Lead Scoring Case Study

Group Members

1. Aman Roy
2. Poonam Thakur

Problem Statement

- X Education sells online courses to industry professionals.
- Although X Education receives a lot of leads, it has an extremely low lead conversion rate. For instance, Let's say they get 100 leads a day, but only 30 of them are actually converted.
- The goal of the business is to find the most promising leads, commonly referred to as "Hot Leads”, in order to increase the efficiency of this process.
- If they are able to locate this group of leads, the lead conversion rate ought to increase since the sales staff will be concentrating more on contacting the potential leads rather than calling each individual.

Business Objectives

- X education wants to know about the most promising leads.
- Their aim is to create a model that detects hot leads for that purpose.
- Deployment of the model for the future use cases that’ll be beneficial for the business in generating more revenue.

Solution Approach

Data cleaning and data manipulation.

1. Checking and handling duplicate data.
2. Checking and handling NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis as well as rows if the no. of missing values are low in a column.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

EDA

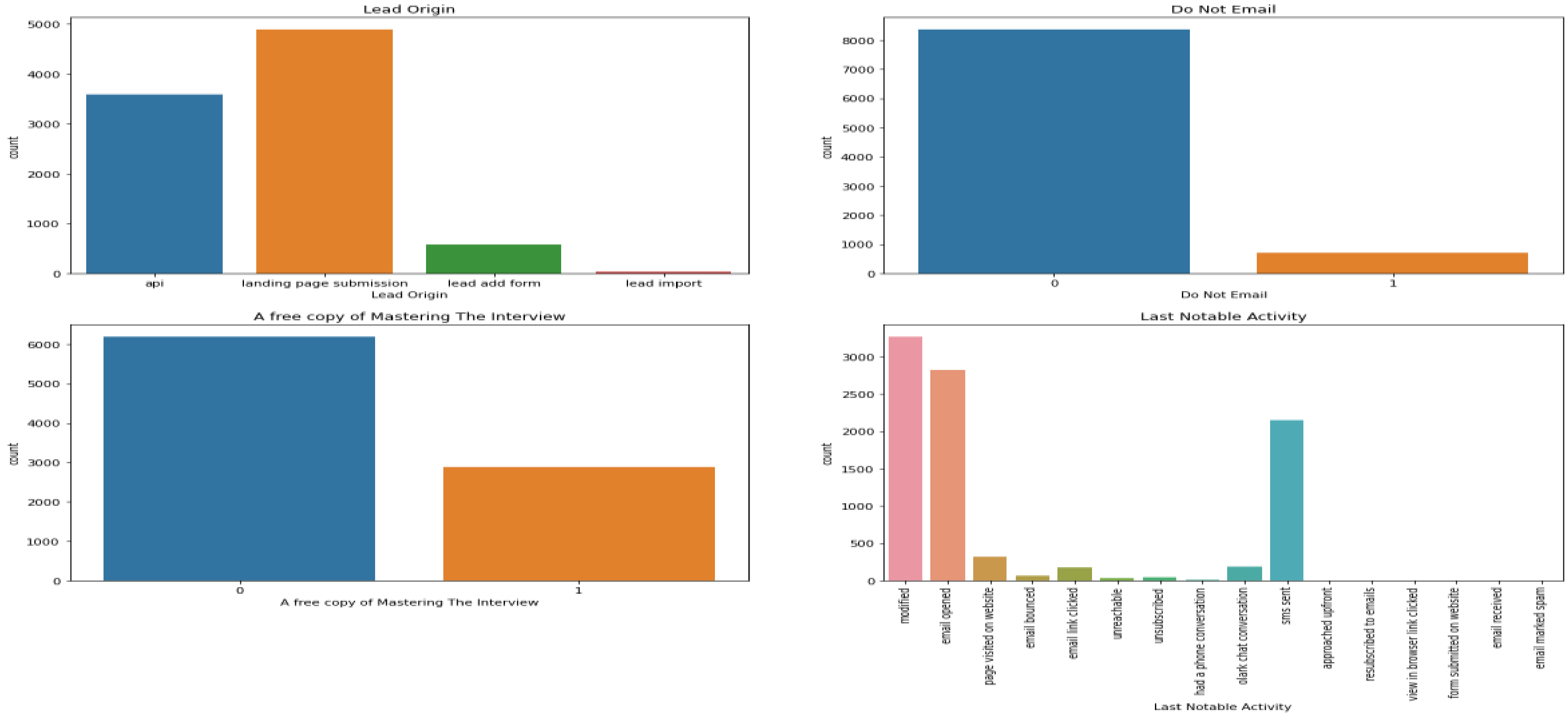
1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- 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

