Lead Scoring

By- Priyanka Nayar

Problem Statement

- X Education sells online courses to industry professionals.
- The company acquires a significant number of leads, but their lead conversion rate is very poor. For example, out of 100 acquired leads, only about 30 of them are converted.
- To improve the efficiency of the conversion process, the company aims to identify the most potential leads, referred to as 'Hot Leads.‘
- By successfully identifying this set of leads, the lead conversion rate is expected to increase. This is because the sales team can focus more on communicating with the potential leads rather than making calls to everyone.

Objective

- X education aims to identify the most promising leads.
- Build a model that can accurately identify hot leads.
- Deploy the model for future use.

Solution Methodology

Data cleaning and data manipulation:

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns with a large amount of missing values and are not useful for analysis.
- Impute missing values if necessary.
- Check and handle outliers in data.

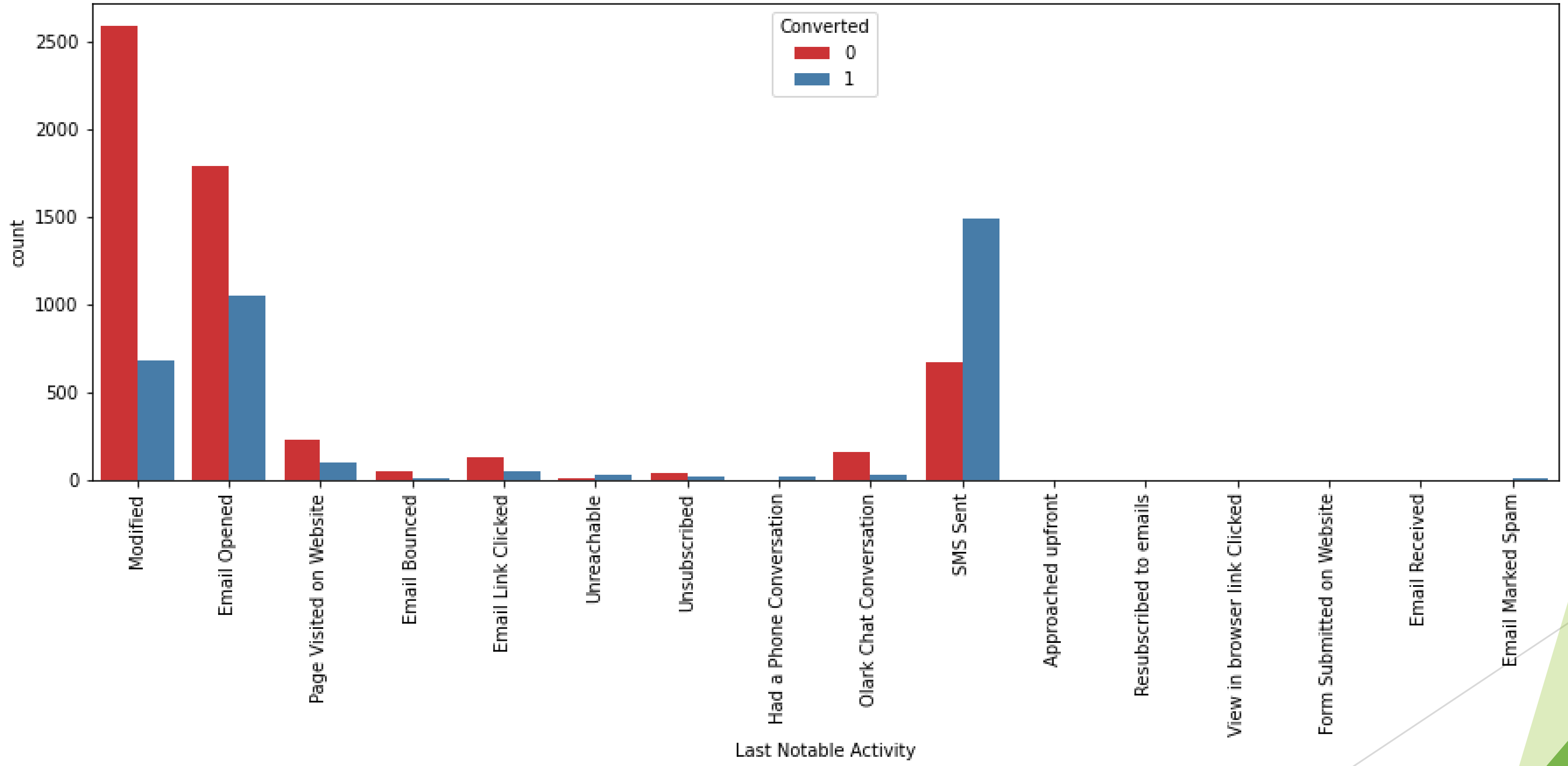
EDA:

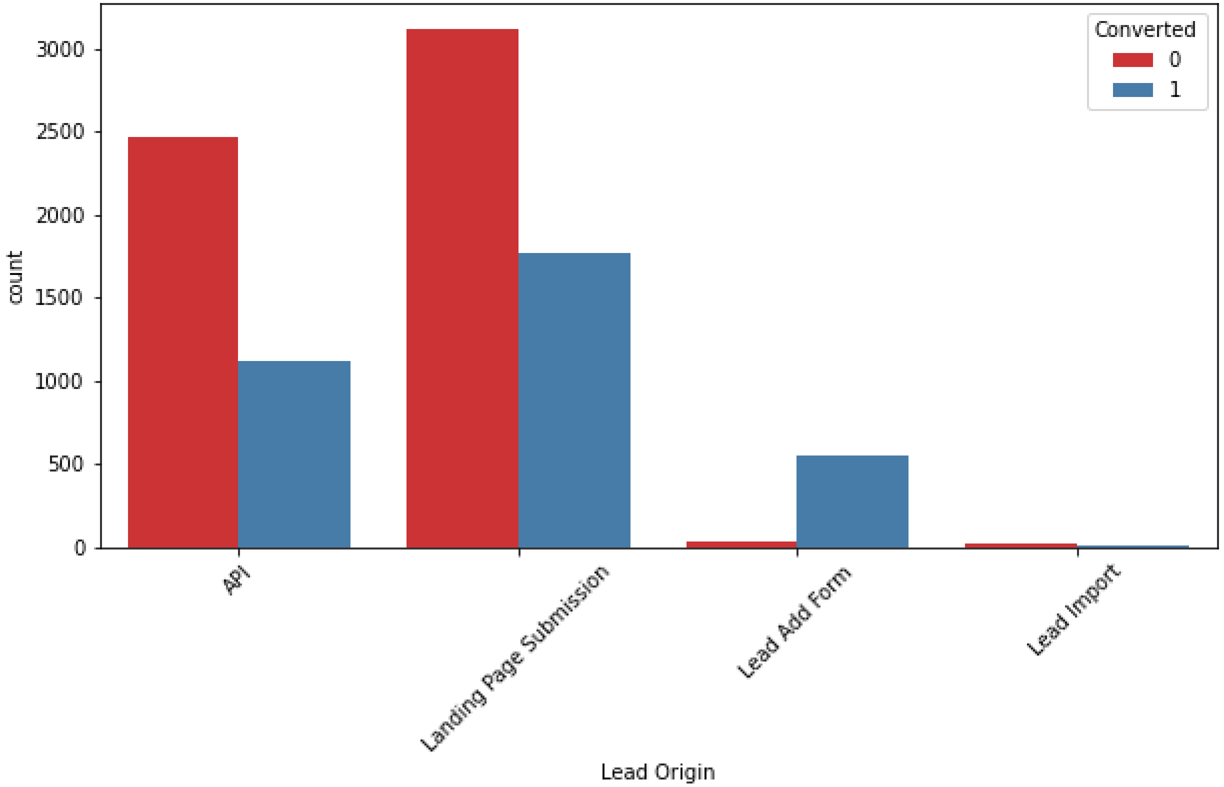
- Perform univariate data analysis: count the values, analyse the variable's distribution, etc.
- Conduct bivariate data analysis: calculate correlation coefficients and analyse patterns between variables.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: Use logistic regression for model building and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

