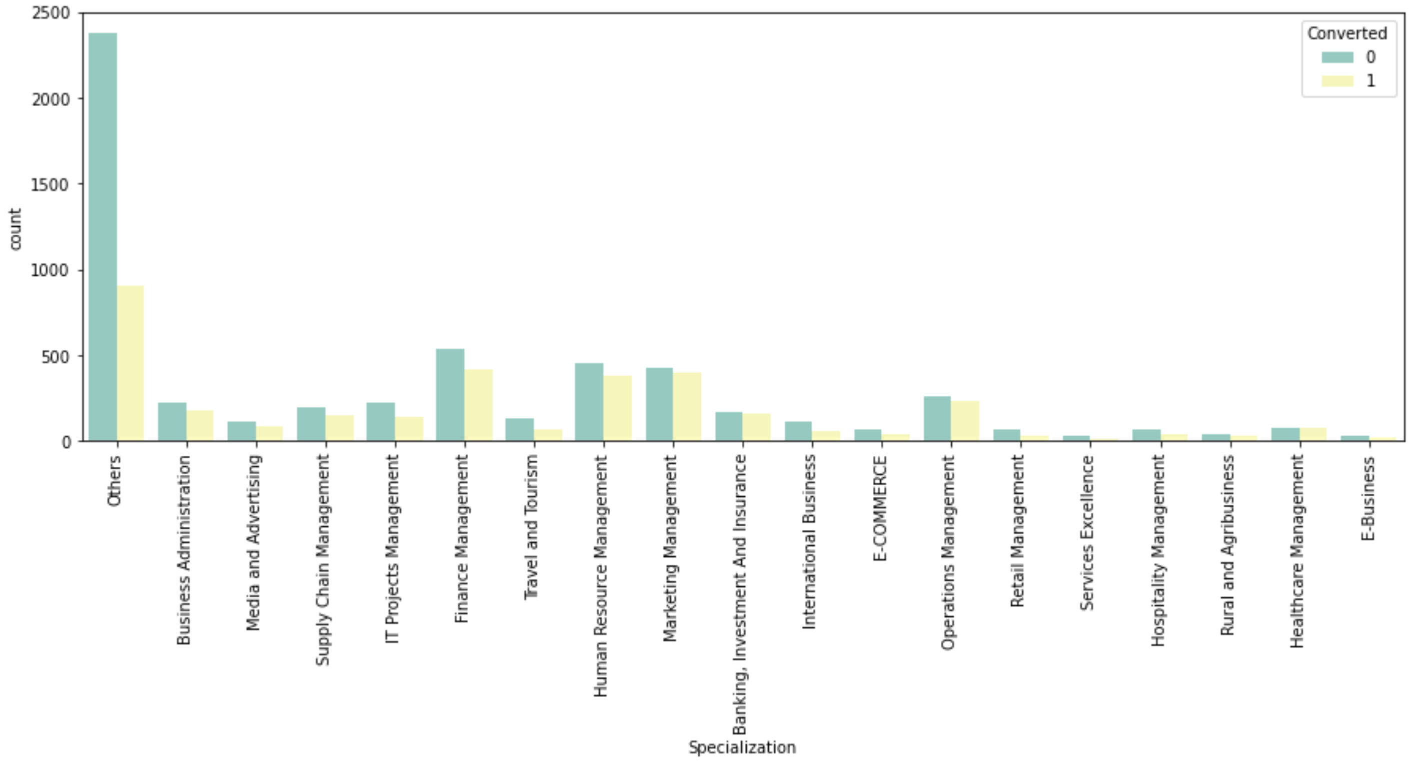
Findings:

- Most of the leads' last activity is opening their emails.
- Nearly 60% of leads with last activity as SMS Sent is converted.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Specialization



