LEAD SCORE CASE STUDY

Group members

- Hemanth Kumar Endla
- Tejas Katre
- Priyadarshini swain

PROBLEM STATEMENT:

- Education offers online courses for purchase for business professionals.
- Although education receives a lot of leads, it has an extremely low lead conversion rate. For instance, only approximately 30 of 100 leads they could gather in a day might actually be 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.
- The lead conversion rate should increase if they are successful in locating this group of leads because the sales staff will be spending more time speaking with potential leads rather than calling everyone.

Business Objective

- Education is interested in the most promising leads.
- They want to develop a model that detects the hot leads for that.
- Deployment of the model for the future use

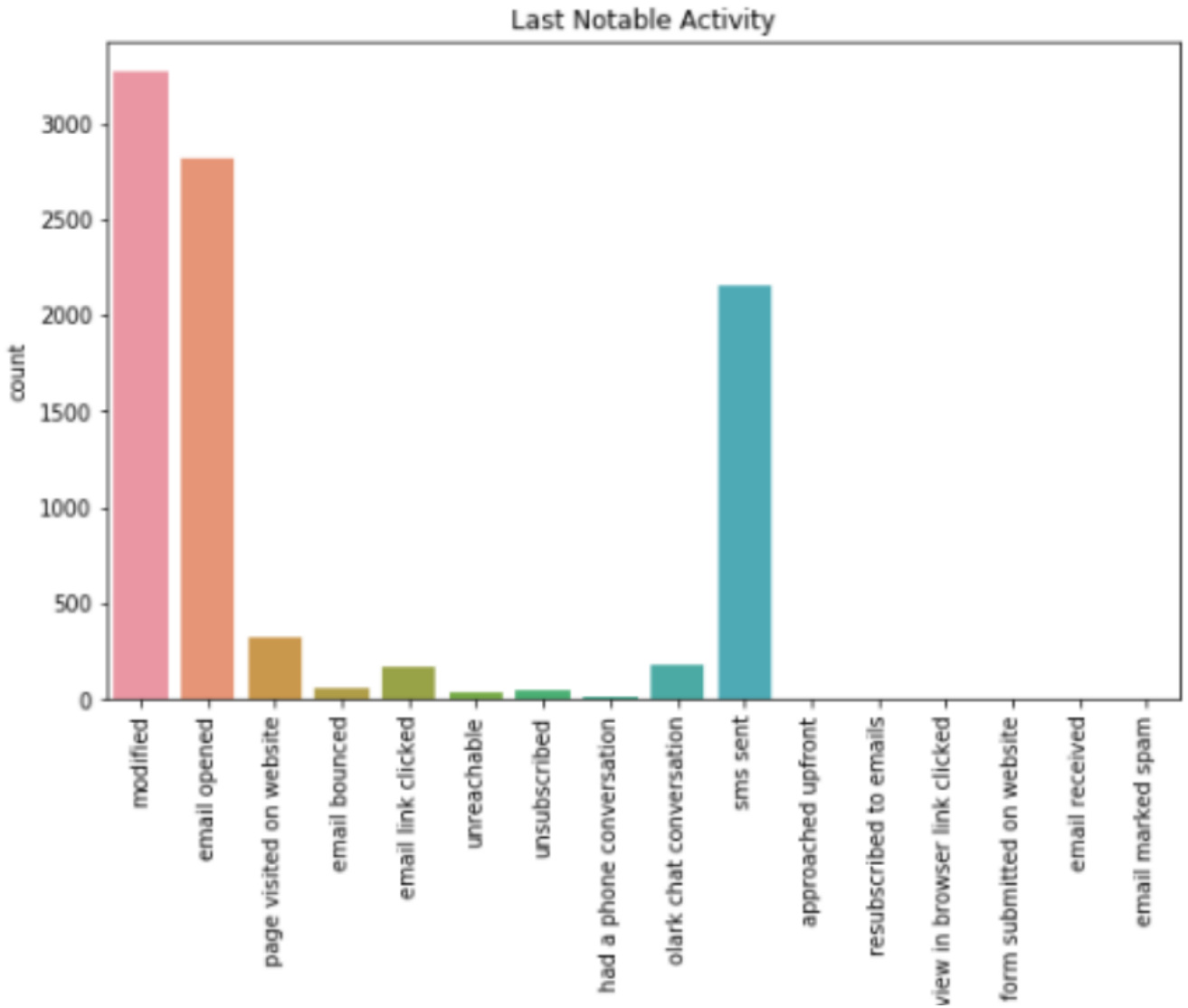
SOLUTION METHODOLOGY

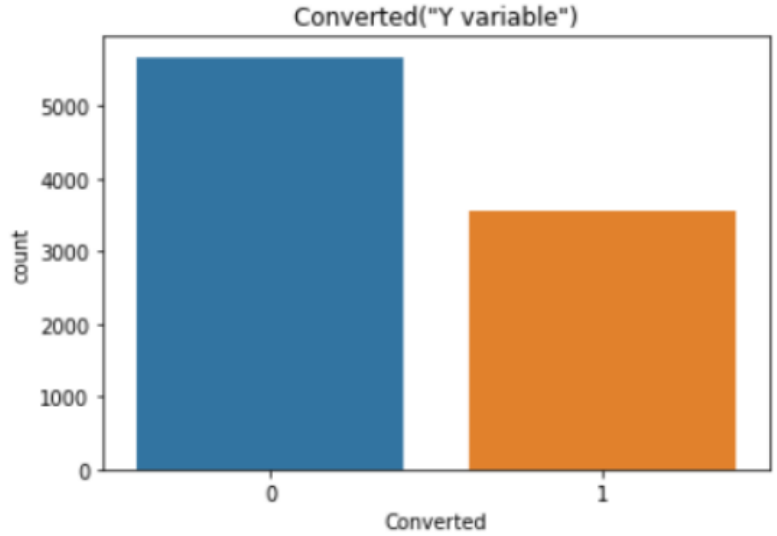
- Data cleaning and data manipulation.
  - Inspect and deal with duplicate data.
  - Inspect and handle missing and NA values.
  - Remove columns from the analysis if they have a significant number of missing values.
  - If necessary, value impugnation.
  - Examine and deal with data outliers.
- EDA
  - Analysis of univariate data: value count, variable distribution, etc.
  - Bivariate data analysis: patterns between the variables and correlation coefficients, etc.
- Scaling Feature & Dummy the data's variables and encoding.
- Logistic regression is a classification technique that is used to create models and make predictions.
- Model presentation
- Model validation
- Conclusions, and recommendations

