Lead Scoring Case Study

Submitted by:

Avita Kumar

Atulya Pattamatta

Rajnish Kumar

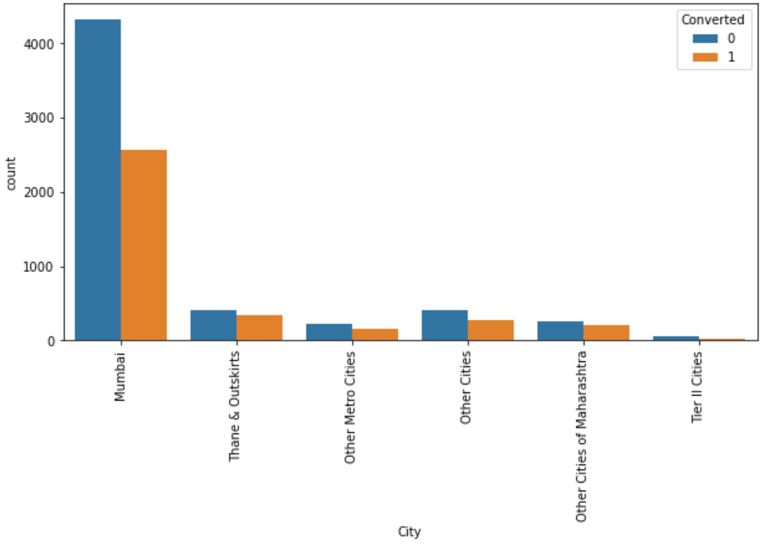
Problem Statement

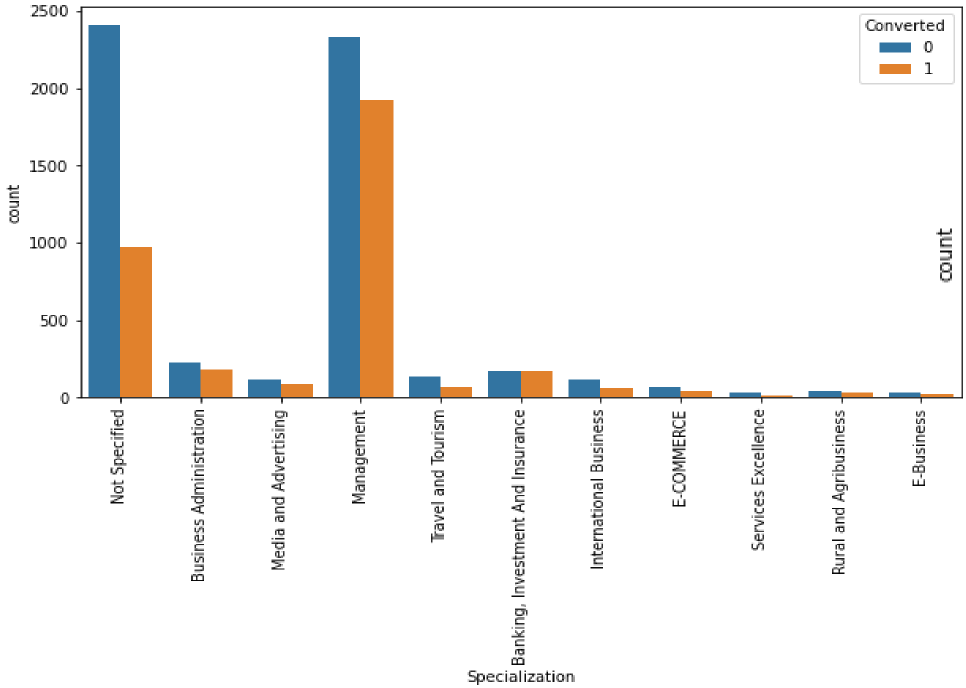
- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- They are trying to find ways to convert more leads (potential customers/ hot leads) to join their courses. The data gives information like how these leads reached the site, time spent on site and conversion rate etc.
- They want to build a model to assign Lead score to potential leads and based on this score increase their conversion rate