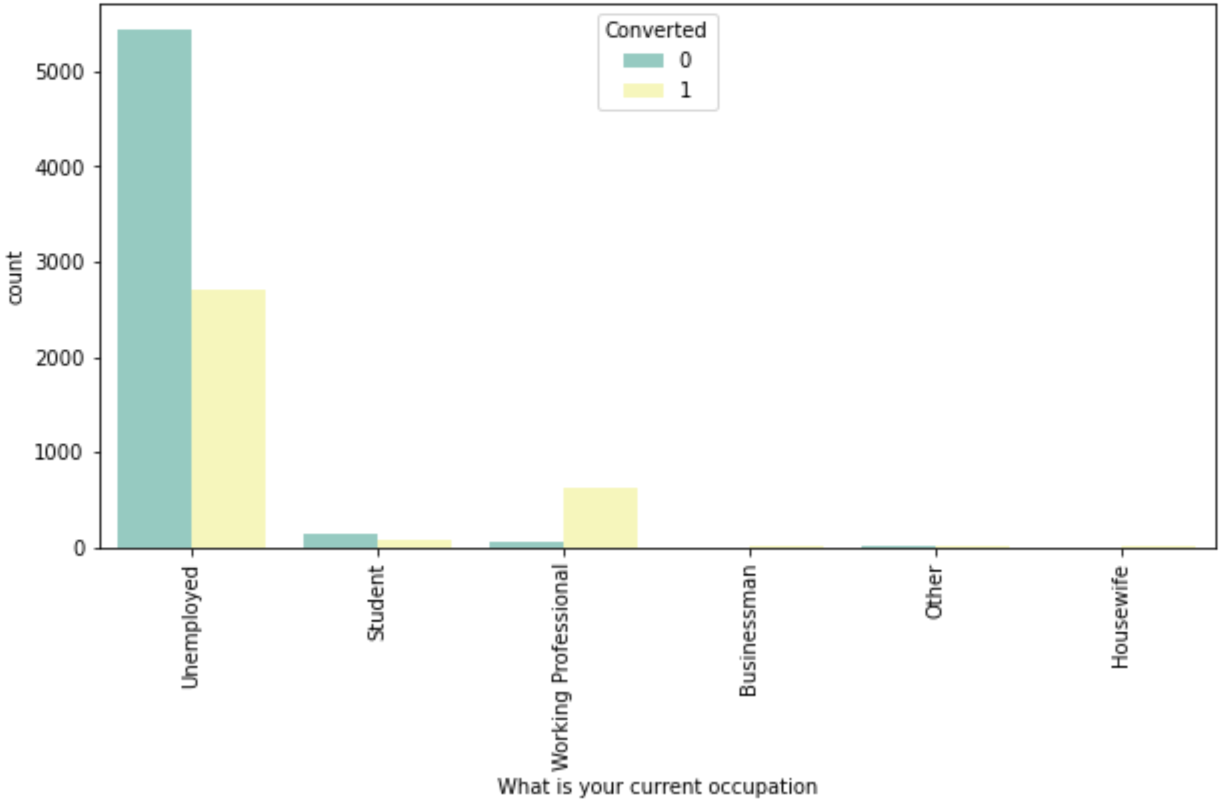
Findings:

- X Education should pay more attention to Specializations with high conversion. However, there is no significant difference between these parameters.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

