Lead Scoring Case Study- X Education

Group Members:

Sandeep L

Tirumala Vedavyas A

ABOUT- X Education

❑ X Education , An education company named sells online courses to industry professionals
Many interested professionals land on their website.
❑ The company markets its courses on several websites like Google. Once these people land on the website, they might browse the courses or fill up a form for

Problem Statement

❑ When these people fill up a form providing their email address or phone number, they are classified to be a lead

❑ 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%.

❑ Identification of Hot Leads to focus more on them and thus enhancing the conversion ratio for X Education

❑ The leads that are most likely to convert into paying customers.

❑ 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
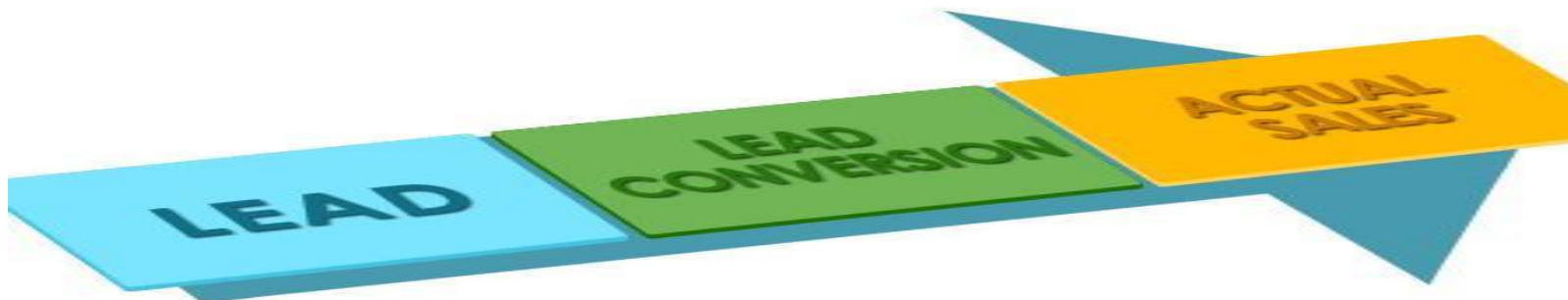
Lead – Conversion Process

Proposed Solution

| Selection of Hot Leads | Communicating with Hot Leads | Conversion of Hot Leads |

❑ **Leads Clustering**
We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

❑ **Focus Communication**
Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

❑ **Increase conversion**
Since we focussed on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

Solution- Selection of hot leads

❏ For our Problem Solution, the crucial part is to accurately identify hot leads.

❏ The more accurate we obtain the hot lead, the more chance we get of higher conversion ratio.

❏ Since we have a target of 80% conversion rate, we would want to obtain a high accuracy in obtaining hot leads.

Implementation

| Data Gathering | Data Cleaning | Performing EDA | Data Preparation | Model Building |

1. Loading & Observing the past data provided by the Company.

2. Duplicate removal, null value treatment, unnecessary column elimination, etc.

3. Univariate for numerical and categorical columns.

4. Outlier Treatment, Feature-Standardization.

5. Performing pre-requisites for RFE and Logistic Regression.

Feature Selection → Model Building → Model Improvemnet → Final Model → Verifying with PCA

6. Selection of top 15 features using RFE.

7. Model building using RFE for selected columns.

8. Reduction of columns and Model re-building.

9. Final Model Analysis and performance on Test Data.

Plots (Visualization)

Lead Quality
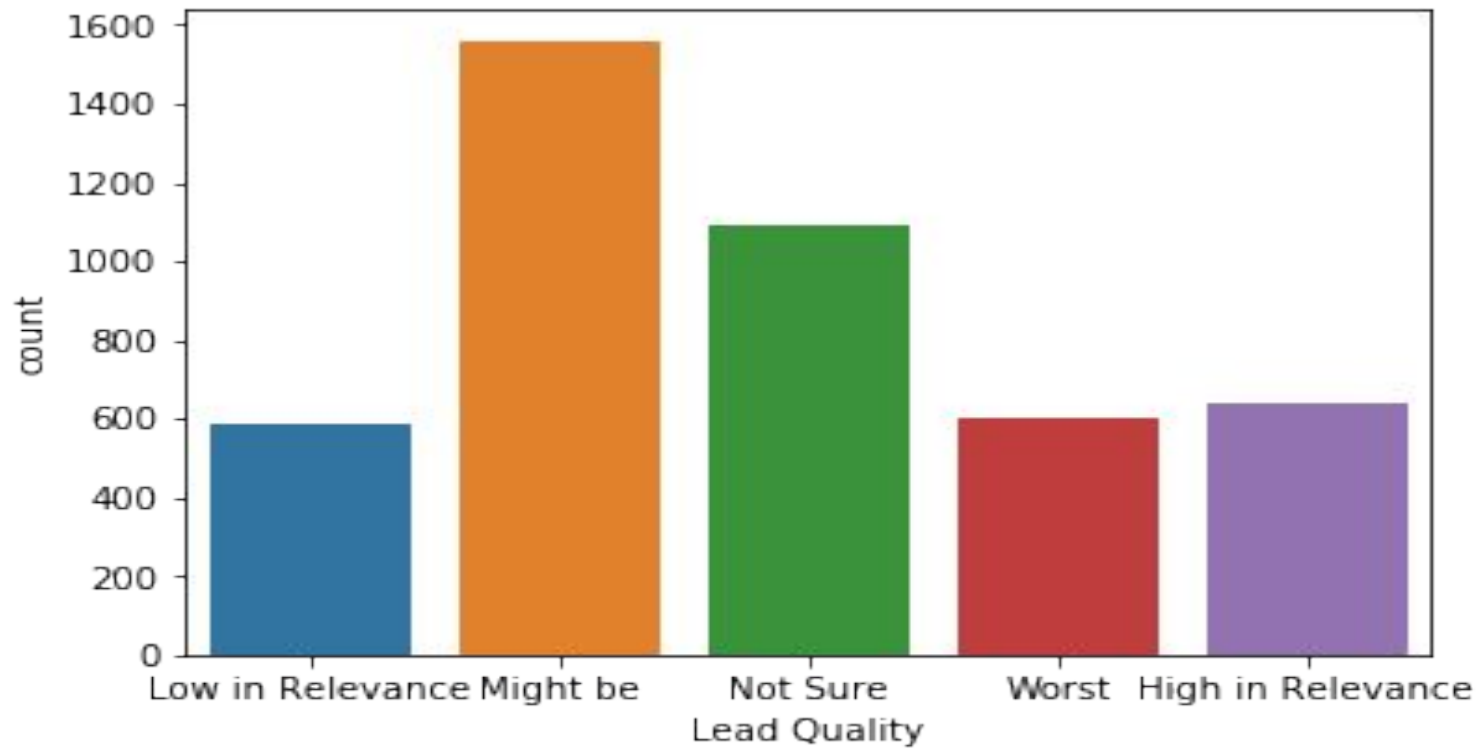
Insight: As Lead quality is based on the intuition of employee, so if left blank we can impute the value Not Sure inplace NaN value.

Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score
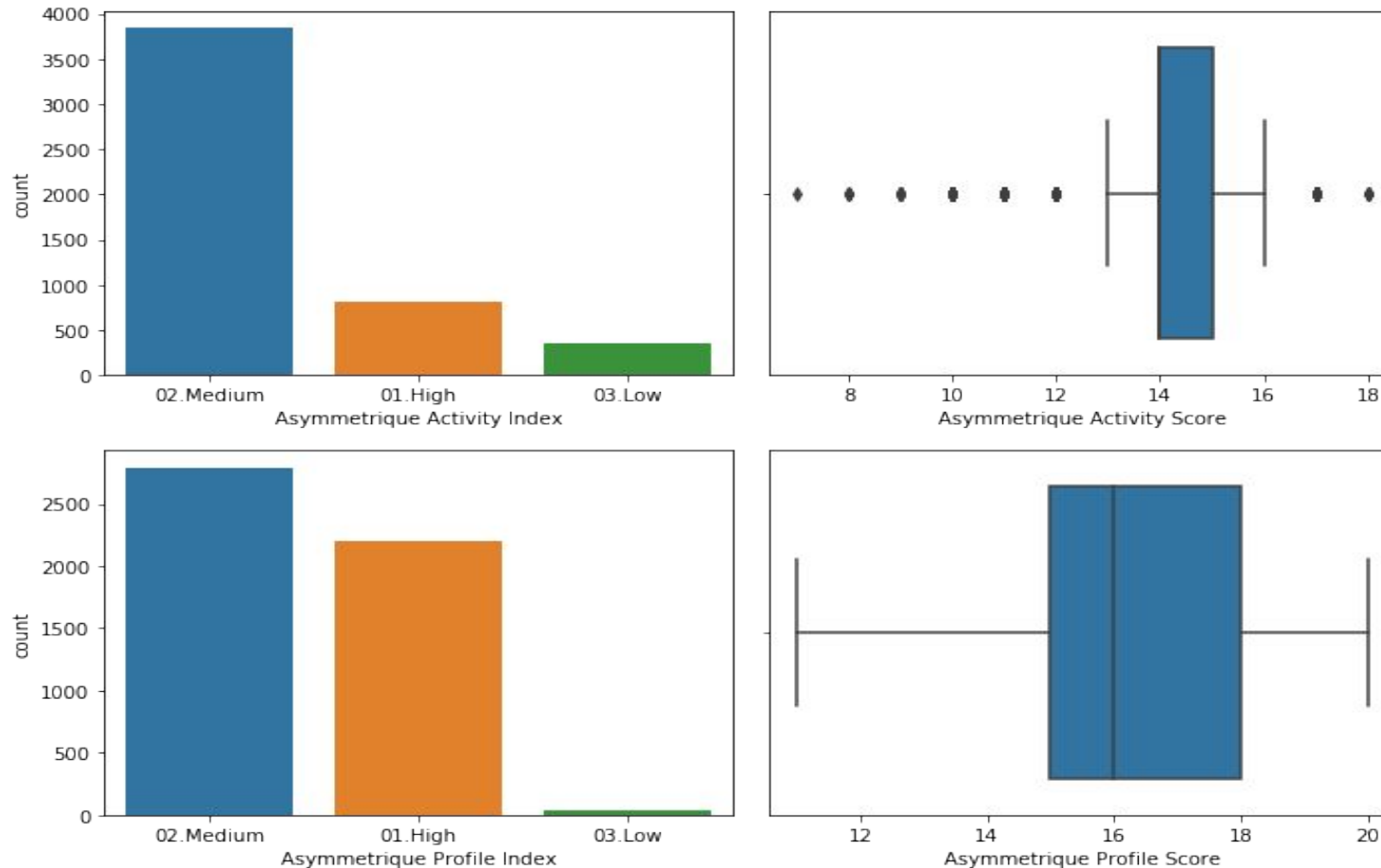
Insight: There is too much variation in these parameters, so it's better not to impute any value in it. Also, all these columns contain more than 45% null values. Hence, we can drop these columns.
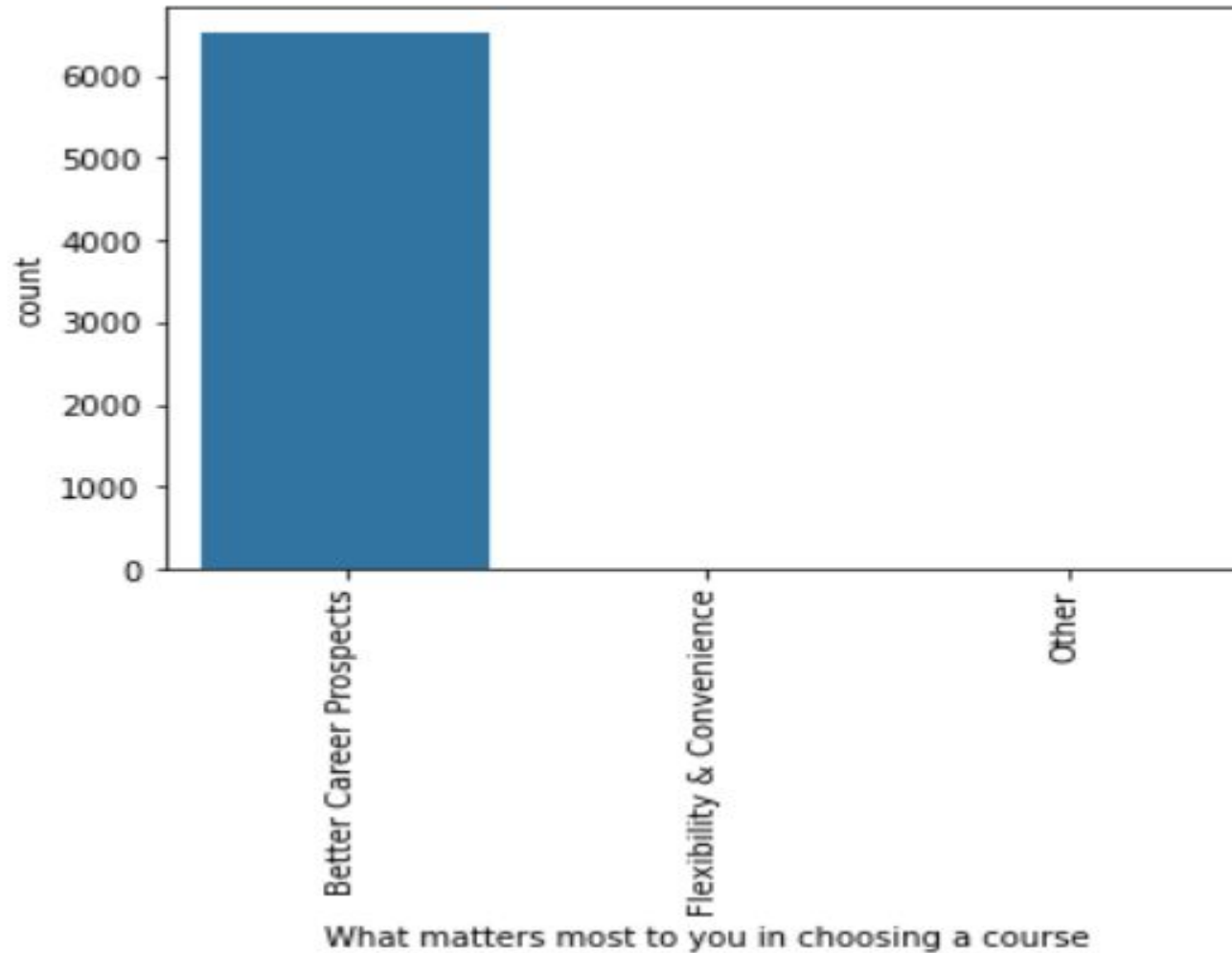
City
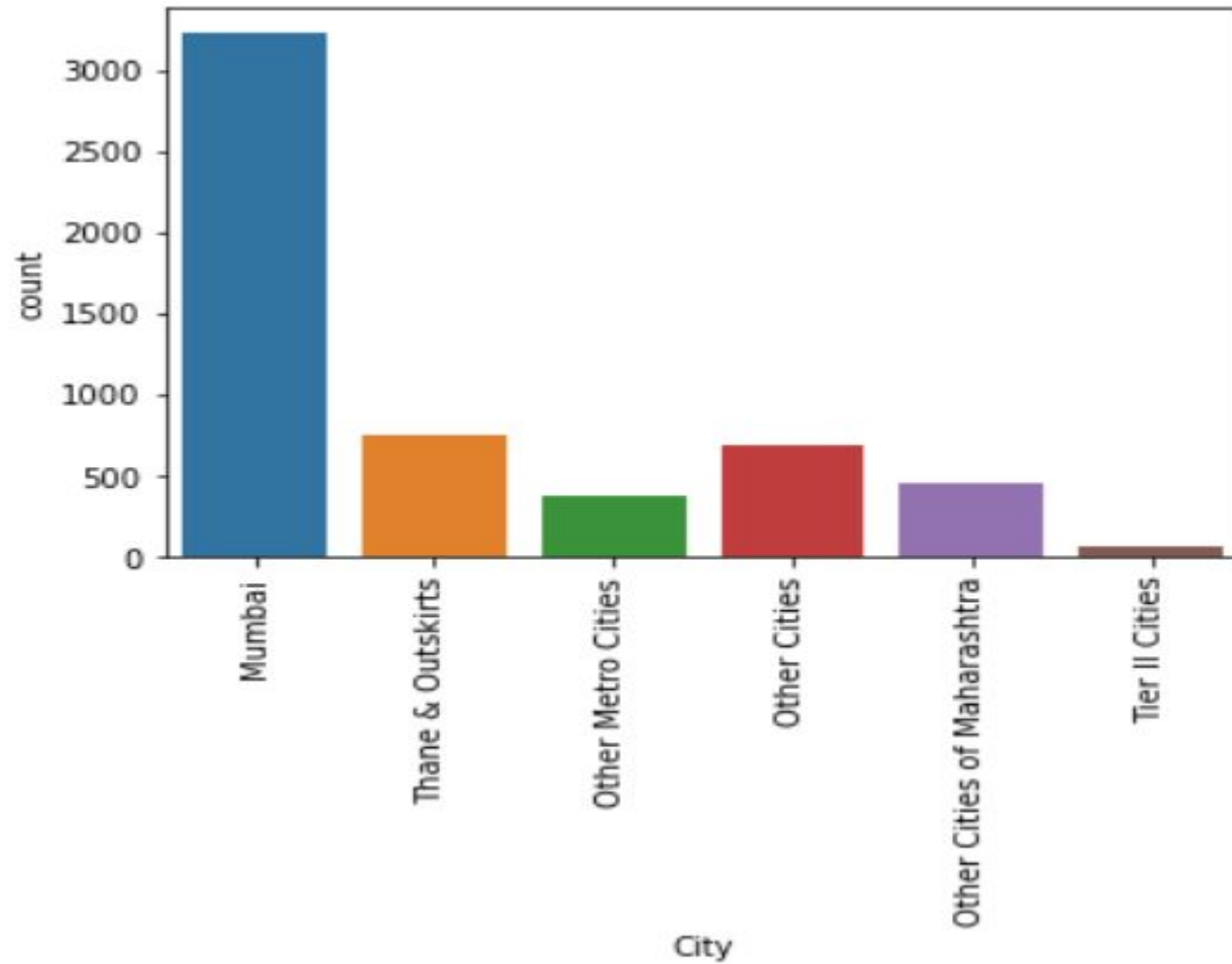
Insight: Around 60% of the data is Mumbai, so we can impute Mumbai in the missing values.

