Lead Scoring Case Study

By:-

Arjun Mukerji

Harinandan Praveen

Abhishek Sahebrao Bolke

Problem Statement

- X Education sells online courses to industry professionals.
- 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.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify 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

- EDA

1. Check and handle duplicate data
2. Check and handle NAN values and missing values
3. Drop columns, if it contains large amount of missing values and are not useful for the analysis
4. Imputation of the values, if necessary
5. Check and handle Outliers in data

- Feature Scaling and Dummy Variables
- Classification Techniques: Logistic Regression used for the model making and prediction

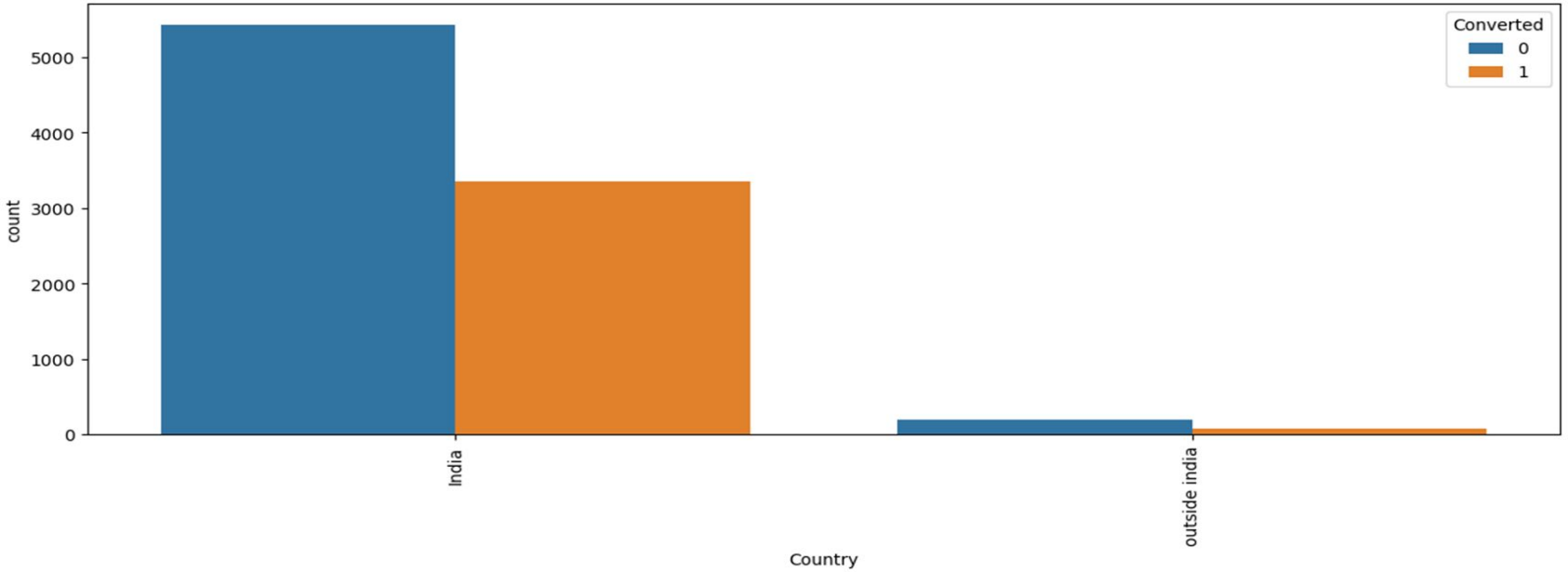
Solution Methodology (Continued)

