LEAD SCORE CASE STUDY

1. KRITHIKA M
2. SRIDHAR S

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 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:

- Univariate data analysis: value count, distribution of variables, etc..
- Bivariate data analysis: correlation coefficient and pattern between the variables,etc..

Feature Scaling and 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 rows = 37 , total number of columns =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 .

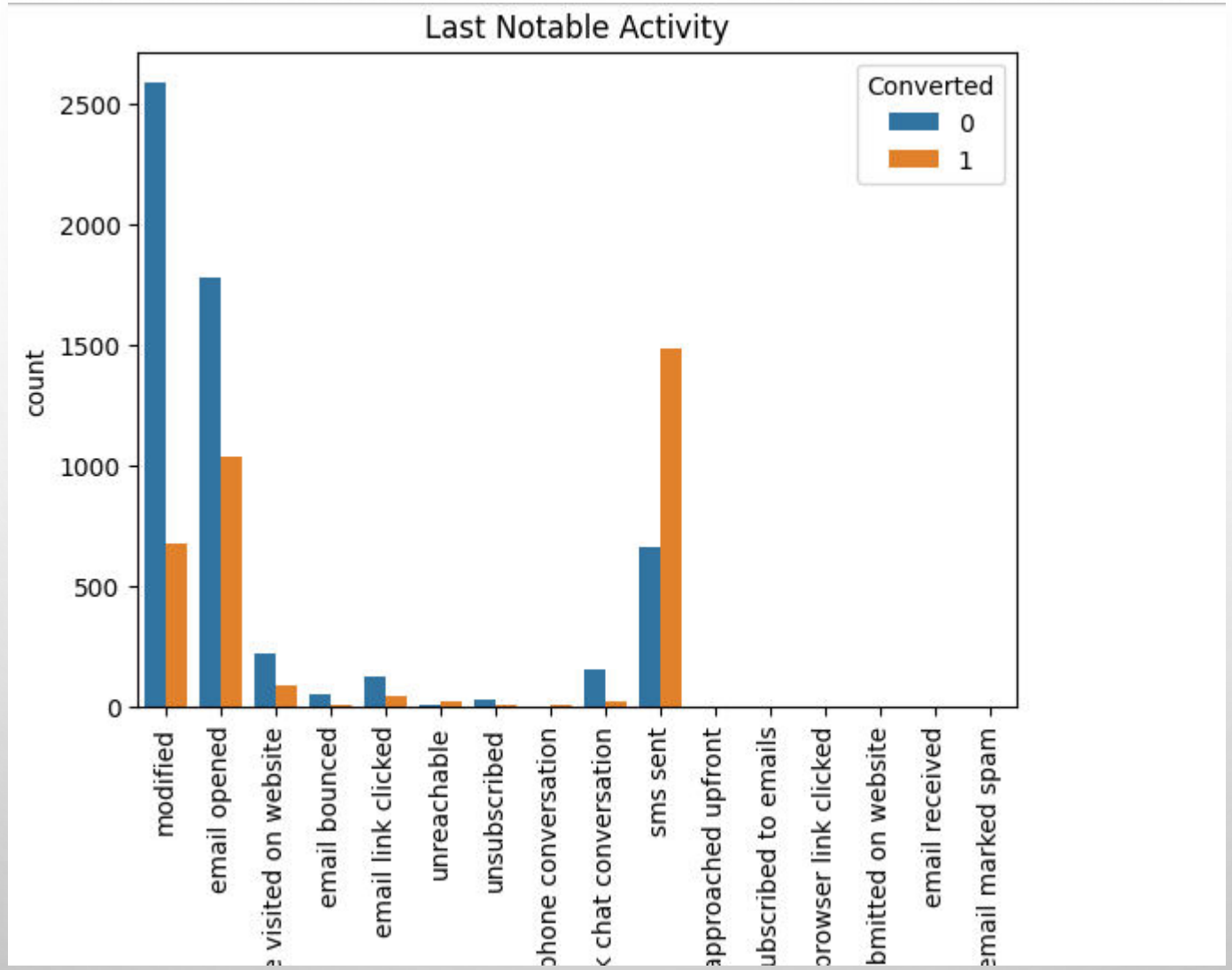
Removing the “Prospect ID ” and “Lead Number ” which is not necessary for the analysis.