Tags
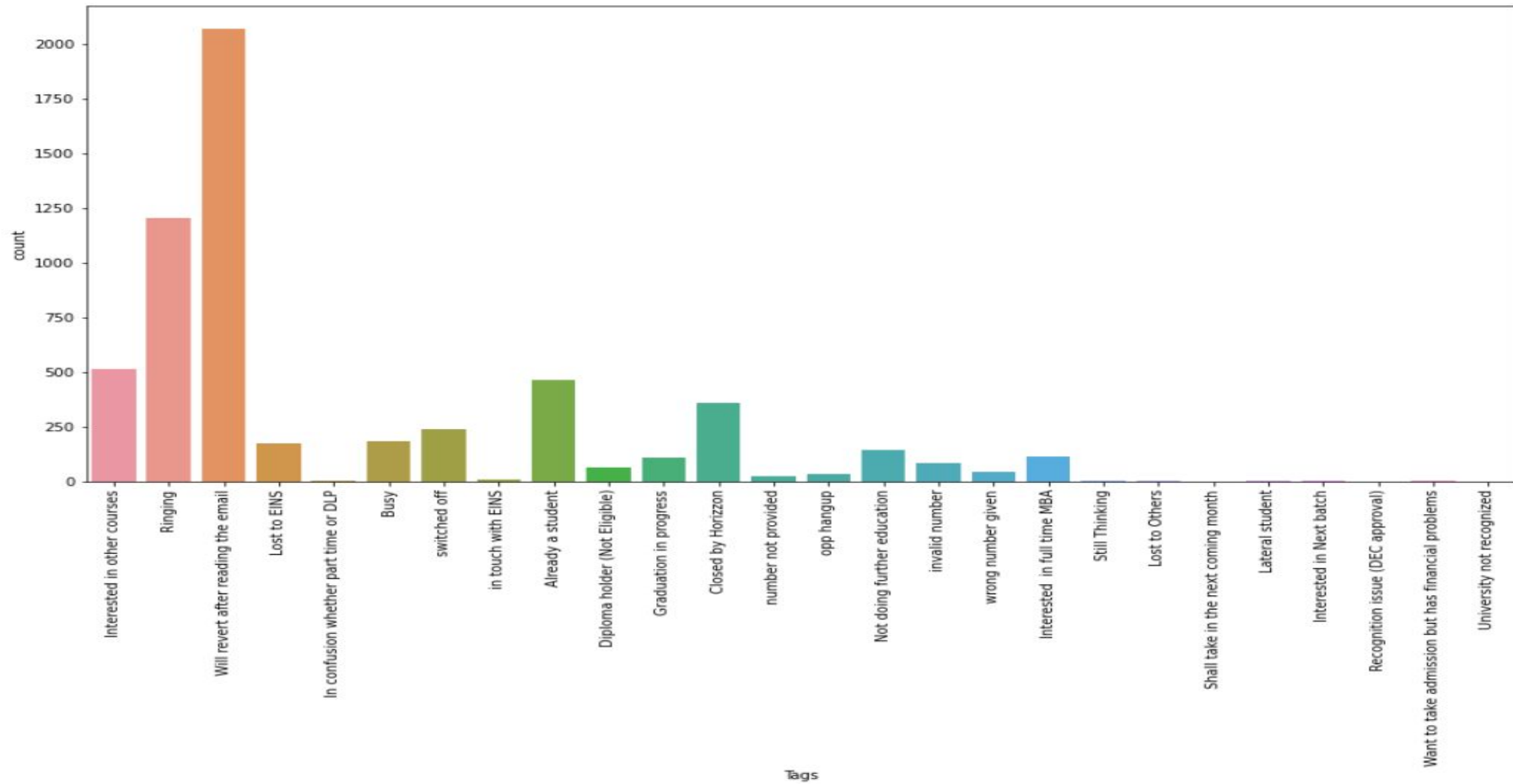
Insight: Missing values in the tags column can be replaced by Will revert after reading the email value as it the most common value.
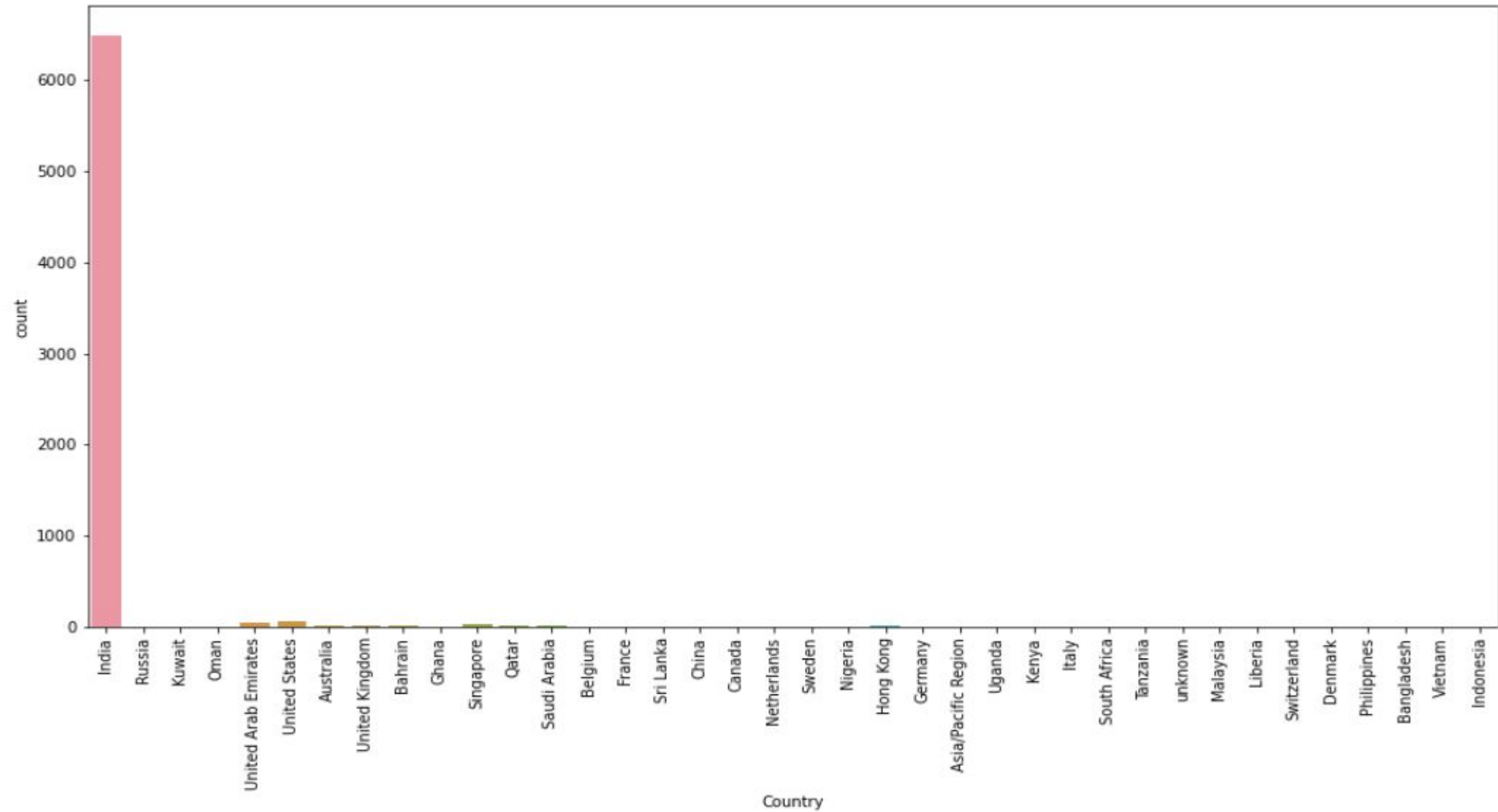
What matters most to you in choosing a course

Insight: Missing values in this column can be replaced by Better Career Prospects as it is the most frequent value.

What is your current occupation
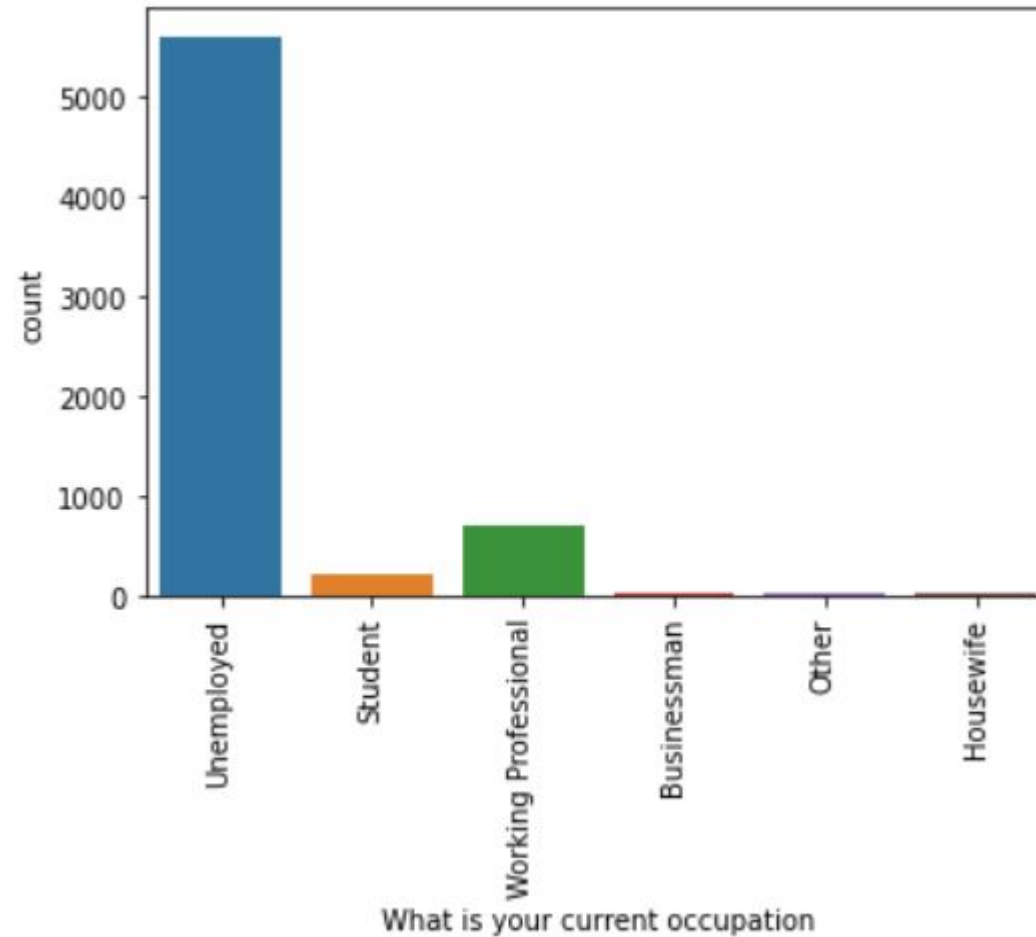
Insight: Nearly 86% entries are of Unemployed. So, we can replace Unemployed in place of missing values.
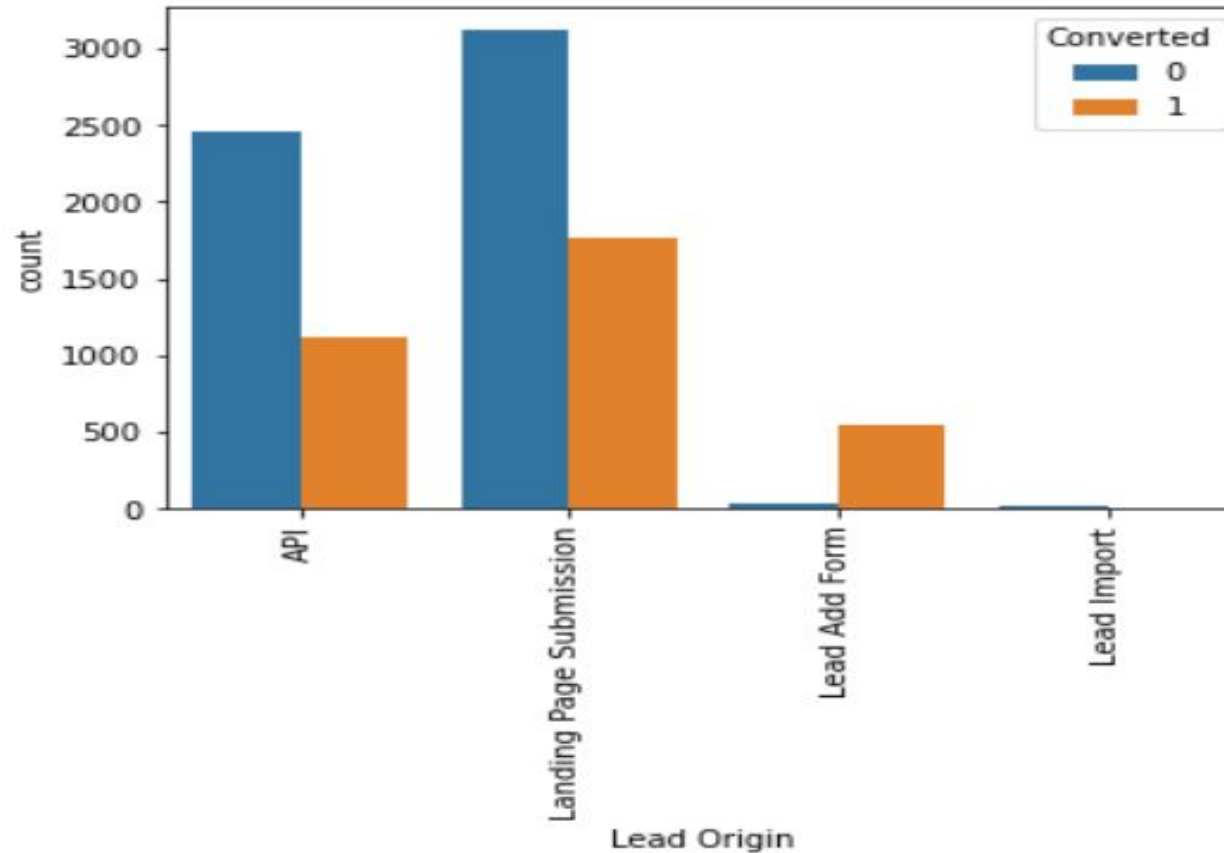
Country

Insight: Country is India for most values so we can replace the same in missing values.

