Lead Scoring Case Study

- GROUP MEMBERS (DS C46 JULY 2022) :

1. GOVIND SINGH MEHRA
2. JENITH MEHTA
3. SHUBHAM DIXIT

PROBLEM STATEMENT

X Education has appointed you to help them select the most promising leads, i.e. 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

1.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.

2.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.

APPROACH FOR THE BUSINESS SOLUTION

A. Data Preparation, cleaning and data manipulation.

   1. Data reading and understanding.
   2. Check missing values.
   3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
   4. Handle and Treatment of outliers in data.

B. Exploratory Data Analysis (EDA).

   1. Univariate analysis.
   2. Bivariate analysis.
   3. Multivariate analysis

C. Create the dummies for the variables and splits the data into train and test dataset and scaling the feature

D. Logistic Regression Algorithm helpful to make model building and prediction.

E. RFE(Recursive feature Elimination) use for selecting the best features.

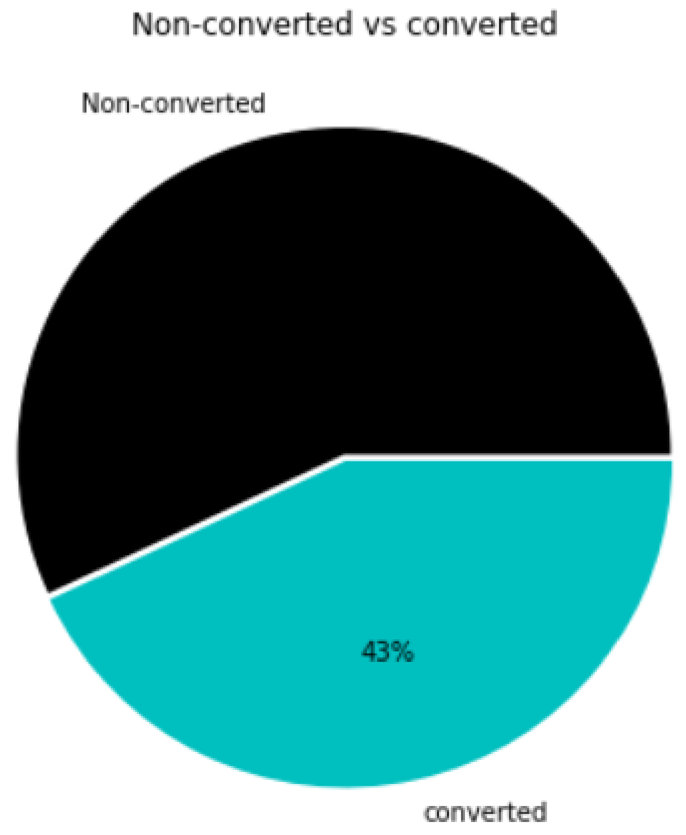
F. Validate the test data

EXPLORATORY DATA ANALYSIS

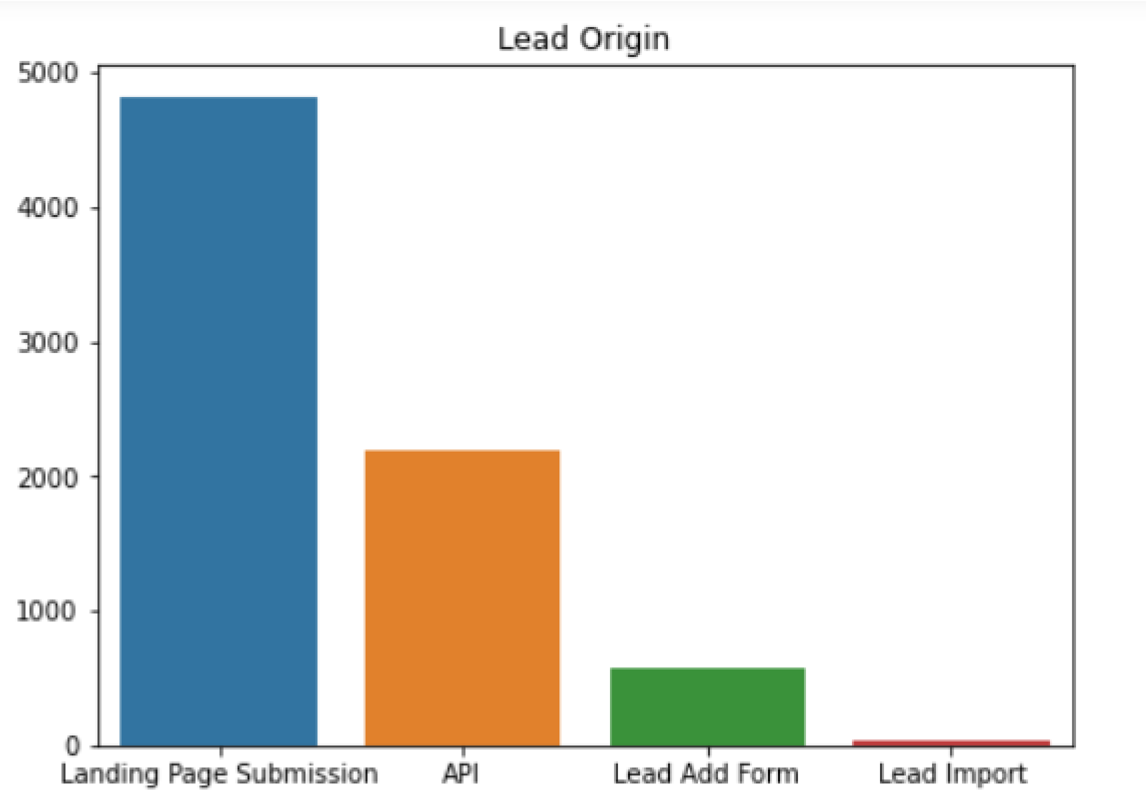
UNIVARIATE ANALYSIS

Converted is the target variable, which tell us whether lead successfully converted (1) or not (0).

As per the graph 43% lead successfully converted.