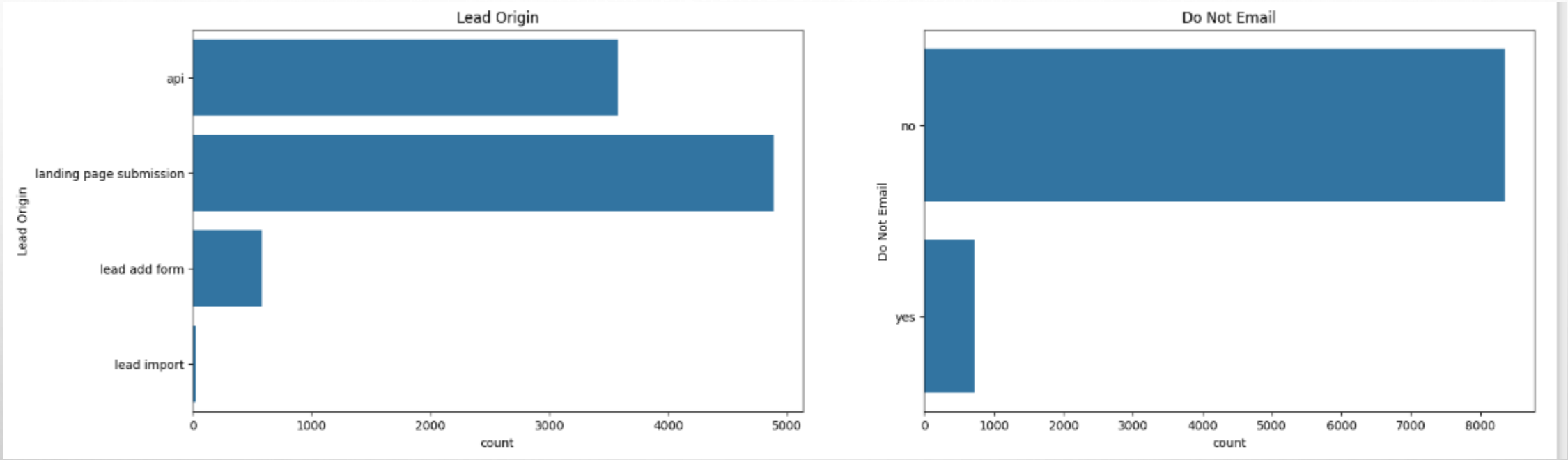
After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are :

“Do not call”, ”what matters most to you in choosing course”, ”Search”, “Newspaper Article”, ”X Education forums”, “Newspaper”, ”Digital Advertisement”, etc..

