Lead Scoring

ANUBHA MISHRA

Problem Statement

1. X Education sells online courses to industry professionals.
2. X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
3. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
4. If they are successfully in identifying this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.
  - ➢ Check and handle duplicate data.
  - ➢Check and handle NA values and missing values.
  - ➢ Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - ➢ Imputation of the values, if necessary.
  - ➢ Check and handle outliers in data.
- ➢EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

➢Feature Scaling & Dummy Variables and encoding of the data.

➢Classification technique:

➢logistic regression used for the model making and prediction.

➢Validation of the model.

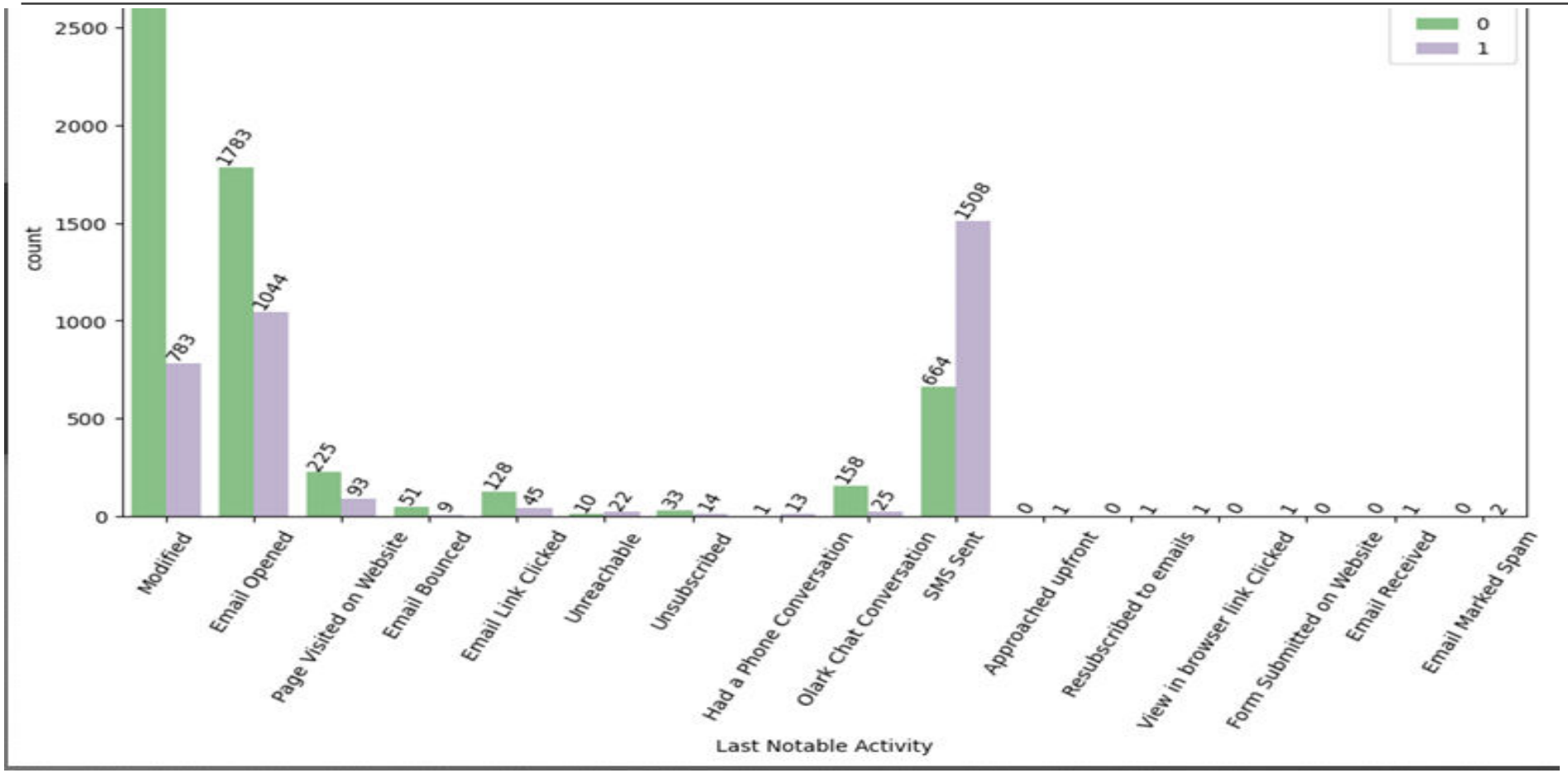
➢Model presentation.