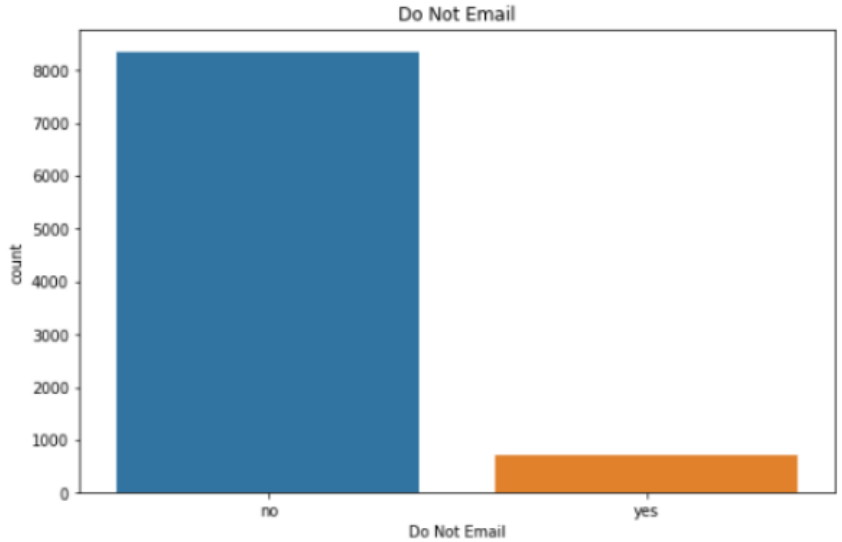
DATA MANIPULATION

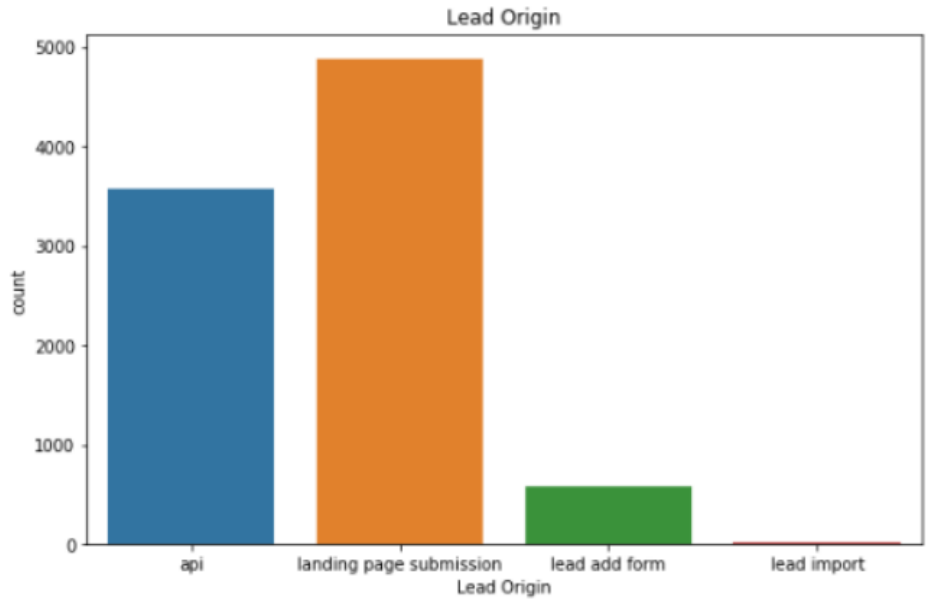
- There are 37 total rows and 9240 total columns.
- Features have a single purpose, such as "Magazine," "Receive More Updates About Our Courses," and "Update me on Supply.“
- The phrases "I agree to pay the sum by check," "I want updates on DM content," and others have been removed.
- Eliminating the "Prospect ID" and "Lead Number" fields that are superfluous to the investigation.
- The features that we have omitted include "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," etc. after examining the value counts for some of the object type variables.
- Dropping the sections like "How did you hear about X Education" and "Lead Profile" that have more than 35% of their values missing

