Lead Scoring Case Study

By
T Sai Kamal Chand
Sampath
Chandan Singh

Process Summary

- **Data Cleaning and Preparation**
- **EDA**
- **Dummy Variables**
- **Dataset split into Train and Test**
- **Feature Scaling**
- **Model Building**
- **Model Evaluation and Prediction on Test dataset**
- **Precision – Recall**
- **Conclusion and Findings**

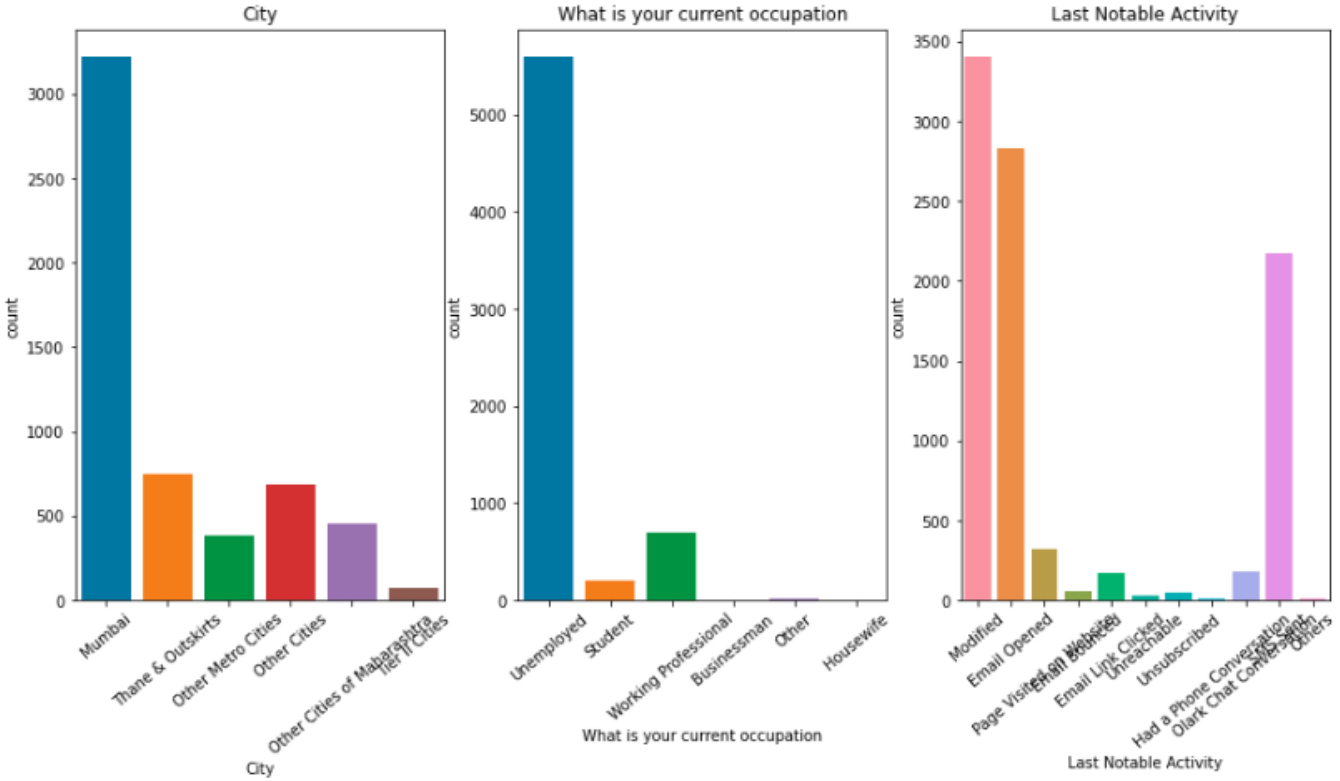
Data Cleaning and Preparation

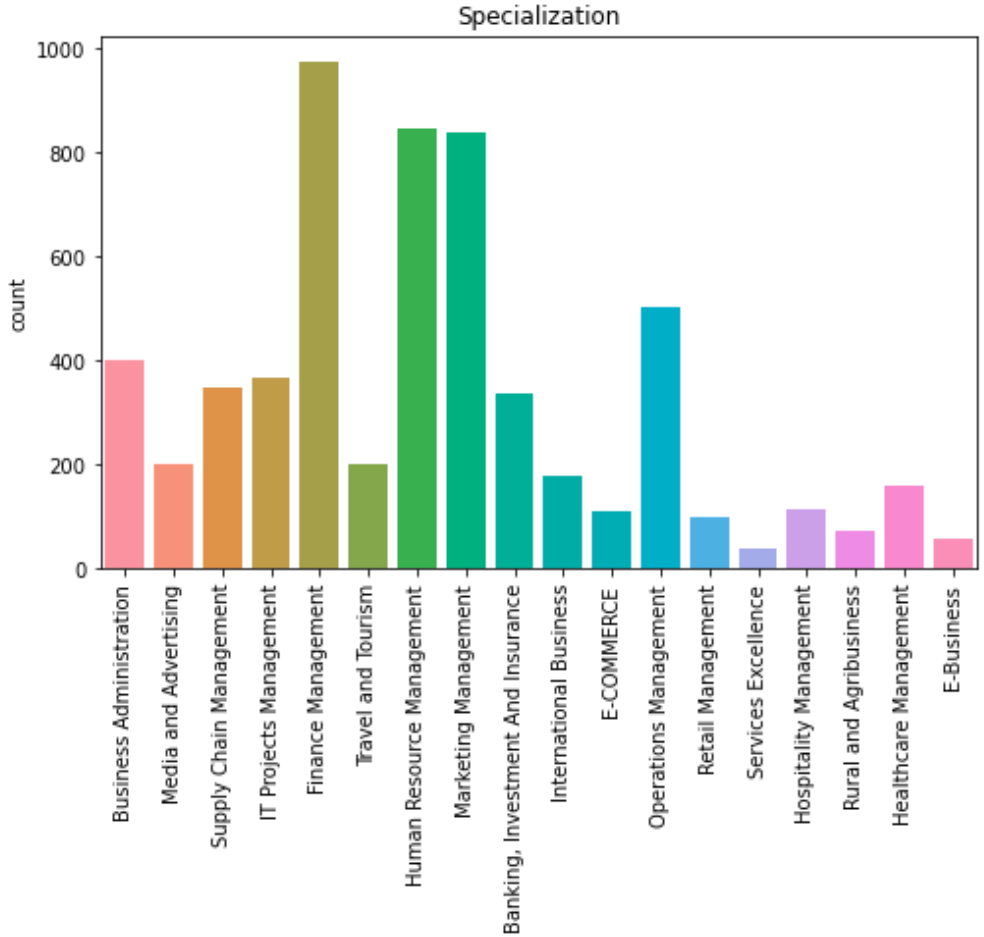