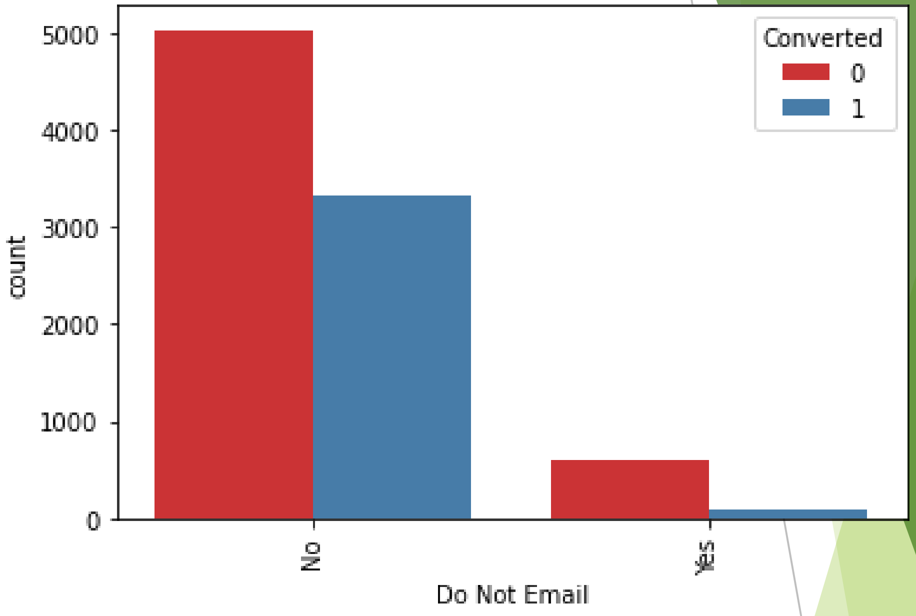
Data Manipulation

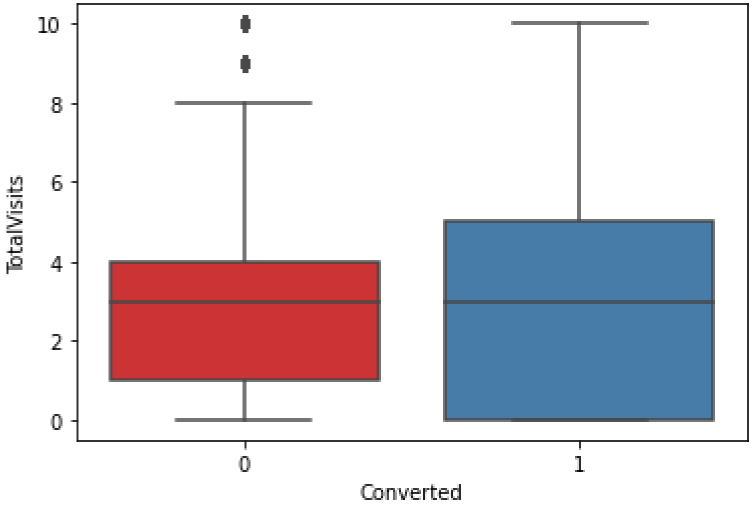
- The dataset contains 37 rows and 9240 columns.
- Certain single value features like "Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," and "I agree to pay the amount through cheque" have been removed.
- The columns "Prospect ID" and "Lead Number" have been dropped as they are not necessary for the analysis.
- Some object type variables have been examined for value counts, and certain features with low variance have been dropped. These features include "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," and "Digital Advertisement."
- Columns with more than 35% missing values, such as "How did you hear about X Education" and "Lead Profile," have been dropped.