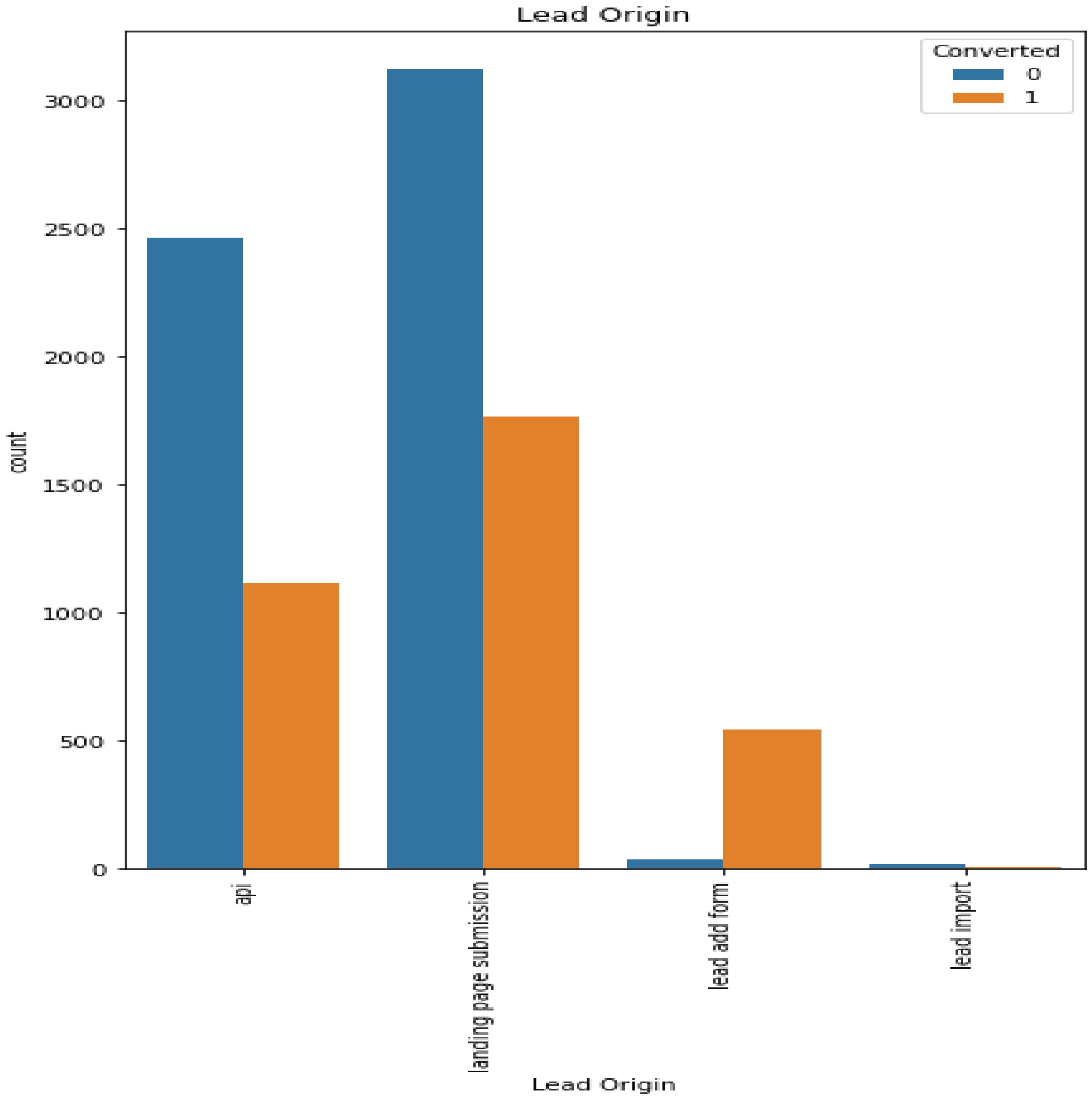
Data Manipulation

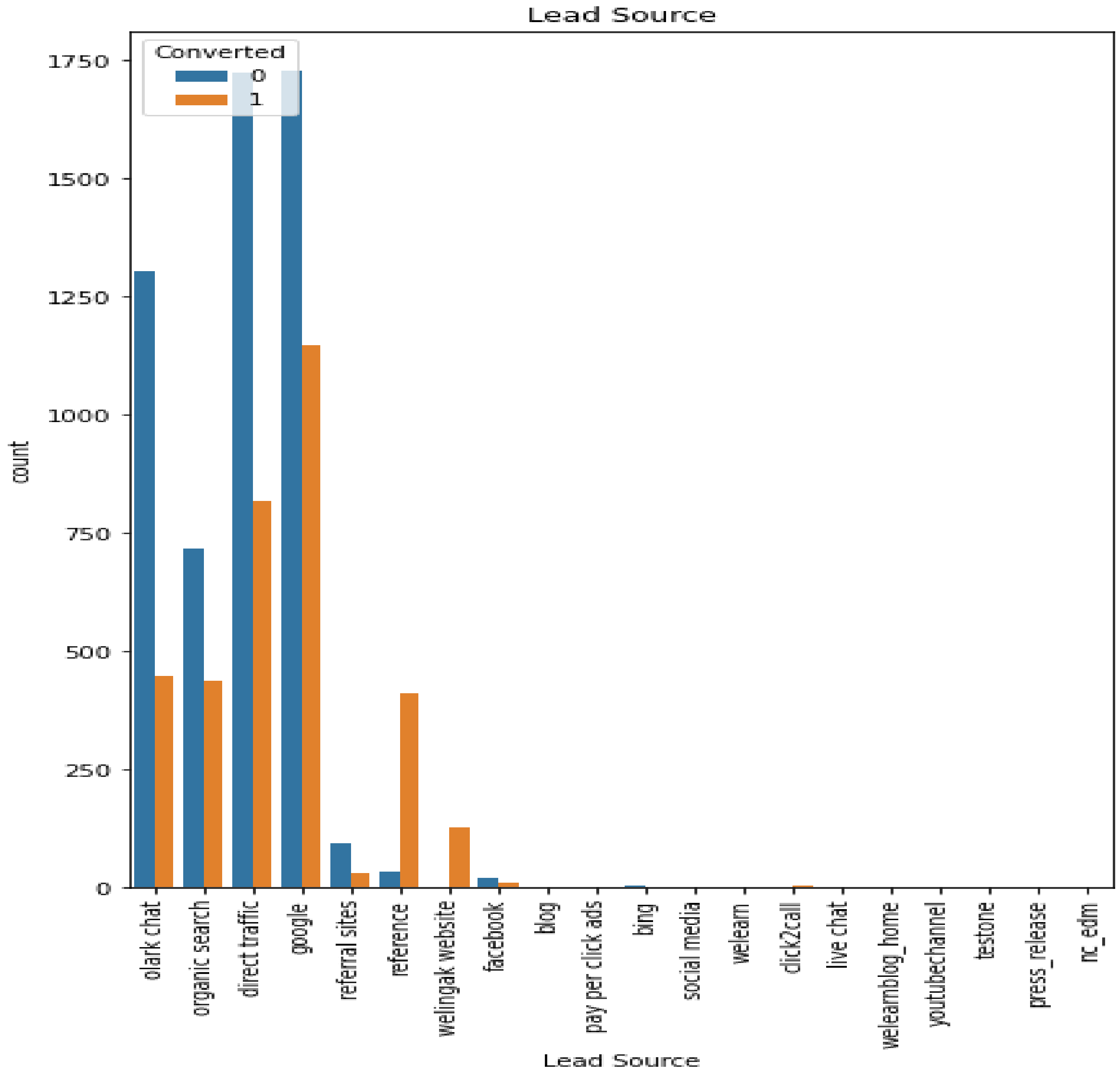
- There are 9240 total rows and 11 total columns.
- Single-value features such as "Magazine," "Receive More Updates About Our Courses," "Update me on Supply," "Chain Content," "Get updates on DM Content," and "I Agree to Pay the Amount Through Cheque", etc. have been removed.
- Removing the "Lead Number" and "Prospect ID" that are not required for the analysis.
- Following a value count check for several object type variables, some features were discovered that lack sufficient variation and have been discarded . These features are: "Do Not Call," "What factors are most important to you when you select a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," etc.
- Removing the columns with missing values in excess of 35%, such as "How did you learn about X Education," "Lead Profile," "Country," etc.