Univariate Analysis

Lead Origin

Insight: API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable. Lead Add Form has more than 90% conversion rate but count of lead are not very high. Lead Import are very less in count.

Inference: To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Lead Source
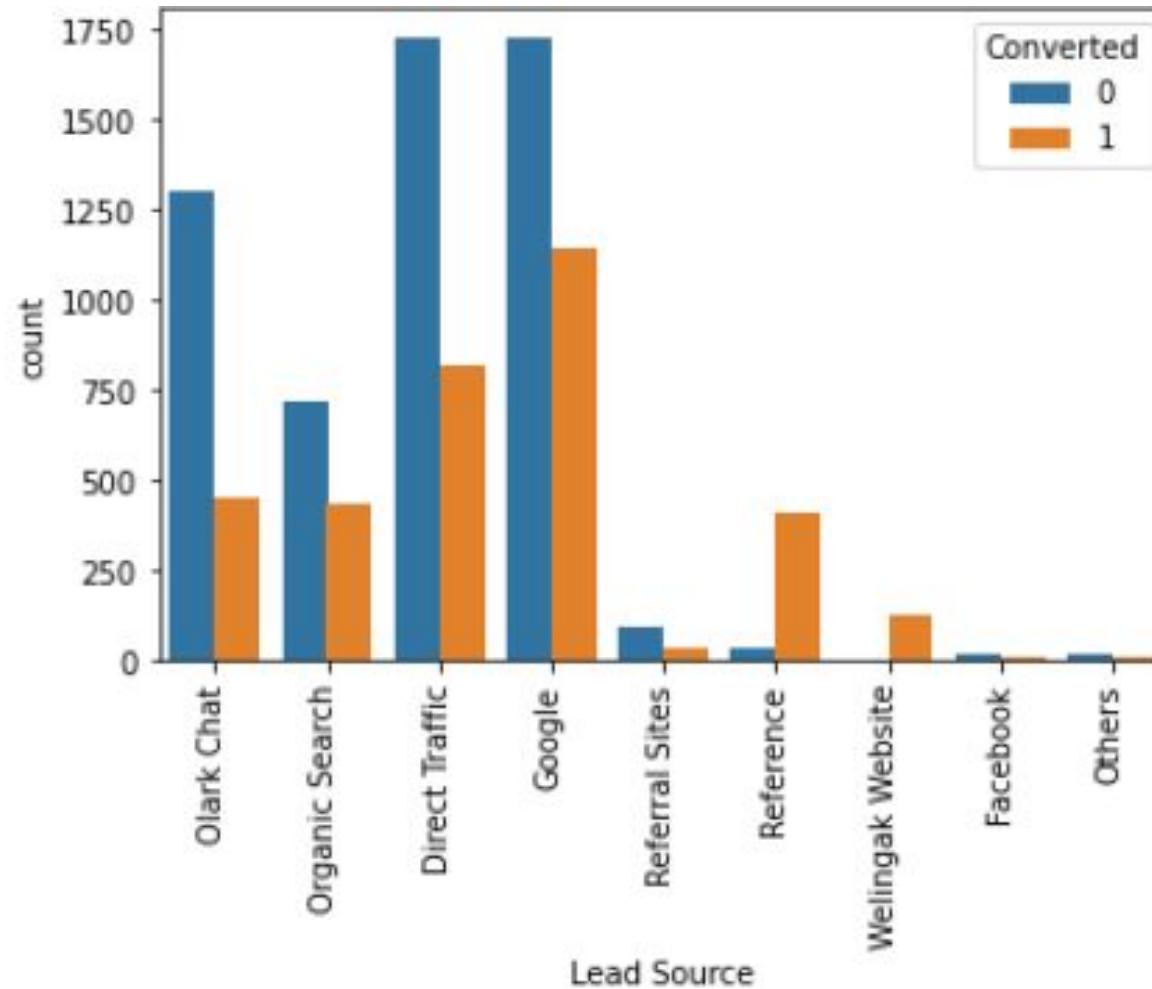
Insight:
Google and Direct traffic generates maximum number of leads.
Conversion Rate of reference leads and leads through welingkar website are high.
Inference: To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google leads and generate more leads from reference and welingak website.
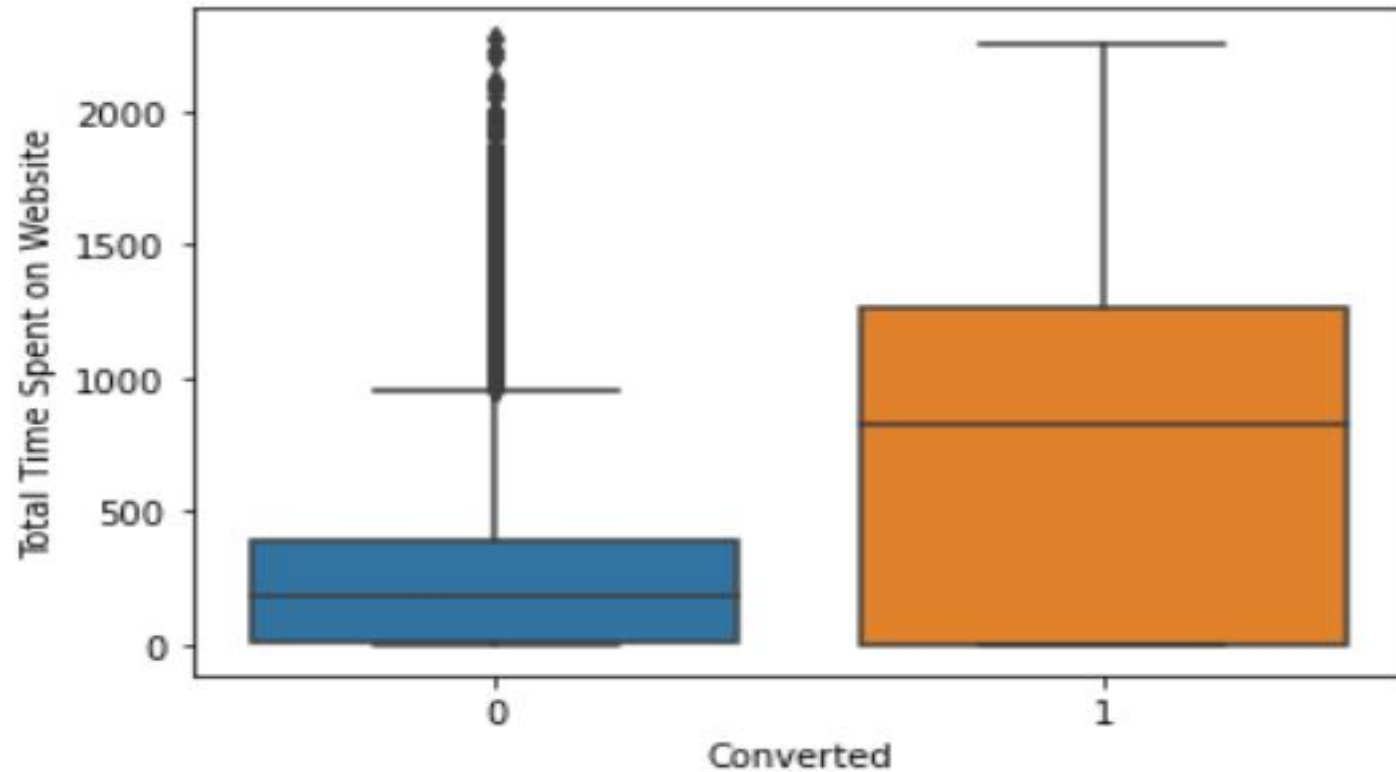
Total Visits
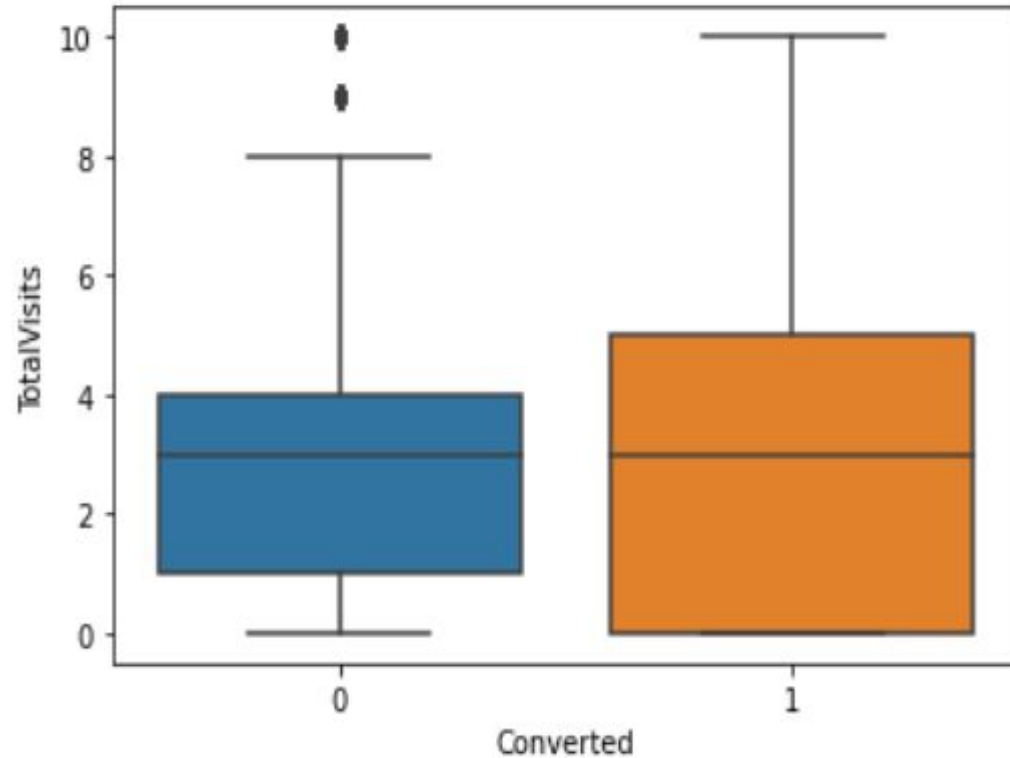
Page views per visit
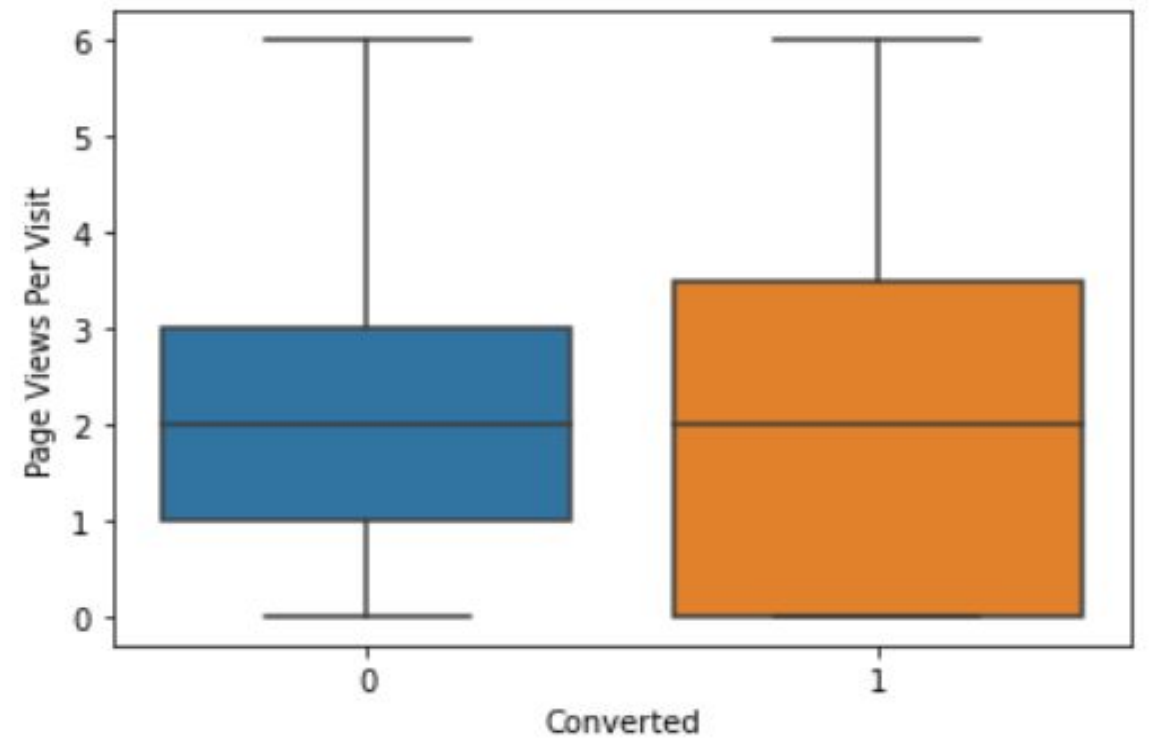
Insight: Median for converted and not converted leads is the same.
Inference: Nothing conclusive can be said on the basis of Total Visits,Page views per visit.
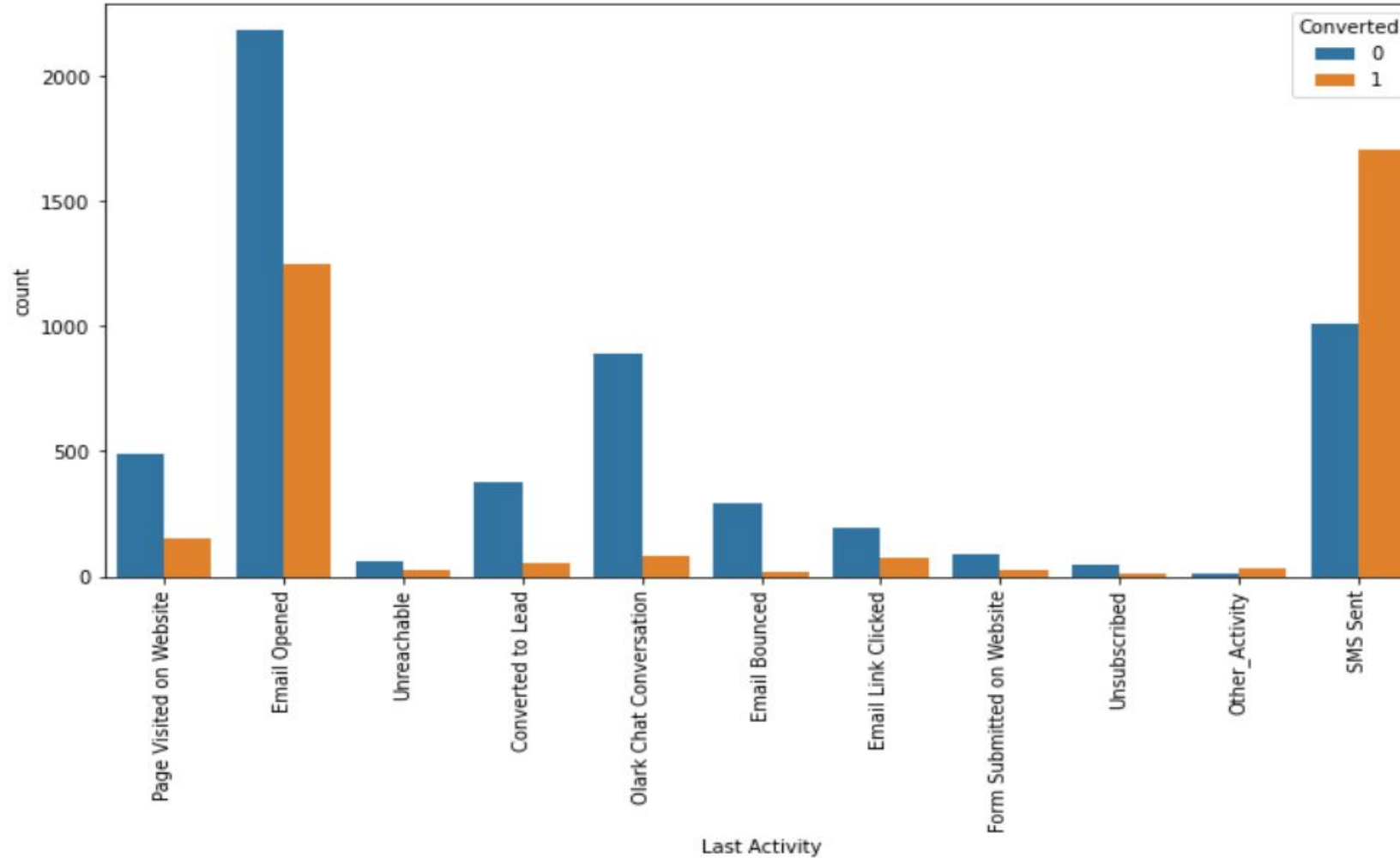
Total time spent on website