➢Conclusions and recommendations.

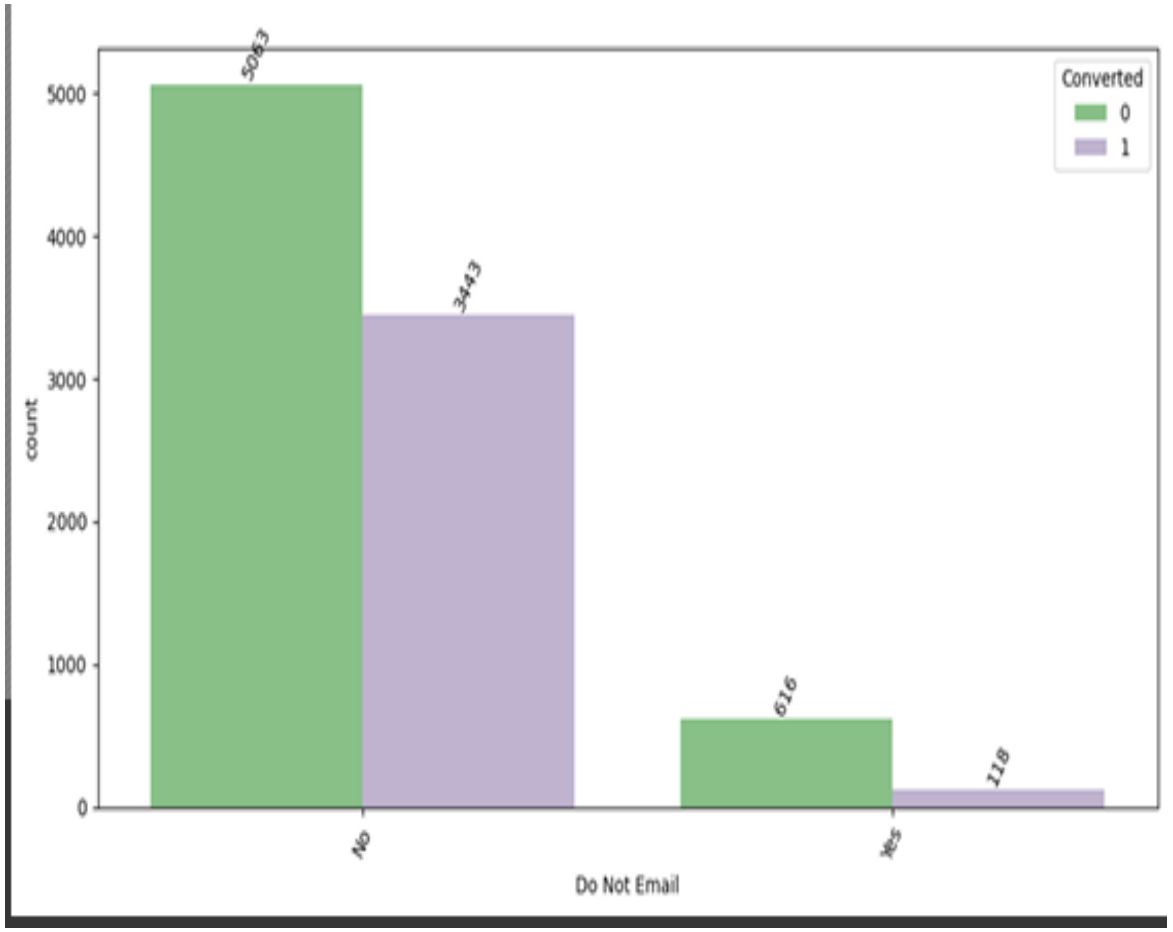
Data manipulation

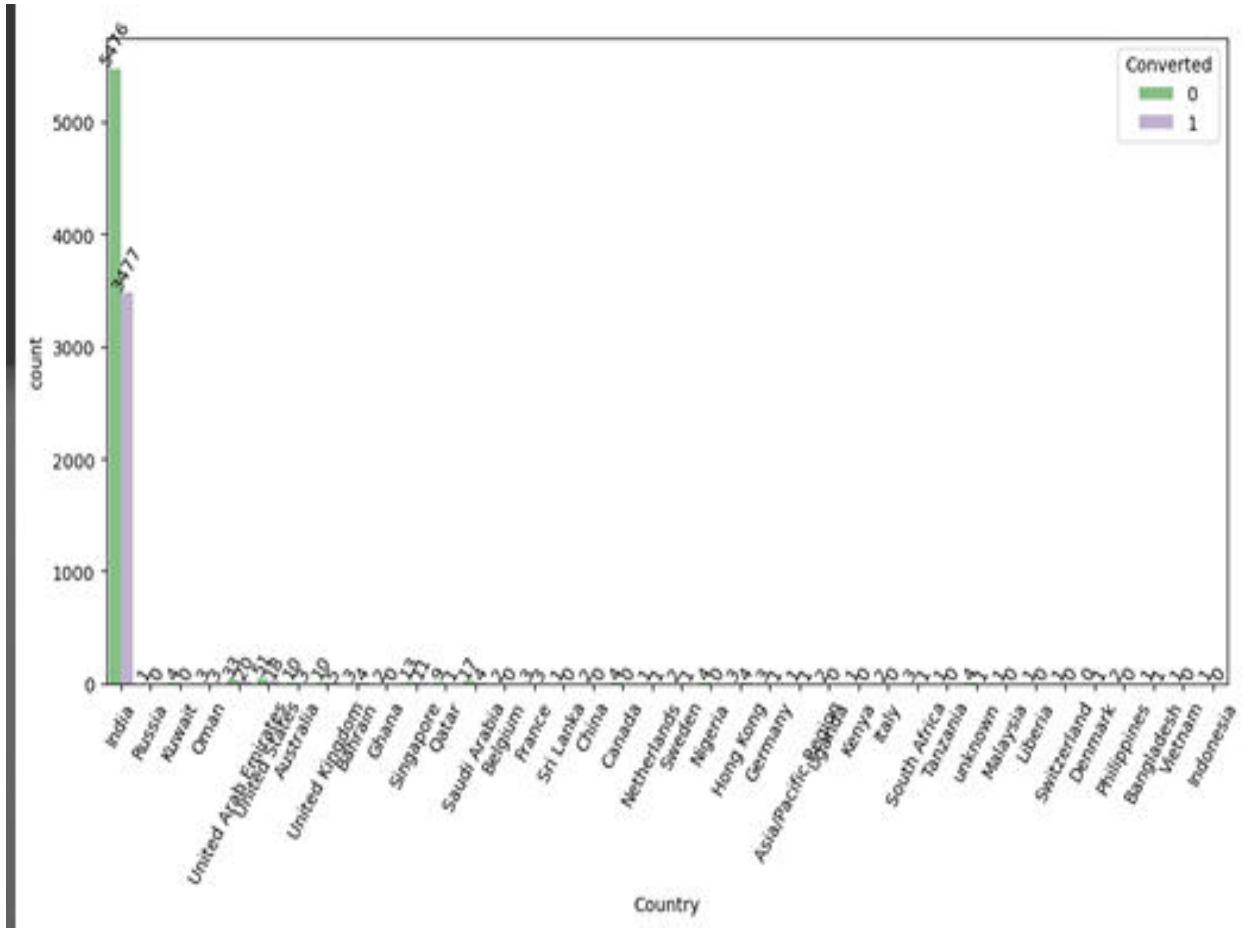
- Total Number of Columns=37, Total Number of Rows =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

