LEAD SCORING CASE STUDY

Group Members
1. Kunal More
2. Tushar Pandit

PROBLEM STATEMENT

- The Education Company named as X sells online courses to industry professionals.
- The company X, although getting a large number of leads, the lead conversion rate at X Education is very poor appx 30%.
- 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 the most promising leads.
- To find out they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.
- Build a logistic regression model to assign a lead score between 0 to 100 to each of the leads which can be used by the client to identify leads.
- A higher score would mean that the lead is hot that is most likely to convert, whereas a low score would mean that the lead is cold and will mostly not get converted.

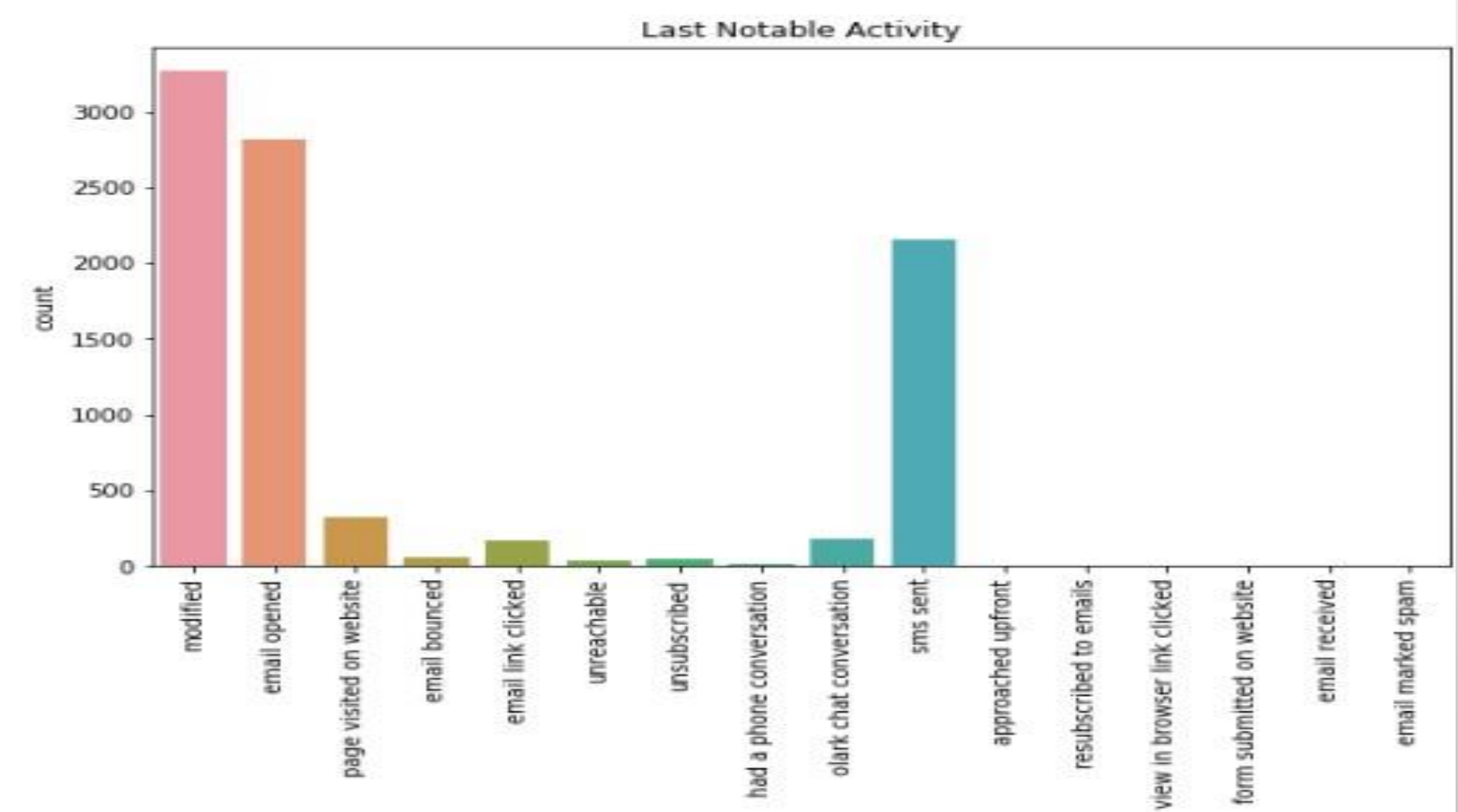
SOLUTION METHODOLOGY

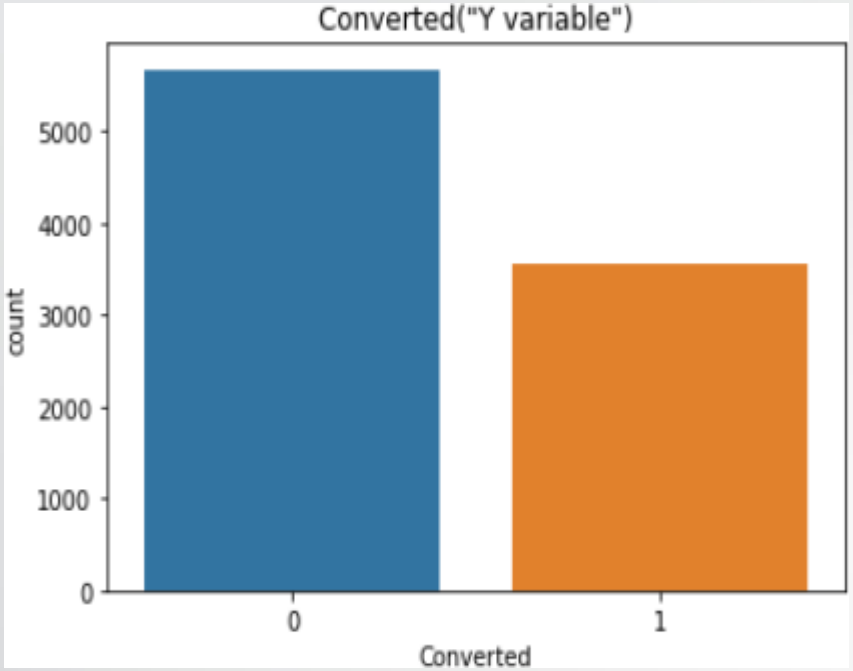
- Data Cleaning and Data Manipulation.
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in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

