Leading Score Case Study

Presented by:
Praveen Thota
Pravin Shinde

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%.

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 higher lead score have a higher conversion chance and the customers with 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

There are quite a few goals for this case study.

- 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.

- 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

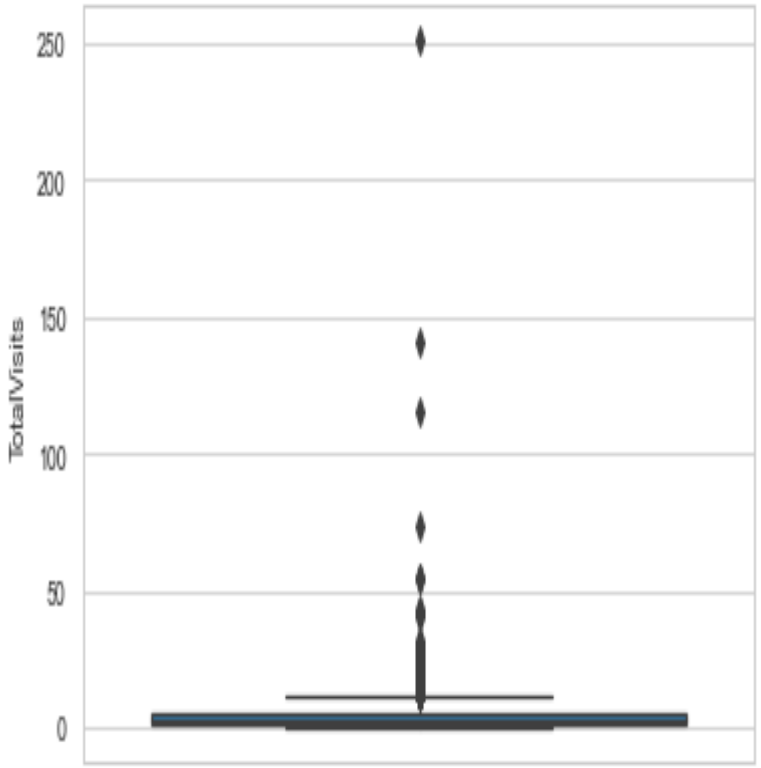
- Loading Libraries and Inspecting Data Frame
- Reading & Understanding the Data
- Data Cleaning and Data Preparation
- EDA
- Feature Scaling
- Train Test split data
- Prepare the data for modelling
- Build model
- Model Evaluation
- Accuracy, specificity, sensitivity
- Precision, Recall and F1 score
- Making predictions on the test set
- Conclusion

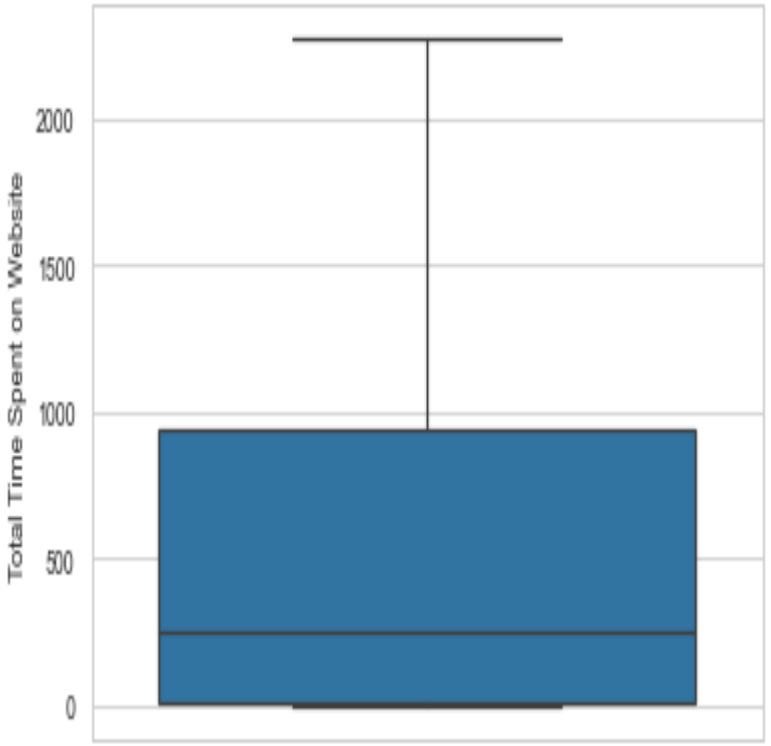
Data Sourcing, Cleaning & Preparation

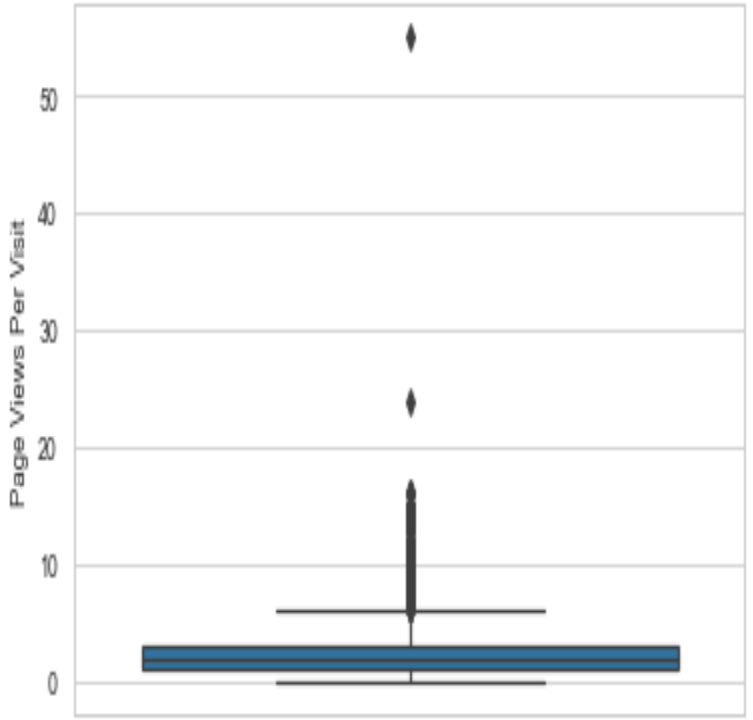
- Importing libraries
- Read the data from the csv file
- Inspecting the data frame
- Handling null values & Removing cols with high null values data
- Assigning an Unique Category to NULL/SELECT values (Unknown values)
- Dropping the columns which have only unique values (Magazine etc.,)
- Deleting the columns 'Asymmetrique Activity Score' & 'Asymmetrique Profile Score' as they will be represented by their corresponding index columns as per correlation
- Imputing null values with mode() and median()
- Removing rows with minor null values in a particular column (Lead Source)
- Outlier treatment
- Removing redundant columns ‘Prospect ID’ and ‘Lead Number’
- Removing outlier values based on the Interquartile distance for some of the continuous variables (TotalVisits, Page Views Per Visit)