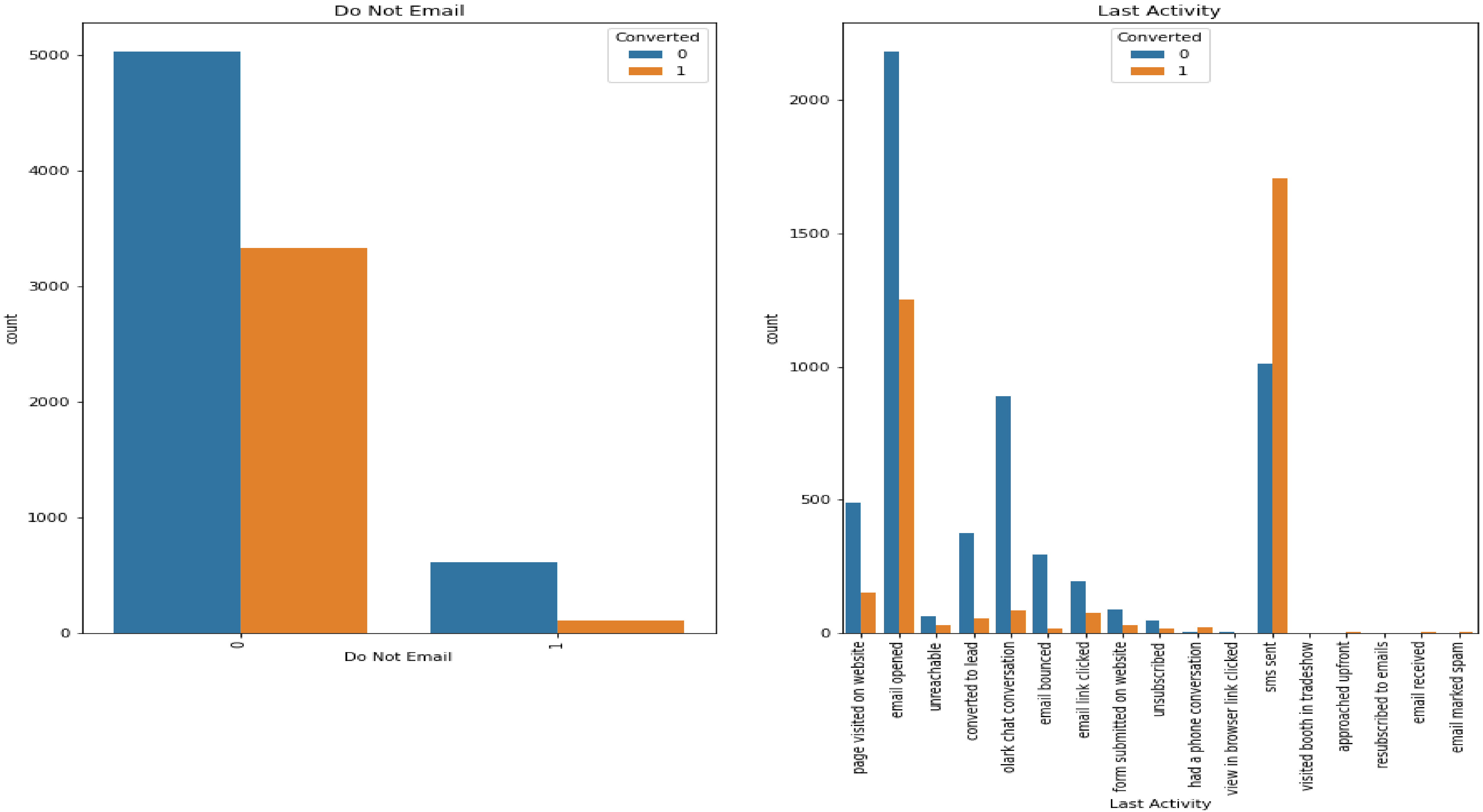
EDA


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin
Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
Do Not Email
A free copy of Mastering The Interview
count
6000
5000
4000
3000
2000
1000
0
0
1
A free copy of Mastering The Interview
Last Notable Activity
count
3000
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