Insight:
Leads spending more time on the website are more likely to be converted.
Inference: Website should be made more engaging to make leads spend more time.
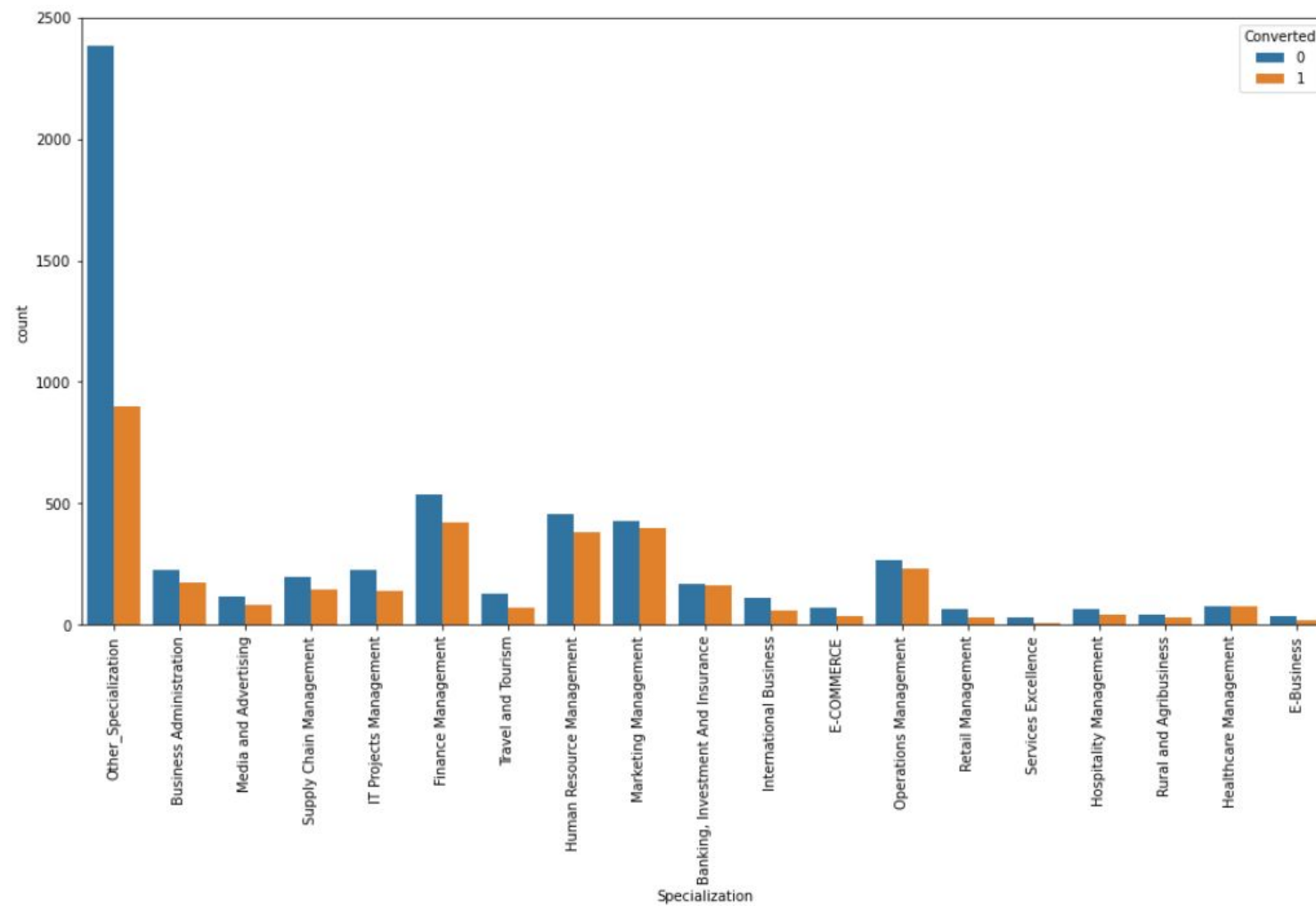
Last Activity