What is your current occupation



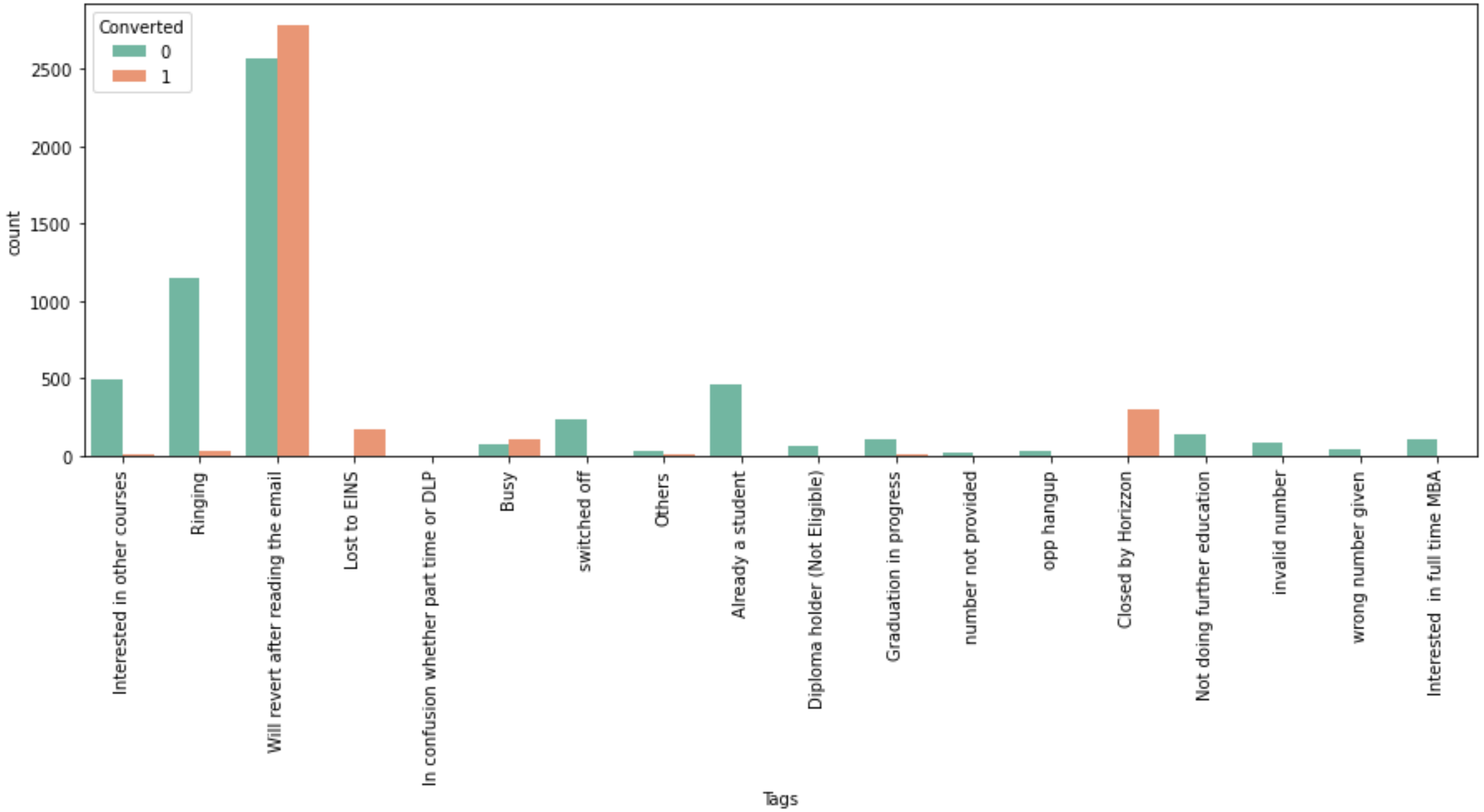
Findings:

- Working Professionals are more likely to be converted comparing to other categories.
- Unemployed leads dominant the dataset; however, it has around solely 30-35% conversion rate.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Tags