- There are 9240 rows and 37 columns.
- Dropped the columns with NULL values that are more than 40% and some redundant variables with unique values of 1 and other variables where only one value is predominantly seen.
- Changed the values with “Select” to NAs.

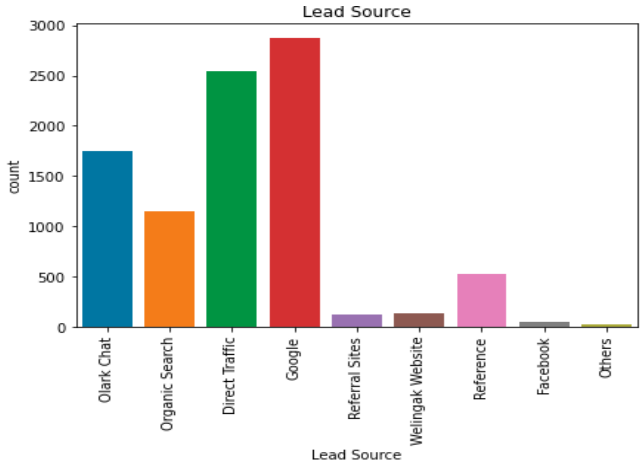
Data Visualization - EDA

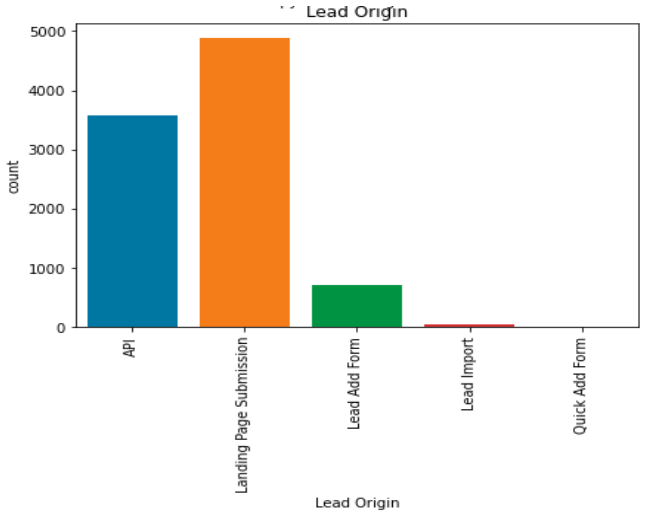
- Categorical and Numerical Variables
  - Univariate Analysis
  - Bivariate Analysis