Inference:
Most of the leads have their Email opened as their last activity.
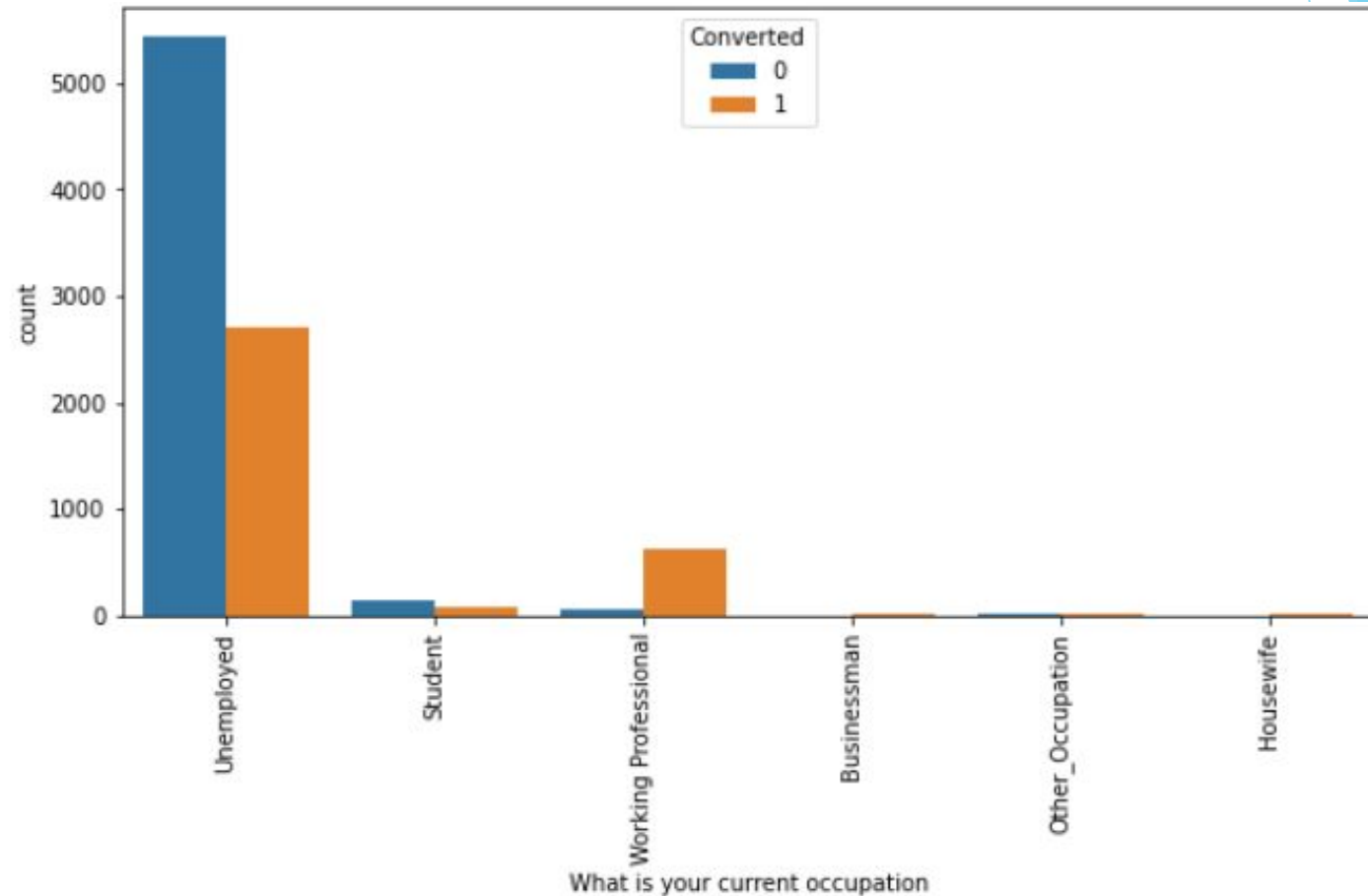Conversion rate for leads with last activity as SMS Sent is almost 60%.

Specialization

Inference: Focus should be more on the Specialization with high conversion rate.
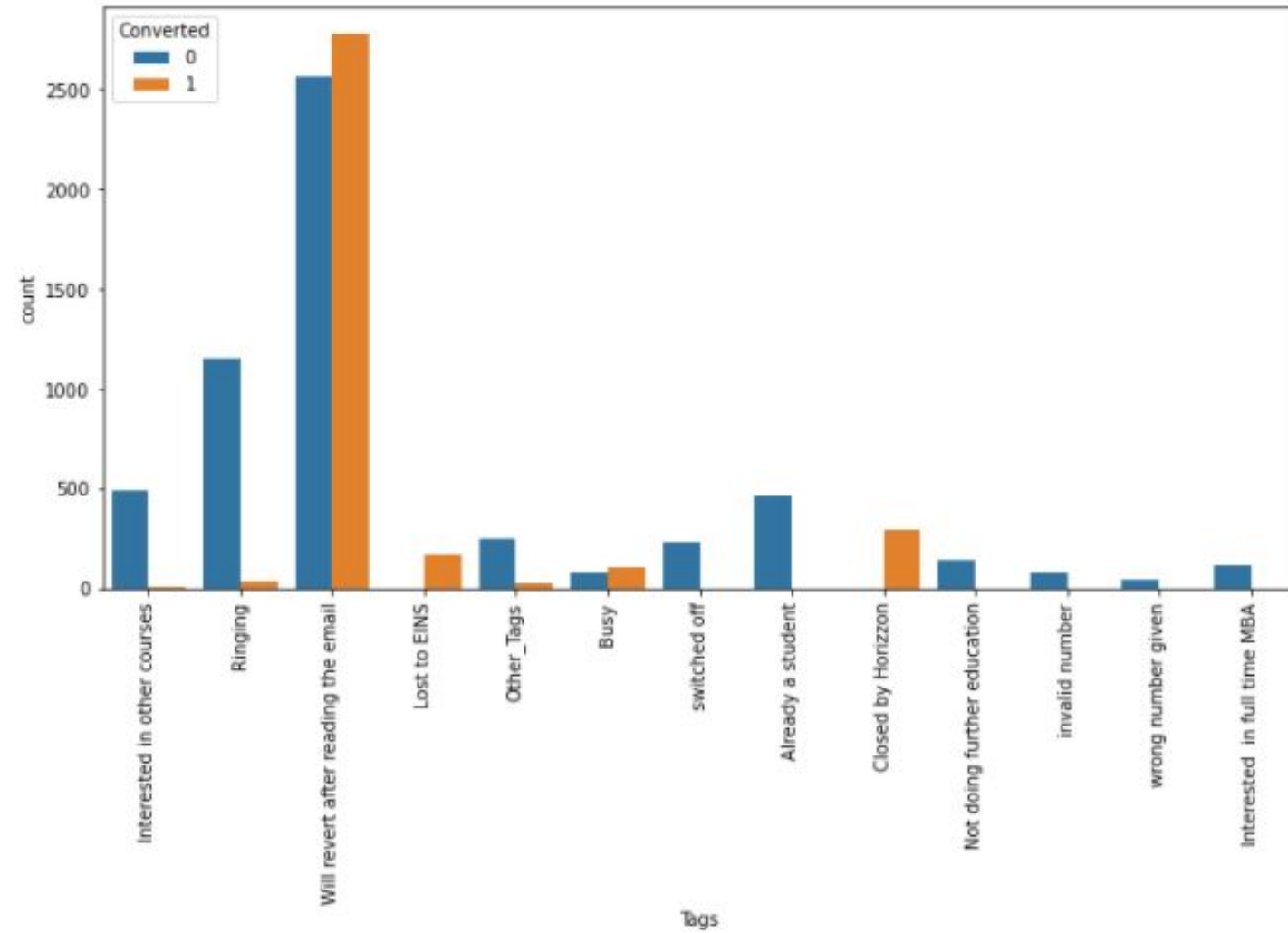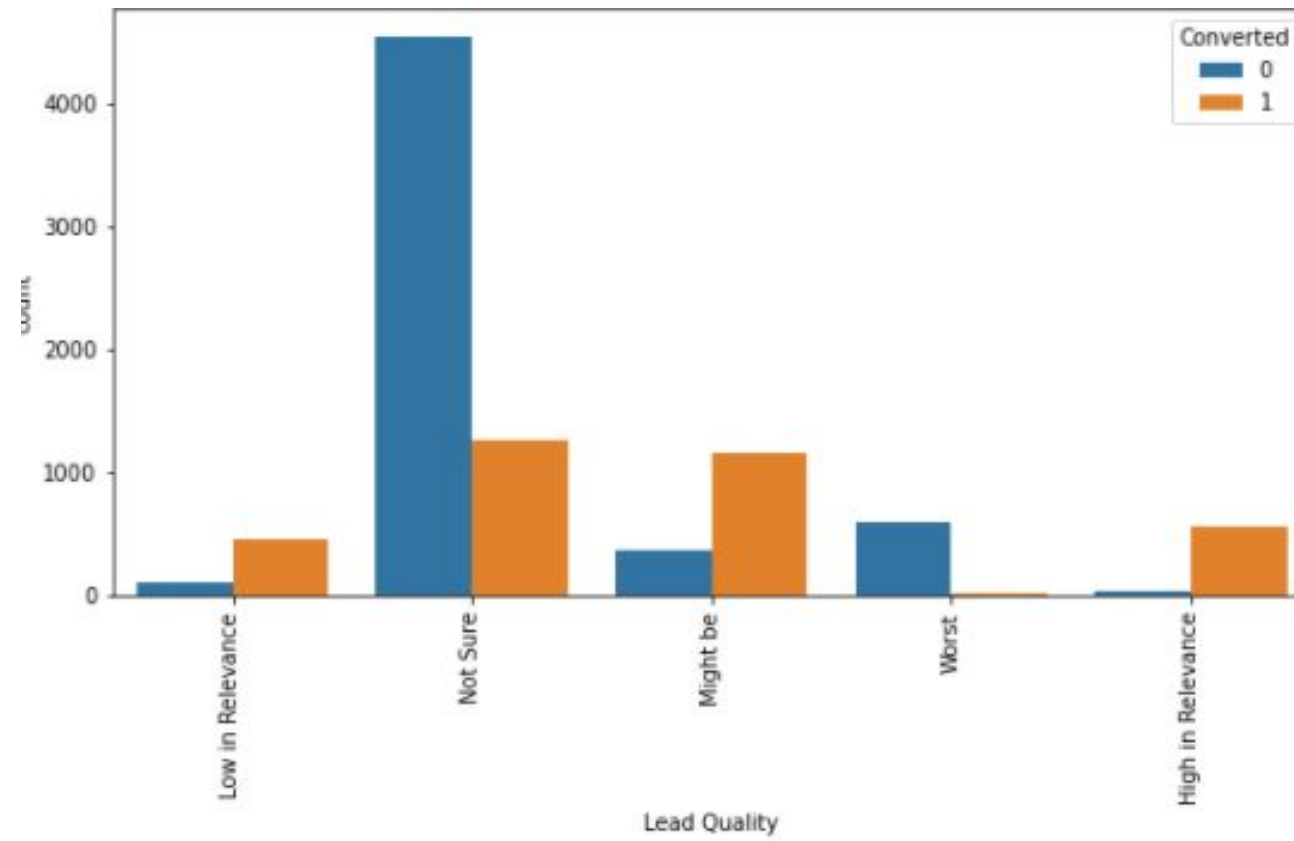
Occupation