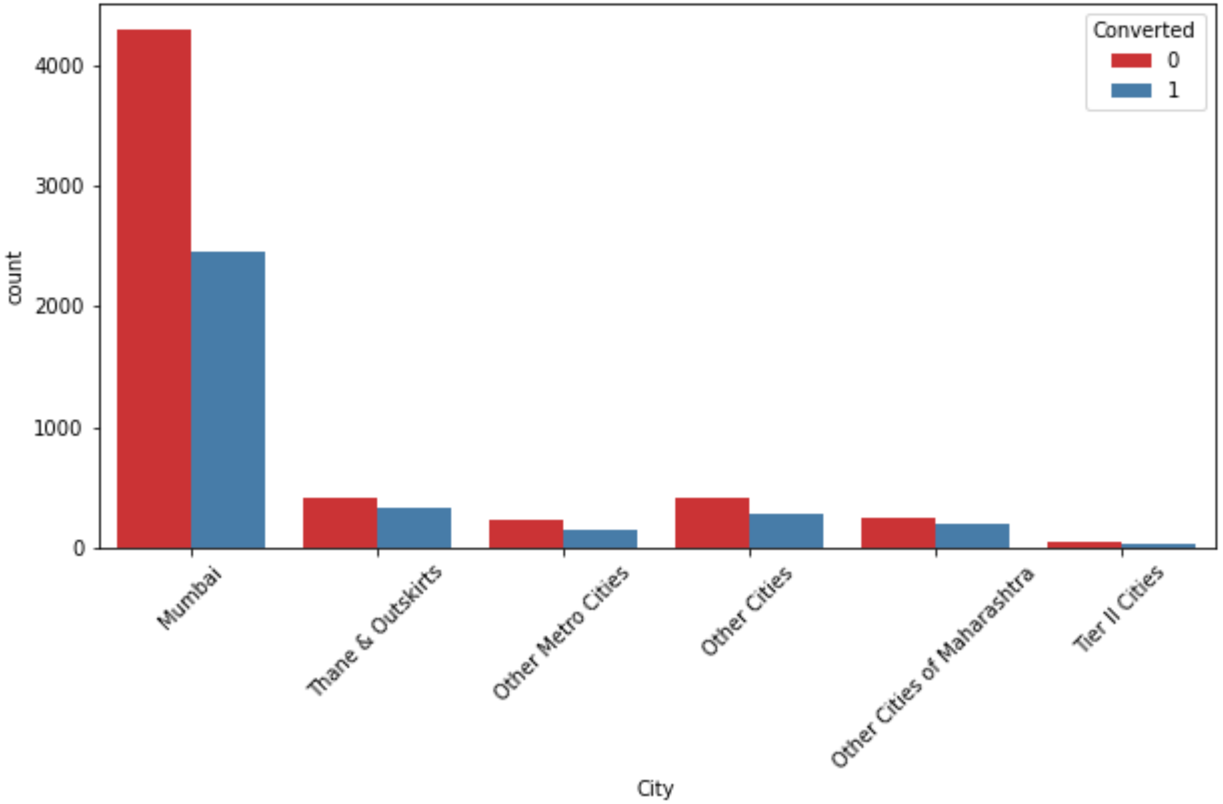
Findings:

- This is not available for model building as they are notes by sales team for tracking. Thus, we will drop this column before building the model.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

City



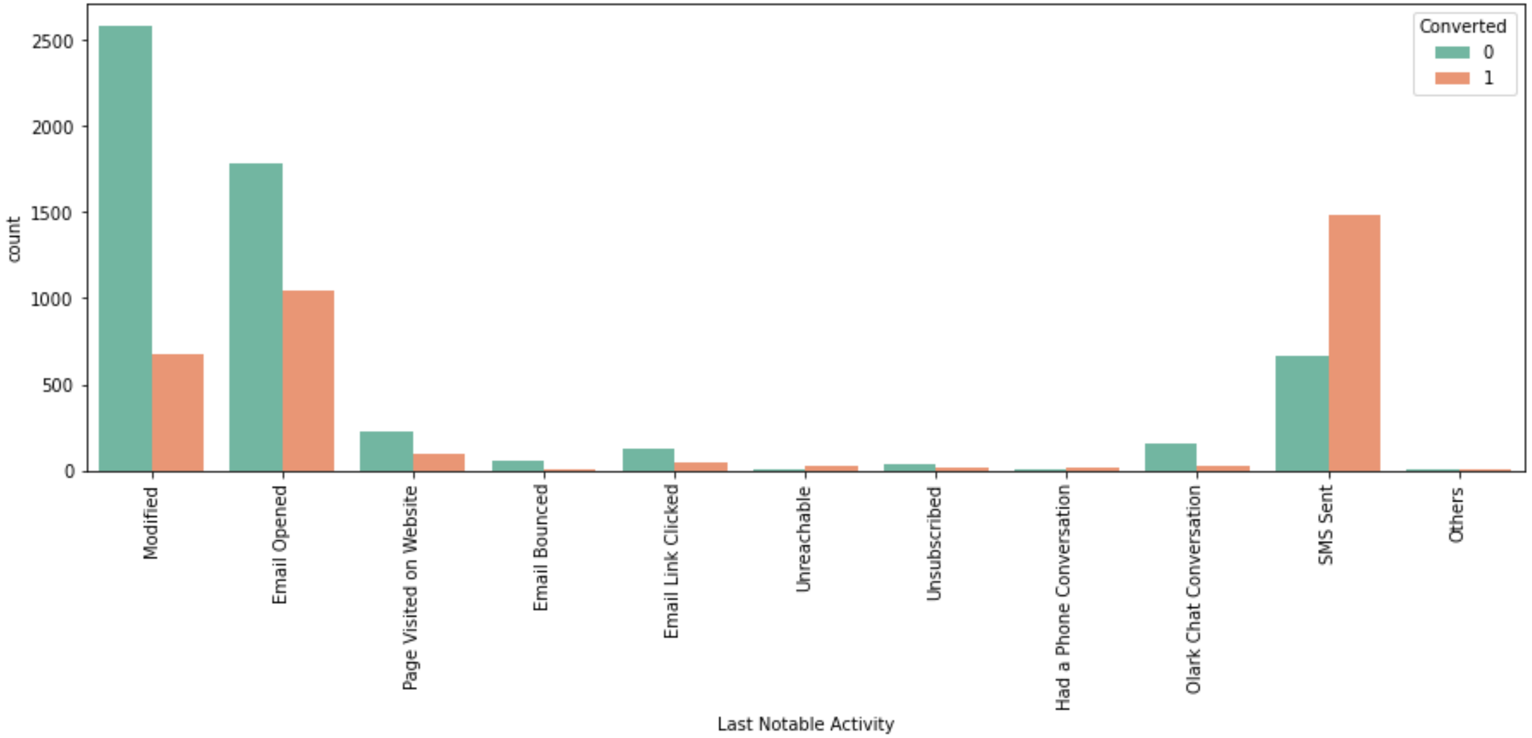
Findings:

- Most leads are from Mumbai with nearly 40% conversion rate.

III. Exploratory Data Analysis

3.1. Univariate Analysis and Bivariate Analysis

Last Notable Activity



Findings:

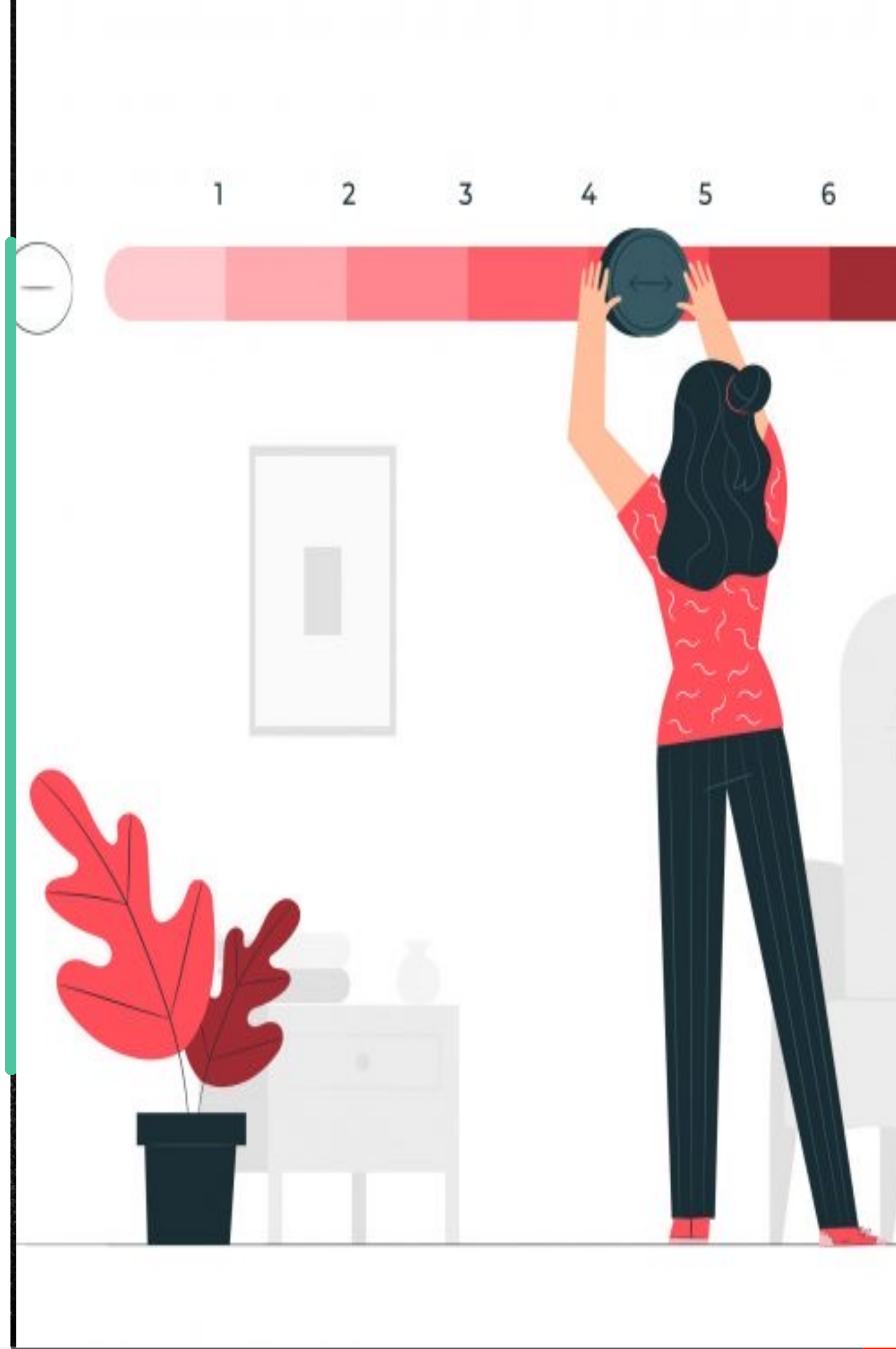
- Most leads came from Modified or Opened Email as their last notable activities. Leads which received SMS has high conversion rate at about 65%.
- From the above-mentioned Univariate Analysis and Bivariate Analysis, we notice that some columns do not have statistic values. Therefore, we can drop them for our best model in later analysis.