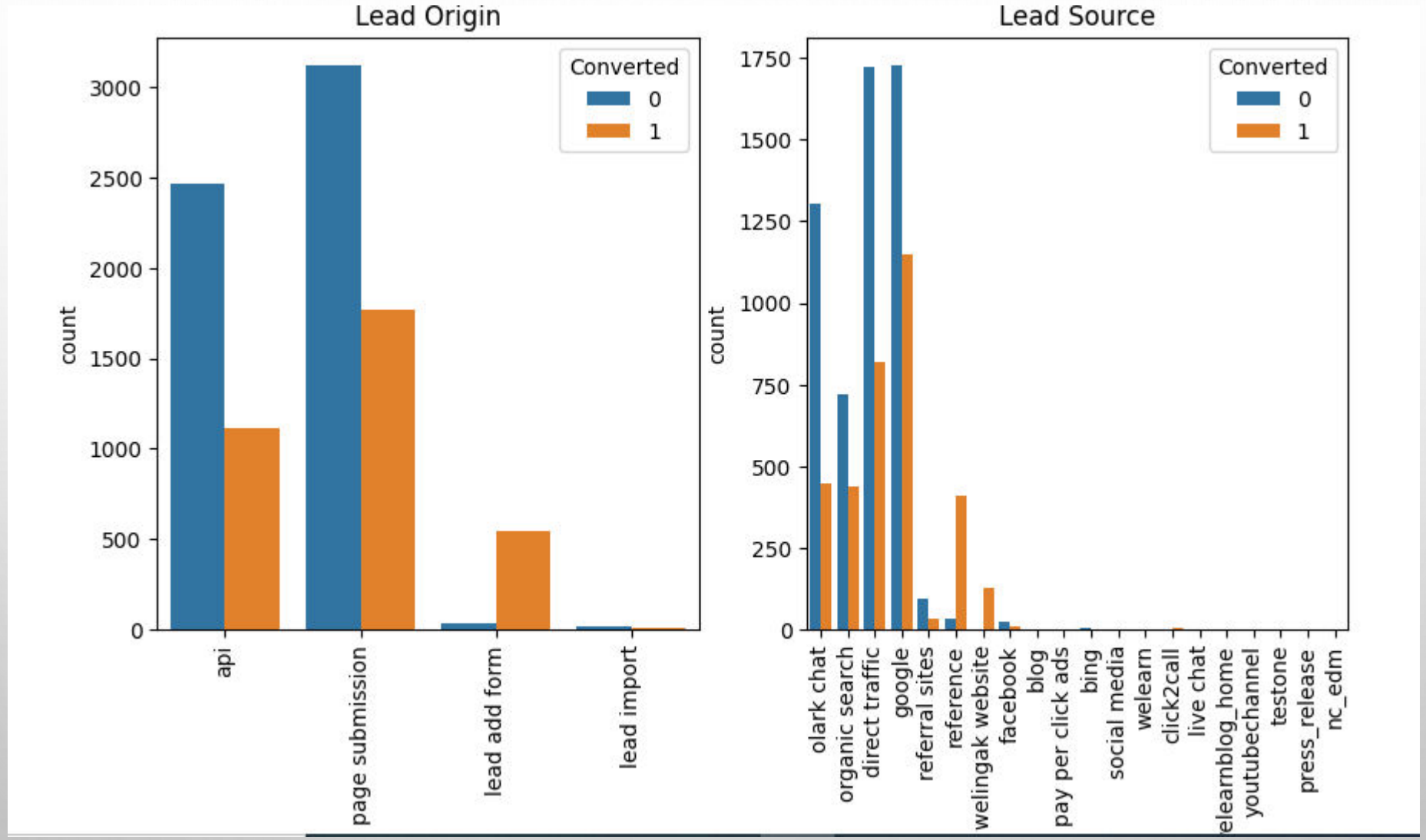
Dropping the columns having more than 35% as missing values such as ‘how did you hear about ‘X Education’ and ‘Lead Profile’.