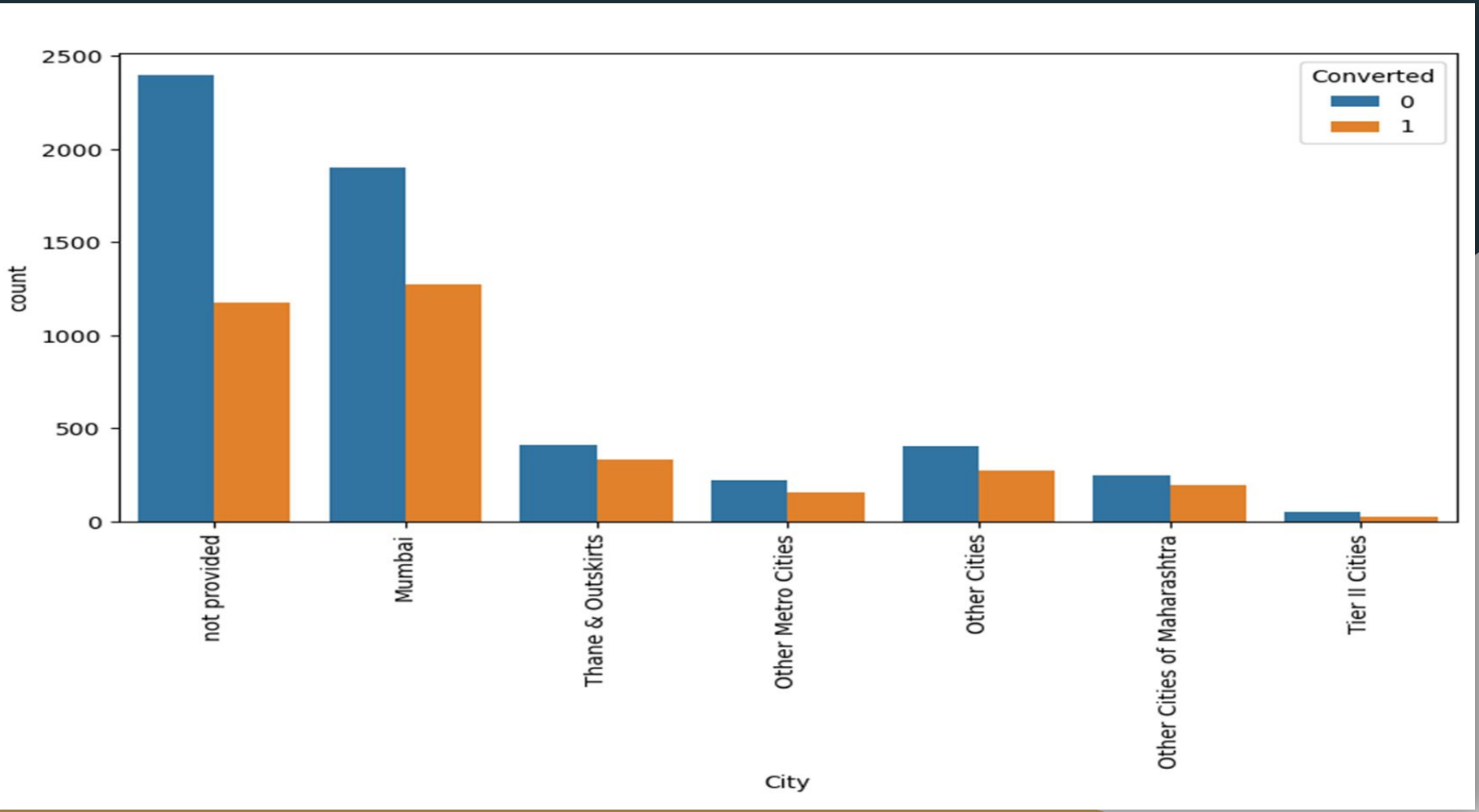
- Validation of the Model
- Model presentation
- Conclusions and Recommendations