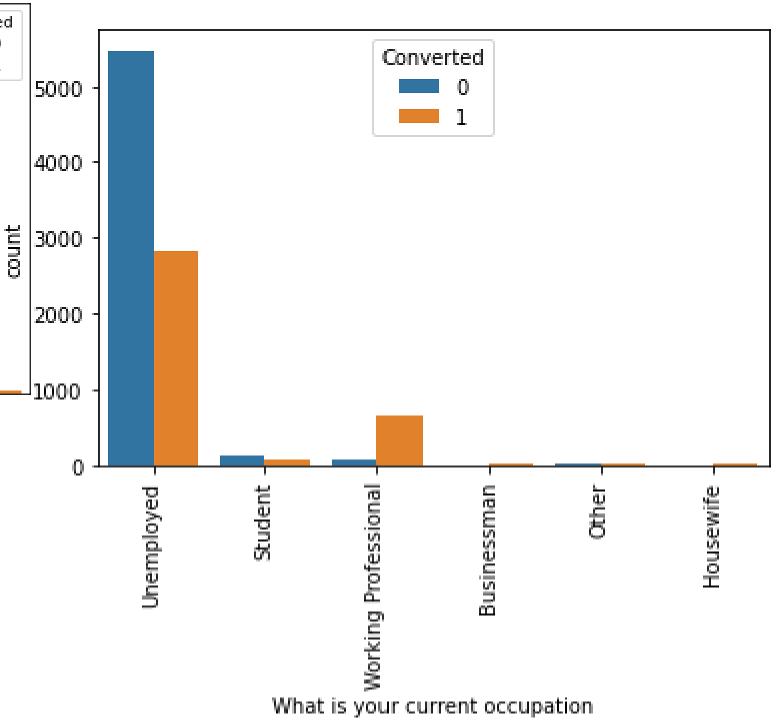
Strategy and steps:

- Clean and prepare Data
- Exploratory Data Analysis
- Split data into train and Test
- Feature scaling
- Build Logistic regression model and calculate Lead Score
- Evaluation of model with metrics Like accuracy, sensitivity, specificity, precision and recall
- Prediction on Test data

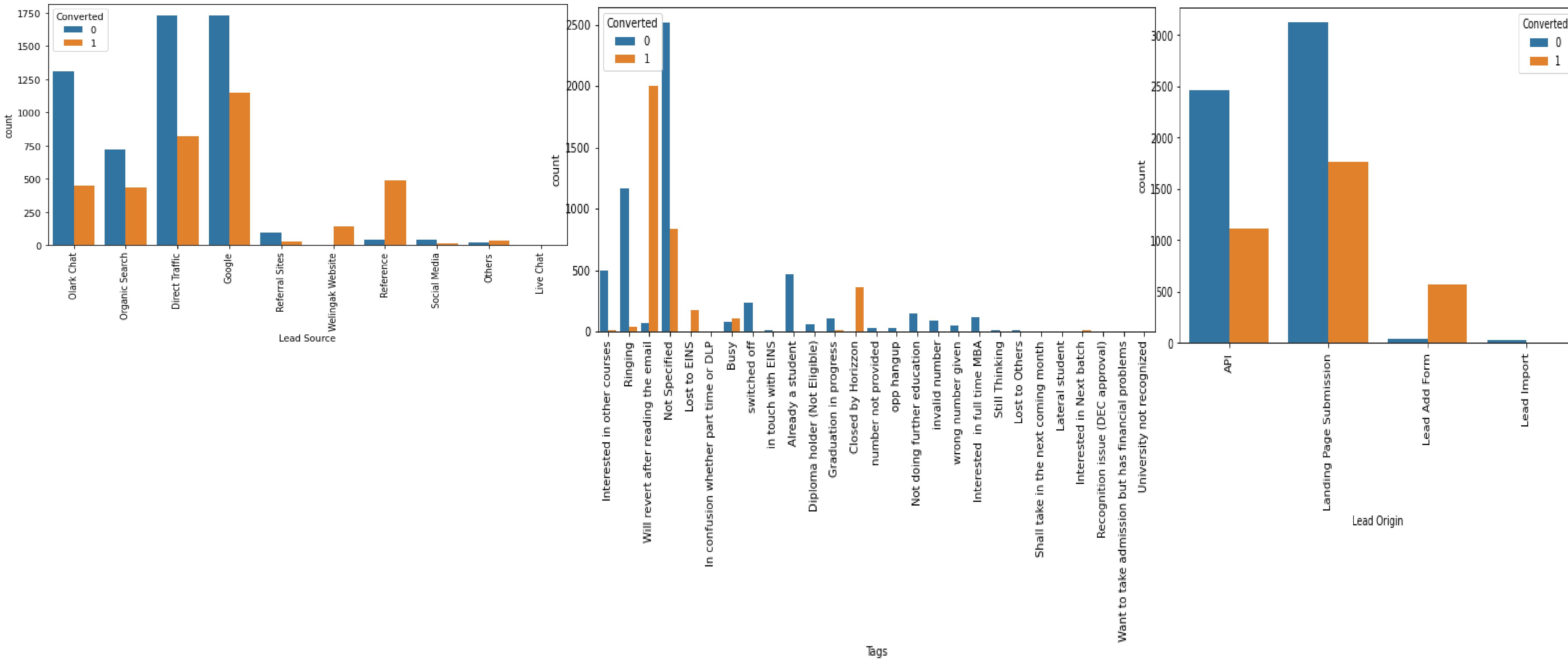
Conversion rates based on EDA of Categorical variables