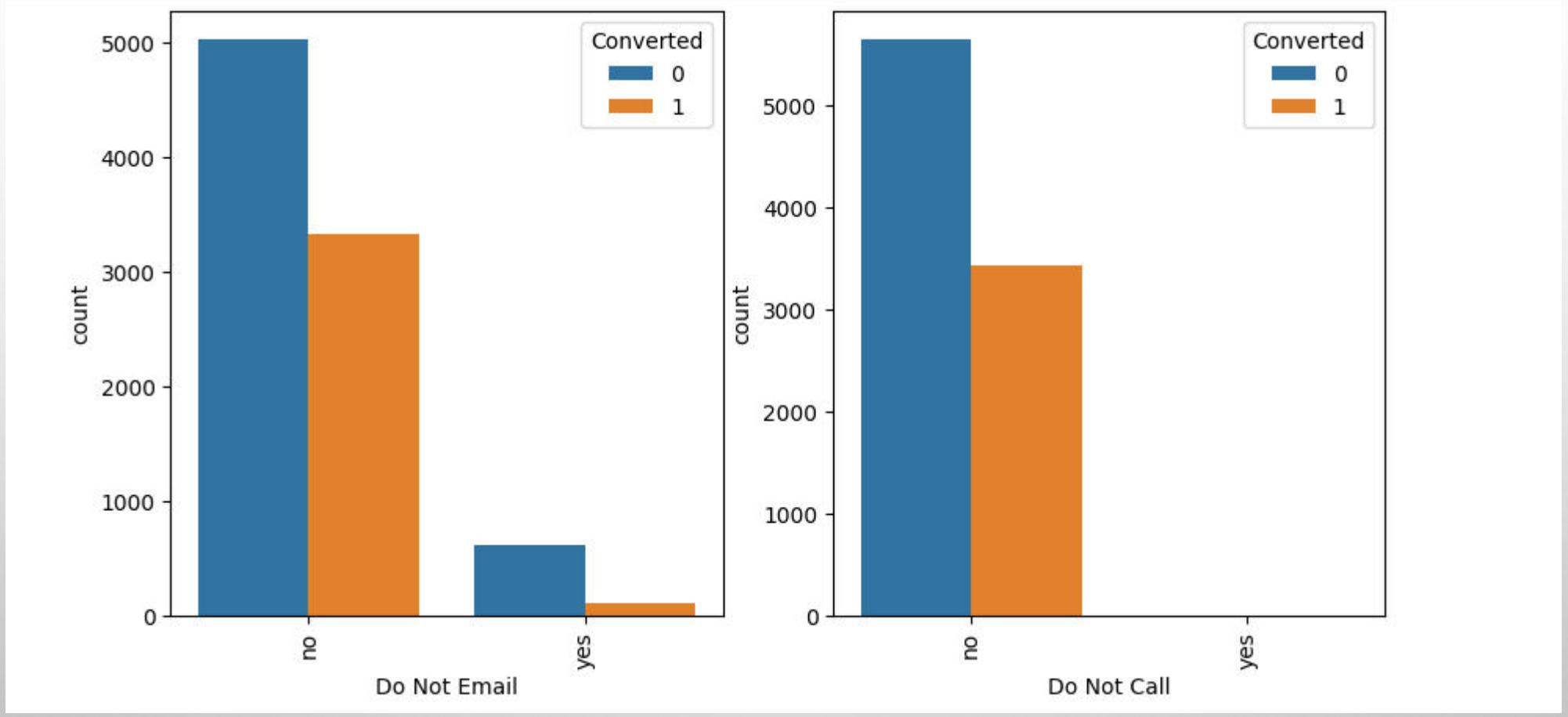
EDA

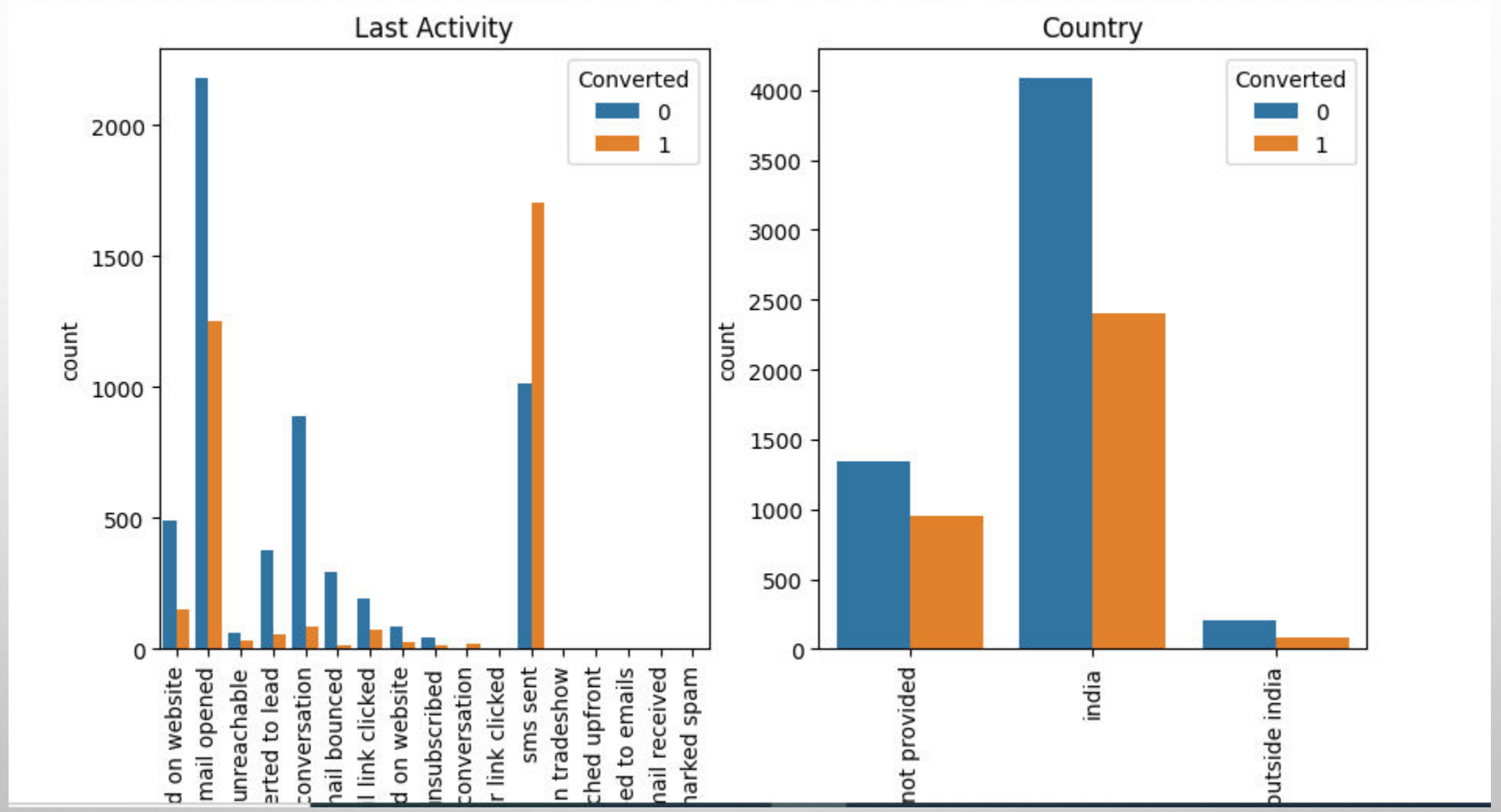

Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
modified
email opened
visited on website
email bounced
email link clicked
unreachable
unsubscribed
phone conversation
chat conversation
sms sent
approached upfront
ubscribed to emails
rowser link clicked
bmitted on website
email received
email marked spam


Lead Origin
api
landing page submission
lead add form
lead import
Lead Origin
0
1000
2000
3000
4000
5000
count
Do Not Email
no
yes
Do Not Email
0
1000
2000
3000
4000
5000
6000
7000
8000
count

CATEGORICAL VARIABLE RELATION




Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Call