EDA Diagram

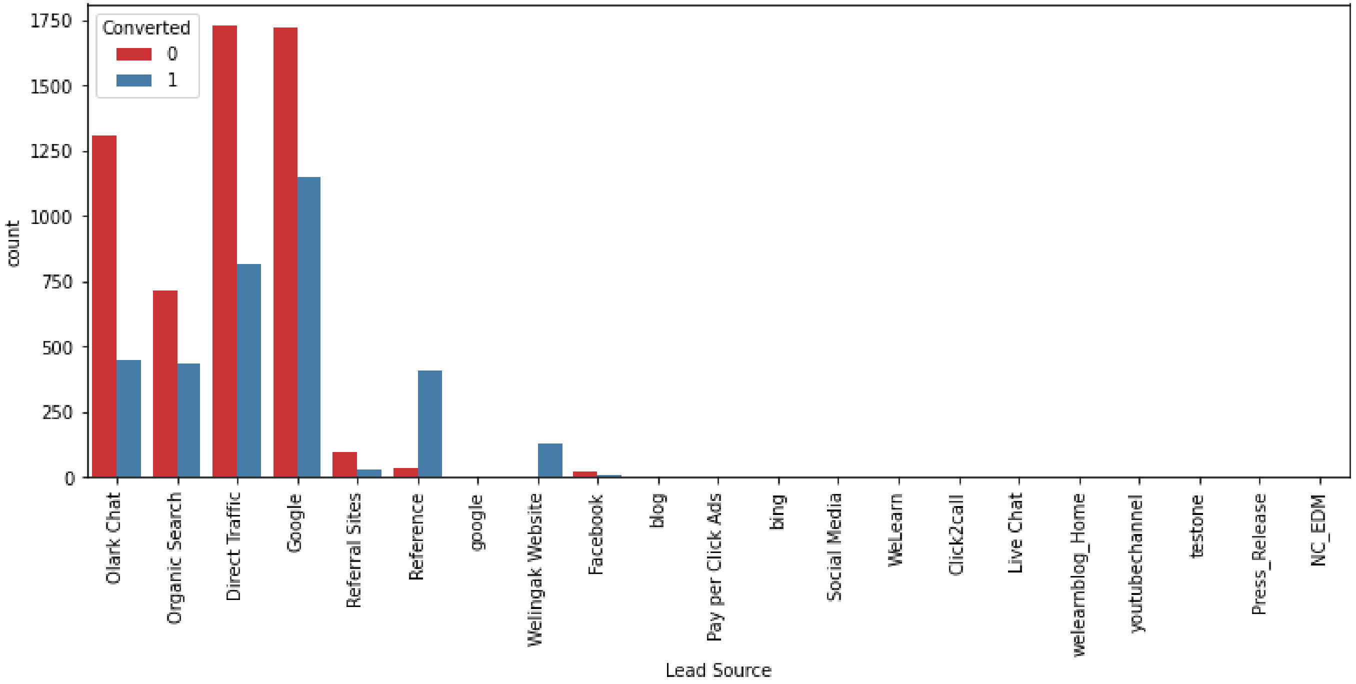

Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

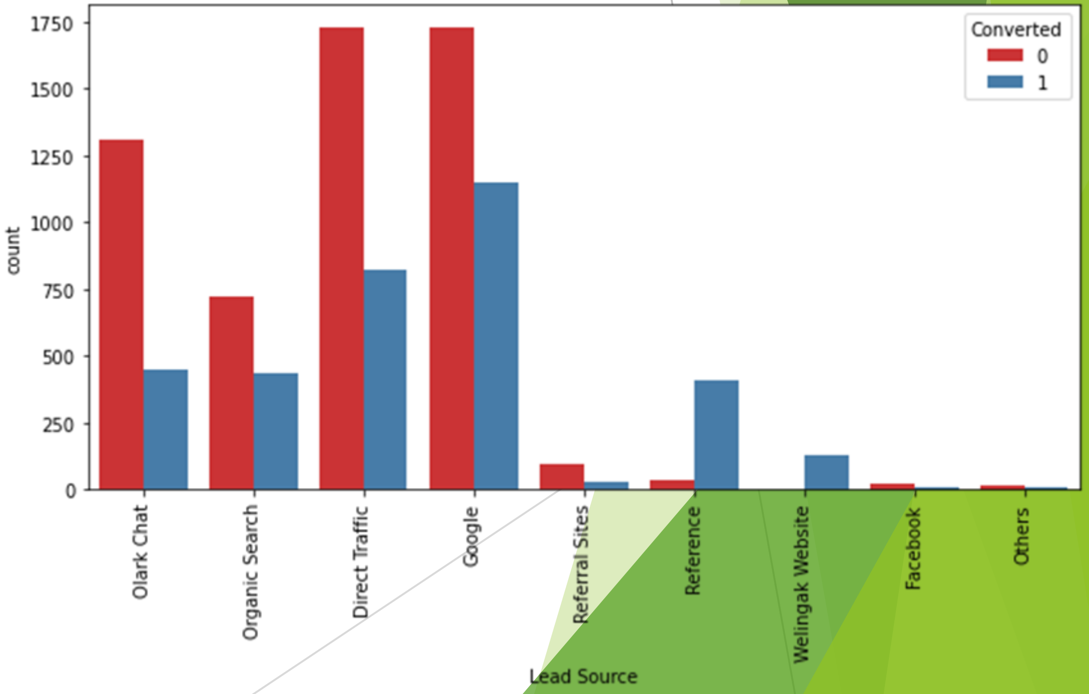

Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

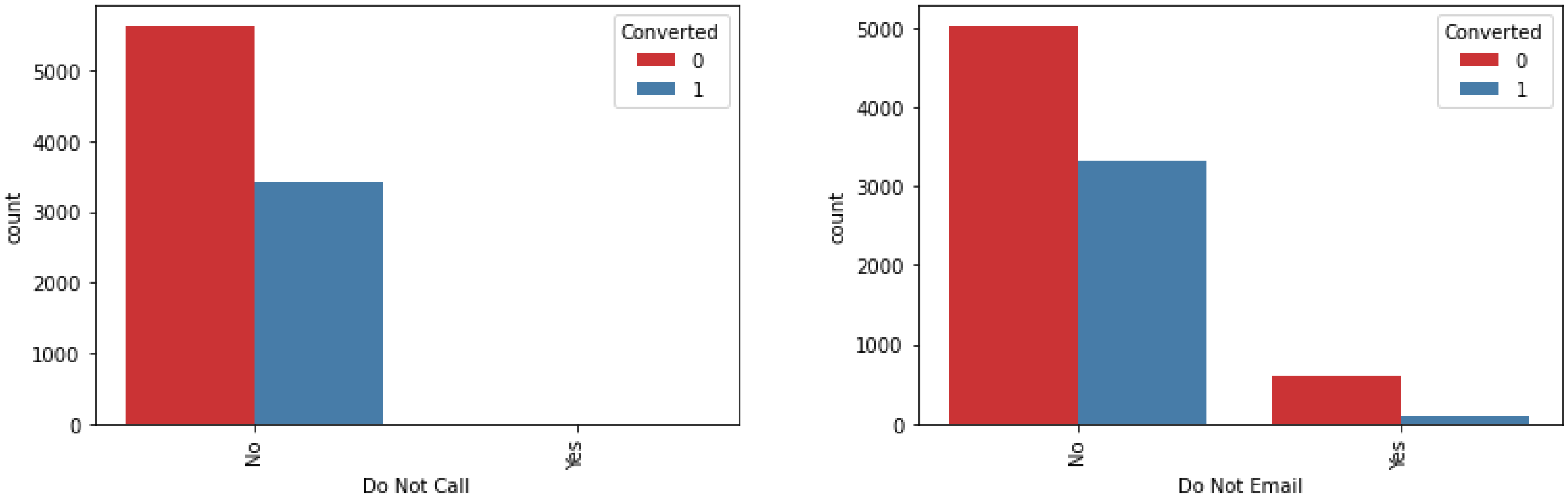

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

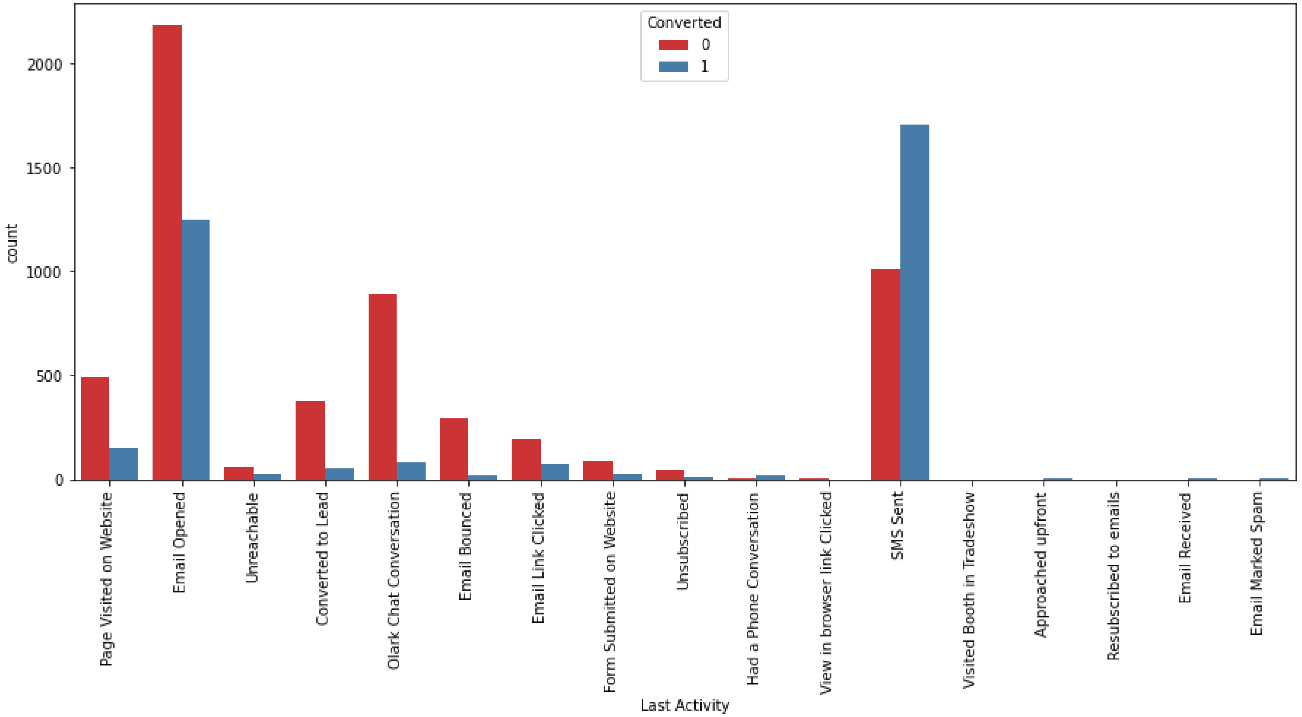

TotalVisits