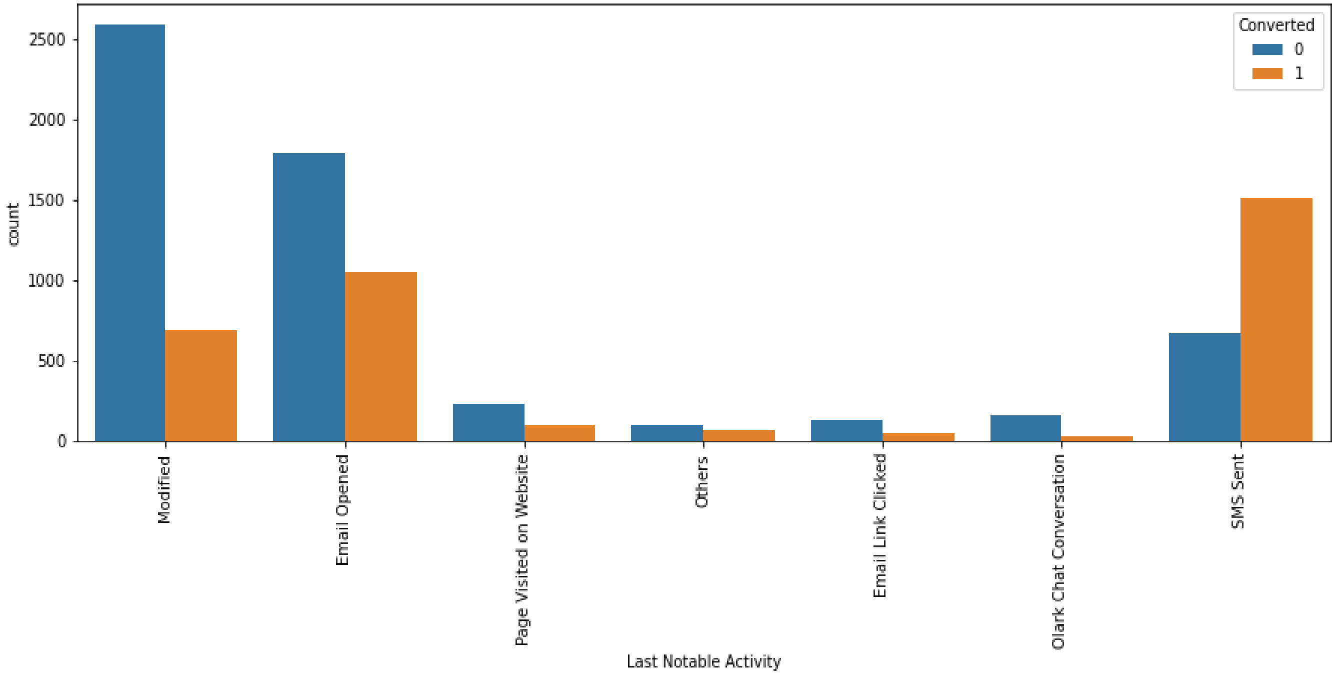




Conversion rate is high for Unemployed occupation, Management Specialization, Mumbai city.



High conversion if Lead Origin is Lead Add form, Tags is Lost to EINS, Lead Source is Wellingak website or reference



If Last Notable activity is SMS Sent conversion is higher

Variables affecting Conversion based on Model

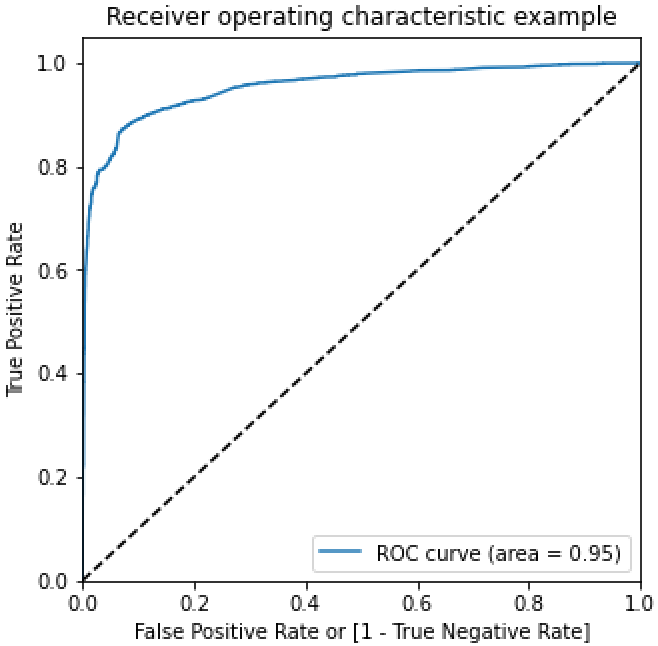
Important variables:

i. Lead Origin_Landing Page Submission,
ii. Tags_Will revert after reading the email
iii. Last Activity_SMS Sent
iv. Lead Origin_Lead Add Form
v. Tags_Not Specified
vi. Last Notable Activity_Modified
vii. Lead Source_Welingak Website
viii. Tags_Ringing
ix. Total Time Spent on Website
x. Tags_Busy
xi. Tags_Lost to EINS
xii. Specialization_Travel and Tourism
xiii. Last Notable Activity_Olark Chat Conversation

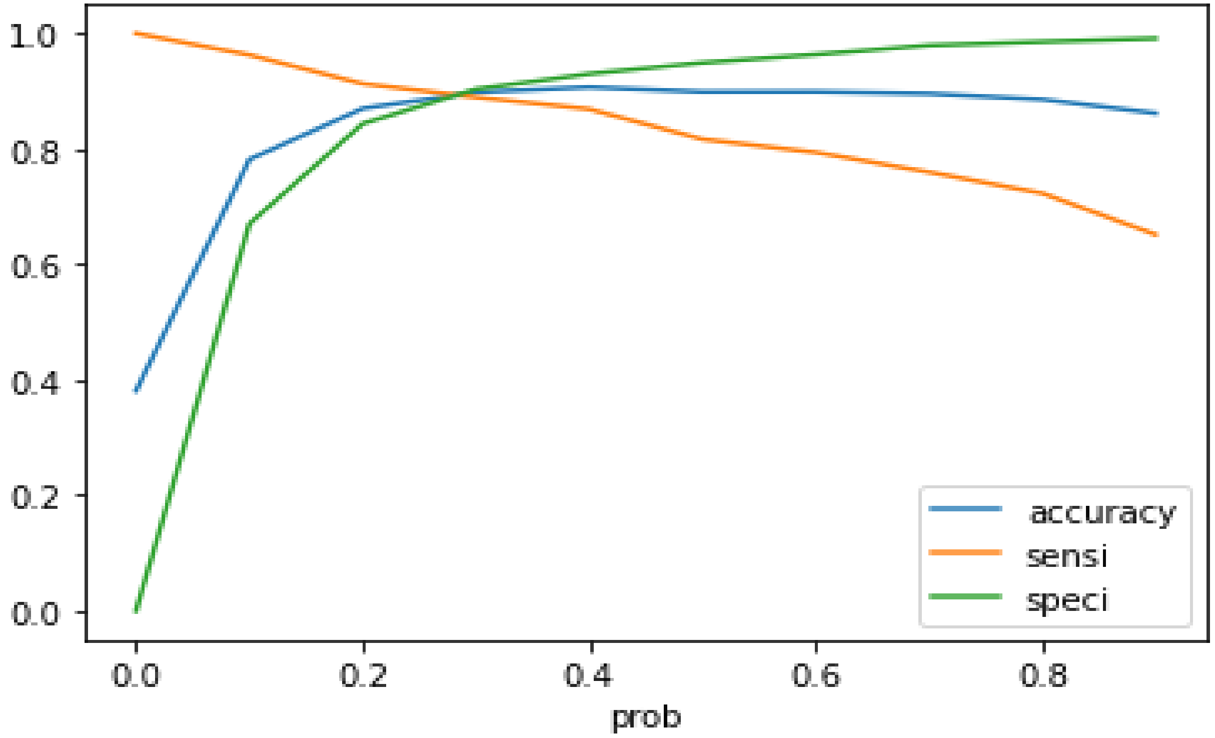
Other Steps

Dummy variables: Dummy variables were created
Train test Split: Data was split in 70% and 30% for train and tests respectively.
Scaling: Standard scaler was used for numeric variables.
Model building: Top 15 variables were selected using RFE. Other variables were removed manually looking at their p-value and VIF.

ROC Curve



Optimal cutoff of probability at 0.3 for Lead Scores



Results:

Train Data: Accuracy : 89.80%

Sensitivity : 88.94%

Specificity : 90.33%

Test Data: Accuracy : 88.94%

Sensitivity : 87.89%

Specificity : 89.57%

Recall rate of 81.65% on test data and train data. So model seems to be performing reasonably ok.