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
d on website
mail opened
unreachable
erted to lead
Conversation
hail bounced
l link clicked
d on website
nsubscribed
Conversation
r link clicked
sms sent
n tradeshow
ched upfront
ed to emails
hail received
harked spam
Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india

DATA CONVERSION

Numerical variables are Normalized.

Dummy variables are created for object type variables.

Total rows for Analysis :8792

Total columns for Analysis : 43

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

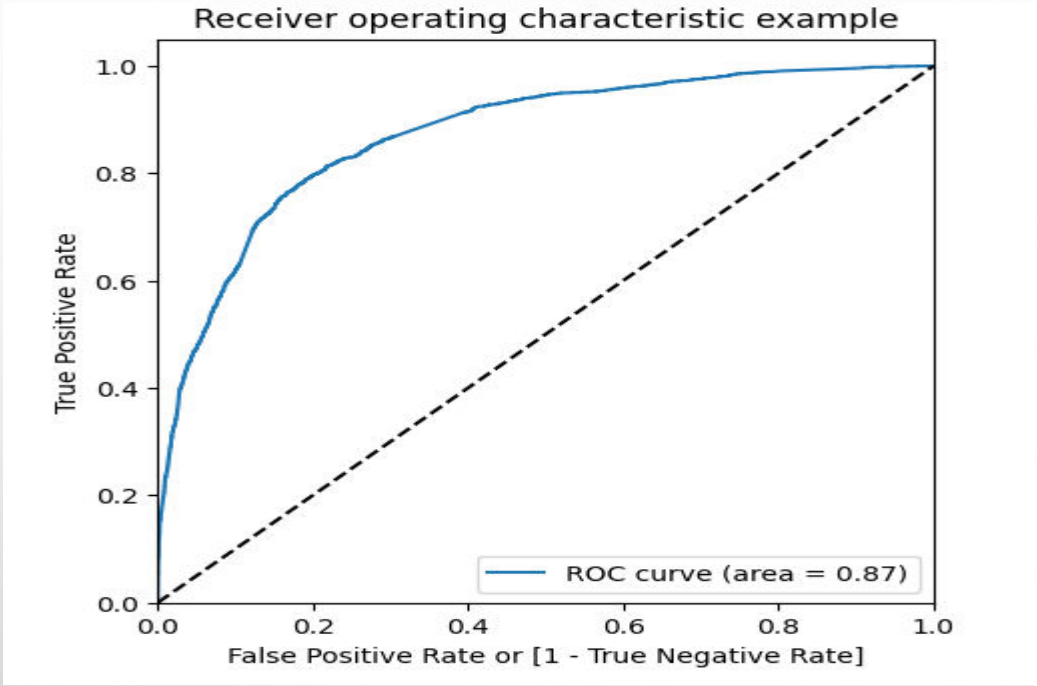
Running RFE with 15 variables as output.