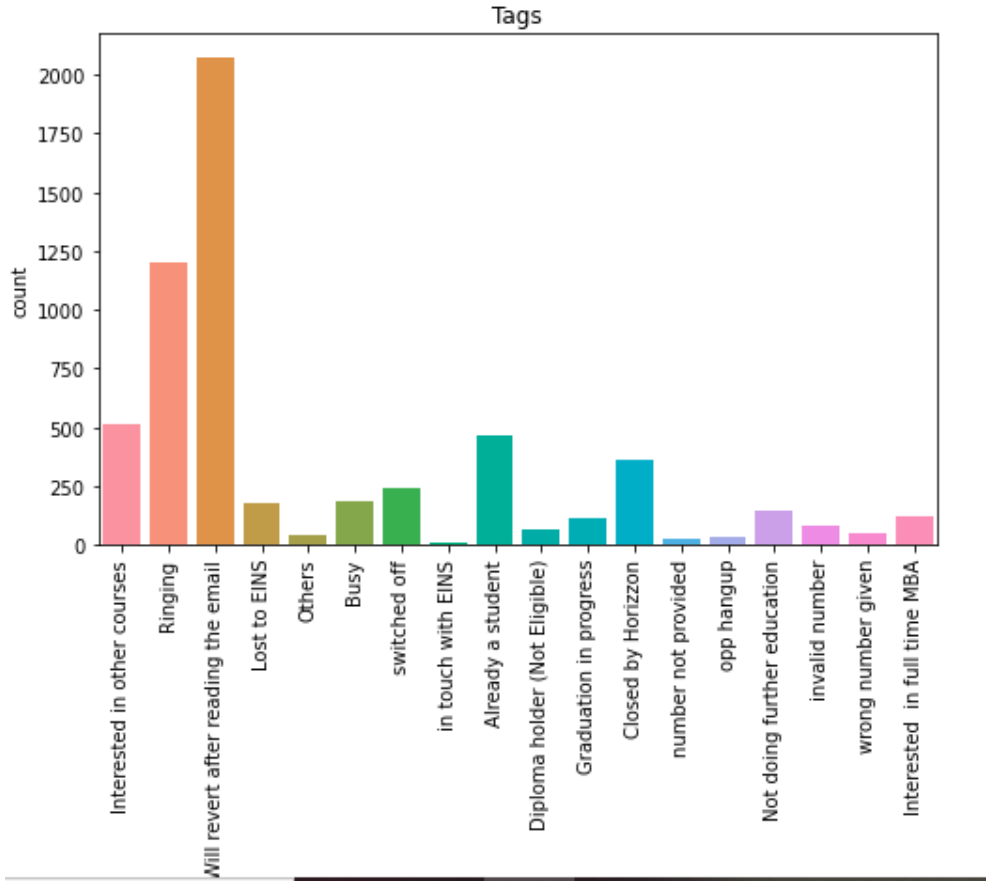
Univariate Analysis

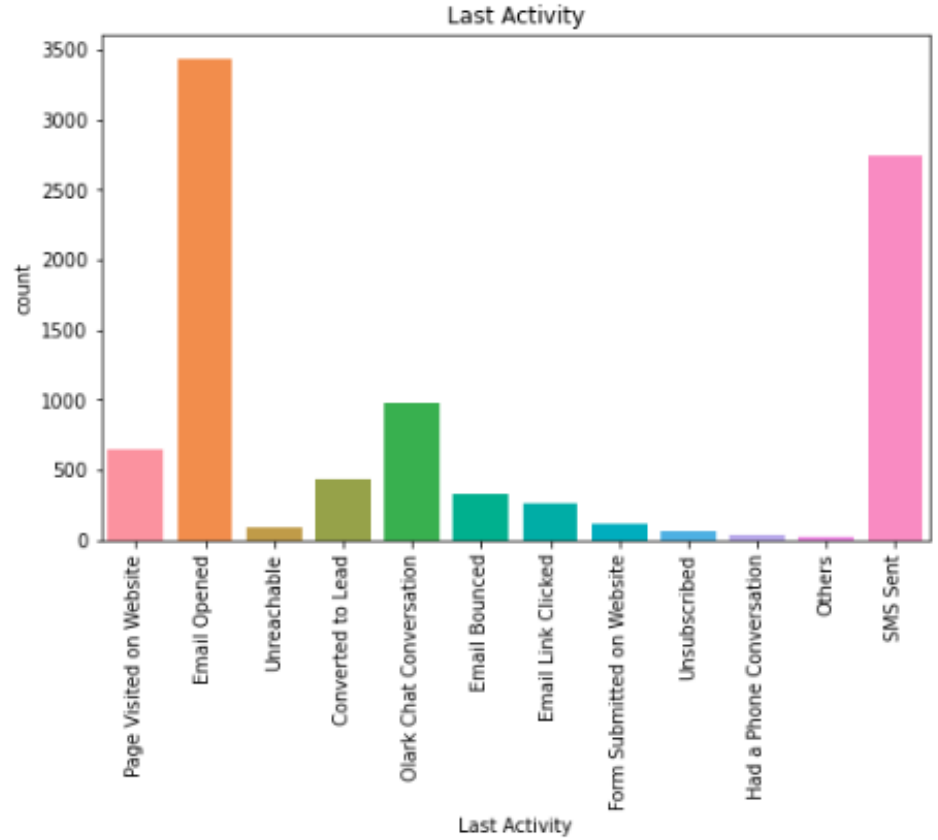


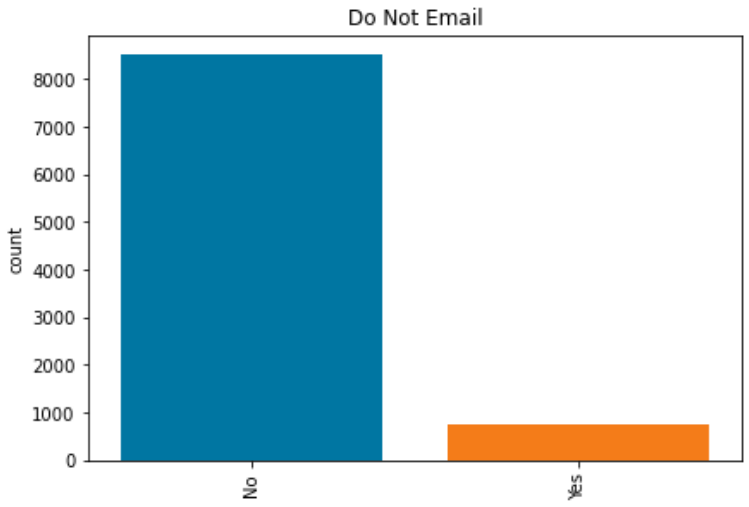




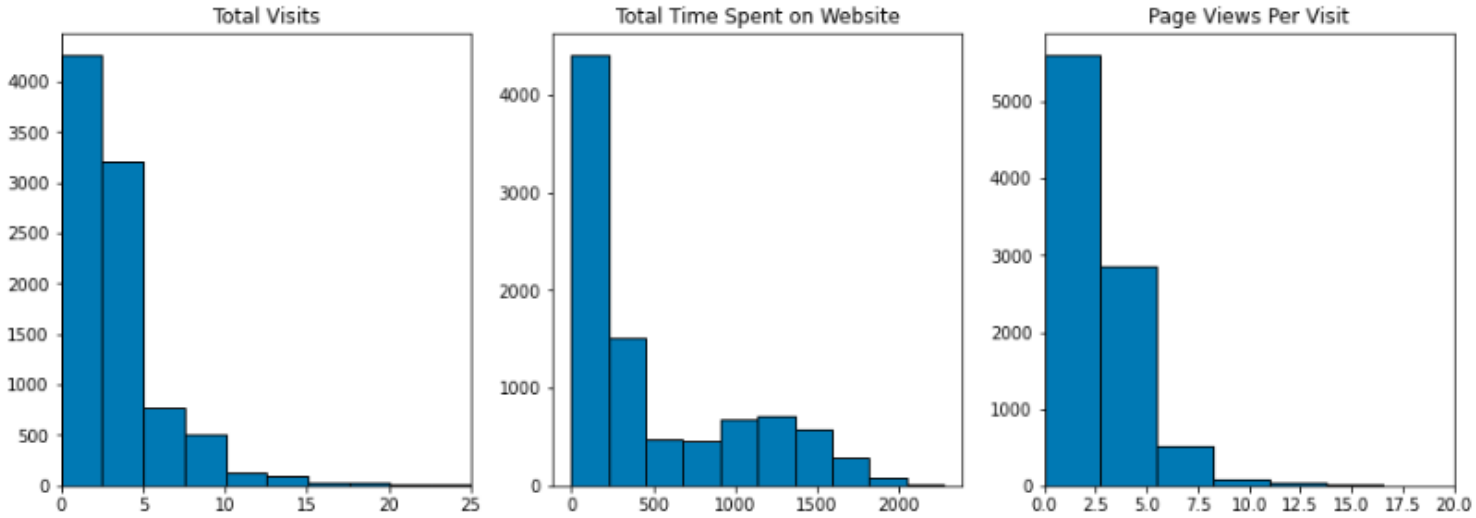



Tags
count
2000
1750
1500
1250
1000
750
500
250