Inference:
Working Professionals going for the course have high chances of joining it.
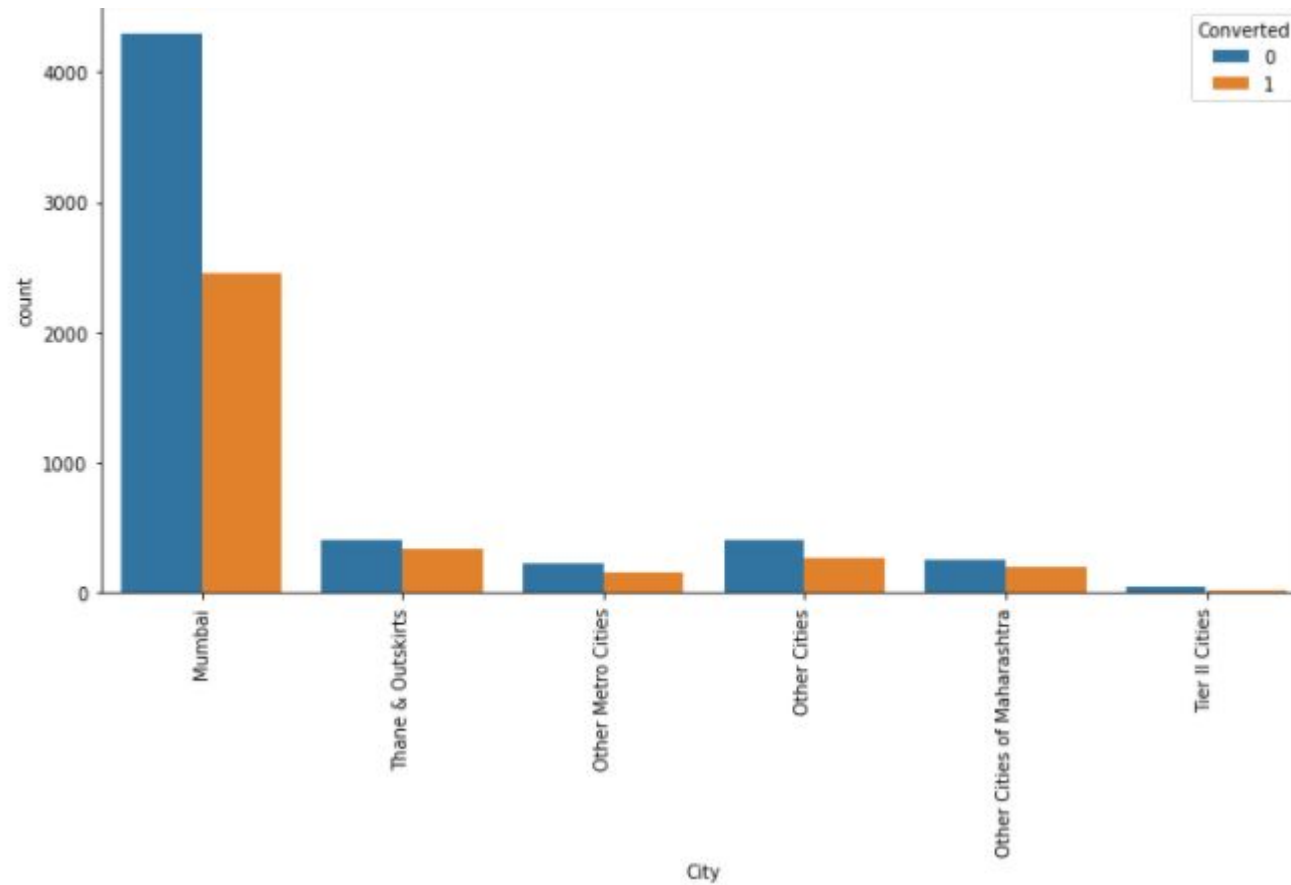Unemployed leads are the most in numbers but has around 30-35% conversion rate.
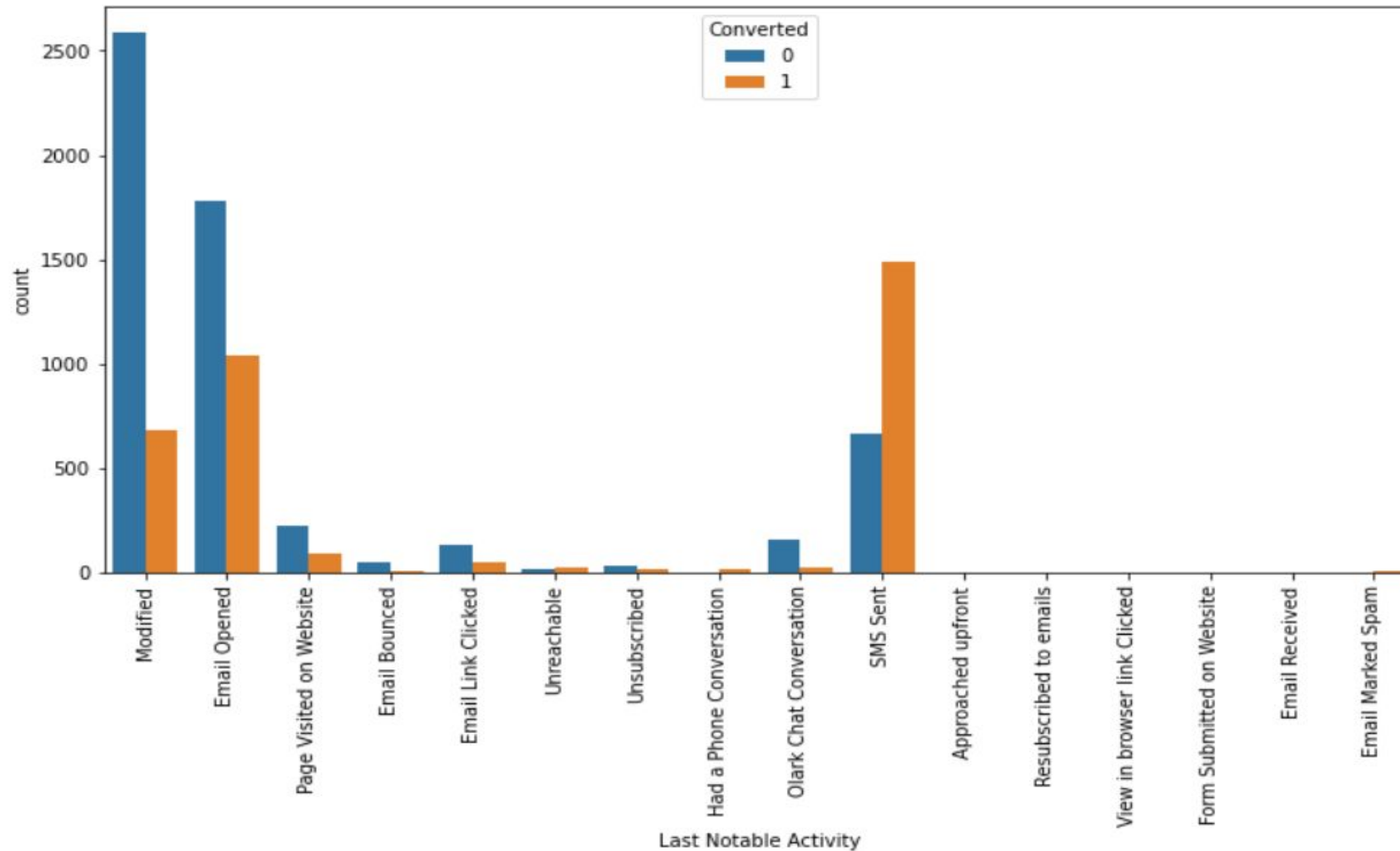
Tags
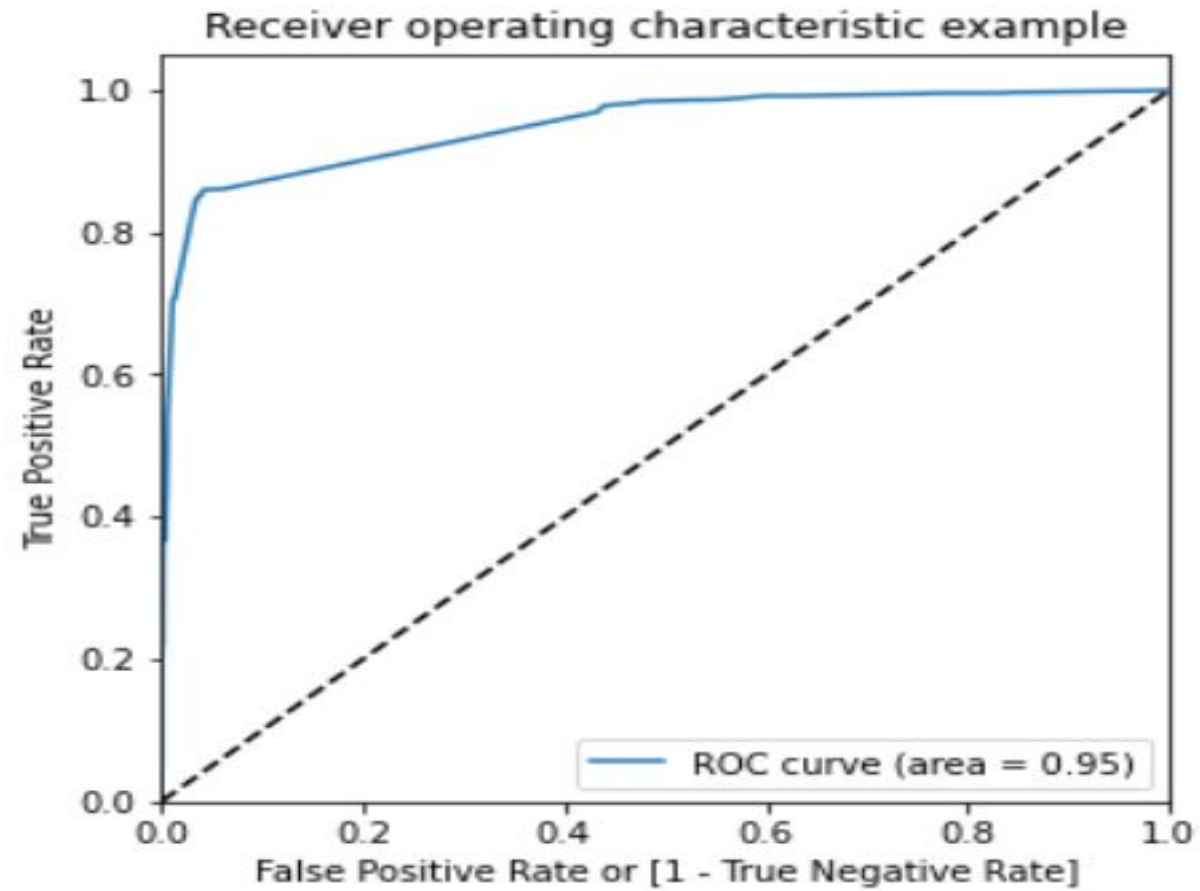
Lead Quality