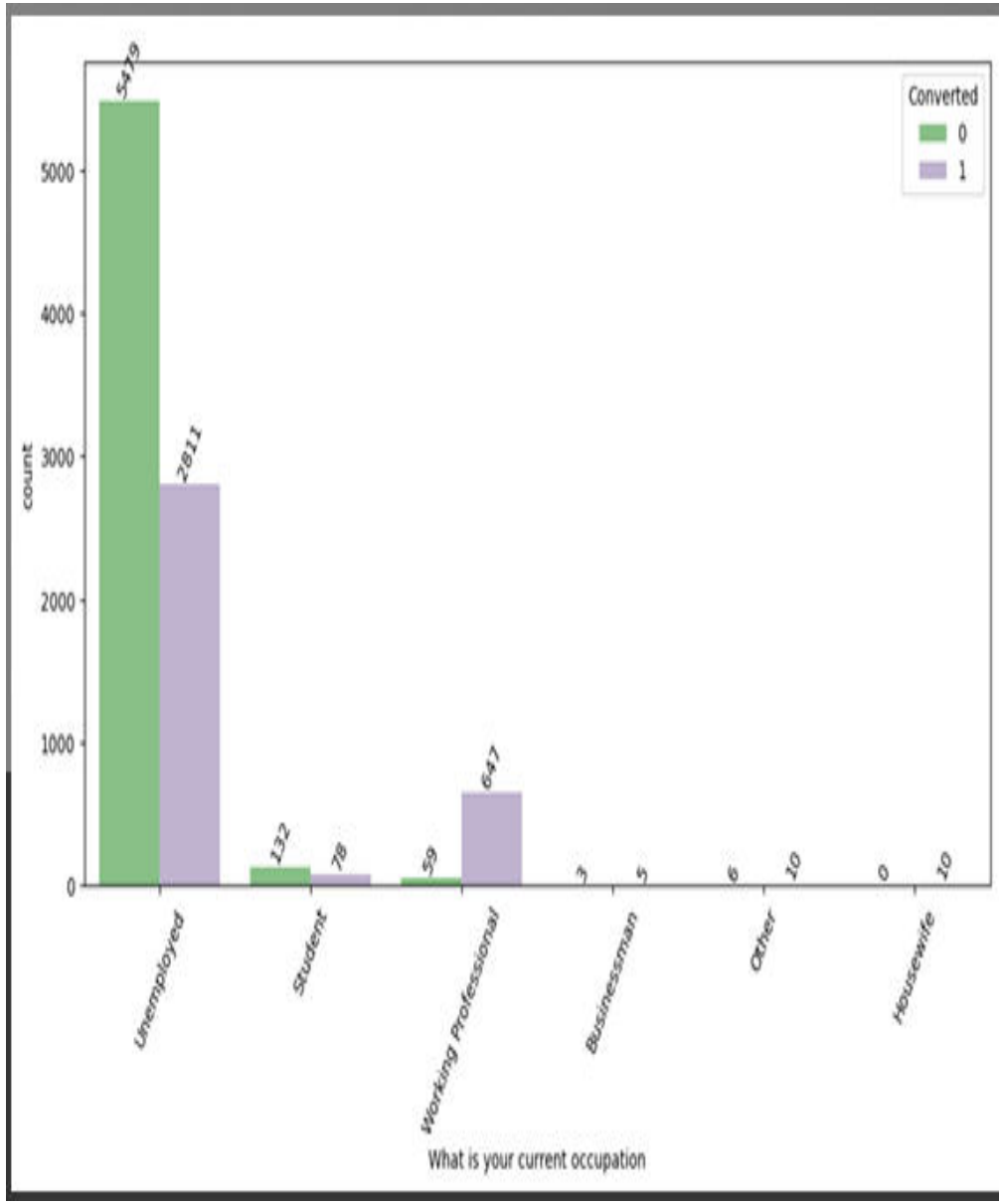
EDA

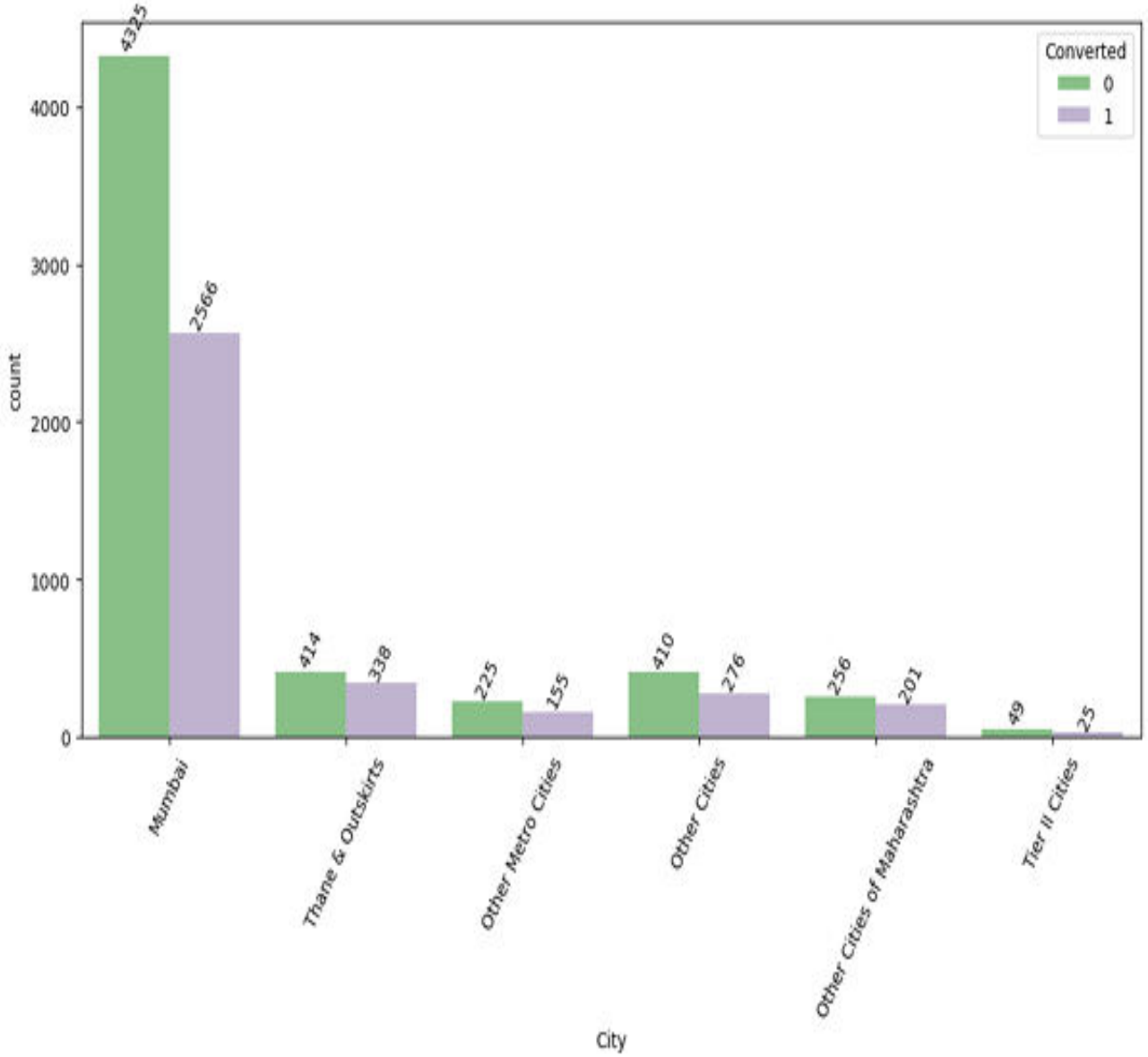

0
1
count
2500
2000
1500
1000
500
0
783
1783
1044
225
93
51
9
128
45
10
22
33
14
1
13
158
25
664
1508
0
1
0
1
1
0
1
0
0
1
0
2
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