10
8
6
4
2
0
0
1
Converted

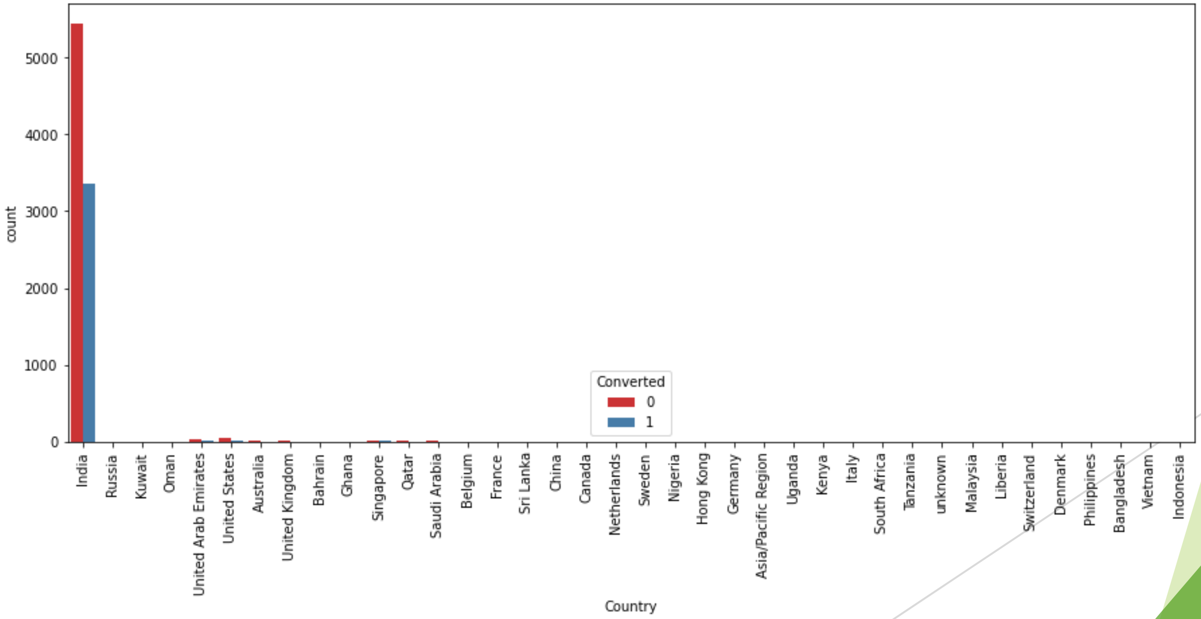
Categorical Variable Relation






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


Converted
0
1
count
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
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


count
5000
4000
3000
2000
1000
0
Converted
0
1
India
Russia
Kuwait
Oman
United Arab Emirates
United States
Australia
United Kingdom
Bahrain
Ghana
Singapore
Qatar
Saudi Arabia
Belgium
France
Sri Lanka
China
Canada
Netherlands
Sweden
Nigeria
Hong Kong
Germany
Asia/Pacific Region
Uganda
Kenya
Italy
South Africa
Tanzania
unknown
Malaysia
Liberia
Switzerland
Denmark
Philippines
Bangladesh
Vietnam
Indonesia
Country