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
Others
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA

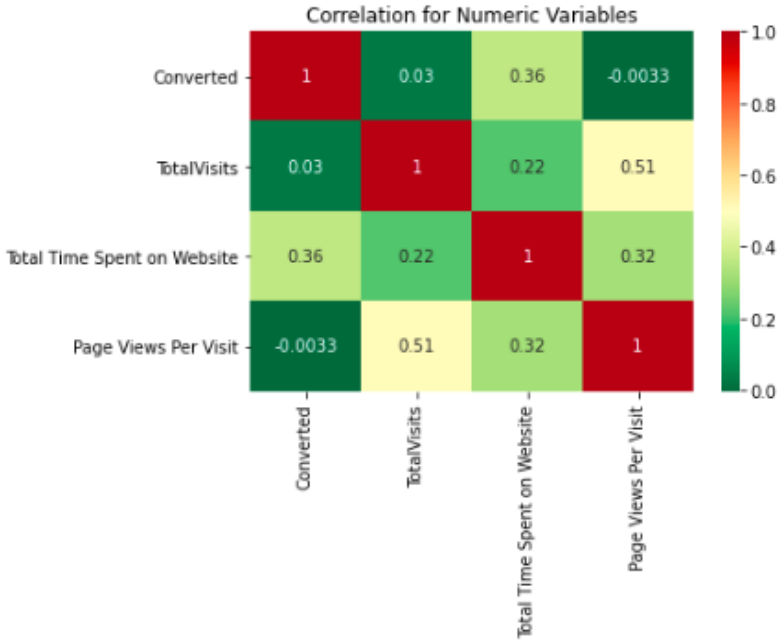

Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
Others
SMS Sent
Last Activity