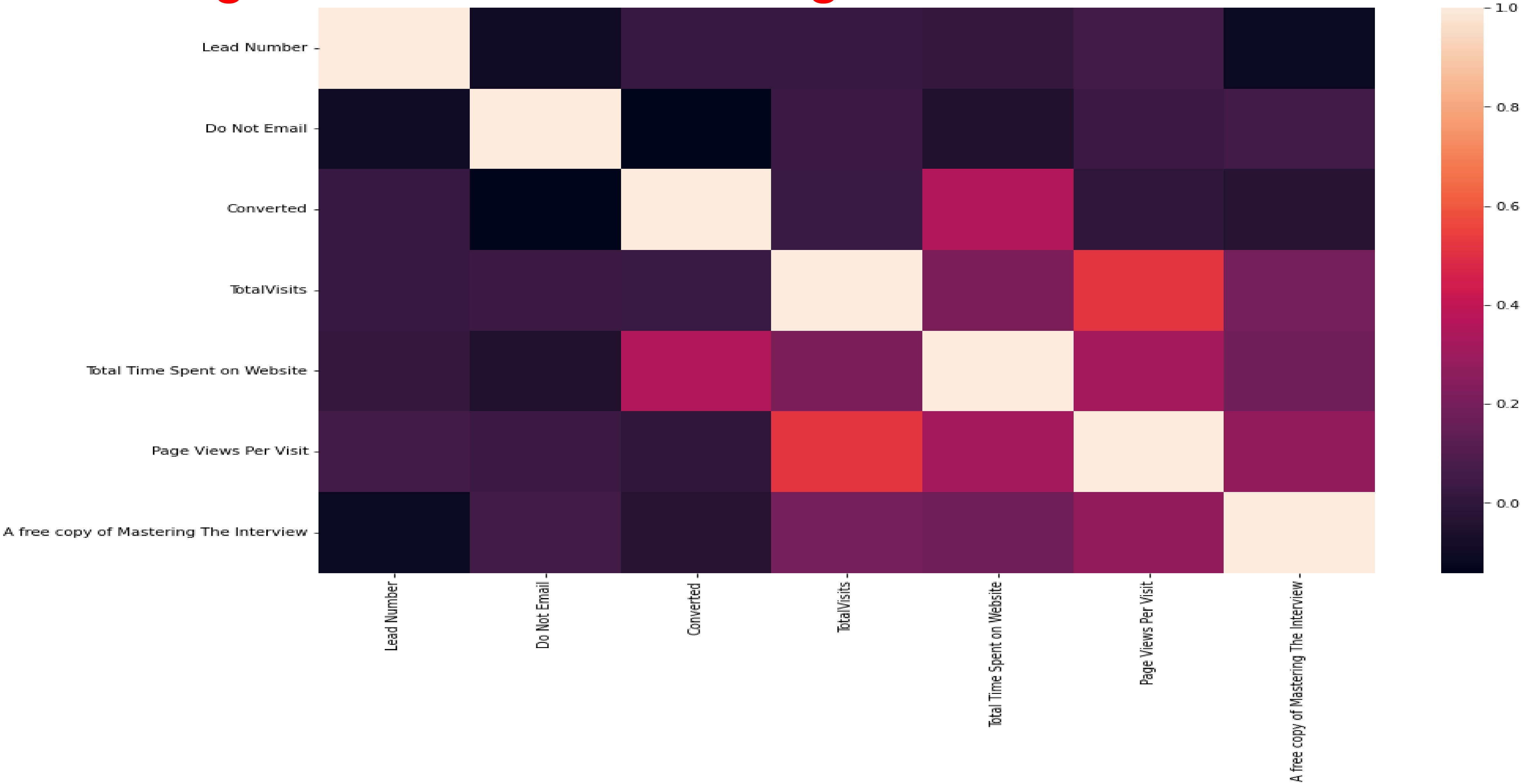
Categorical Variable Relation






Do Not Email
Converted
🟦 0
🟧 1
count
5000
4000
3000
2000
1000
0
0
1
Do Not Email
Last Activity
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