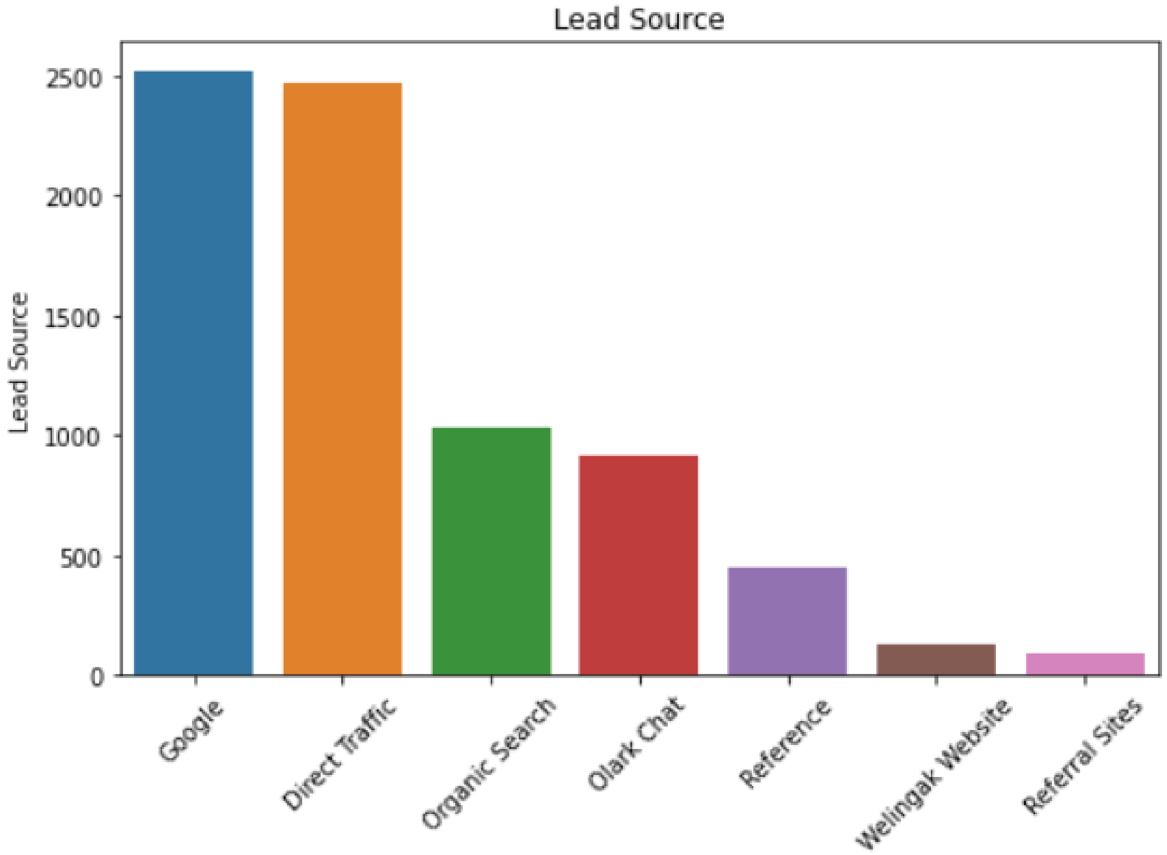


Lead origin



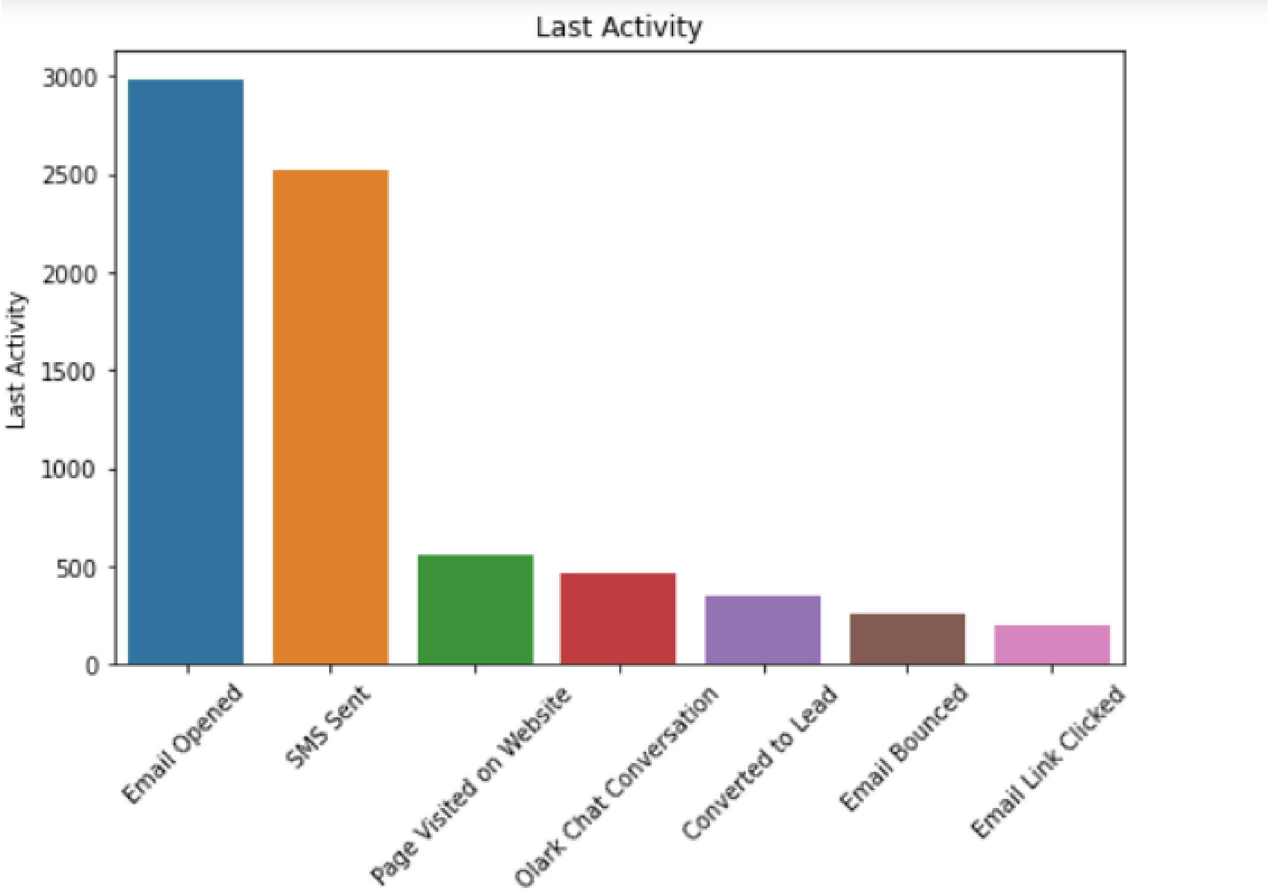
- Most of the lead originating from the Landing page submission followed by API