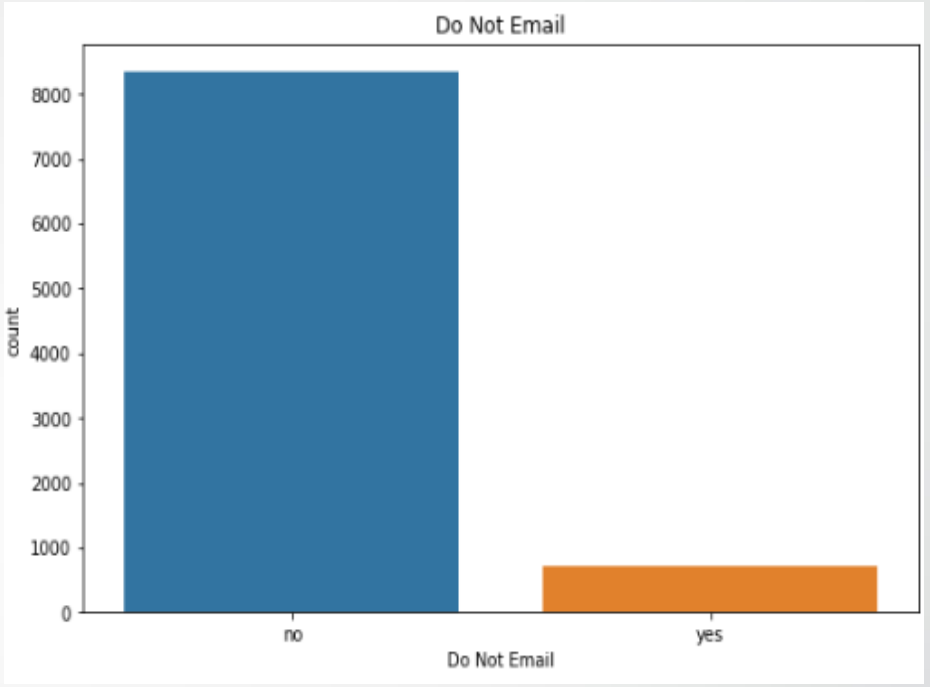
DATA MANIPULATION

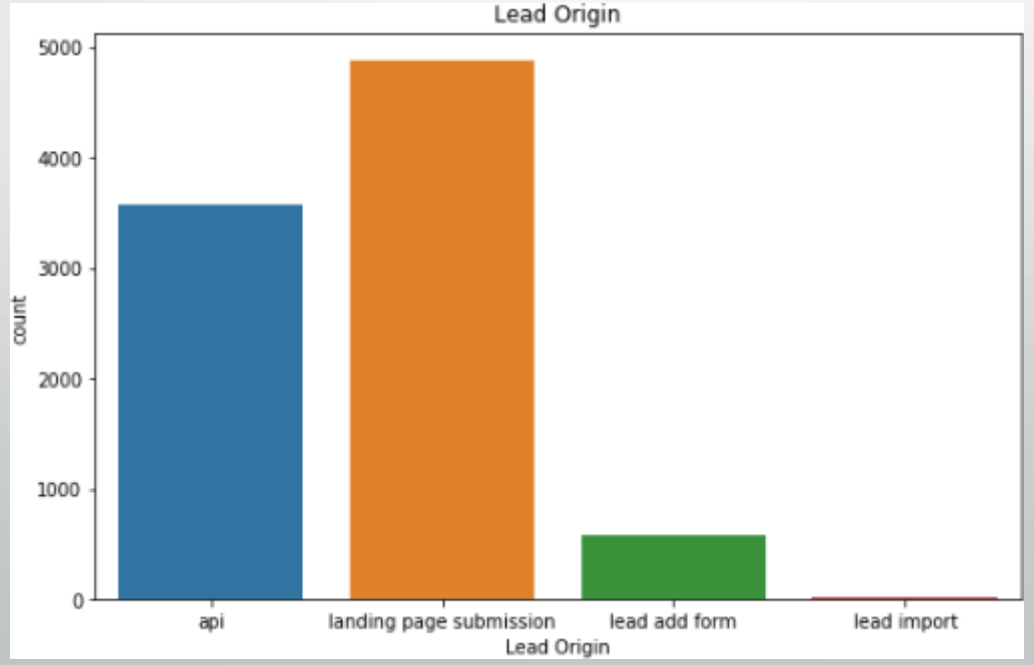
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped. “Prospect ID” and “Lead Number” also have been removed.
- 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”, “Newspaper, “Digital Adventisement” etc
- Dropping the columns having more than 35% as missing value such as ‘Tags’ and ‘Lead Profile’.