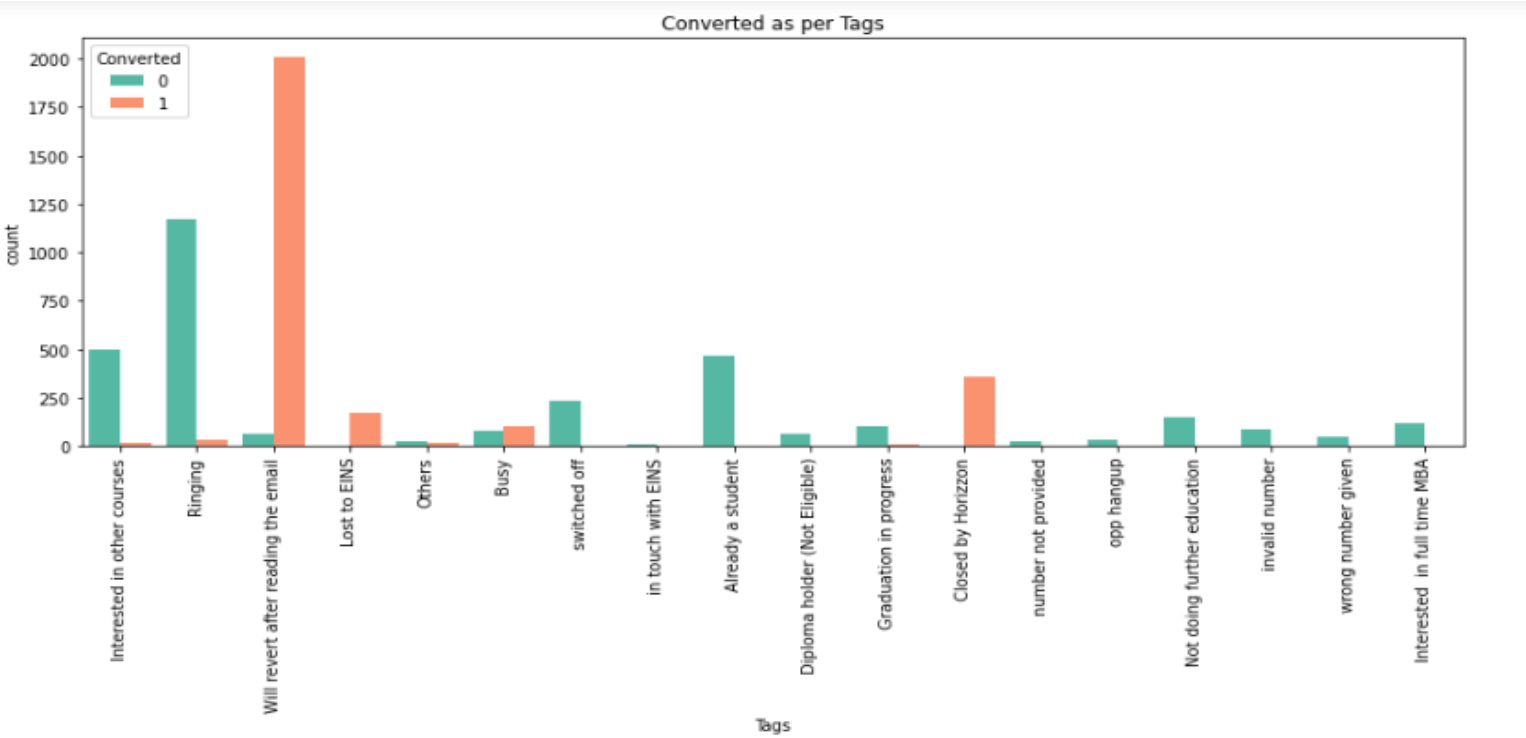

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes

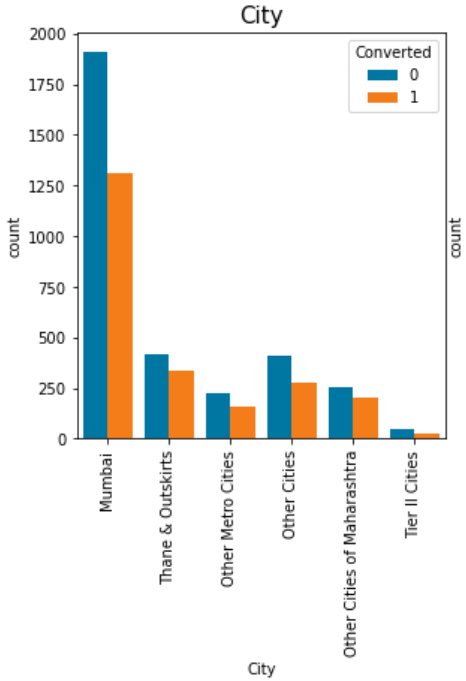

A free copy of Mastering The Interview
count
6000
5000
4000
3000
2000
1000
0
No
Yes

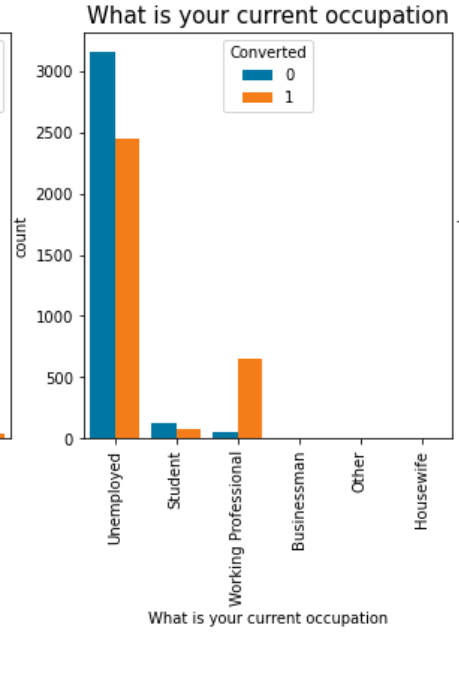


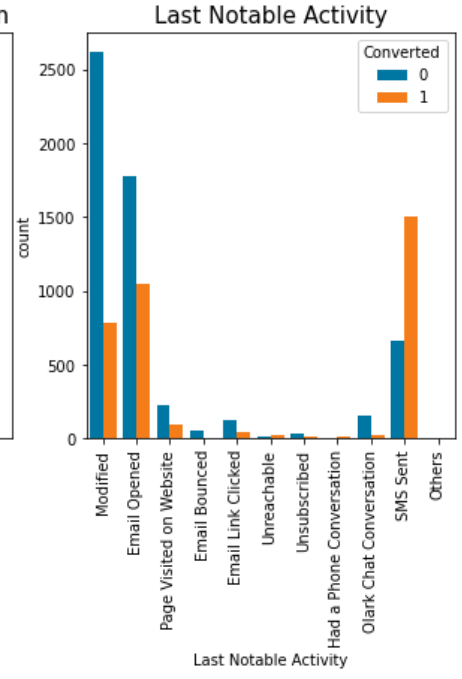
Bivariate Analysis and Correlation Matrix