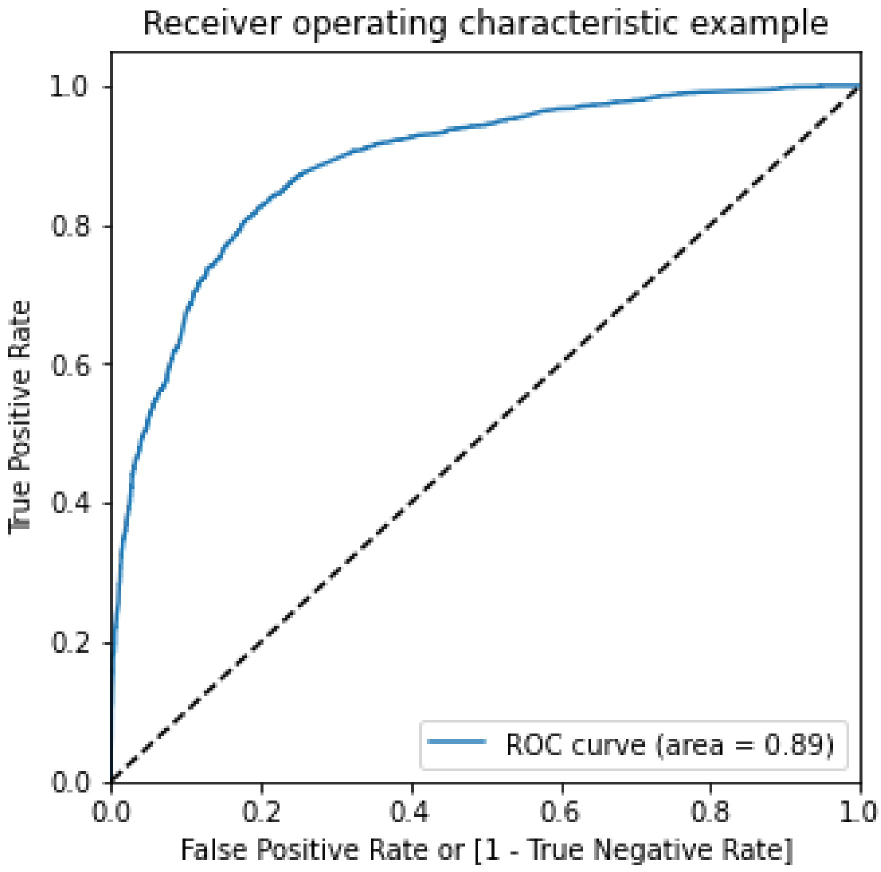
Data Conversation

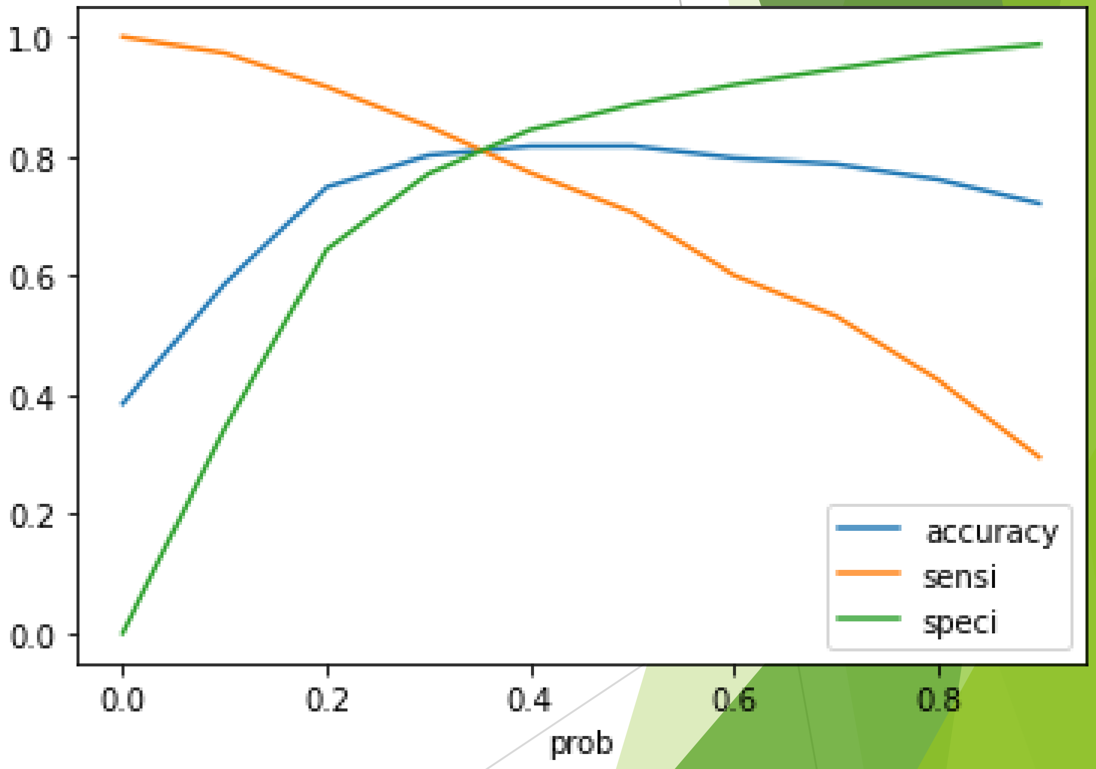
- Numerical Variables are Normalised.
- Dummy Variables are created for object type variables.
- Total Rows for Analysis: 9074
- Total Columns for Analysis: 100

Model Building

- The initial step in regression is to split the data into training and testing sets
- RFE (Recursive Feature Elimination) is used for feature selection, with 20 variables selected as output.
- Since the Pvalues of all variables is 0 and VIF values are low for all the variables, model-9 is our final model. We have 12 variables in our final model.
- Predictions are made and we got sensitivity of 70% and this was mainly because of the cut-off point of 0.5 that we had arbitrarily chosen. Now, this cut-off point had to be optimized in order to get a decent value of sensitivity and for this we will use the ROC curve

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)


0.0
0.2
0.4
0.6
0.8
1.0
prob
accuracy
sensi
speci

Finding of ROC Curve

- The optimal cut-off point is the probability value at which we achieve a balance between sensitivity and specificity.
- Based on the second graph, it can be observed that the optimal cut-off point is around 0.5.

Conclusion

- The variables that have the greatest impact on potential buyers are ranked in descending order: total time spent on the website, total number of visits, lead source (a. Google, b. Direct traffic, c. Organic search, d. Welingak website), and last activity (a. SMS, b. Olark chat conversation).
- X Education has a high chance of success because they meet the criteria that matter the most to potential buyers.
- The company should make calls to lead sources, working professionals, people who spend more time on the website, last activity , SMS sent
- Should not make calls to Do not email customers who have replied as yes, some from the last activity ho negatively respond and with specialization. These may not get converted