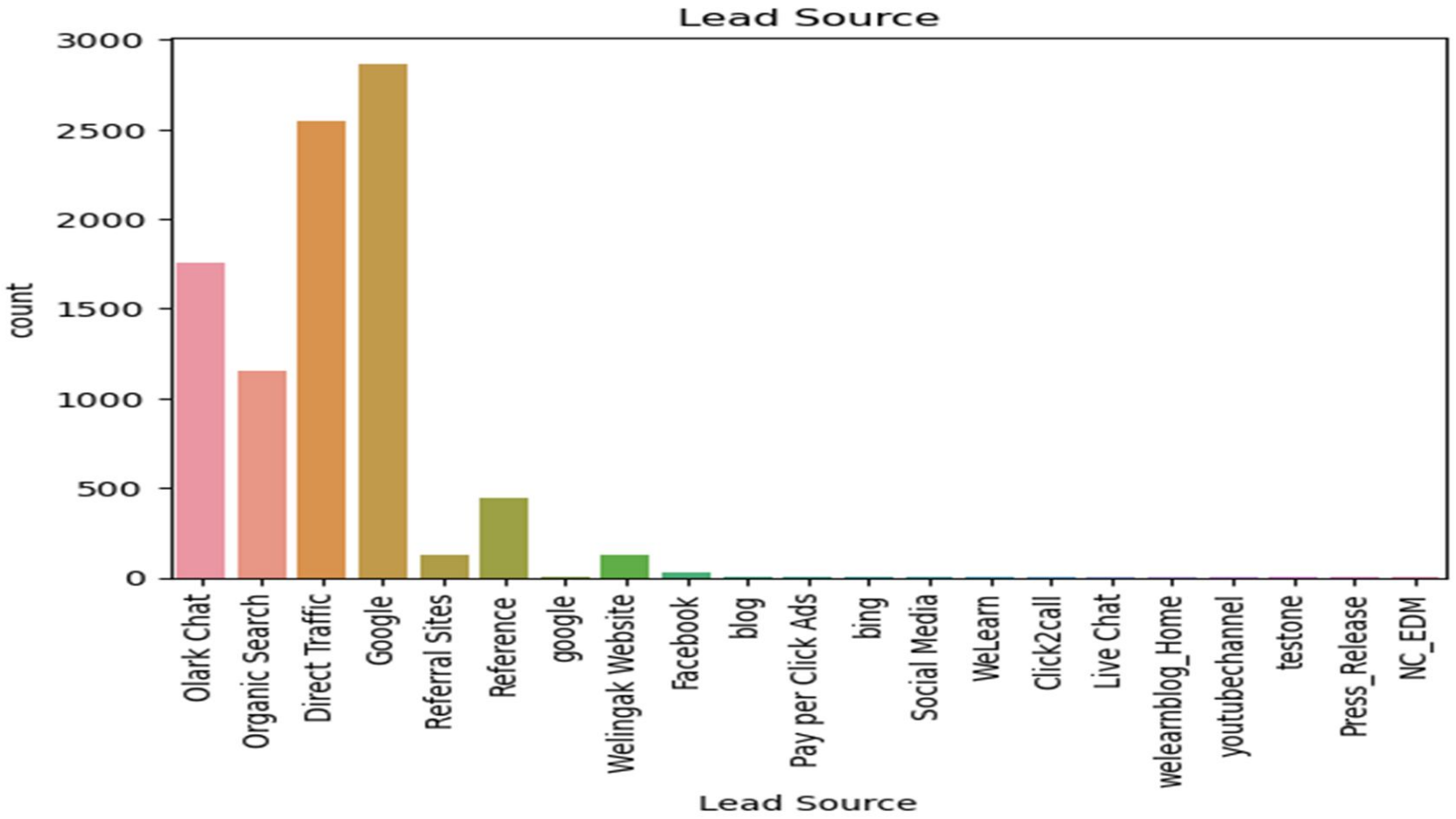
Data Cleaning and Data Manipulation

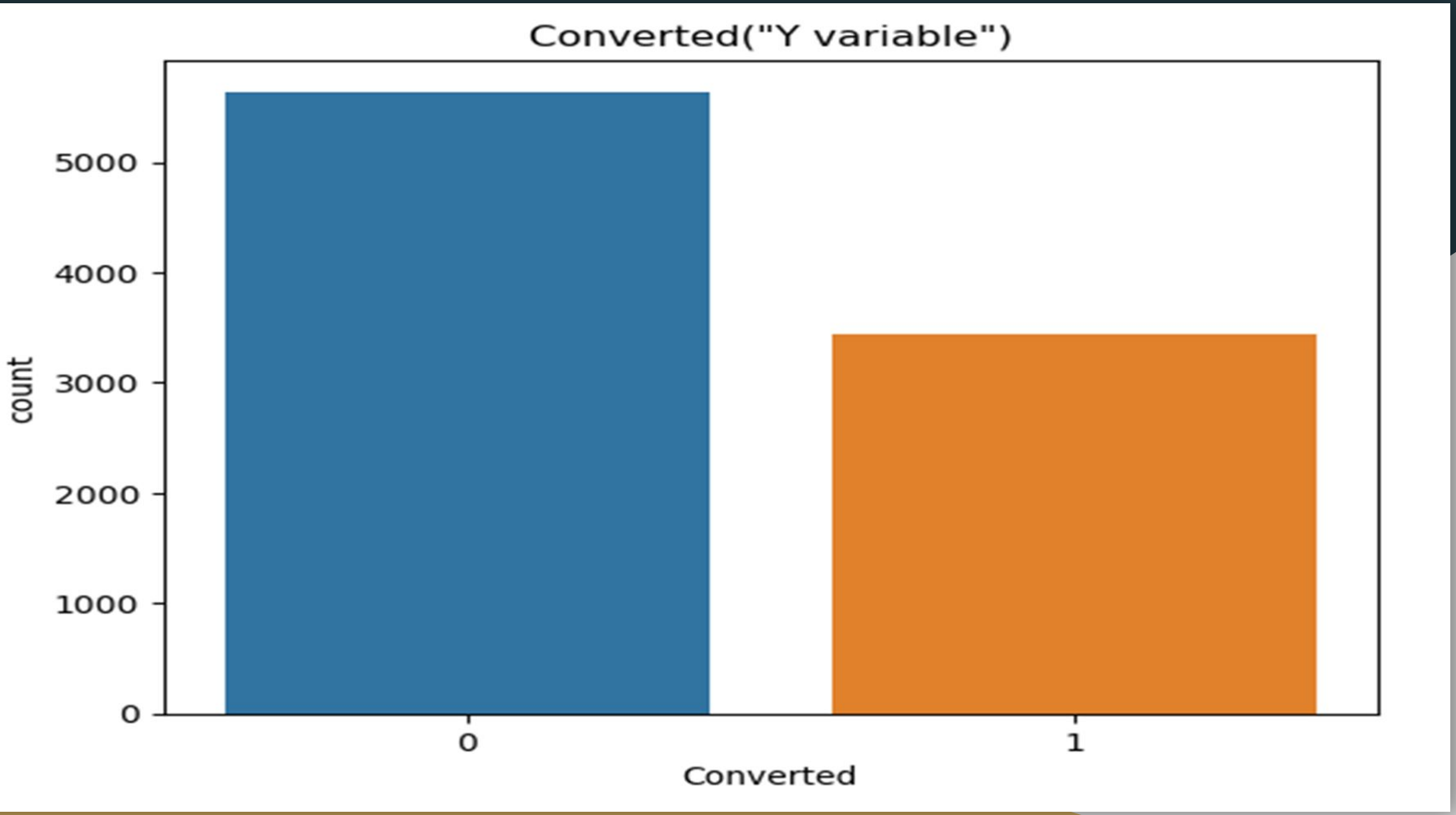
- Total number of Rows = 9240, Total number of Columns=37.
- Removing ‘Lead Number’ and ‘Prospect ID’ which is not necessary for analysis since they all have Unique values.
- Converting ‘Select’ values to ‘NAN’.
- Dropping unique valued Columns.
- Dropping off the Columns with more than 45% missing values.
- Imputing ‘Not Provided’ values with ‘India’ since ‘India’ is the most common occurrence among the non-missing values.