Exploratory Data Analysis

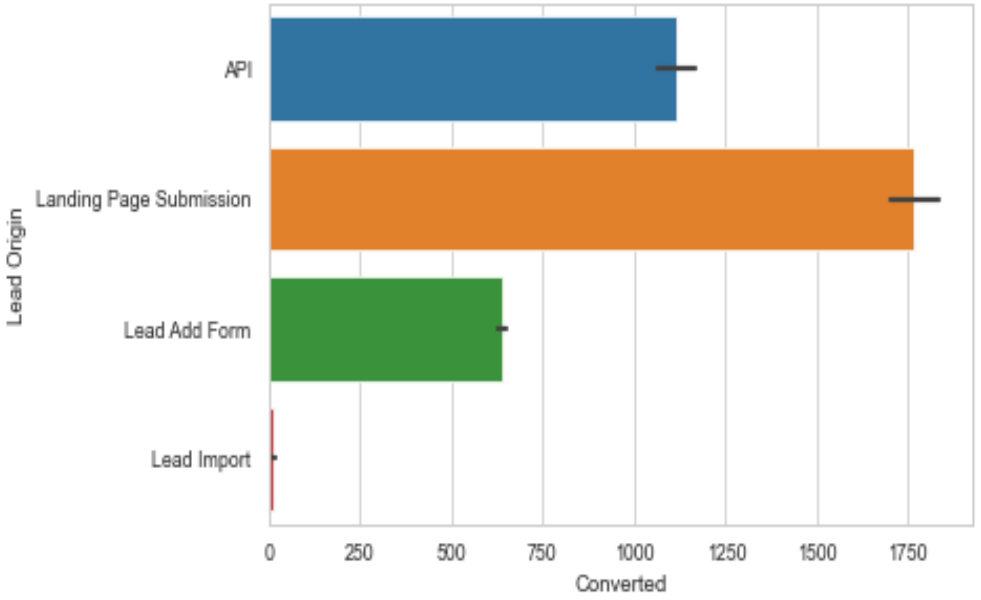
- Box Plot (Checking Outliers):