III. Exploratory Data Analysis

3.2. Data Preparation

In this session, we will:

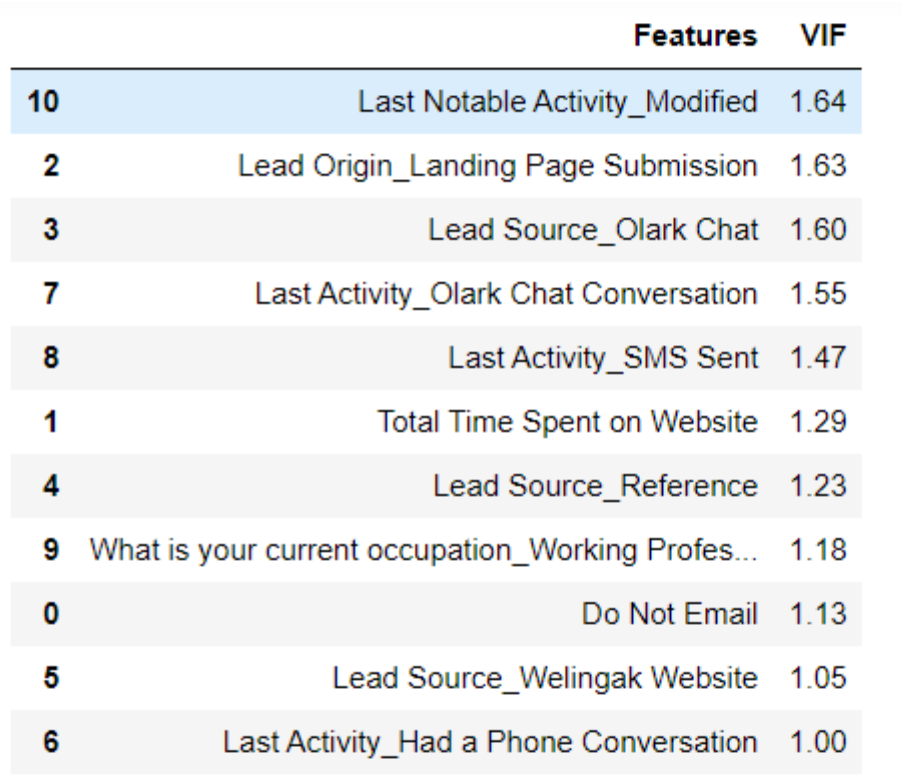
- Converting binary variables: 'Do Not Email', 'Do Not Call’
- Handling Dummy variables: 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'City', 'Last Notable Activity’
- Splitting the data into train and test set
- Rescaling the Features: Normalisation