EDA (Categorical Attributes Analysis)

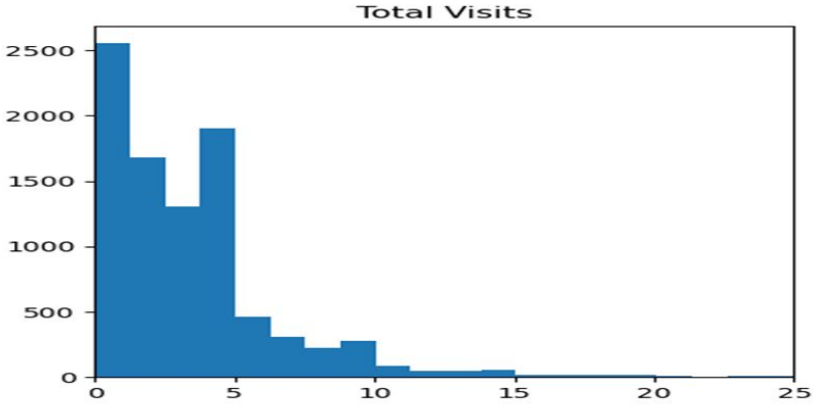


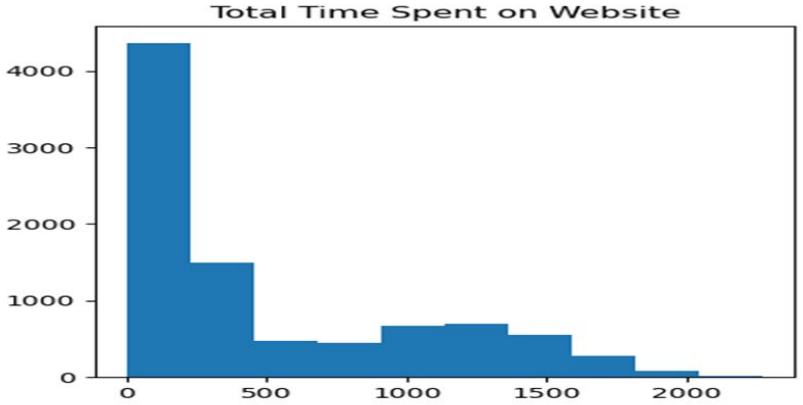

Converted
0
1
count
2500
2000
1500
1000
500
0
not provided
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

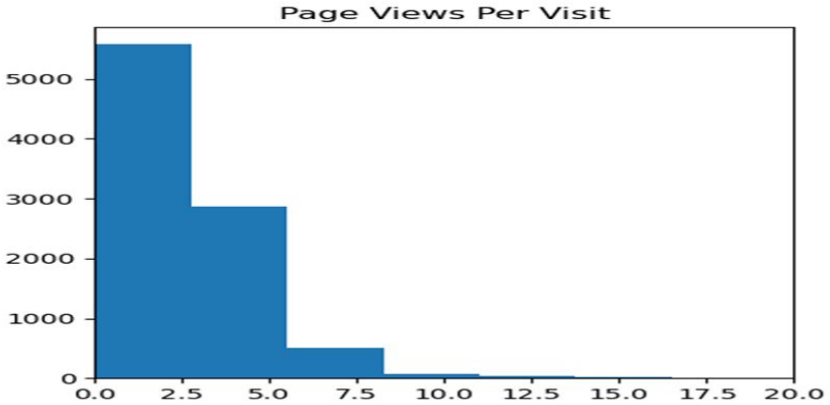

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Converted("Y variable")
5000
4000
3000
2000
1000
0
count
0
1
Converted