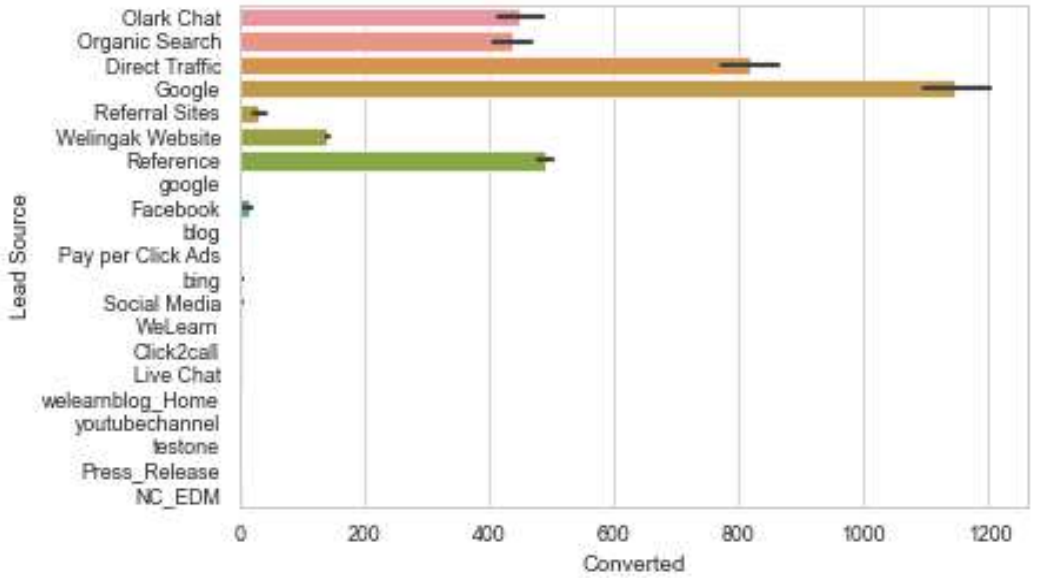


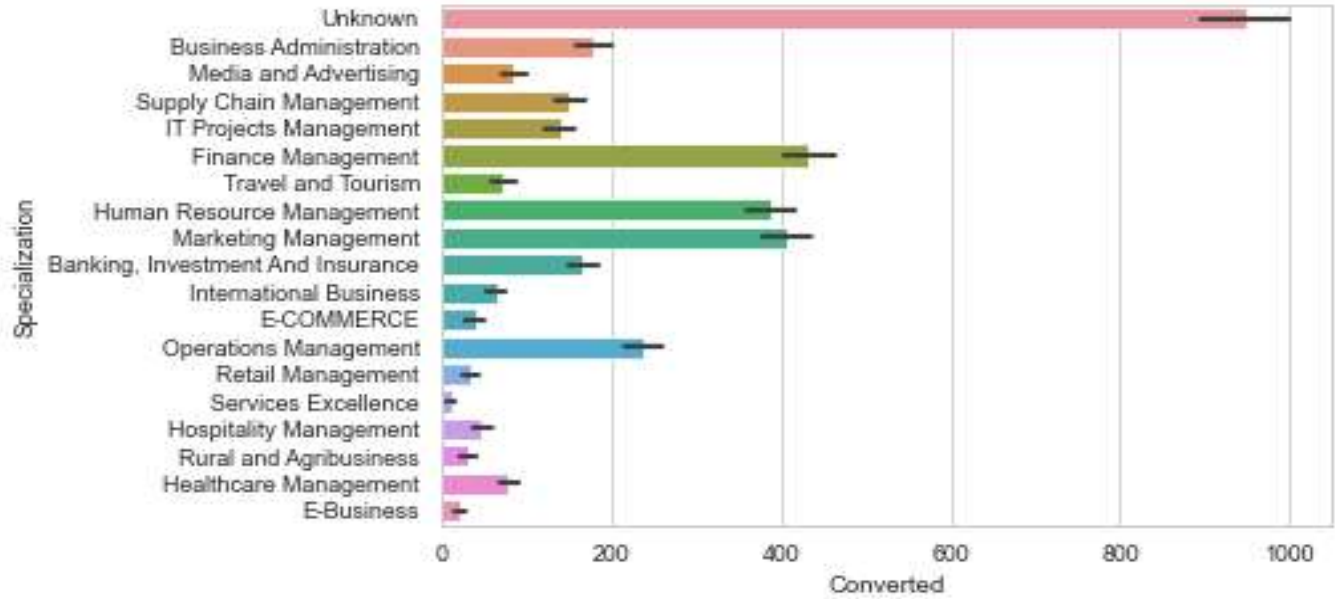


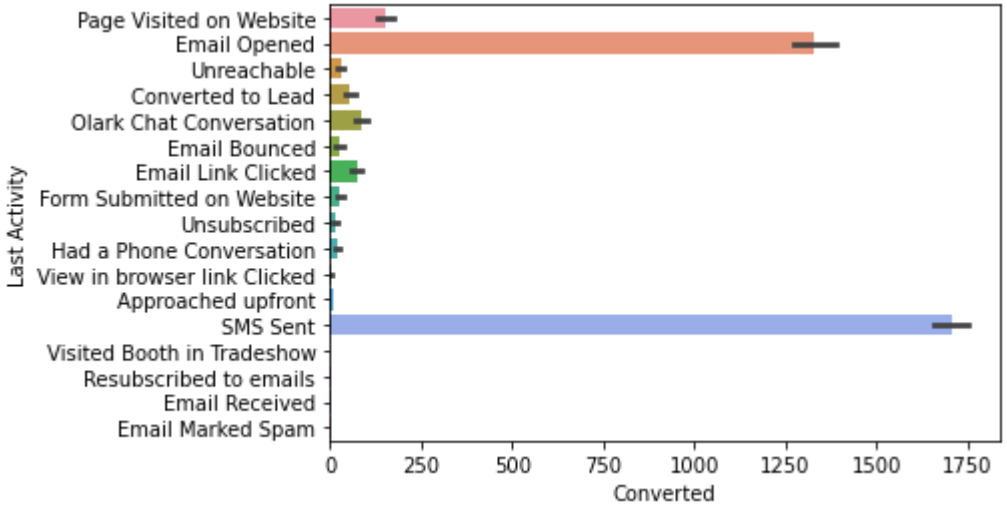


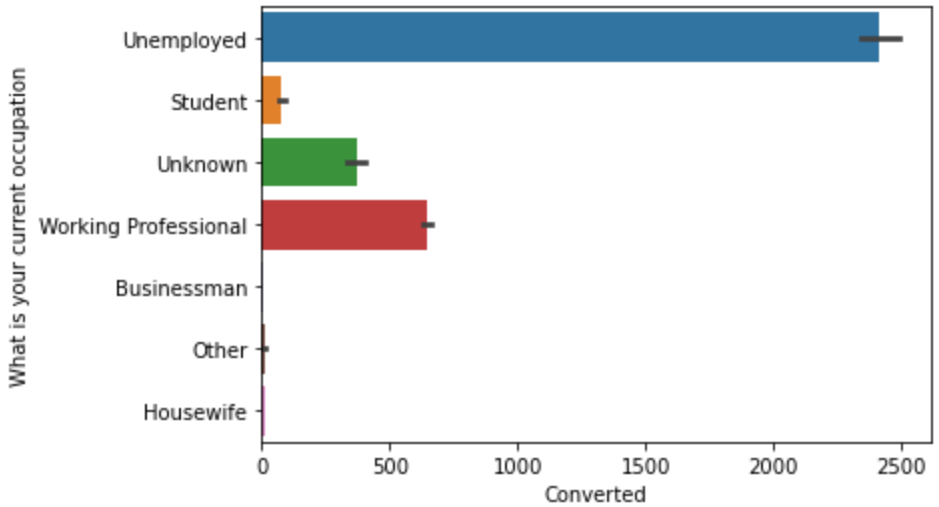
Bar Plots

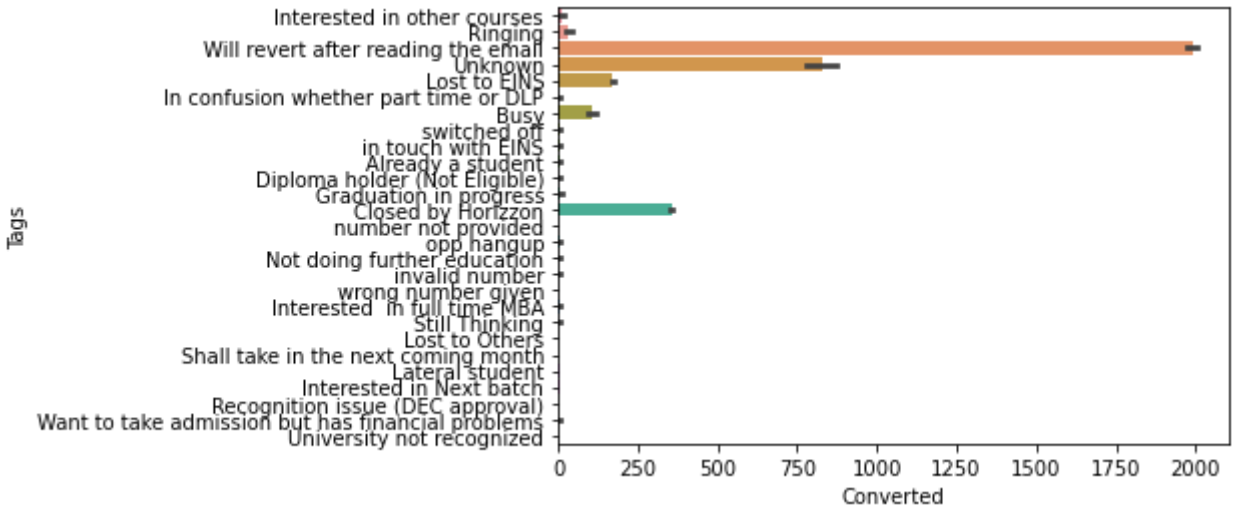

API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin
0
250
500
750
1000
1250
1500
1750
Converted

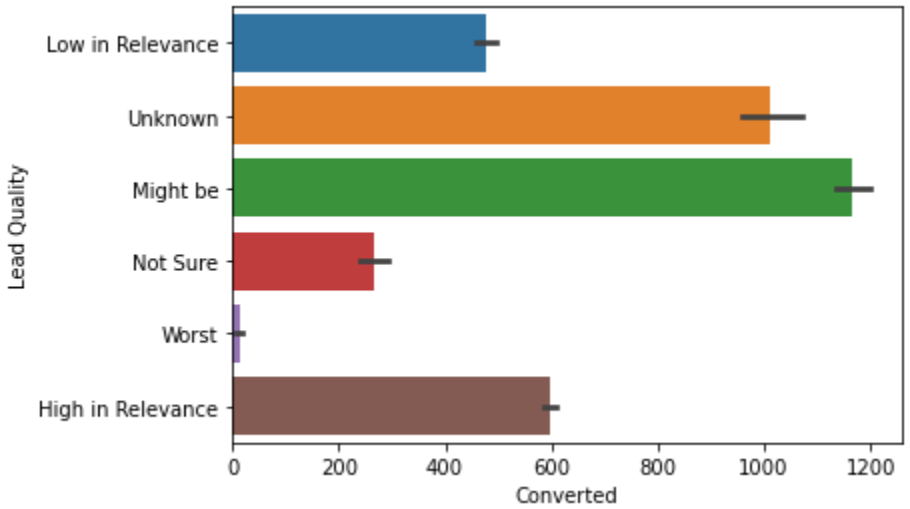

Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
0
200
400
600
800
1000
1200
Converted

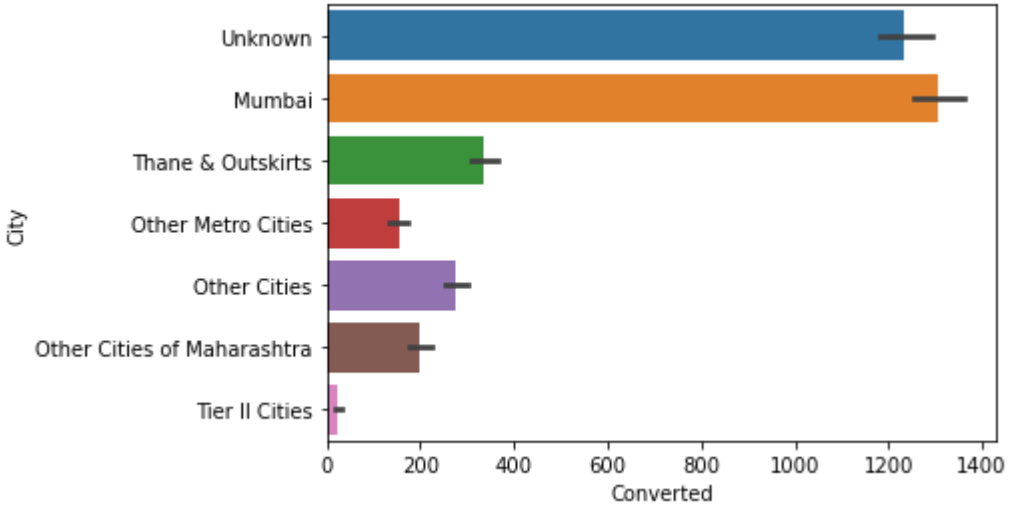

Unknown
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization
0
200
400
600
800
1000
Converted

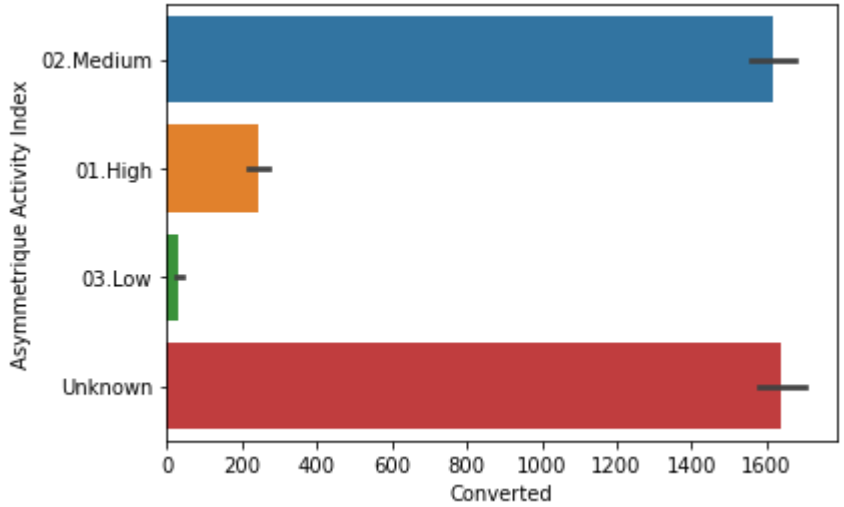

Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity
0
250
500
750
1000
1250
1500
1750
Converted