Lead Source



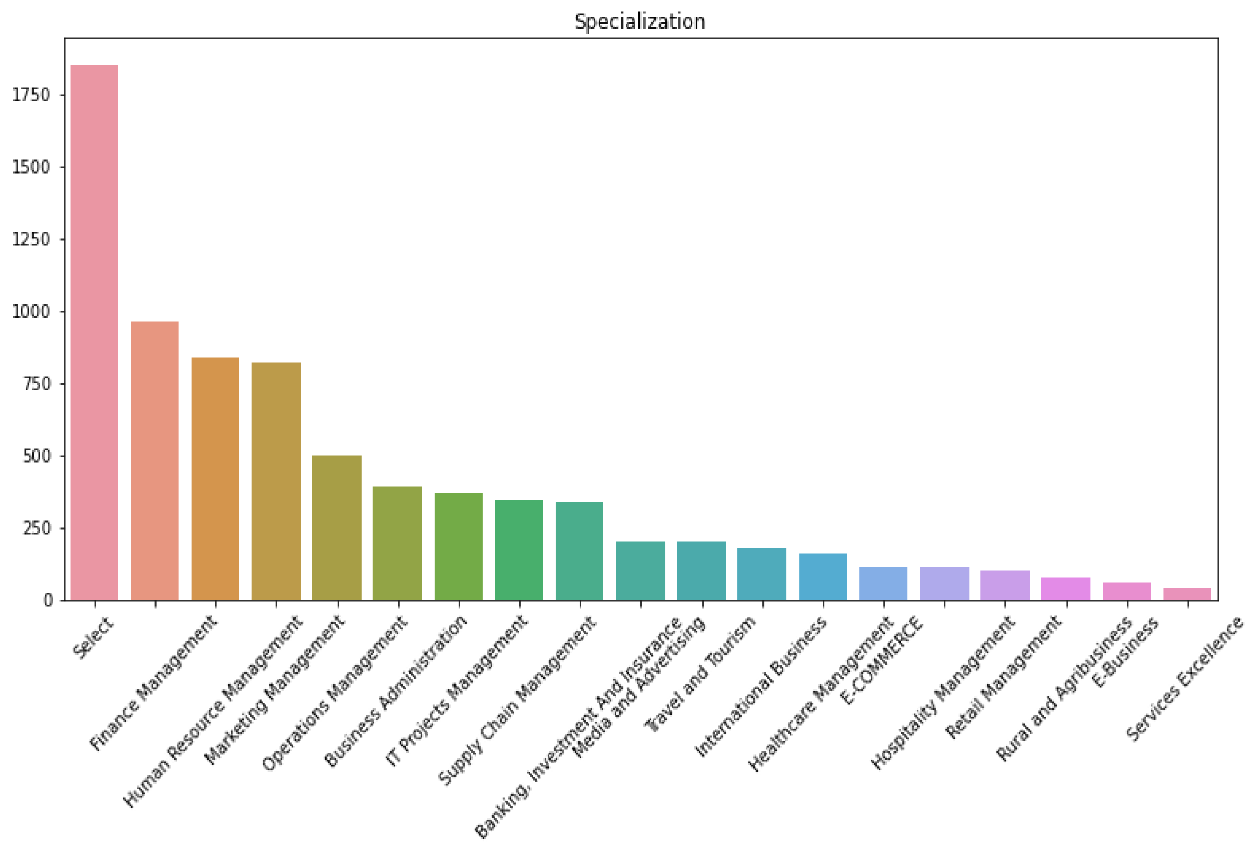
Google found to be the best Lead source.

Last Activity



Email opened followed by SMS are frequently performed by the customer.