IV. Building Model

4.1. SciKit Learn LogisticRegression

We will use the SciKit Learn LogisticRegression function for its compatibility with RFE (Recursive Feature Elimination - a utility from sklearn). After using statsmodel and Checking VIF, we found our best model at 7th attempt as below:



	Features	VIF
10	Last Notable Activity_Modified	1.64
2	Lead Origin_Landing Page Submission	1.63
3	Lead Source_Olark Chat	1.60
7	Last Activity_Olark Chat Conversation	1.55
8	Last Activity_SMS Sent	1.47
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.23
9	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
5	Lead Source_Welingak Website	1.05
6	Last Activity_Had a Phone Conversation	1.00

Findings:

The VIFs and p-values are both within acceptable limits. Hence, we proceed to generate predictions using this model.

- For each term, the p-value tests the null hypothesis that the coefficient is less than 0.05. A low p-value (< 0.05) suggests that the null hypothesis may be rejected.
- Generally, if a VIF is more than 10, we have strong multicollinearity. In our scenario, VIFs < 5 indicate that we are in good shape and may continue with our regression.
- Model 7 is our final model. We have 11 variables in our final model.

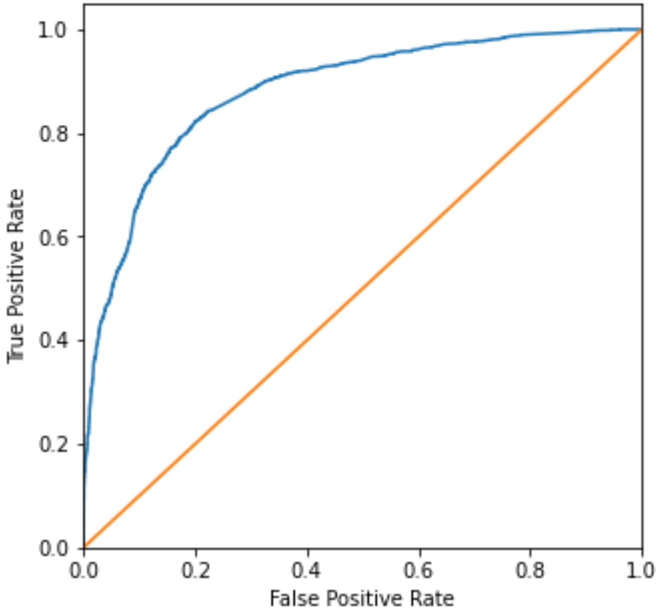
V. Making Predictions

Our specificity was good (~89%) but sensitivity was only well-under 70%. This is due to our initial chosen cut-off point of 0.5.

We use ROC curve to optimize sensitivity as:

- It shows the tradeoff between sensitivity and specificity
- It depicts how accurate the test is