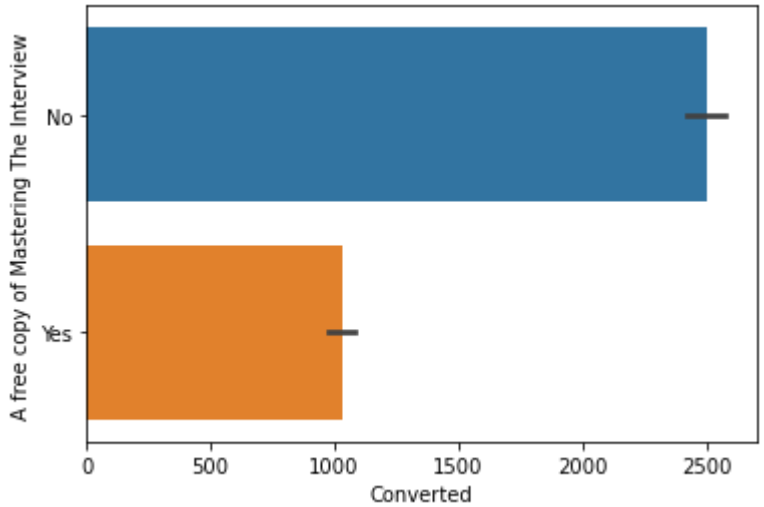

Unemployed
Student
Unknown
Working Professional
Businessman
Other
Housewife
What is your current occupation
0
500
1000
1500
2000
2500
Converted

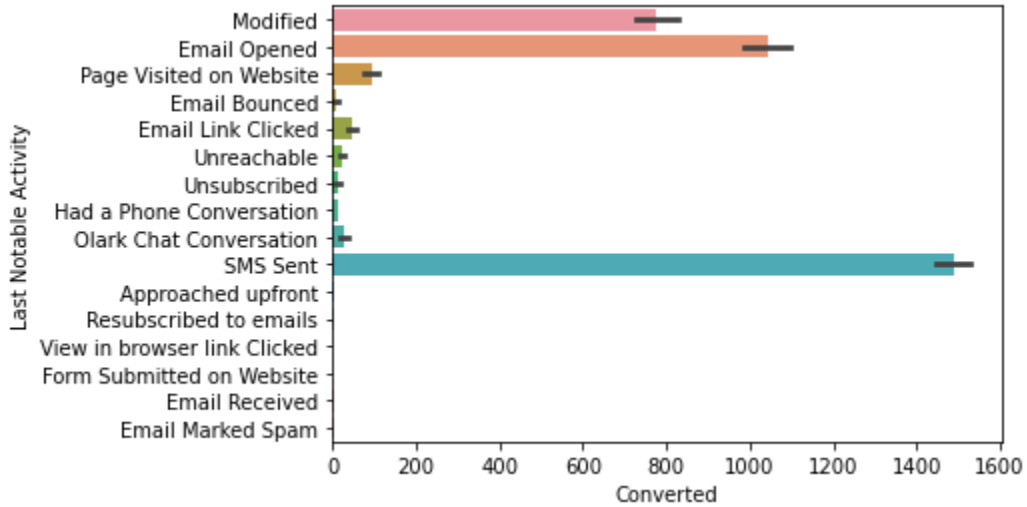

Interested in other courses
Ringing
Will revert after reading the email
Unknown
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags
0
250
500
750
1000
1250
1500
1750
2000
Converted


Low in Relevance
Unknown
Might be
Not Sure
Worst
High in Relevance
Lead Quality
0
200
400
600
800
1000
1200
Converted

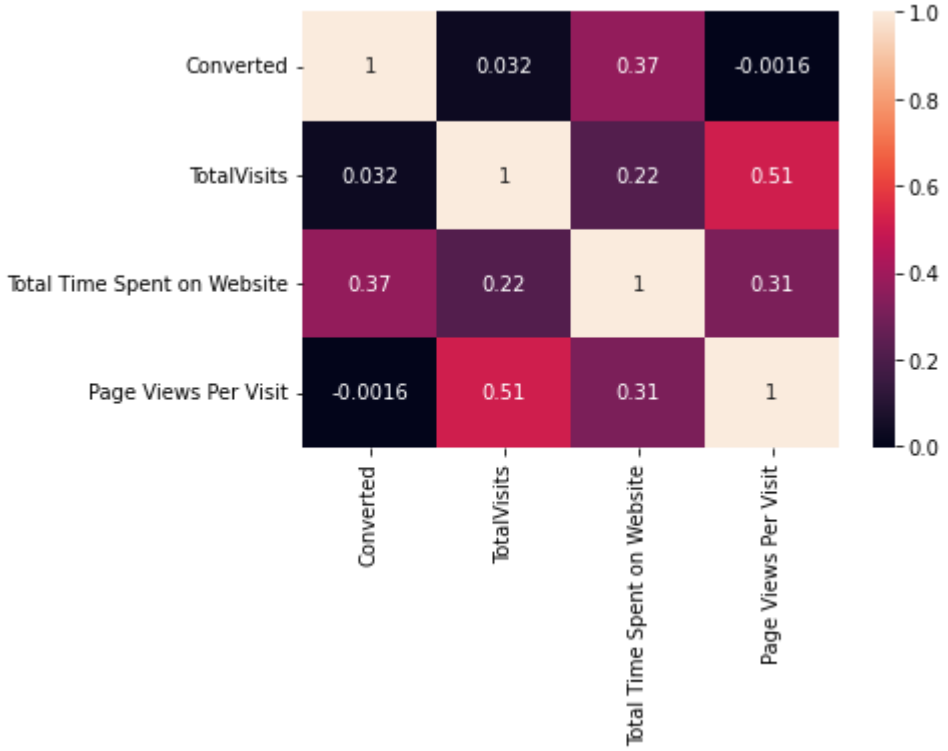

City
Unknown
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
0
200
400
600
800
1000
1200
1400
Converted