City

Inference:
Most leads are from Mumbai with around 30% conversion rate.
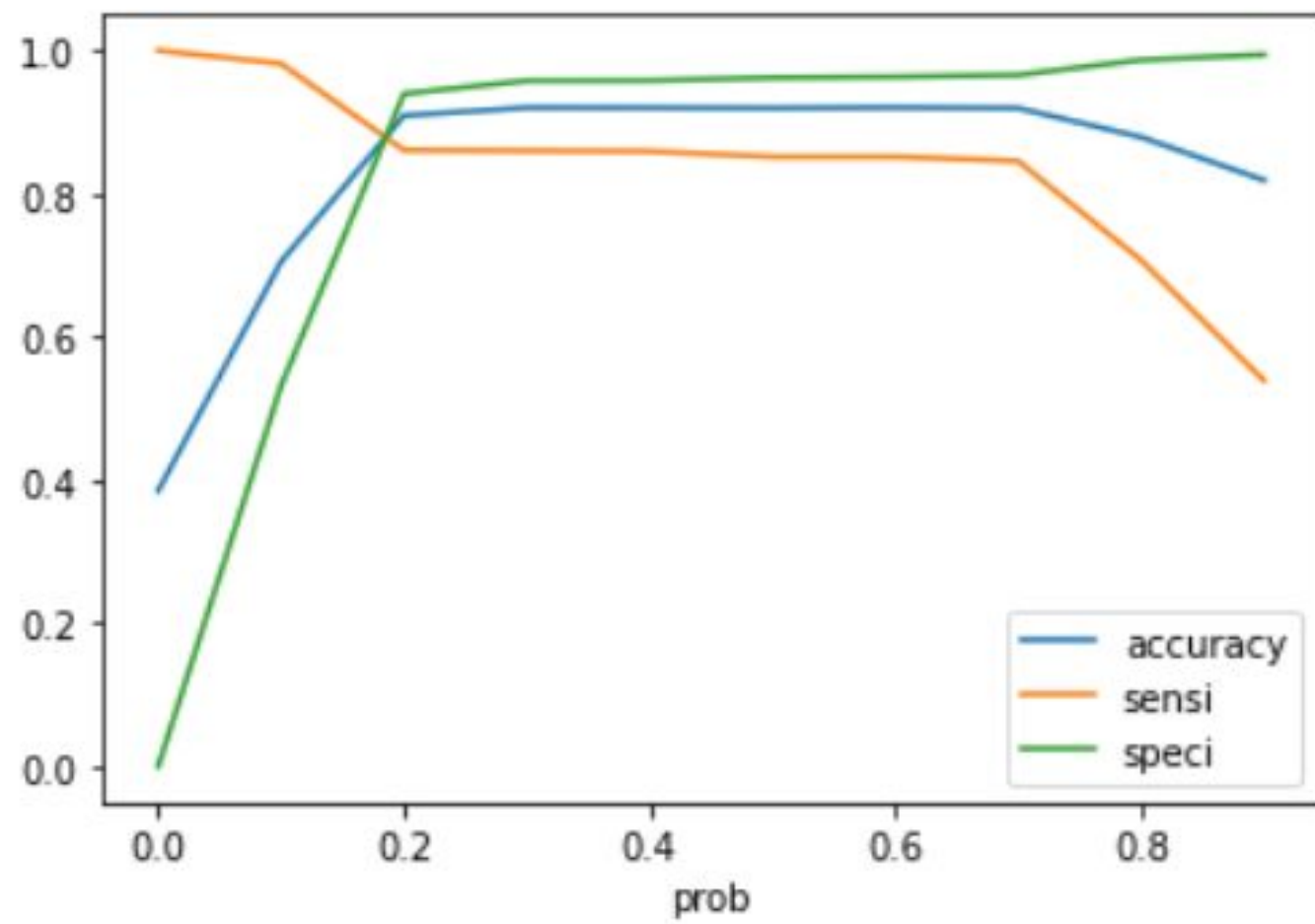
Last Notable Activity

Final Results: Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we can drop these columns for further analysis.
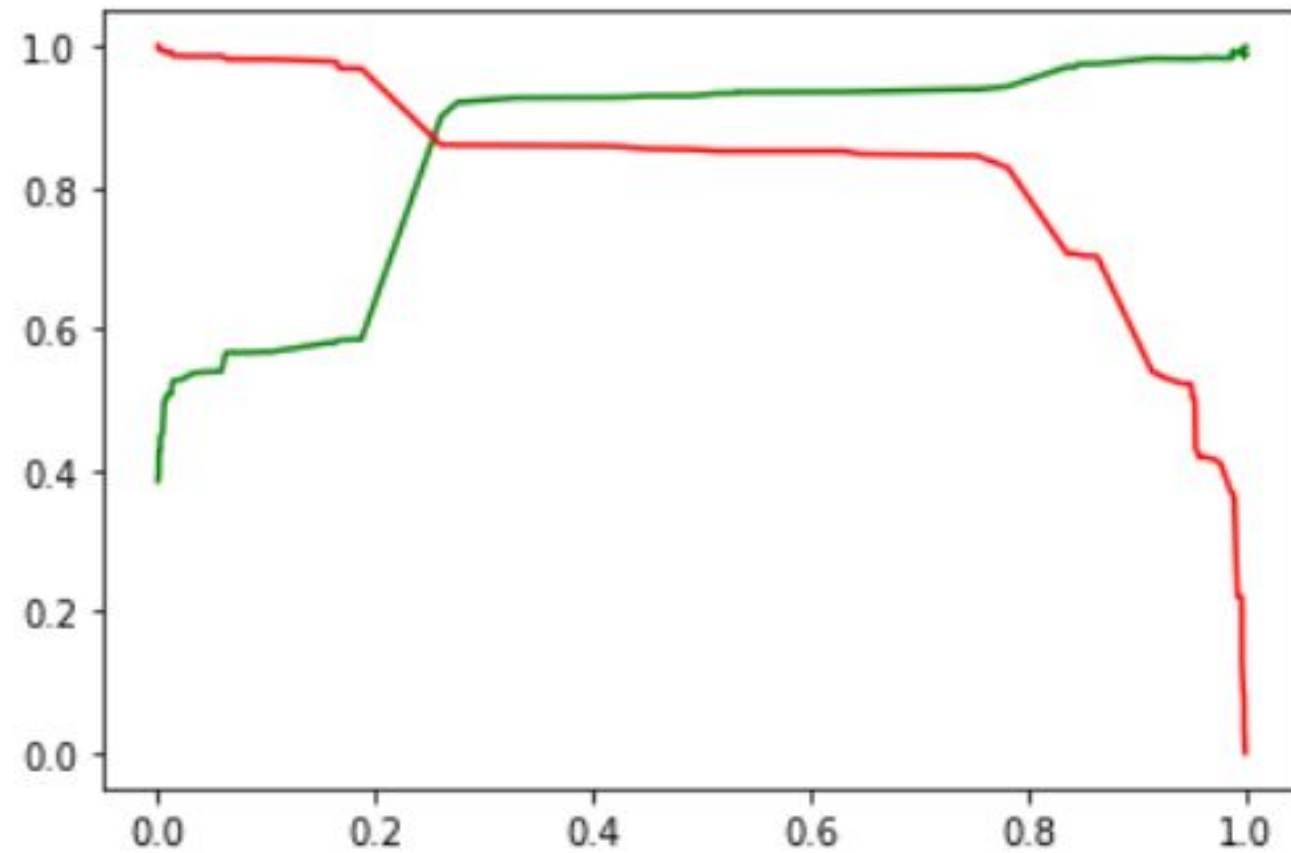
ROC Curve

Optimal Cutoff Point

Precision and recall tradeoff

Conclusion

Model Analysis
Performance of our Final Model

❑ Overall accuracy on Test set: 90.45%.

❑ Sensitivity of our logistic regression model: 84.43%.

❑ Specificity of our logistic regression model: 93.89%.

Business Insights Derived from our Model:

Top 3 variables in our model, that contribute towards lead conversion are:
1. Tags
2. Lead Source
3. Last Notable Activity

Top 3 categorical/dummy variables in our model, that should be focused are:
1. Tags_Lost to EINS (positively impacting)
2. Tags_Closed by Horizzon (positively impacting)
3. Tags_Will revert after reading the email (positively impacting)

Conclusion (LR Model)

Our Logistic Regression Model is decent and accurate enough, with 90.45% Accuracy on Test Set, 84.43% Sensitivity and 93.89% Specificity.
We can vary these parameters by varying the cut-off value and thus predict Hot leads based on scenarios like availability of extra resources and vice-versa.

Conclusion (Recommendation)

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

❑ Increase user engagement on their website since this helps in higher conversion.

❑ Increase on sending SMS notifications since this helps in higher conversion.

❑ Get Total visits increased by advertising etc. since this helps in higher conversion.

❑ Improve the Lead Quality since this is affecting the conversion negatively.