EDA

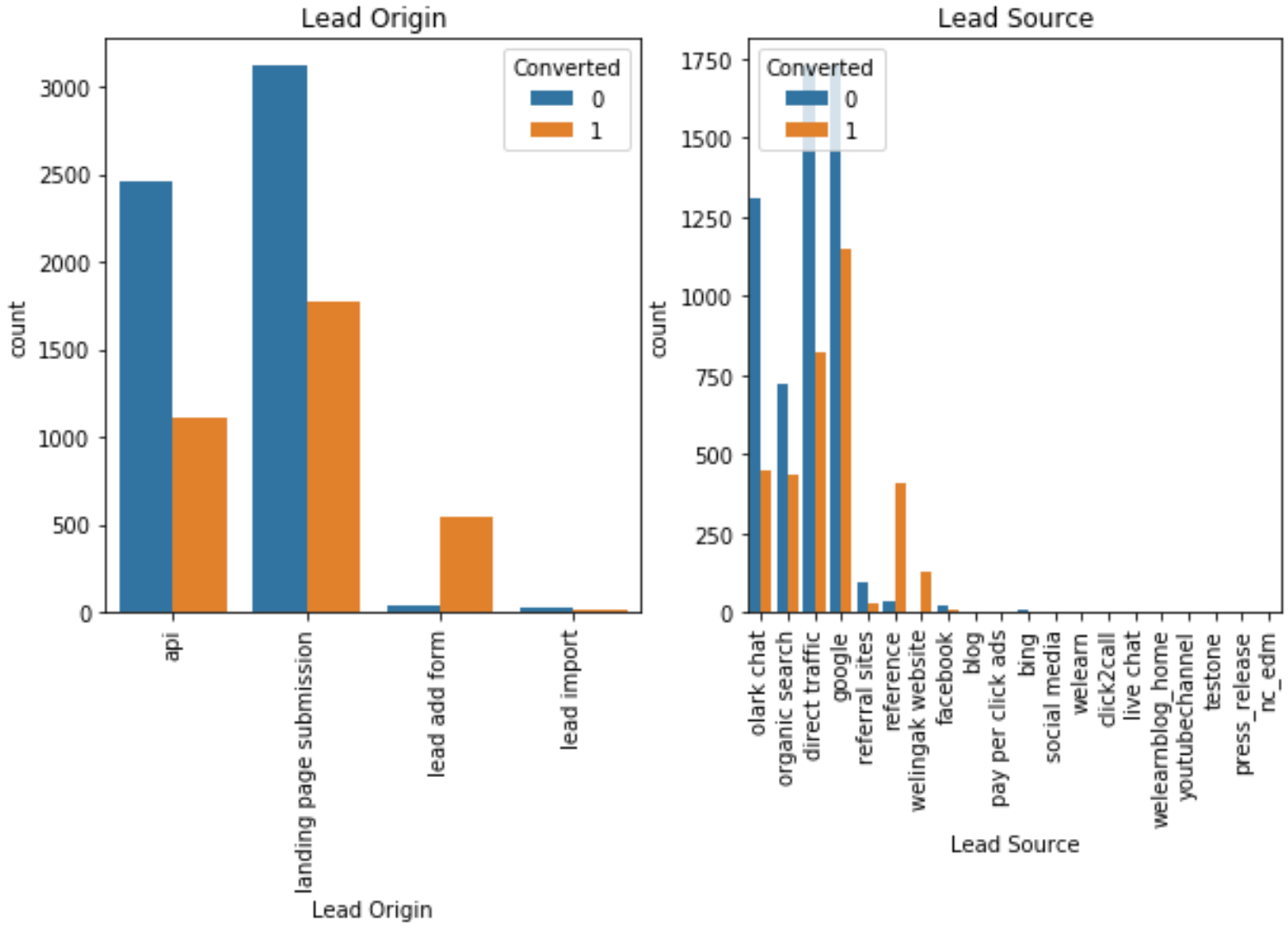

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam

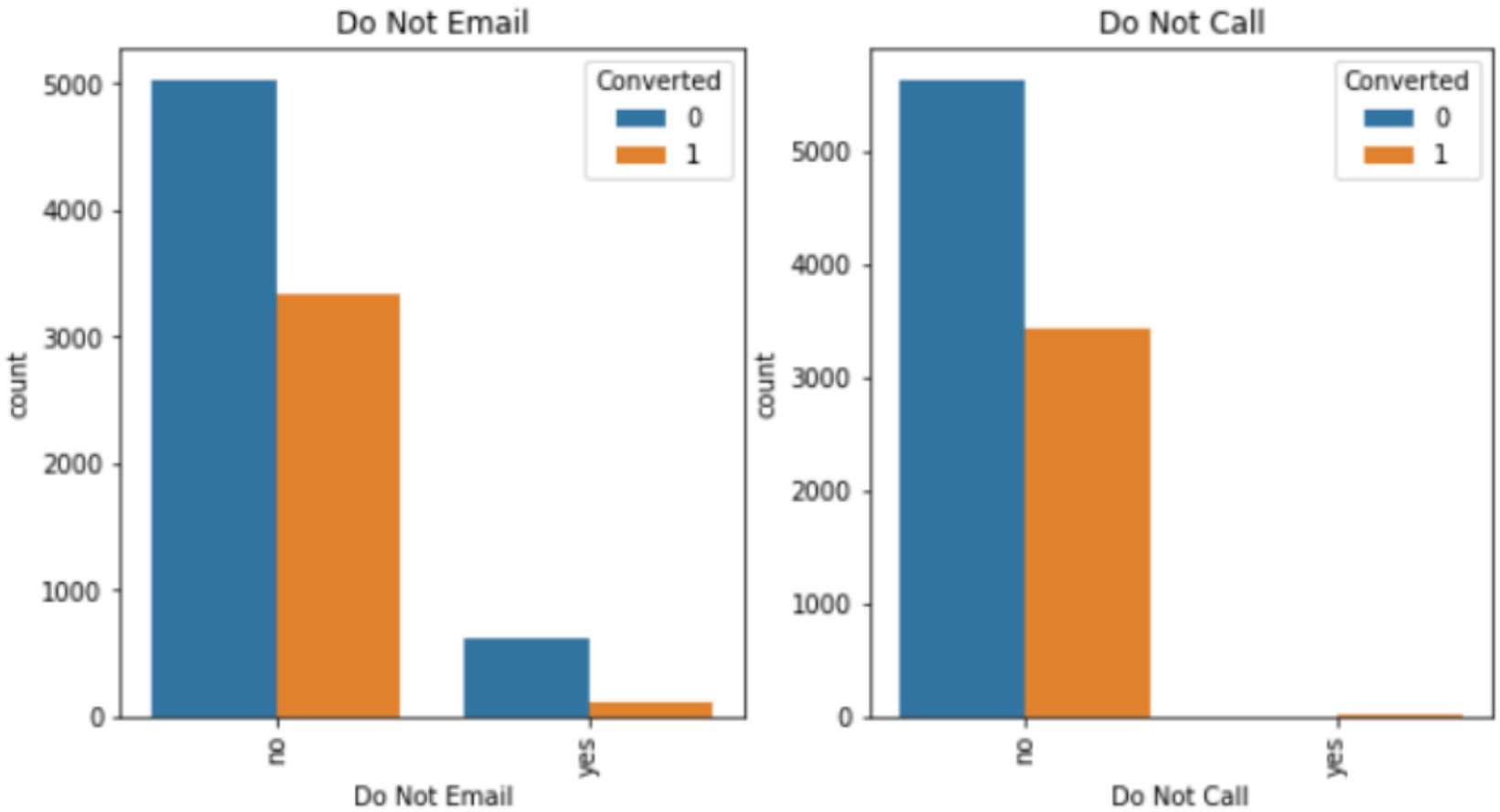

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

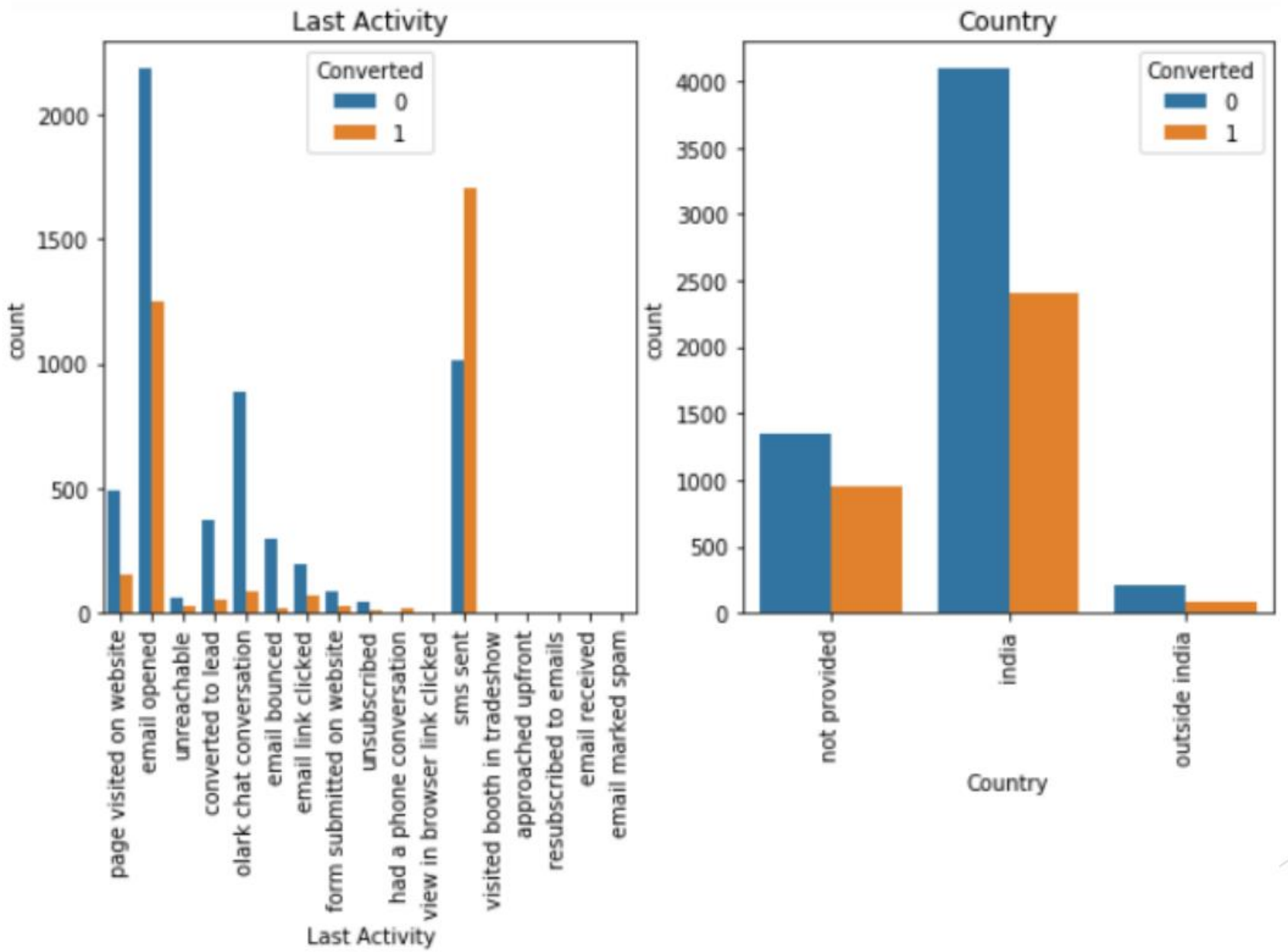

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email


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

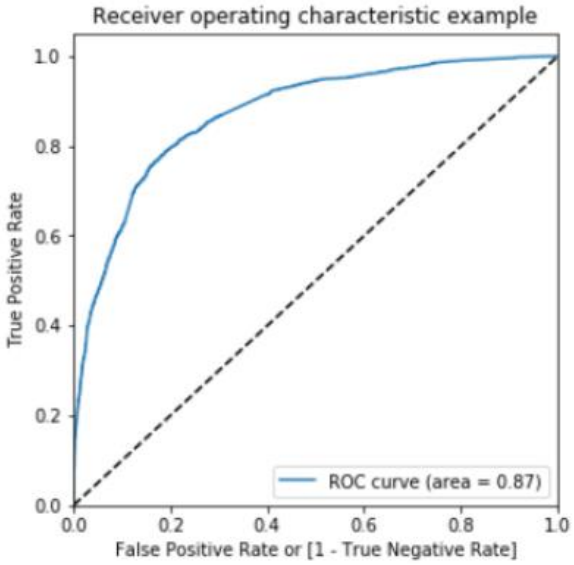