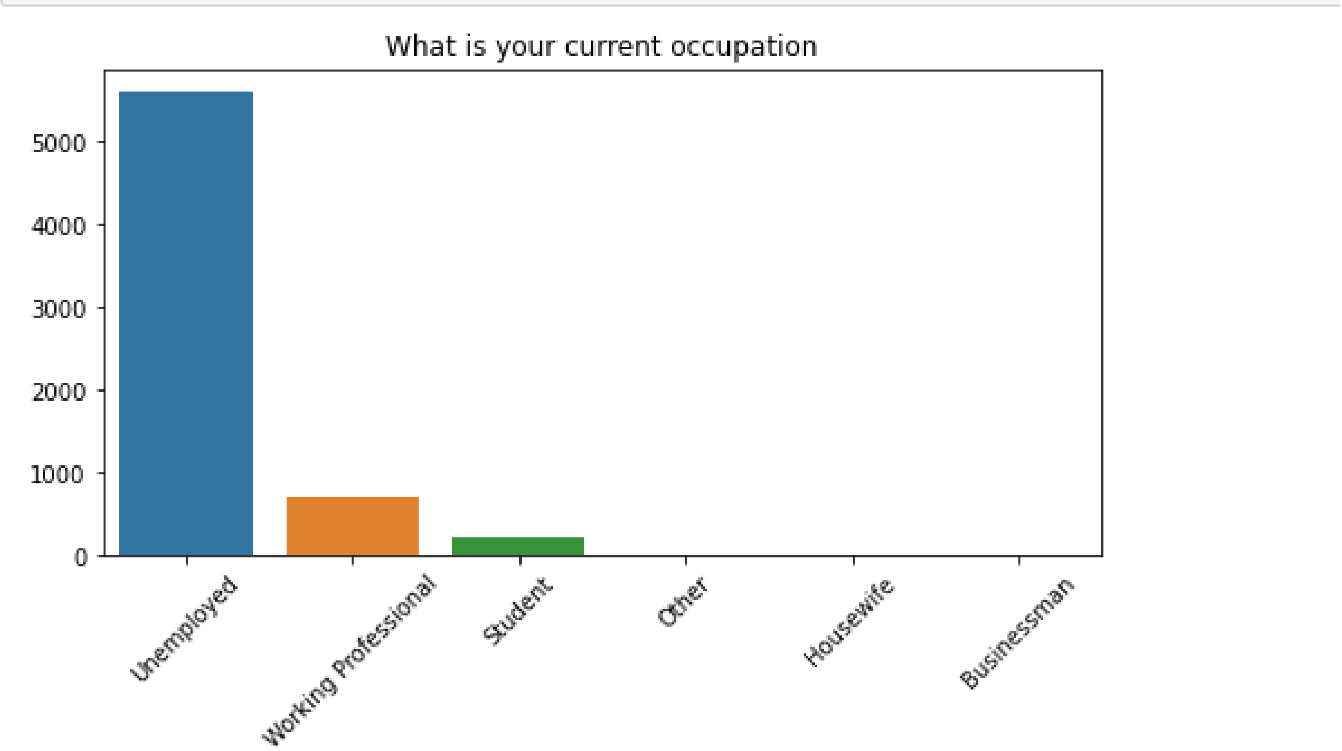
Specialization.



Large Number of customers have not specified their Specialization.