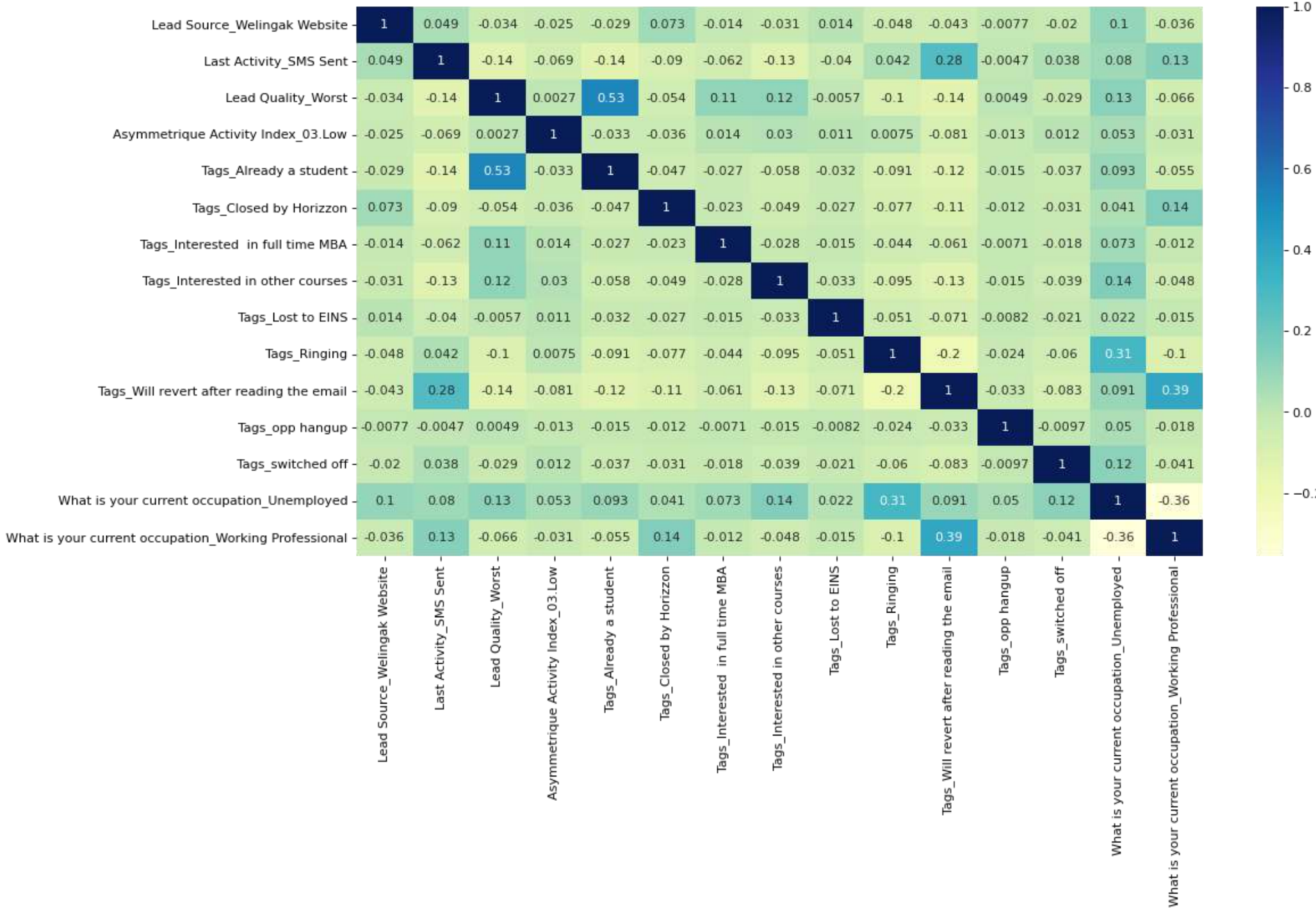

Asymmetrique Activity Index
02.Medium
01.High
03.Low
Unknown
0
200
400
600
800
1000
1200
1400
1600
Converted


A free copy of Mastering The Interview
No
Yes
0
500
1000
1500
2000
2500
Converted


Last Notable Activity
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
0
200
400
600
800
1000
1200
1400
1600
Converted