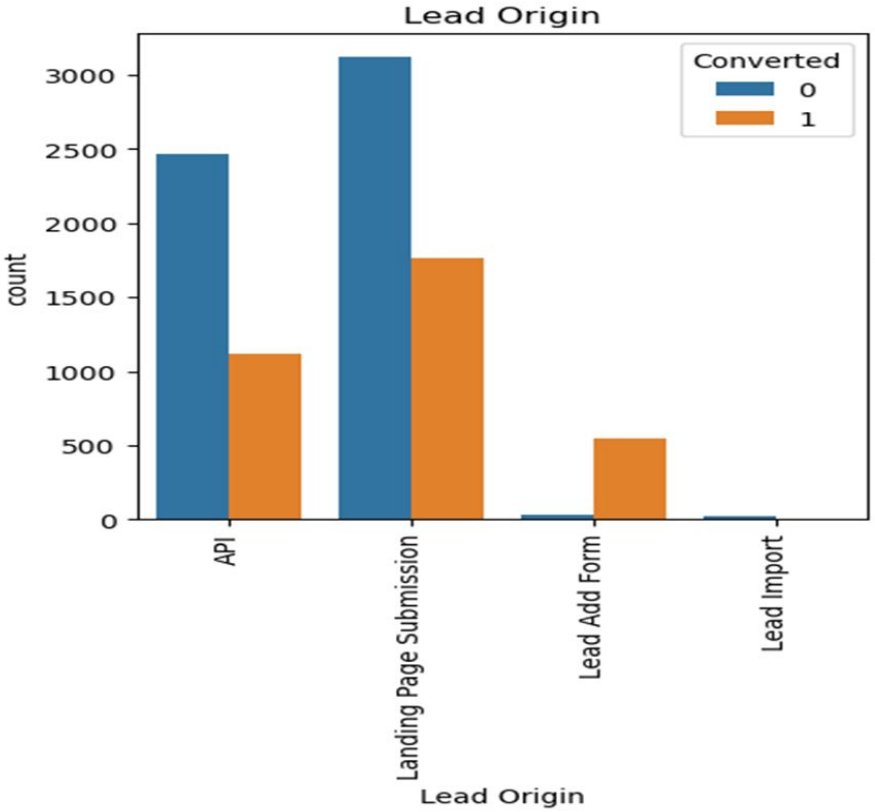
EDA (Numerical Variables)