EXPLORATORY DATA ANALYSIS

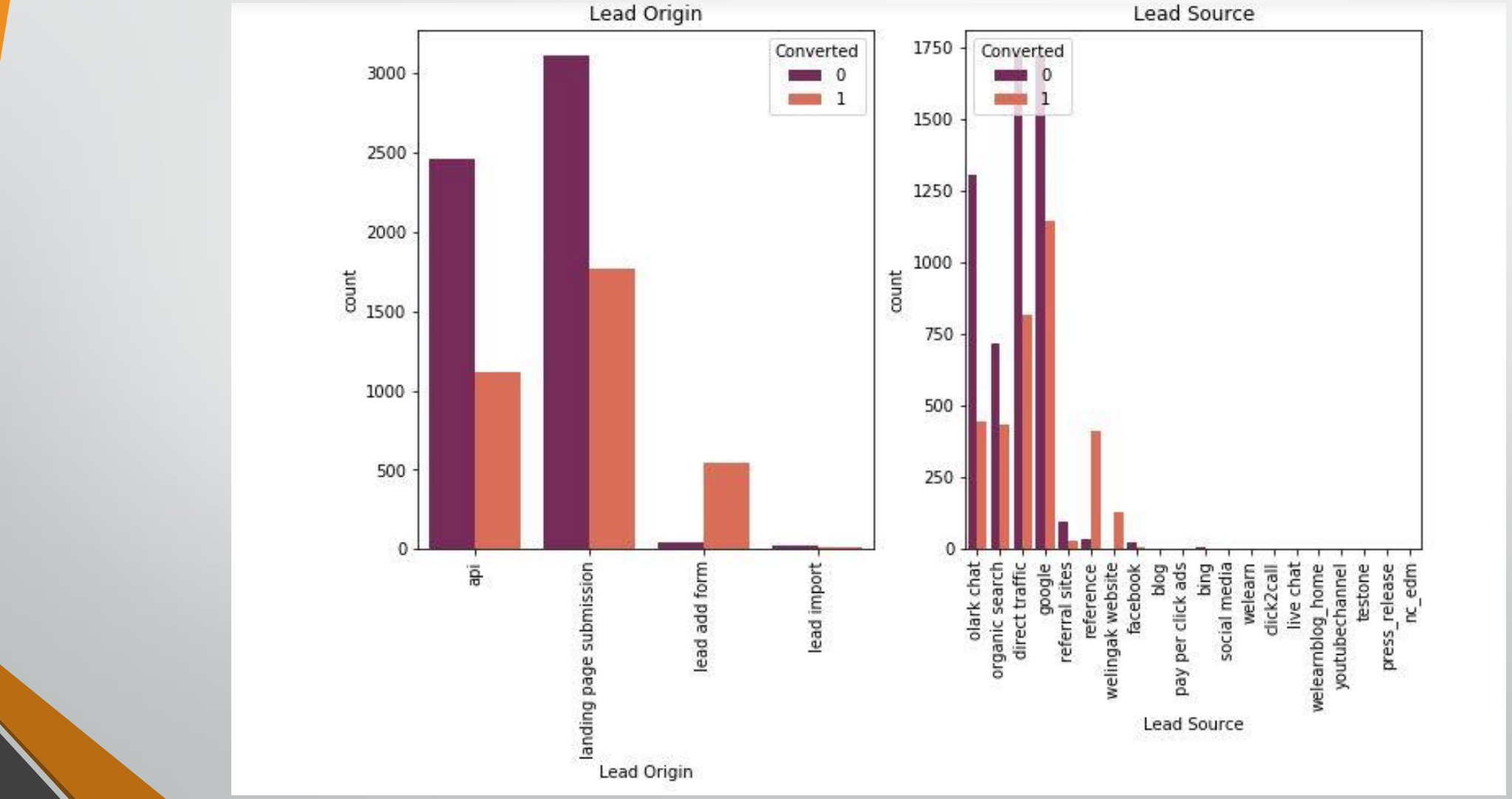


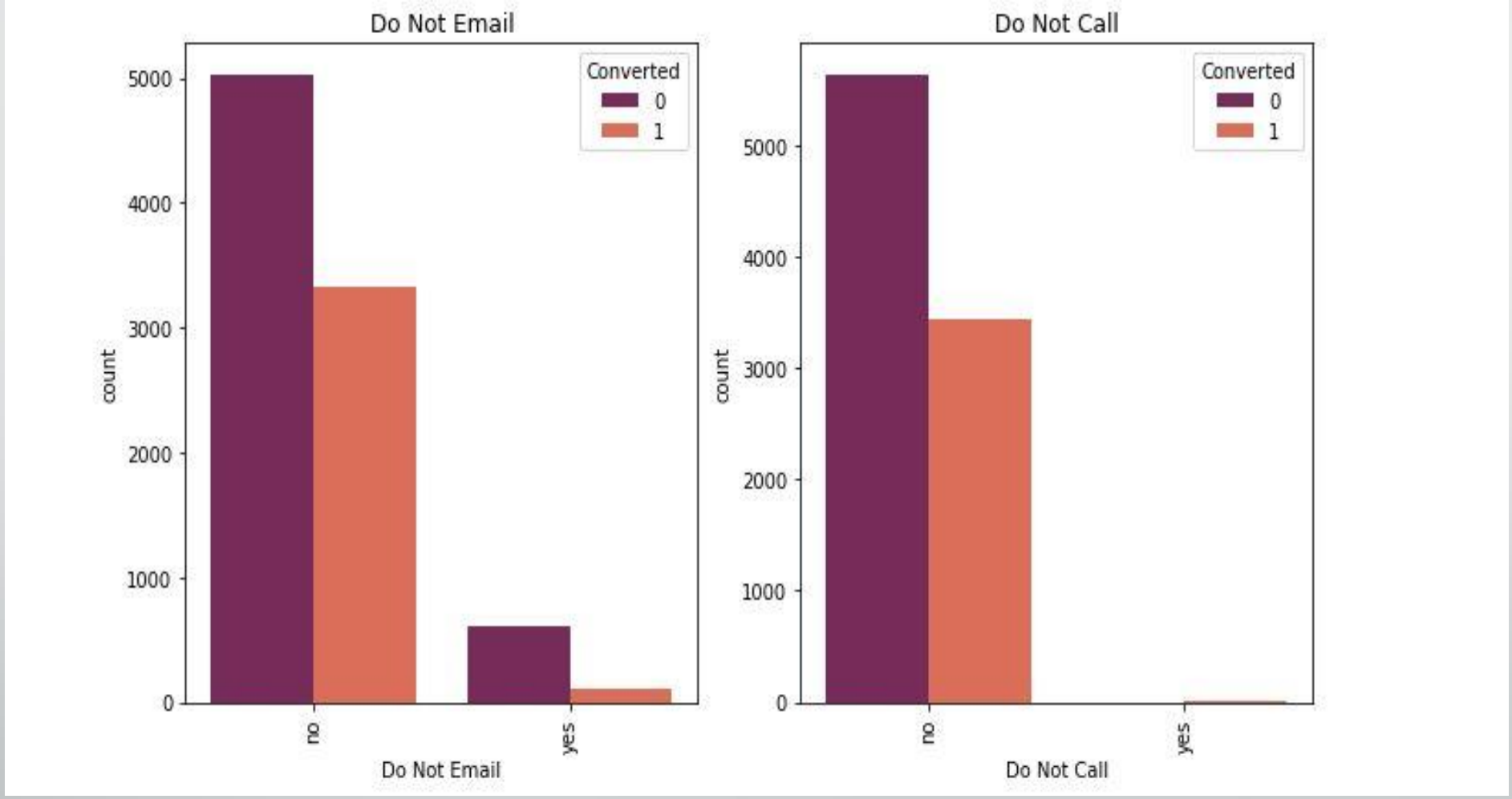

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