Correlation among numeric variables




Lead Source_Welingak Website 1 0.049 -0.034 -0.025 -0.029 0.073 -0.014 -0.031 0.014 -0.048 -0.043 -0.0077 -0.02 0.1 -0.036
Last Activity_SMS Sent 0.049 1 -0.14 -0.069 -0.14 -0.09 -0.062 -0.13 -0.04 0.042 0.28 -0.0047 0.038 0.08 0.13
Lead Quality_Worst -0.034 -0.14 1 0.0027 0.53 -0.054 0.11 0.12 -0.0057 -0.1 -0.14 0.0049 -0.029 0.13 -0.066
Asymmetrique Activity Index_03.Low -0.025 -0.069 0.0027 1 -0.033 -0.036 0.014 0.03 0.011 0.0075 -0.081 -0.013 0.012 0.053 -0.031
Tags_Already a student -0.029 -0.14 0.53 -0.033 1 -0.047 -0.027 -0.058 -0.032 -0.091 -0.12 -0.015 -0.037 0.093 -0.055
Tags_Closed by Horizzon 0.073 -0.09 -0.054 -0.036 -0.047 1 -0.023 -0.049 -0.027 -0.077 -0.11 -0.012 -0.031 0.041 0.14
Tags_Interested in full time MBA -0.014 -0.062 0.11 0.014 -0.027 -0.023 1 -0.028 -0.015 -0.044 -0.061 -0.0071 -0.018 0.073 -0.012
Tags_Interested in other courses -0.031 -0.13 0.12 0.03 -0.058 -0.049 -0.028 1 -0.033 -0.095 -0.13 -0.015 -0.039 0.14 -0.048
Tags_Lost to EINS 0.014 -0.04 -0.0057 0.011 -0.032 -0.027 -0.015 -0.033 1 -0.051 -0.071 -0.0082 -0.021 0.022 -0.015
Tags_Ringing -0.048 0.042 -0.1 0.0075 -0.091 -0.077 -0.044 -0.095 -0.051 1 -0.2 -0.024 -0.06 0.31 -0.1
Tags_Will revert after reading the email -0.043 0.28 -0.14 -0.081 -0.12 -0.11 -0.061 -0.13 -0.071 -0.2 1 -0.033 -0.083 0.091 0.39
Tags_opp hangup -0.0077 -0.0047 0.0049 -0.013 -0.015 -0.012 -0.0071 -0.015 -0.0082 -0.024 -0.033 1 -0.0097 0.05 -0.018
Tags_switched off -0.02 0.038 -0.029 0.012 -0.037 -0.031 -0.018 -0.039 -0.021 -0.06 -0.083 -0.0097 1 0.12 -0.041
What is your current occupation_Unemployed 0.1 0.08 0.13 0.053 0.093 0.041 0.073 0.14 0.022 0.31 0.091 0.05 0.12 1 -0.36
What is your current occupation_Working Professional -0.036 0.13 -0.066 -0.031 -0.055 0.14 -0.012 -0.048 -0.015 -0.1 0.39 -0.018 -0.041 -0.36 1
Lead Source_Welingak Website
Last Activity_SMS Sent
Lead Quality_Worst
Asymmetrique Activity Index_03.Low
Tags_Already a student
Tags_Closed by Horizzon
Tags_Interested in full time MBA
Tags_Interested in other courses
Tags_Lost to EINS
Tags_Ringing
Tags_Will revert after reading the email
Tags_opp hangup
Tags_switched off
What is your current occupation_Unemployed
What is your current occupation_Working Professional
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Data Preparation for Logistic Regression

- Converting some binary variables (Yes/No) to 0/1
- Create dummy variables for categorical variables
- *Creating dummy variables for the remaining categorical variables and dropping the level called 'Unknown' which represents null/select values*
- Dropping the repeated features from the data frame

Split Train and Test set & Feature scaling

- Split data set into train and test data set
- Scaling of features for standardization (numerical data)

Model Building

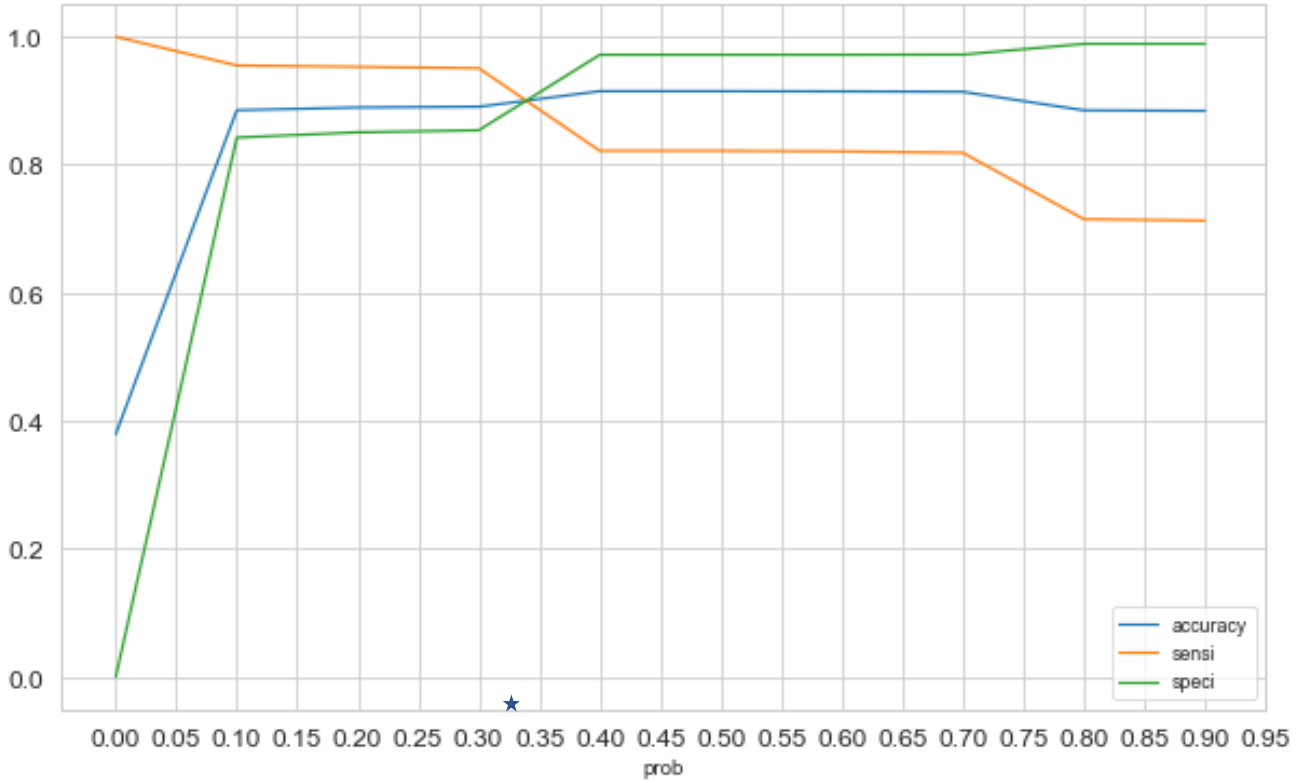
- Feature selection using RFE on the trained data set
- Find out optimal model using Logistic Regression
- Calculate p-values and VIF values for each trained model
- Find out correlation between numeric variables
- Plot the ROC curve and calculate GINI
- Find out the optimal cutoff probability for the data set from default
- Find out the precision and recall trade off (in our case for 80%)
- Calculate accuracy, sensitivity, specificity, precision, recall, f1 score,

Positive predictive value, negative predictive value from confusion matrix etc.