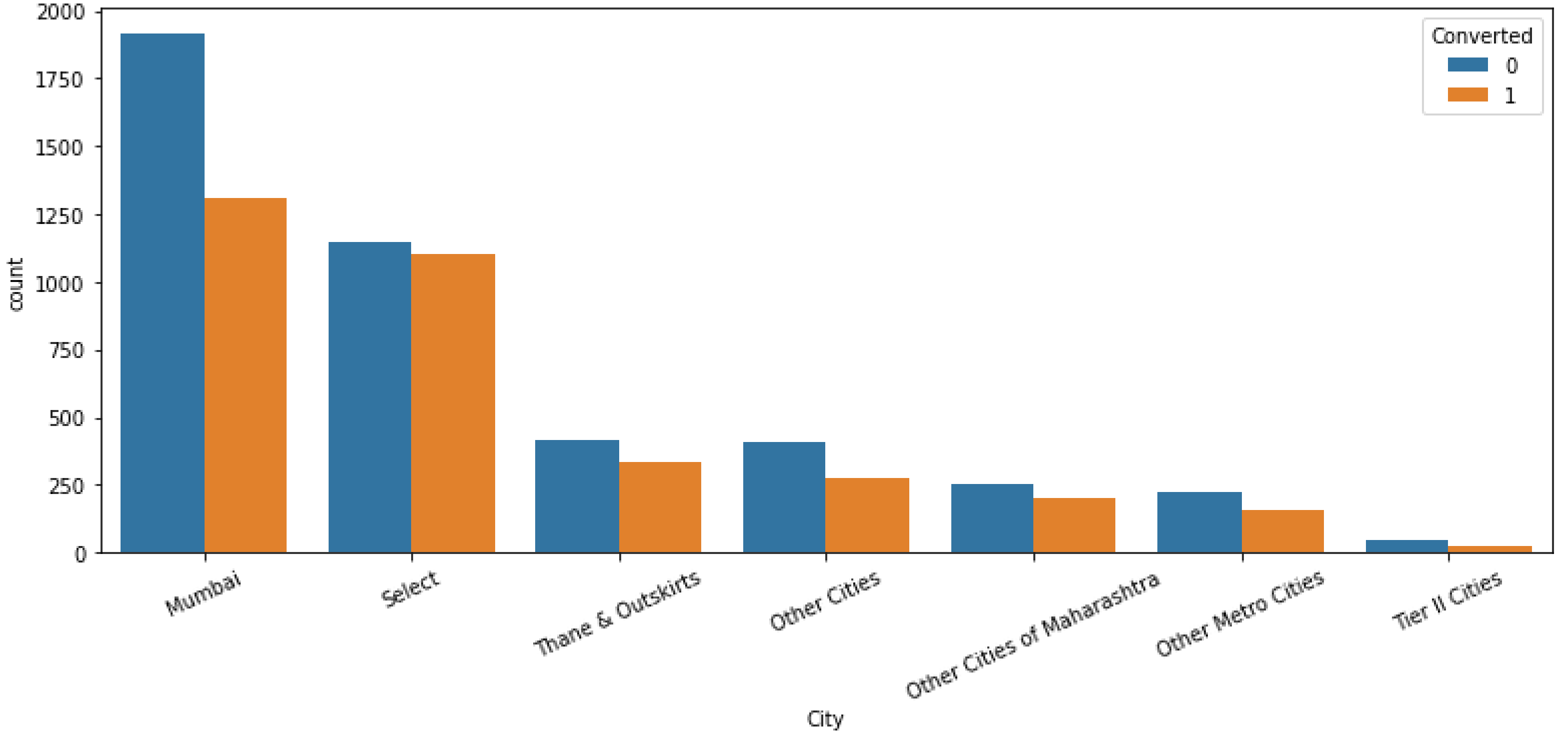
Most of the customers are from management background.