Categorical Features

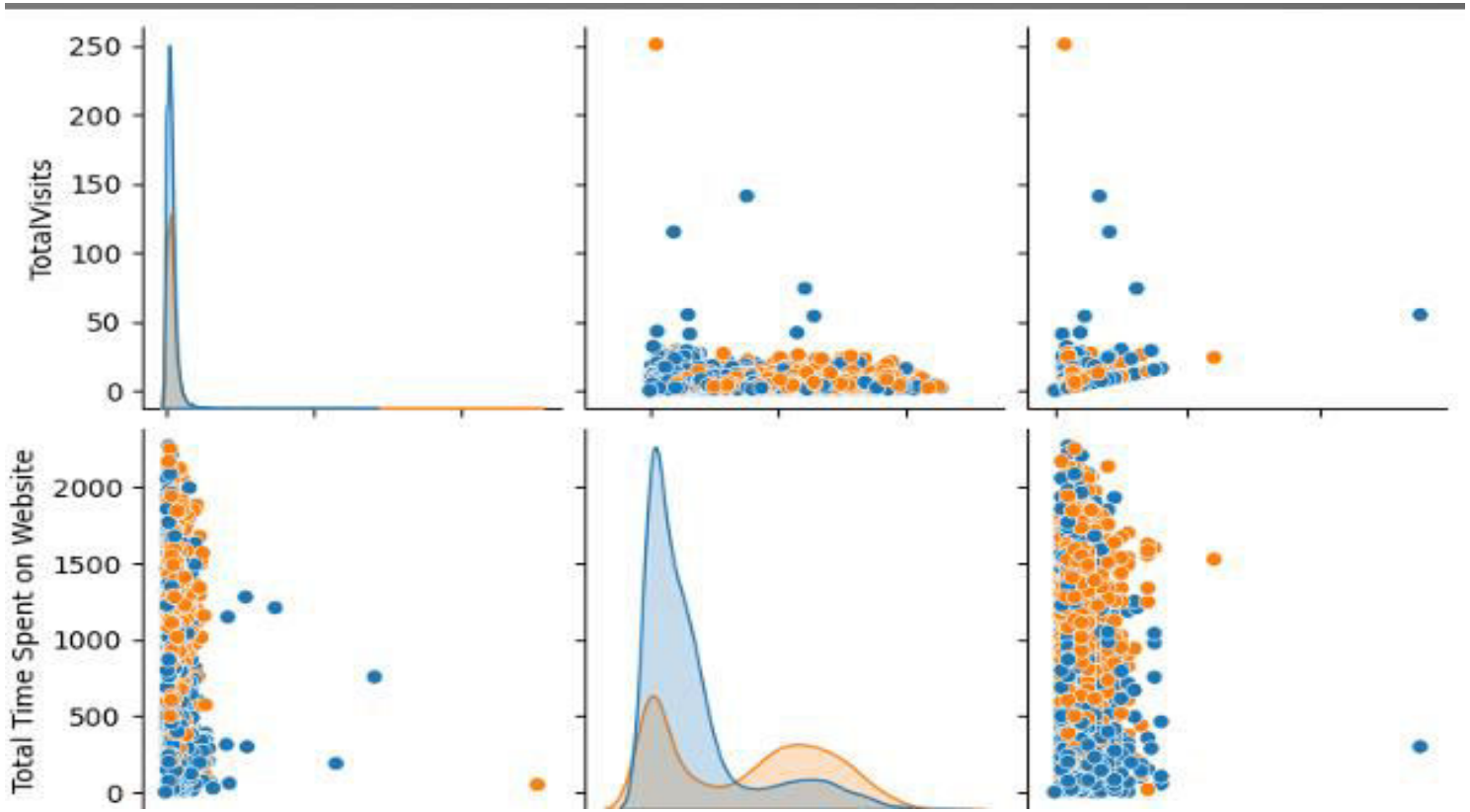




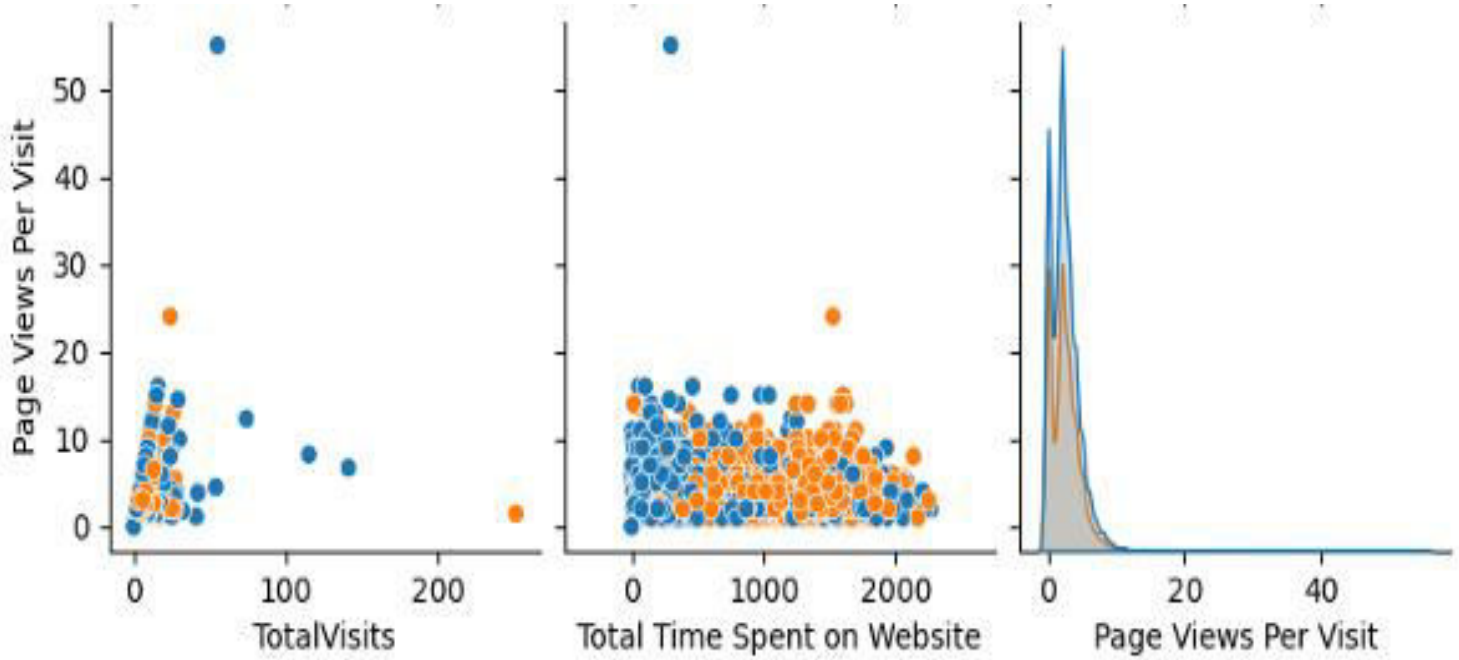

Converted
0
1
count
5000
4000
3000
2000
1000
0
5419
2811
132
78
59
647
3
5
6
10
0
10
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Converted
0
1
count
4000
3000
2000
1000
0
4325
2566
414
338
225
155
410
276
256
201
49
25
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

Numerical features



Numerical features




TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.22