Plotting the ROC Curve

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

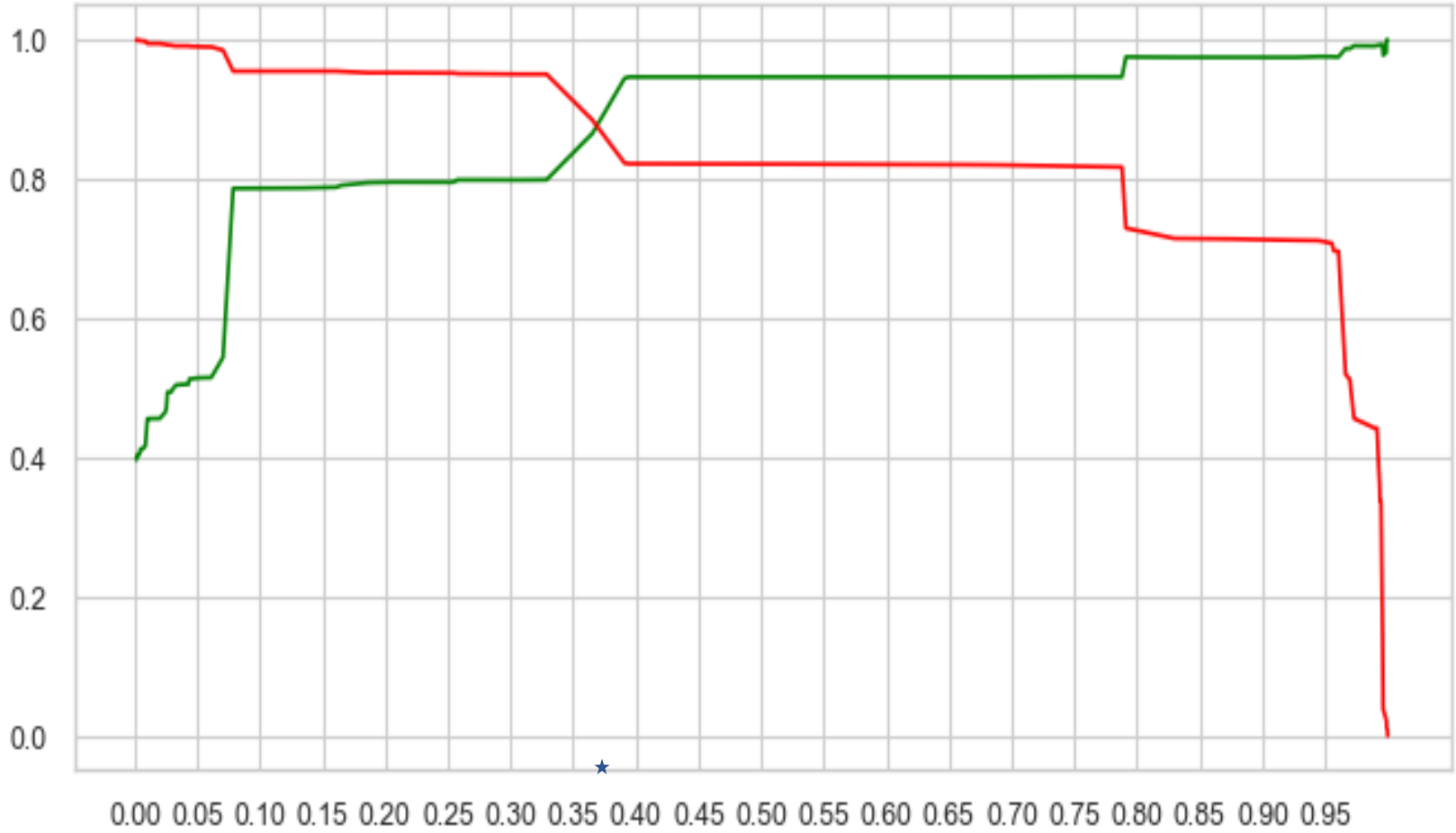
Model Evaluation (Accuracy, Sensitivity, Specificity on Trained data set)



Accuracy – 0.90
Sensitivity – 0.88
Specificity – 0.91

From the curve above, 0.33 is the optimum point to take it as a cutoff probability.

Model Evaluation (Precision and Recall on Trained data set)



Precision – 0.86
Recall – 0.88
F1 score – 0.875

From the precision-recall graph above, we get the optical threshold value as close to .37. However our business requirement here is to have Lead Conversion Rate around 80%.
This is achieved with our earlier threshold value of around 0.33. So we will stick to this value.

Model Evaluation (Accuracy, Sensitivity, Specificity on Test data set)

- Accuracy – 0.90
- Sensitivity – 0.89
- Specificity – 0.90
- Precision – 0.86
- Recall – 0.89

Feature Importance

Positive coefficient values

Negative coefficient values

Summary

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
- After trying several models, we finally chose a model with the following characteristics.
  - All variables have p-value < 0.05.
  - All the features have very low VIF values, meaning, there is `hardly any multicollinearity` among the features. This is also evident from the heat map.
  - The overall accuracy of 0.914 at a probability threshold of 0.33 on the test dataset is also very acceptable.
- While we have checked both sensitivity & specificity as well as precision & recall metrics , we have considered the optimal cut off based on sensitivity & specificity for calculating the final prediction.
- Accuracy, sensitivity & specificity values of test set are approx. closer to values calculated using trained data set.
- Depending on the business requirement, we can increase or decrease the probability threshold which ultimately results in either decrease or increase in sensitivity and increase or decrease in specificity of the model.
- High sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted, while high specificity will ensure that leads that are on the probability of getting Converted are not selected.
- Hence, overall model seems to be good.