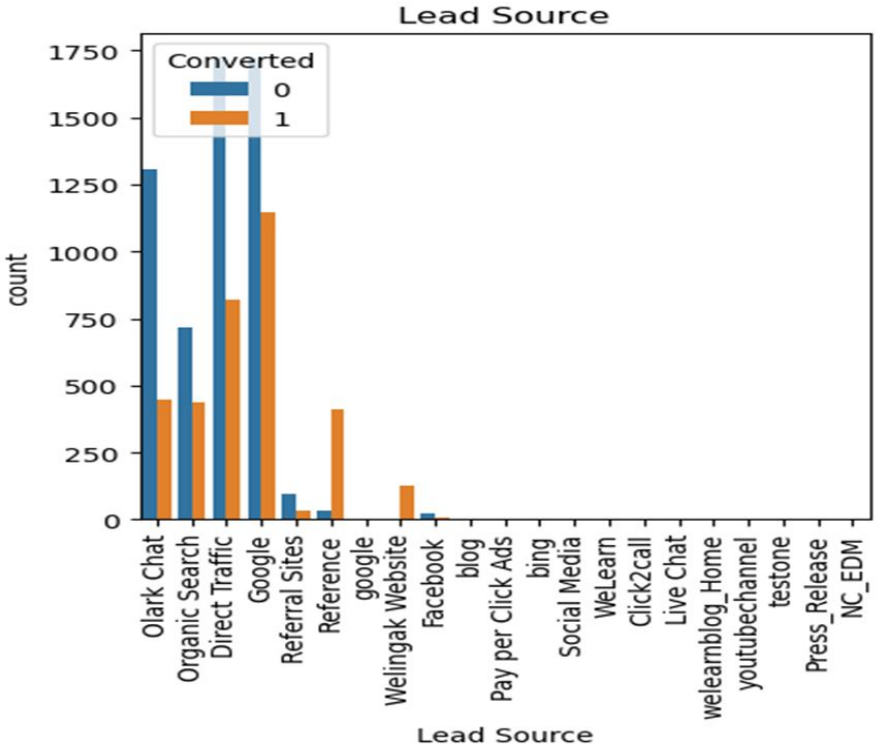


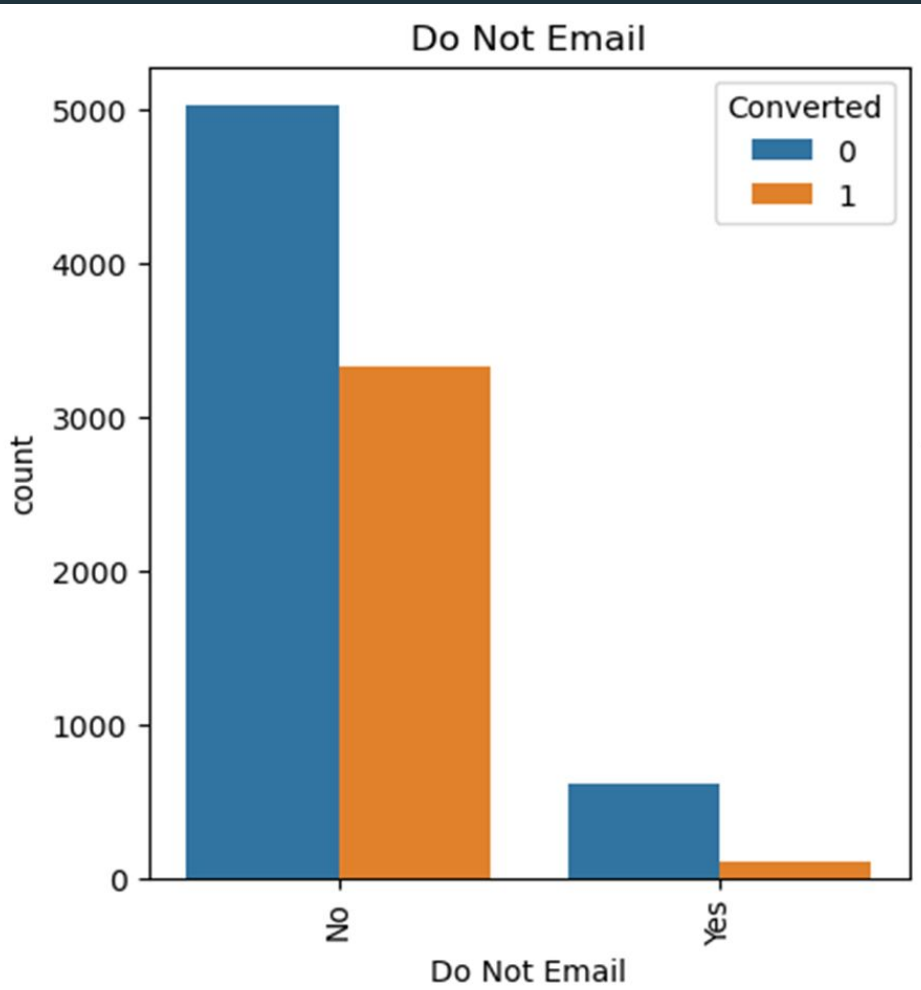


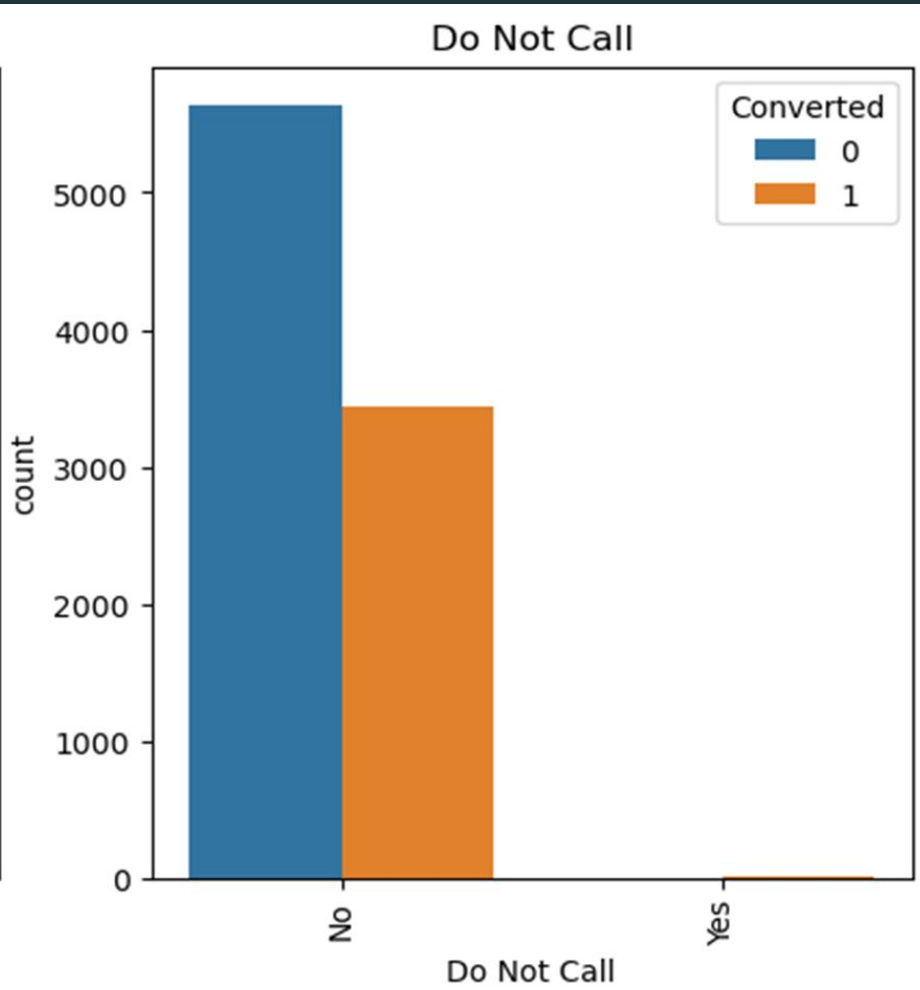


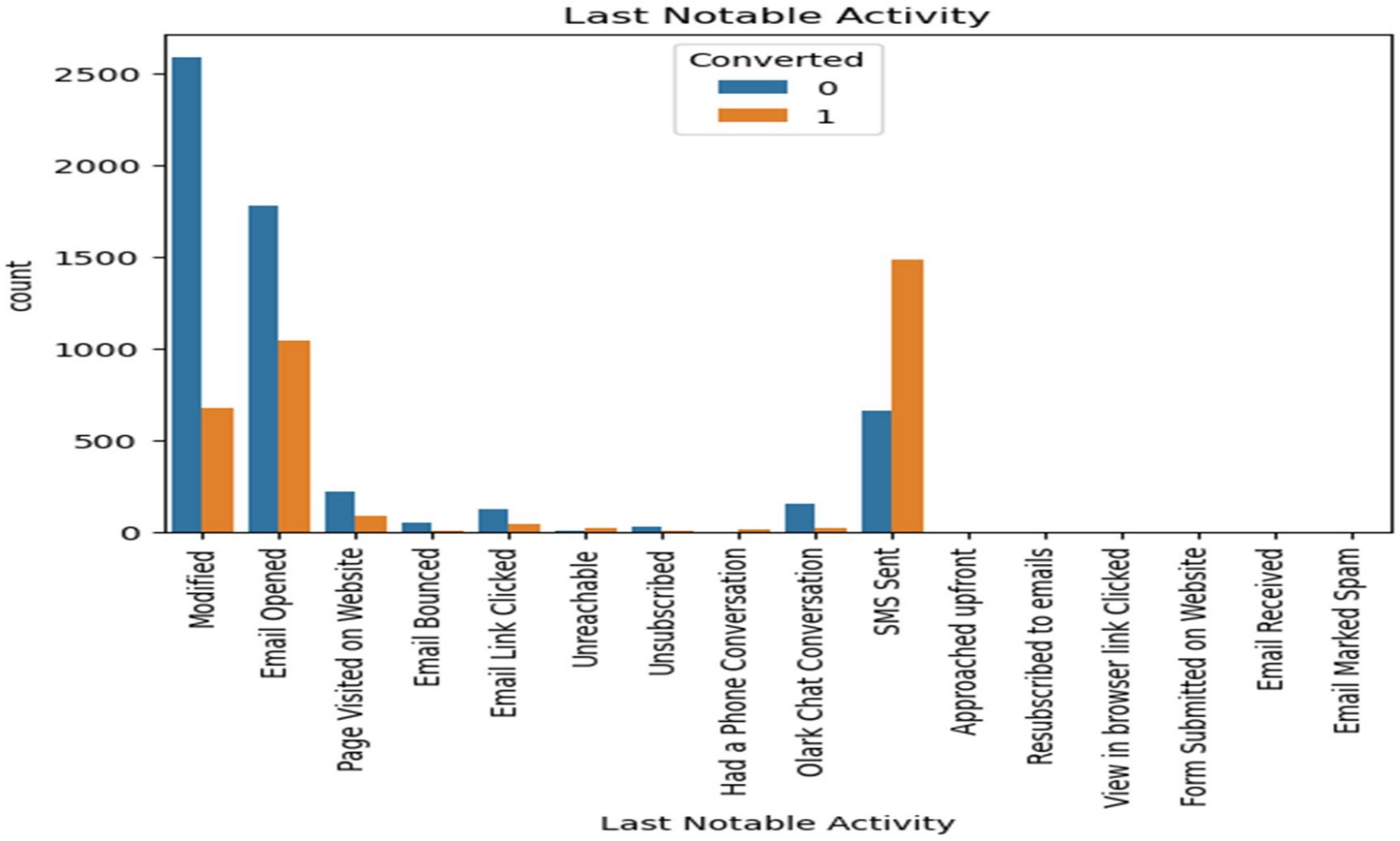
EDA (Categorical Variable Relation)






Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email


Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call


Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
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