0.51
0.22
1
0.31
0.51
0.31
1
TotalVisits
Page Views Per Visit
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

Data conversion

- Numerical Variables are Normalized
- Dummy Variables are created for object type variables

Model Building

➢Splitting the Data into Training and Testing Sets

➢The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

➢Use RFE for Feature Selection

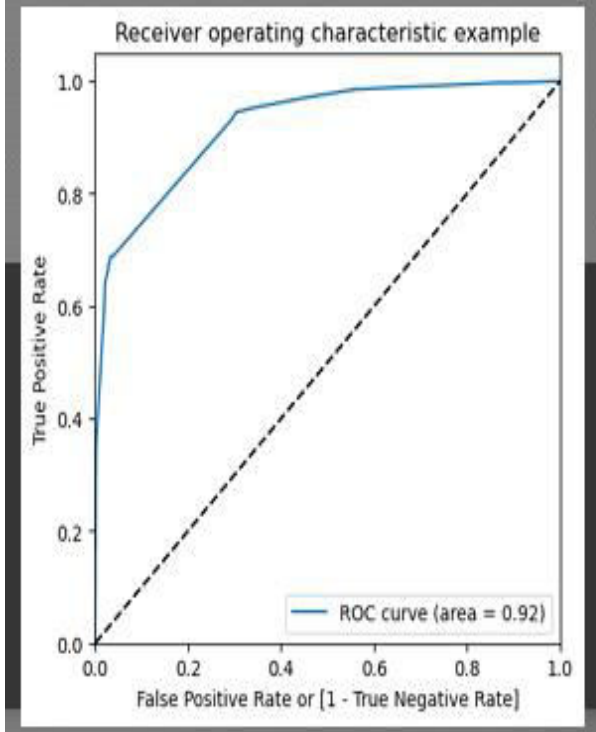
➢Running RFE with 15 variables as output