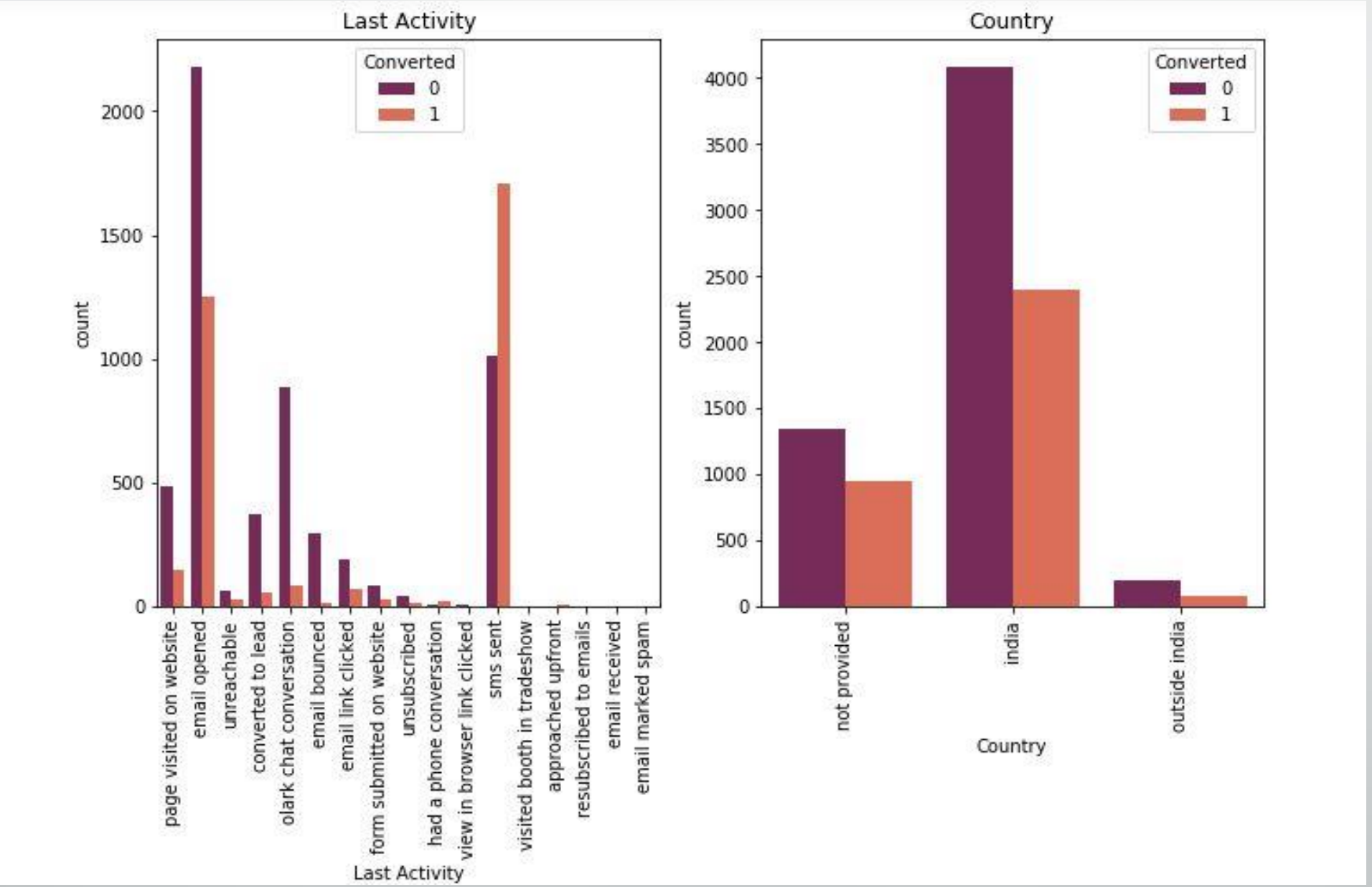

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

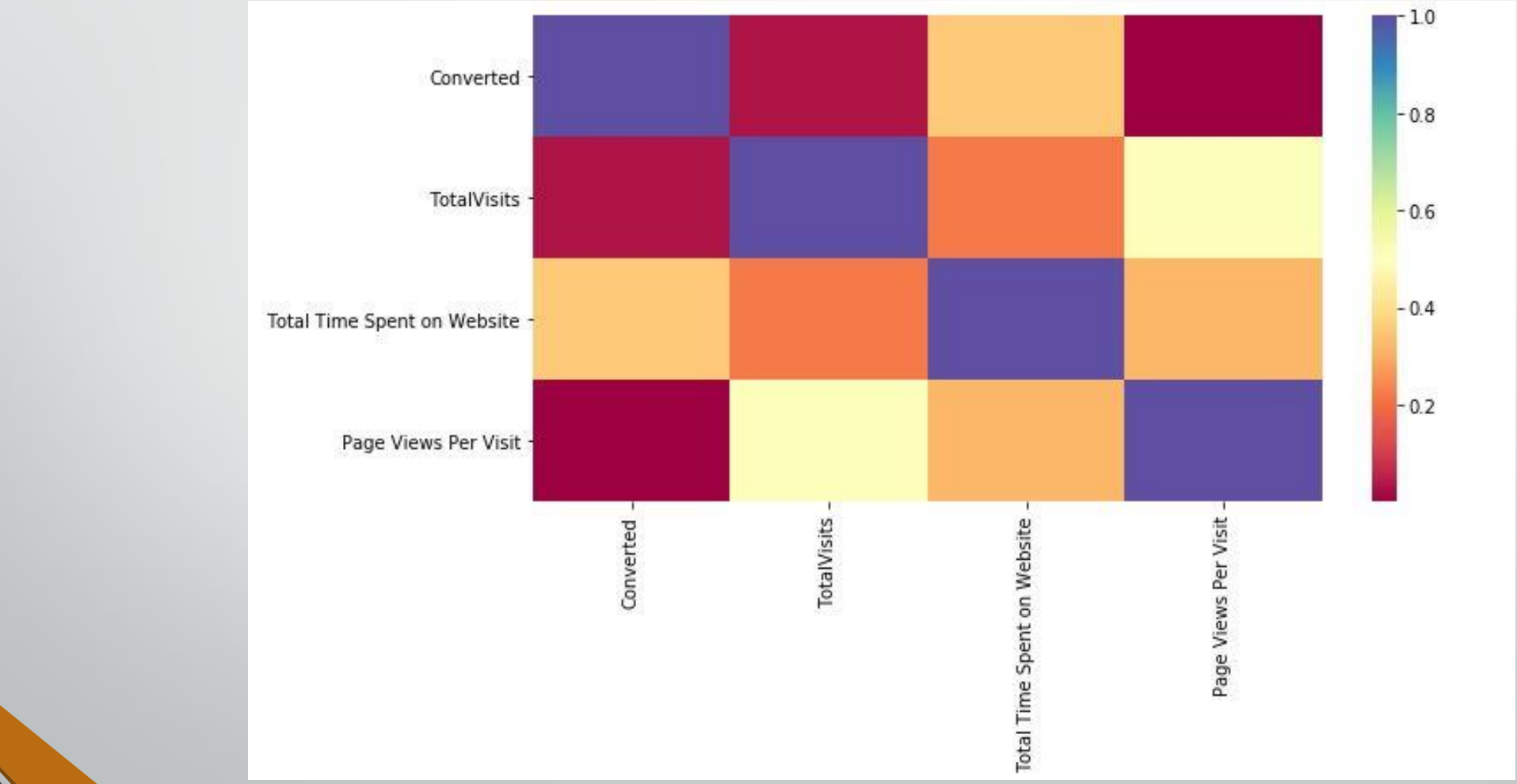
CATEGORICAL VARIABLE RELATION




Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email
Do Not Call
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
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
Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