Checking for correlation among numerical variables



Data Conversion

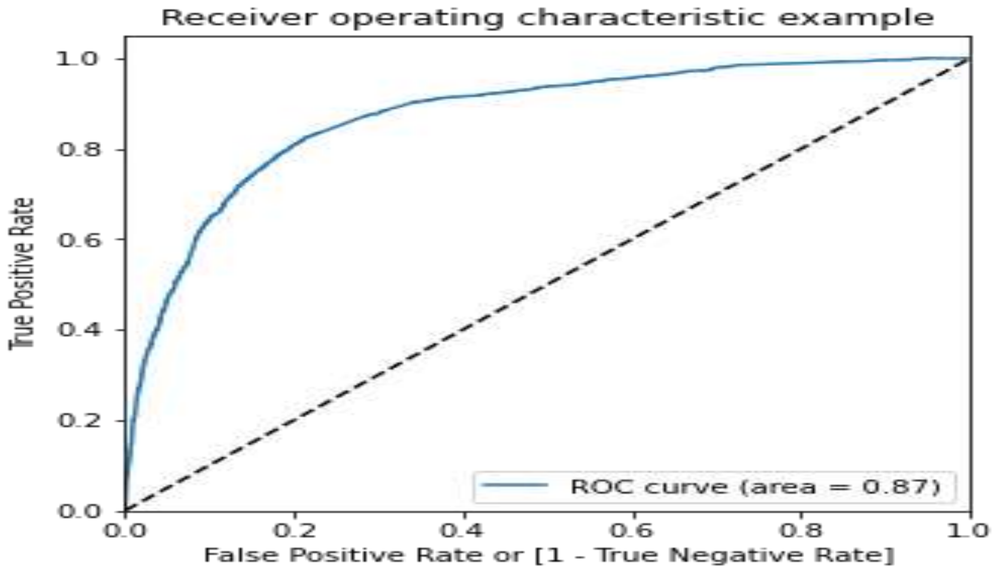
- Numerical Variables are Normalized and outliers are treated.
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9074
- Total Columns for Analysis: 60

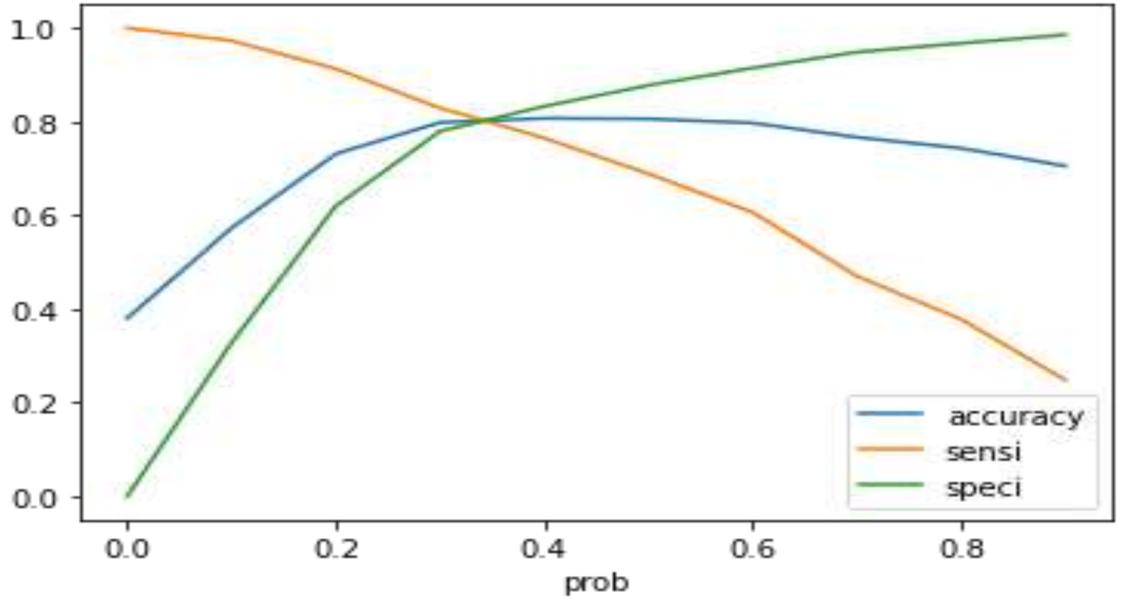
Model Building

- Creating training and test sets from the data.
- A train-test split is the first fundamental stage in the regression process; we have selected a 70:30 split.
- For feature selection, RFE has been used.
- Executing RFE with 15 output variables.
- Creating a model by deleting any variables with a p-value and a vif value greater than 0.05 and 5.
- Predictions on the test data set.81% overall accuracy on Train dataset with a cut off of 0.5.