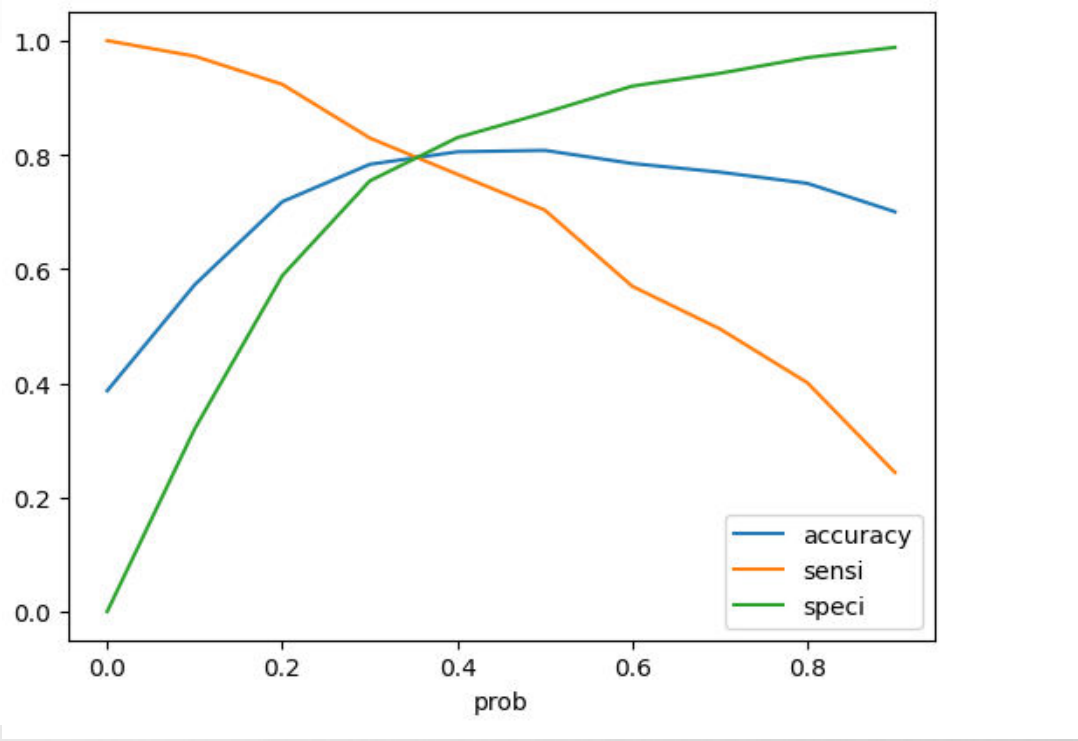
Building model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.

Predictions on test data set.

Overall accuracy 81%

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 .

CONCLUSIONS

It was found that the variables that mattered the most in the potential buyers are (in descending order):

The total time spend on the Website.

Total Number of Visits.

When the lead source was :
Google
Direct Traffic
Organic Search
Welingak website

When the last activity was :
SMS
Olark Chat Conversion

When the lead origin is lead add format.

When the current occupation is as a working professional.