What is your current occupation



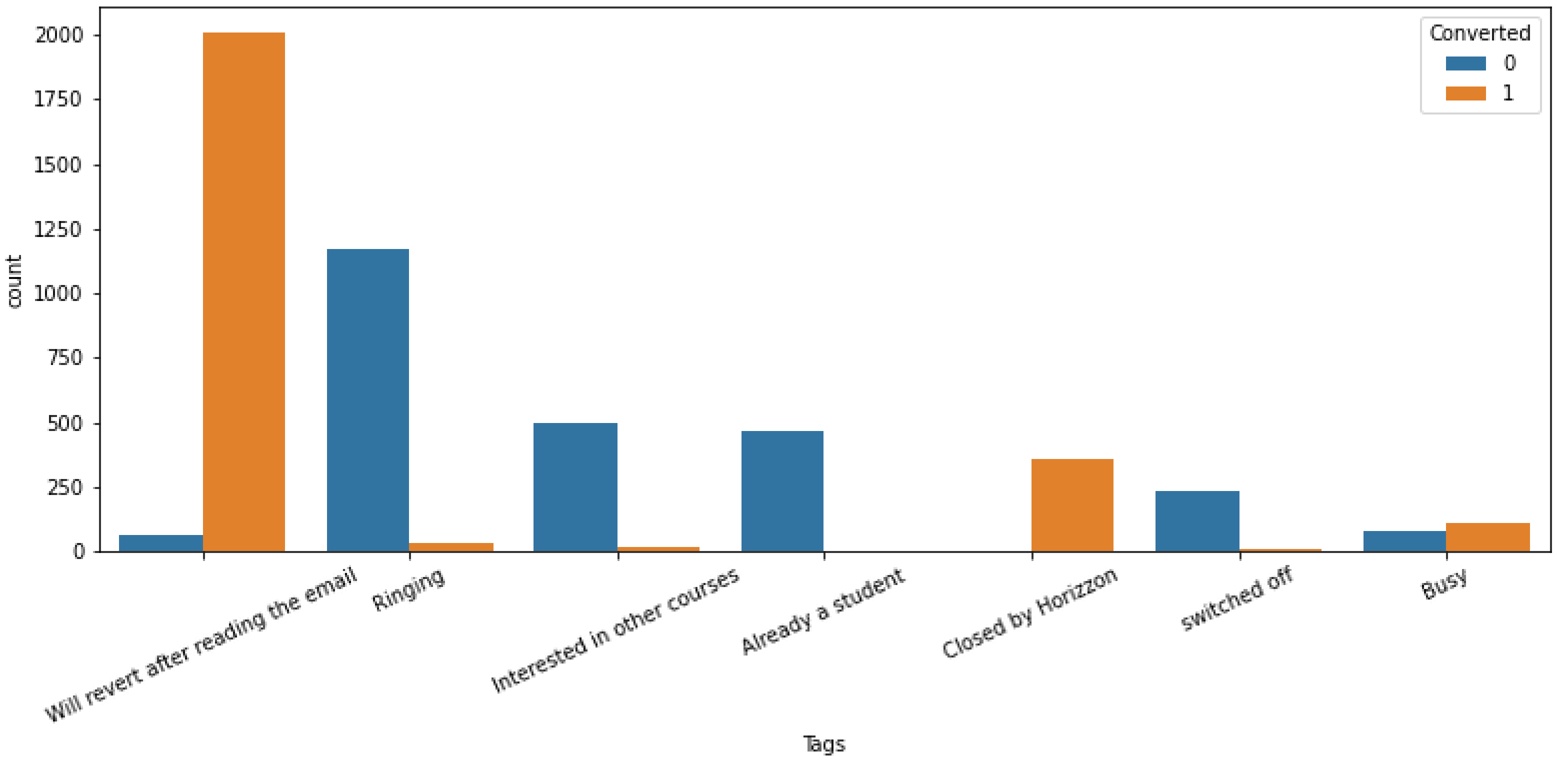
Indicates most of the customer are unemployed followed by working professional.