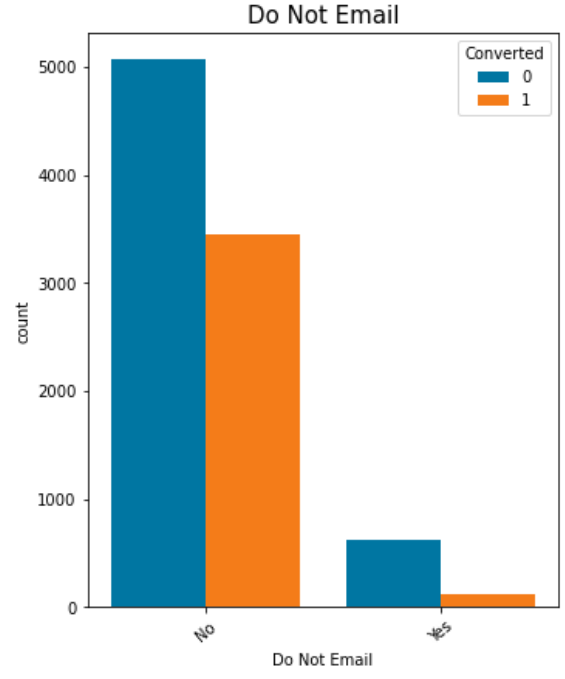


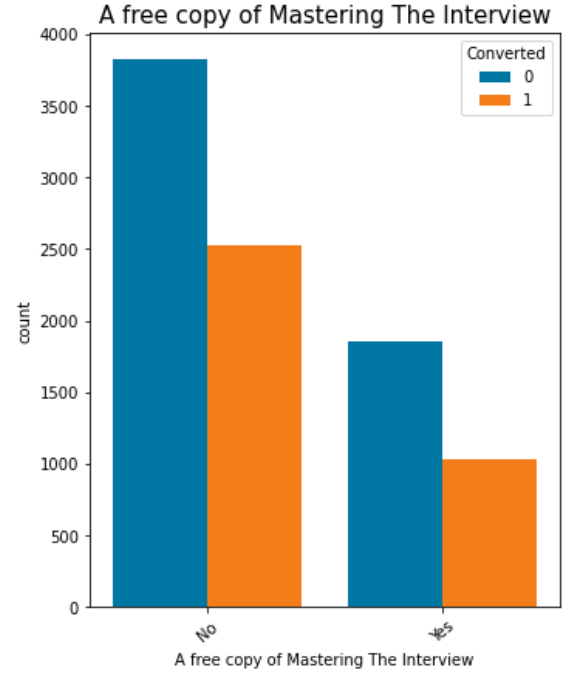


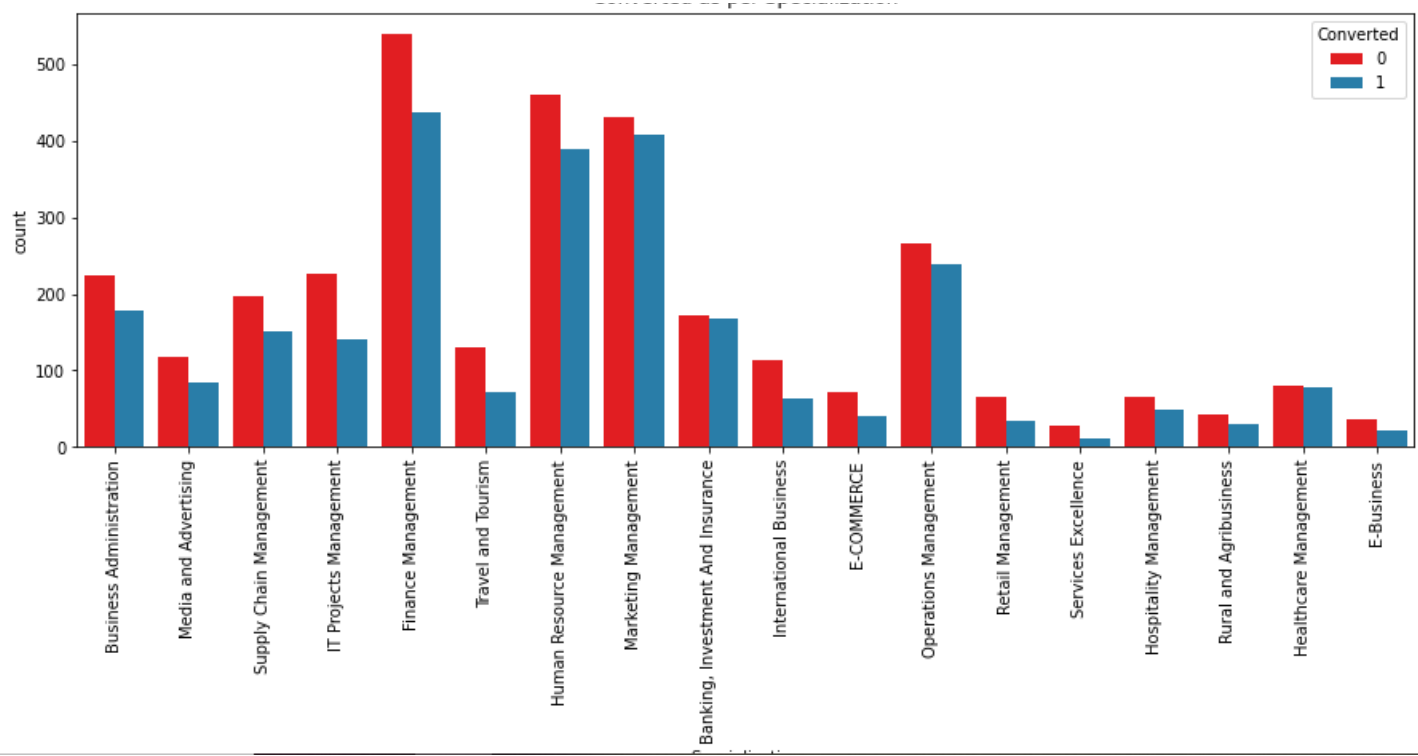

City
Converted
0
1
count
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

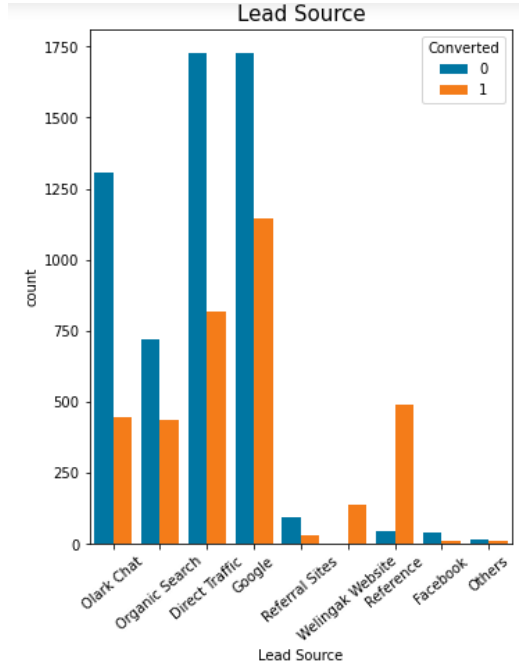

What is your current occupation
Converted
0
1
count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

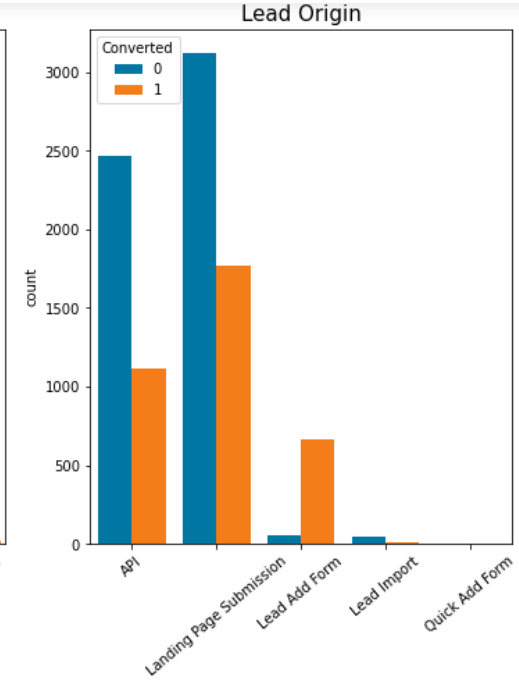

Last Notable Activity
Converted
0
1
count
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Others