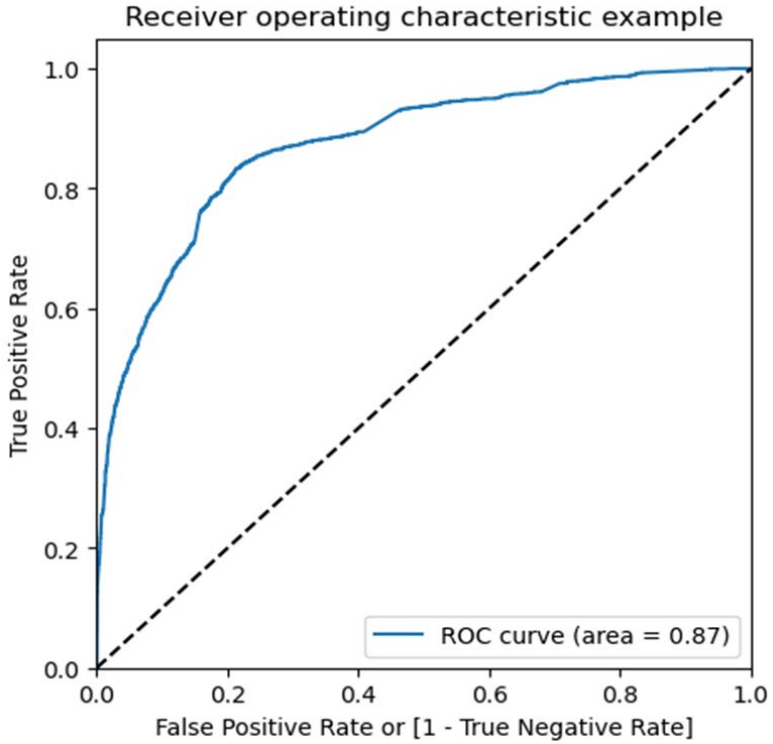
Data Conversion

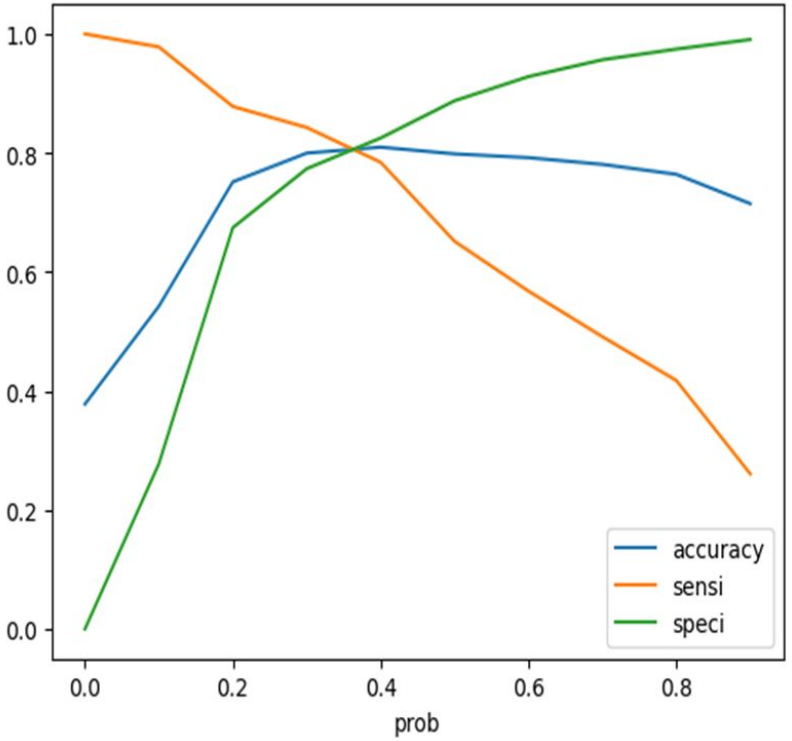
- Numerical Variables are Normalised.
- Dummy Variables are created for Object Type Variables.
- Total Rows for Analysis: 8991
- Total Columns for Analysis: 82

Model Results

- 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 is 78.9%.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.87)
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

ROC Curve (Continued)

- Finding Optimal Cut Off Point.
- Optimal Cut Off Probability is that Probability where we get balanced Sensitivity and Specificity.
- From the second graph, it’s visible that the optimal cut-off is approximately at 0.37.

Conclusion and Recommendations

•Looking at the below variables, we should be able to help X-Education company with finding out hot leads which will be very high potential customers:-

1.TotalVisits

2.The total time spend on the Website

3.Lead Origin_Lead Add Form

4.Lead Source_Direct Traffic

5.Lead Source_Google

6.Lead Source_Welingak Website

Conclusion and Recommendations (Continued)

7.Lead Source_Organic Search

8.Lead Source_Referral Sites

9.Lead Source_Welingak Website

10.Do Not Email_Yes

11.Last Activity_Olark Chat Conversation

12. Last Activity_Email Bounced

•Taking the above details into consideration and using this model, the X-Education company can maximise the sales of their course and also generate an excellent revenue.

THANK YOU!!!