BIVARIATE ANALYSIS
City



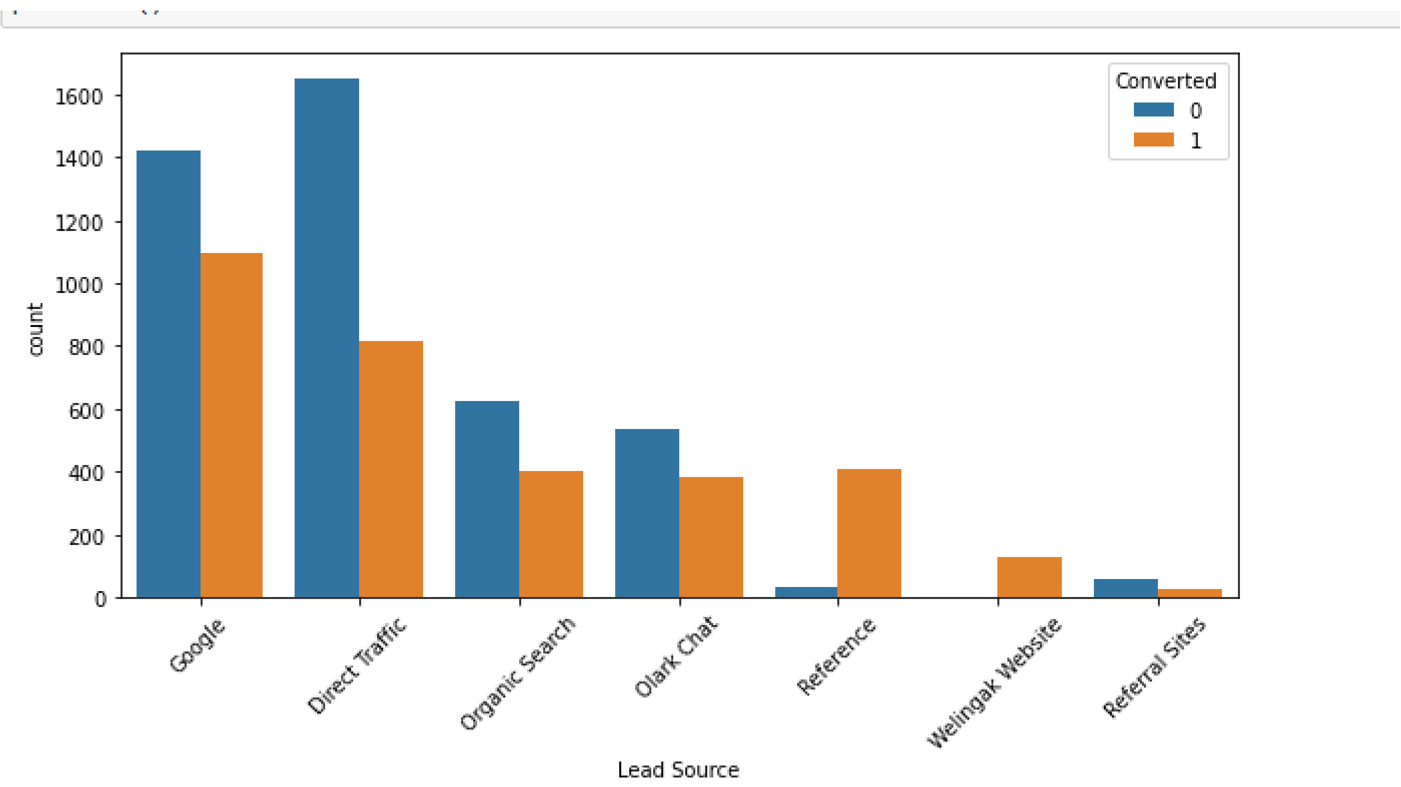
Mumbai and Thane are major places from where the leads are coming. Since the area is limited, company can go for offline marketing like using Billboards, Radio. Also the company should think about expansion to neighbouring cities like Pune and Nagpur.