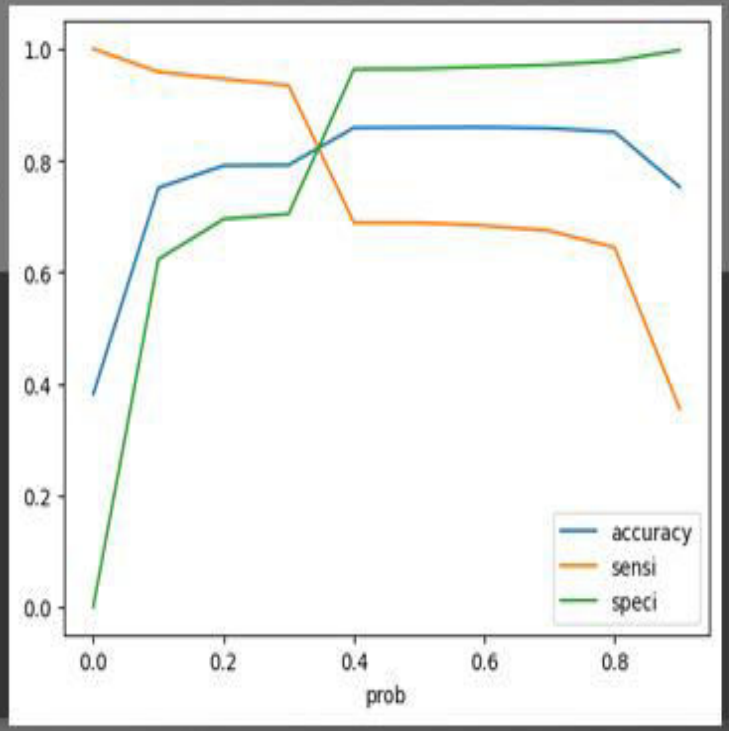
➢Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5

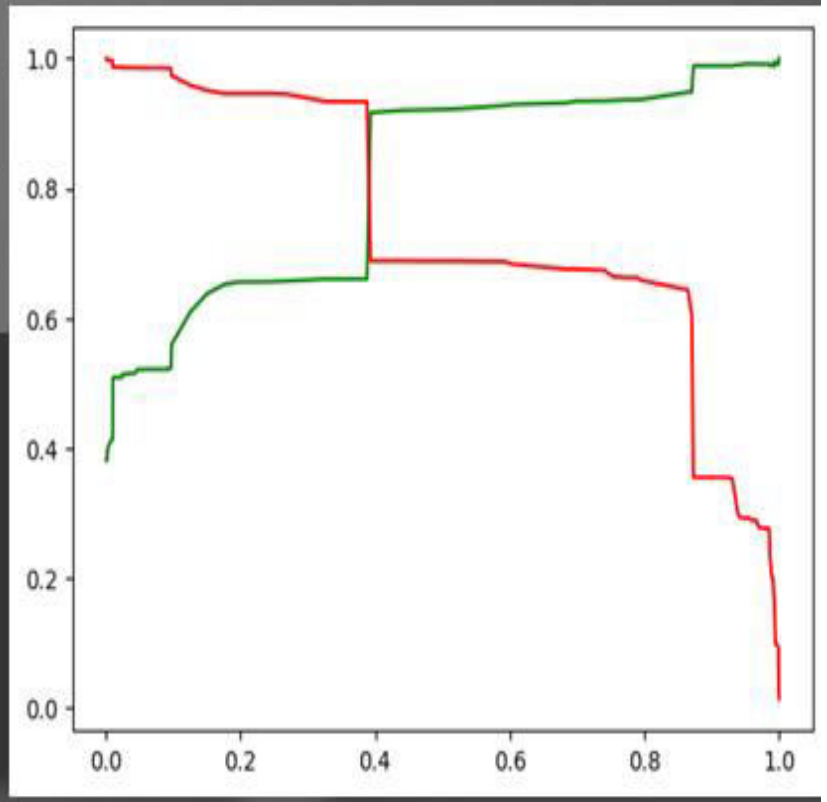
➢Predictions on test data set

➢Overall accuracy 79.43% on test data, and 79.17% on train data set

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0


0.0
0.2
0.4
0.6
0.8
1.0

ROC Curve

Finding Optimal Cut off Point

➢Optimal cut off probability is that

➢Probability where we get balanced sensitivity and specificity.

➢From the second graph it is visible that the optimal cut off is at 0.35.

➢From Precision and recall chart suitable threshold is 0.38 or 0.39

➢Although the values are same at both cutoff 0.35 and 0.38, so kept 0.35

Conclusion

It was found that the variables that mattered the most in the lead converted are

- Lead Origin
- Lead Add Form
- Do Not Email
- Yes
- Last Activity
- Converted to Lead
- Olark Chat Conversation
- What is your current occupation_

- Unemployed
- Working Professional
- Tags
- Busy
- Closed by Horizon
- Lost to EINS
- Will revert after reading the email
- in touch with EINS
- Last Notable Activity
- SMS Sen