Last Notable Activity


Do Not Email
Converted
0
1
count
No
Yes
Do Not Email


A free copy of Mastering The Interview
Converted
0
1
count
No
Yes
A free copy of Mastering The Interview


Converted
0
1
count
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business


Lead Source
Converted
0
1
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
Others
Lead Source

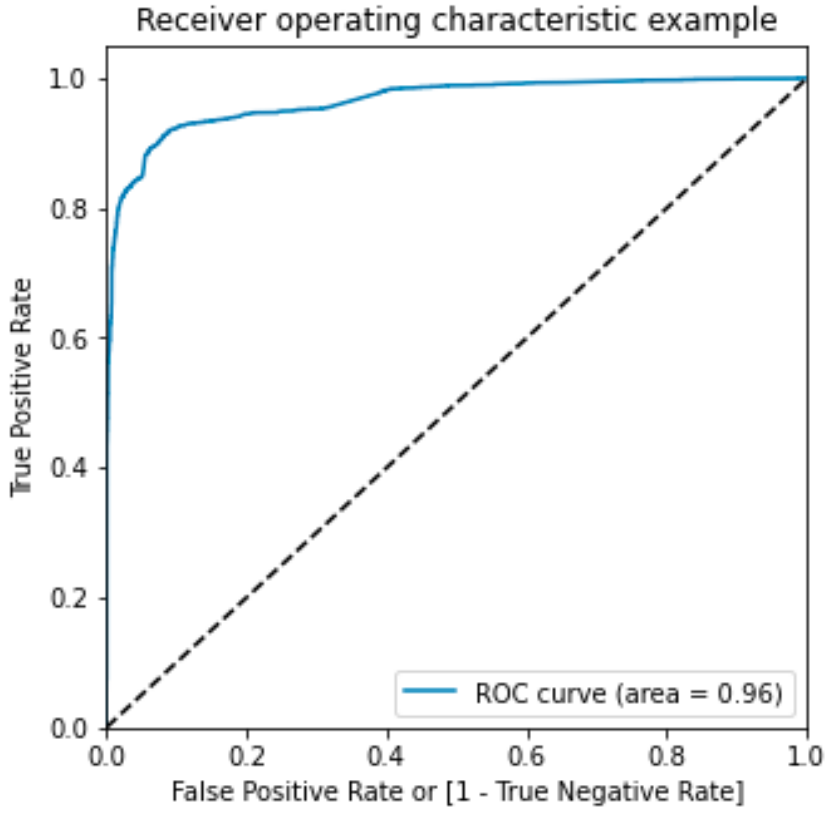

Lead Origin
Converted
0
1
count
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