TAGS



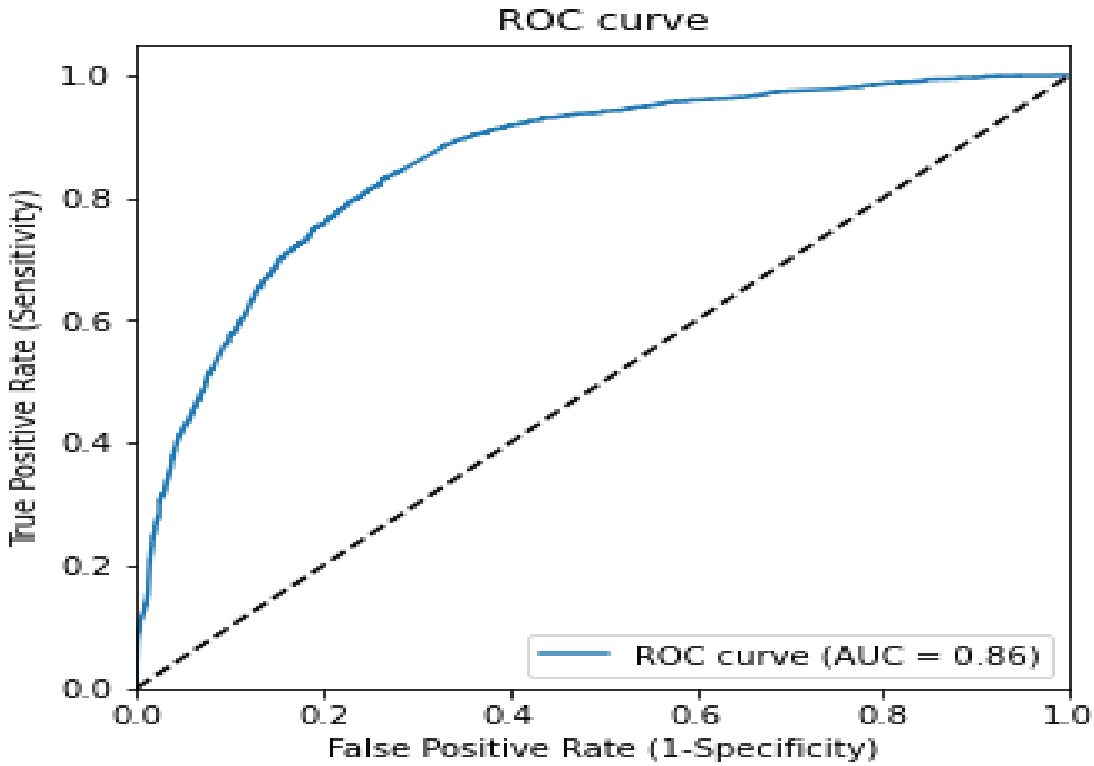
People with Tags 'Will revert after reading the mail' are most probable to convert. Sales team should give good amount of attention and service to these people. At the same time people not picking up phone, represented by 'Ringing' are least probable to convert. They should be given limited span of attention.

LEAD SOURCE



Most number of converted leads are from Google, followed by Direct Traffic. Thus it will make sense for the company to invest in Google advertising.

ROC CURVE


ROC curve
True Positive Rate (Sensitivity)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (1-Specificity)
ROC curve (AUC = 0.86)

CONCLUSION

When the sales team target leads with a score of 70 and above. We expect the team to close almost 80% of such leads. But by this strategy, the team will only be able to target 50% of the hot-leads available in the funnel.

We recommend an alternative strategy to target people with a score above 30. This way, the sales team will be able to close almost 65% of cases and also target 90% of the hot leads present in the funnel. Both these strategies are useful in different scenarios.

TRAIN DATA SET (Cutoff – 0.07)

Precision : 83%
Sensitivity : 48%
Specificity : 93%

TEST DATA SET (Cutoff – 0.07)

Precision : 85%
Sensitivity : 50%
Specificity : 93%