ROC Curve

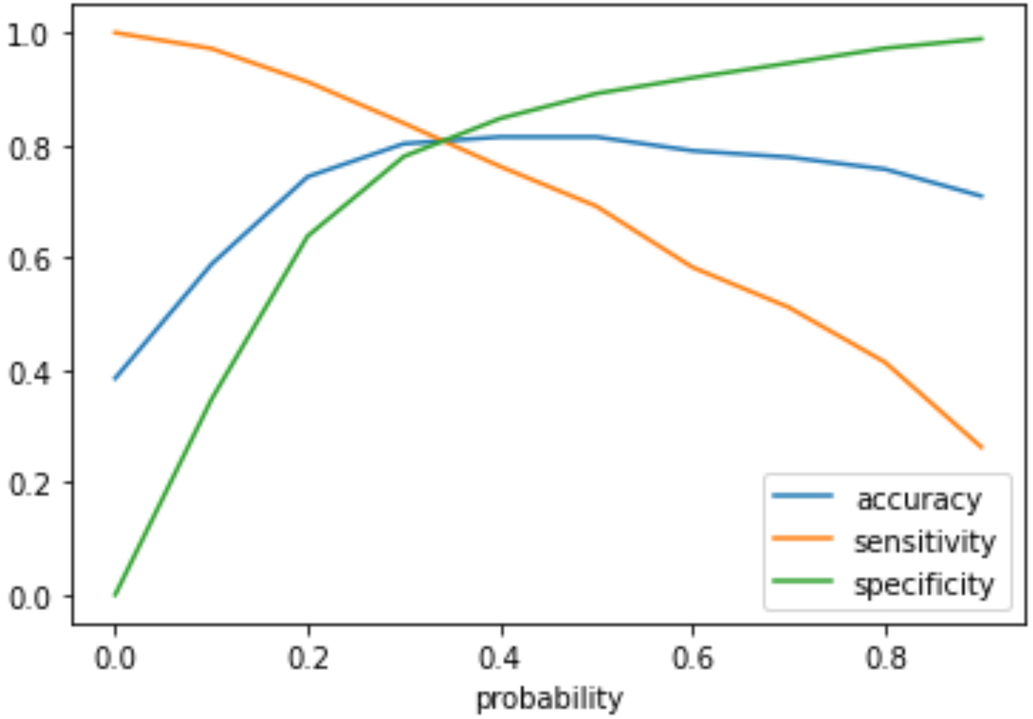


Findings:

- As our ROC curve is close to the left-hand and top border, our model is good, and the test is accurate.

V. Making Predictions

Let's find the optimal cut-off probability where there is balanced sensitivity and specificity.



Findings:

- The chart suggests that the optimal cut-off probability should be about 0.35

VI. Model Evaluation

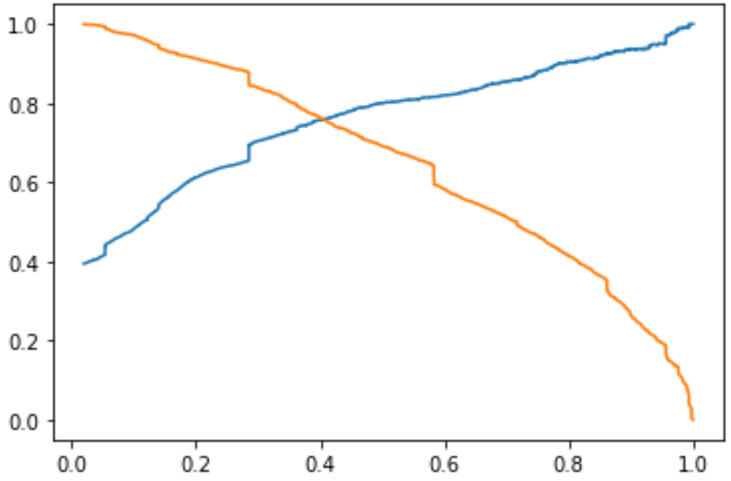
Train Data:

- Accuracy : 80.9 %
- Sensitivity : 80.4 %
- Specificity : 81.2 %

Test Data:

- Accuracy : 80.3 %
- Sensitivity : 79.0 %
- Specificity : 81.3 %

Precision and recall tradeoff



Findings:

- This is a good model which can predict the Conversion Rate well. The CEO has now strong confidence in making strategic decision as the model has given a ballpark of the target lead conversion rate to be around 80%.

VII. Conclusion

- There are 364 hot leads which should be contacted as they are more likely to be converted with the Score >=85

Recommendations:

X Education should target those groups of customers who have high chance to be converted:

- Calls leads coming from the lead sources 'Welingak Websites', 'Reference', and 'Olark Chat'
- Calls leads whose last activity was 'Phone Conversation' and 'SMS Sent'
- Calls leads who are the 'Working Professionals'
- Calls leads who spent more time on the websites ('Total Time Spent on Website')

VII. Conclusion

X Education should NOT focus on those groups of customers who have low chance to be converted:

- Whose lead origin is 'Landing Page Submission'
- Whose last notable activity is 'Modified'
- Whose last activity was 'Olark Chat Conversation'
- Whose do not want to receive emails ('Do Not Email')