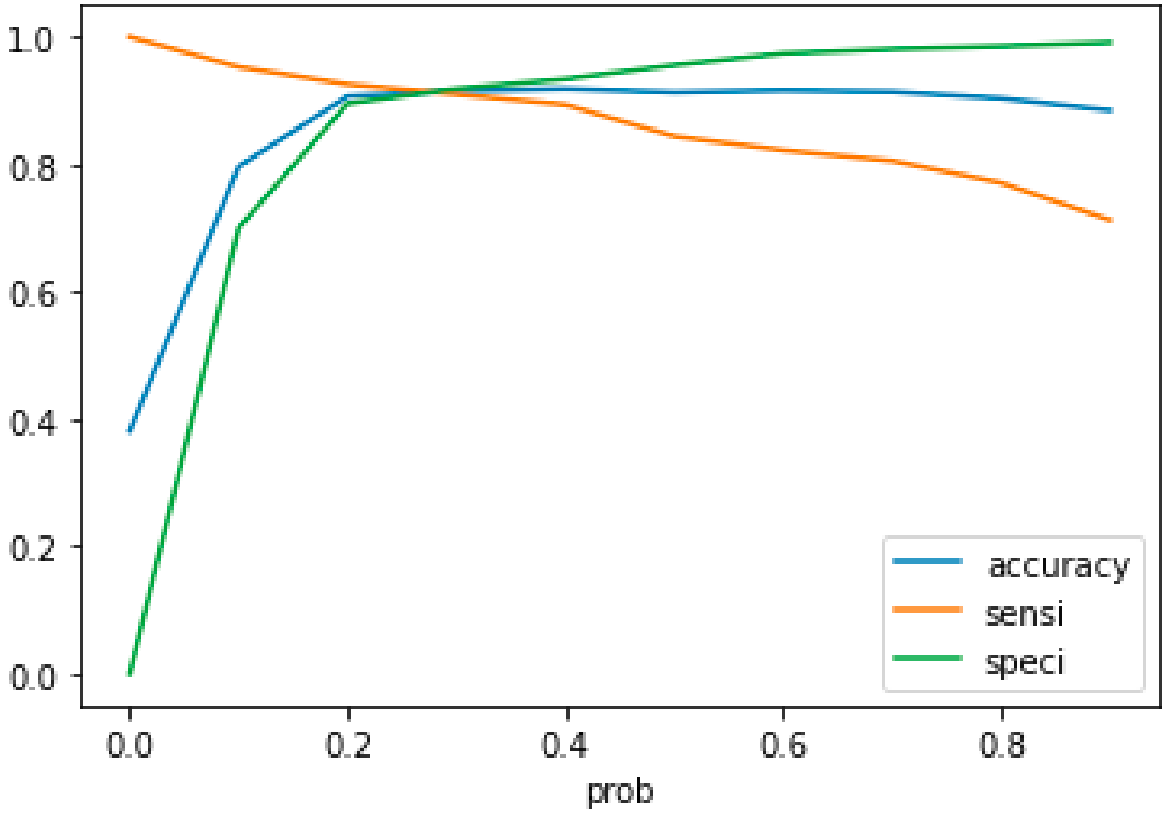
- **Dummy Variables**
- Created dummy variables for all categorical variables.
- **Dataset split into Train and Test**
- The dataset was split into 70% train and 30% test.
- Missing value treatment was done for X train dataset.
- **Feature Scaling**
- Using MinMaxScaler, all the numeric variables were re-scaled.
- **Model Building**
- Using Recursive Feature Elimination (RFE) for selecting 15 top important features.
- Built model with the variables which have VIF < 5 and p-value < 0.05.
- Checked the optimal probabilities and the cutoff points on the train set.

- **Model Evaluation and Prediction on Test dataset**
- Area under ROC Curve is 0.96.
- For the train dataset, with a 0.3 cutoff, Accuracy, Sensitivity, and Specificity 91%.
- Predicted the test dataset with Accuracy (91%), Sensitivity (92.5%), and Specificity (91%).

- **Precision – Recall**
- With cutoff 0.41,
  - Accuracy - 91.9%, sensitivity - 89.2% and specificity - 93.5% on the train dataset.
  - Accuracy - 91.6%, sensitivity - 90.4% and specificity - 92.4% on the test dataset.
  - And, precision is 88%, and recall is 90% on the test dataset.

ROC Curve and Optimal Probabilities





- **Conclusion and Findings**
- The lead score on the test dataset - 90.41%.

- Features with high probability of a lead getting converted are:
  - Tags_Closed by Horizzon
  - Tags_Lost to EINS
  - Lead Source_Welingak Website
  - Tags_Will revert after reading the email.
  - Last Activity_SMS Sent

- Negatively affected variables to avoid. Some of them are
  - Tags_Ringing
  - Tags_switched off
  - Tags_invalid number