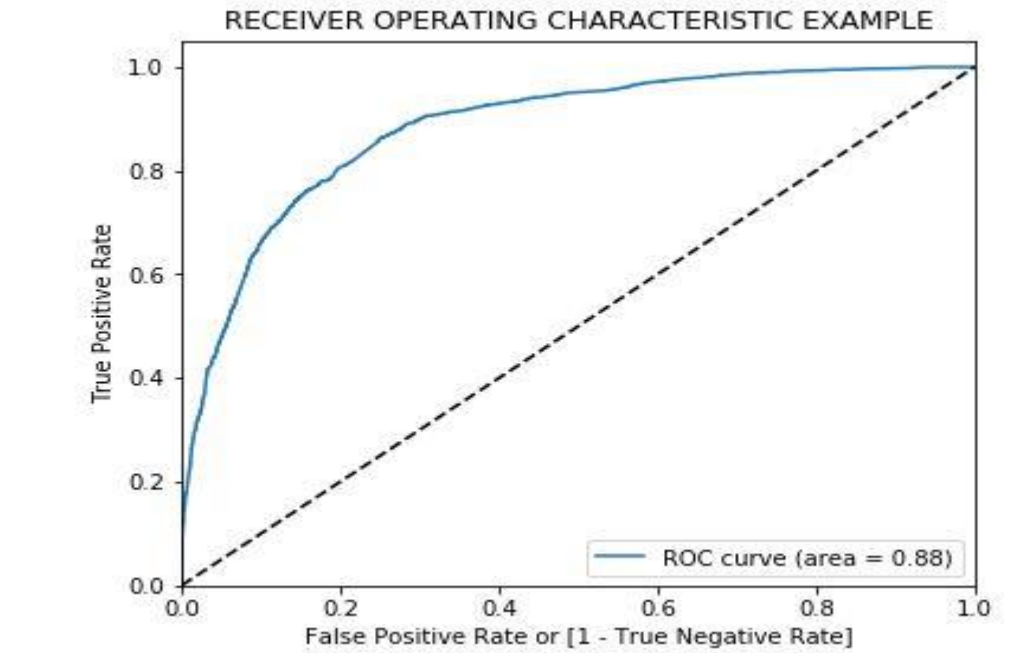
CORRELATION BETWEEN VARIABLES

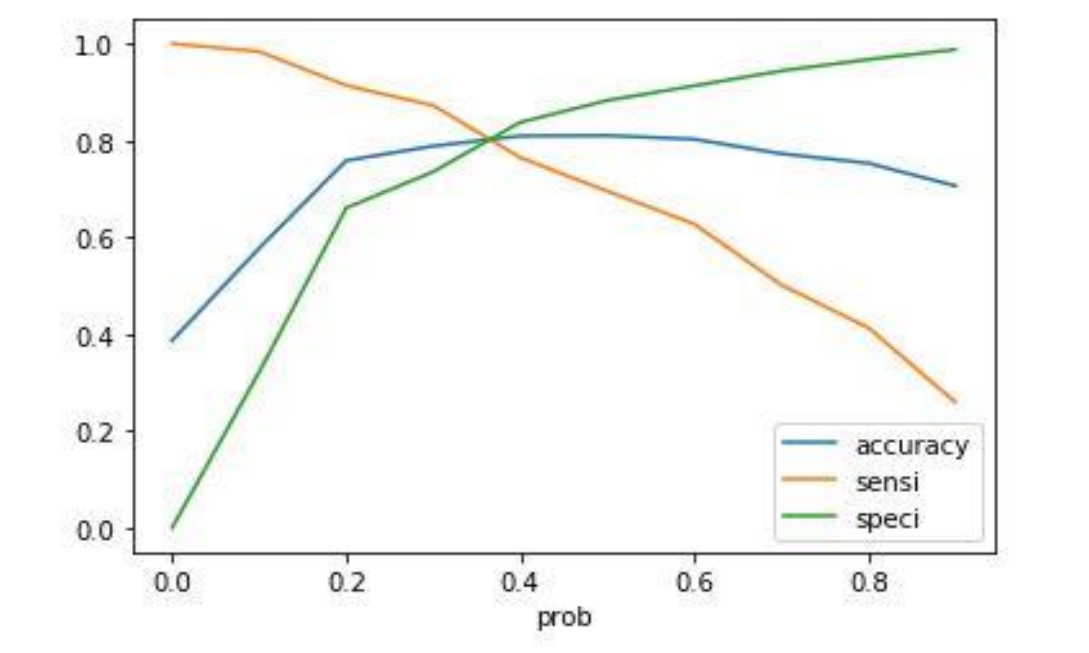


MODEL BUILDING

- The first steps for building the model is to split the Data into Training and Testing Sets
- The ratio chosen here and the standard ratio chosen is 70:30 i.e the data set is split into parts of 70% and 30% for training and testing the model respectively.
- For Feature Selection RFE is used, so first we have 15 variables while we run RFE Running RFE with 15 variables as output
- The model is built continuously by removing the variables whose p-value is greater than 0.5 and VIF value is greater than 5 since it already corresponds to r^2 more than 80%.
- Then Predictions are done on test data set.
- Overall accuracy is about 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.

CONCLUSION

We have found that the variables that mattered the most in the potential buyers are :-

- 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 conversation
- When their current occupation is as a working professional.
- 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.

Apart from these variables proper knowledge should be given to the Company on what has to be done to get the “hot-leads” which were required.