ROC Curve

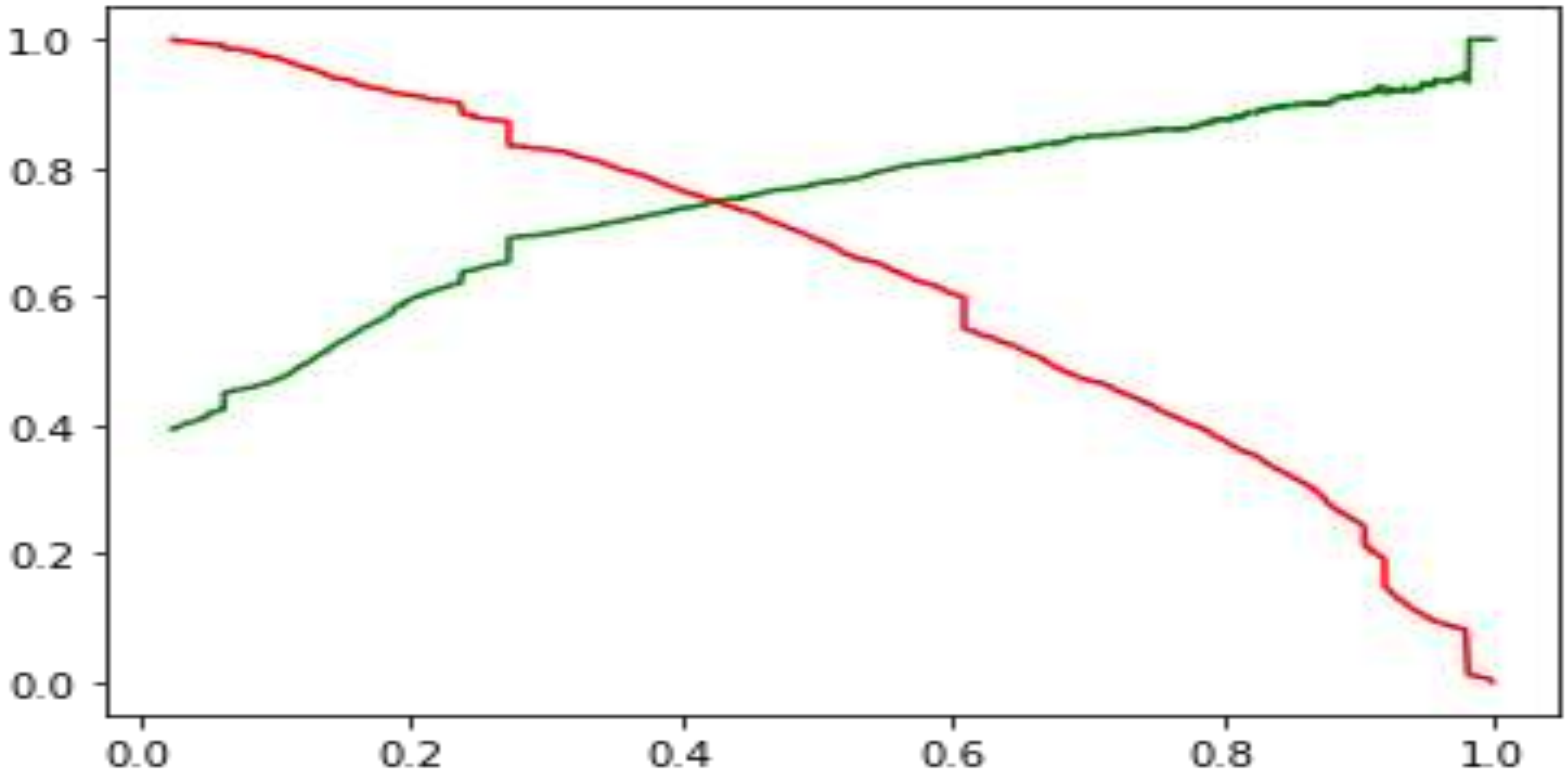
- Finding Optimal Cut off Point.
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3.





Precision and Recall Tradeoff

- The optimal cut off was taken as 0.36 as per the graph and business requirement.



Conclusion

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :
- -Total Time Spent on Website
- -In Lead Origin (Lead Add Form)
- -In Lead Source (Olark Chat)
- -TotalVisits
- -In Last Notable Activity (Email Link Clicked)
- -In Last Notable Activity (Olark Chat Conversation)
- -In Last Notable Activity (Modified)
- -Do Not Email
- -In Last Activity (Olark Chat Conversation)
- -In Last Notable Activity (Email Opened)
- -In Last Notable Activity (Page Visited On Website)
- Keeping these in mind the X Education can flourish as they have a very high
- chance to get almost all the potential buyers to change their mind and buy their courses.