DATA CONVERSION

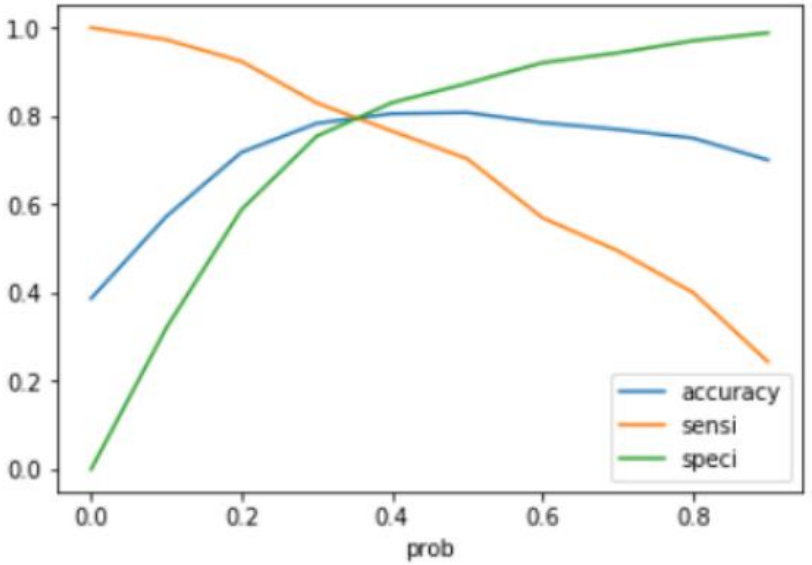
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

MODEL BUILDING

- Splitting the Data into Training and Testing Sets.
- Performing a train-test split is the first fundamental step in regression; we have selected a 70:30 split ratio.
- Use RFE while choosing features.
- Executing RFE with 15 output variables.
- Creating a model by removing any variables with a p-value and a vif value of more than 0.05 and 5.
- Predictions are based on the test data set.
- Overall accuracy is 81%.

ROC CURVE





Finding Optimal Cut off Point

- The probability where we have balanced sensitivity and specificity is the optimal cut off probability.
- The second graph makes clear that 0.35 is the ideal cutoff.

CONCLUSION

According to research, the following factors affected potential purchasers the most (in descending order):

The total time spend on the Website.

Total number of visits.

When the lead source was:

a. Google

b. Direct traffic

c. Organic search

d. Welingak website

When the last activity was:

a. SMS

b. Olark chat conversation

When the lead origin is Lead add format.

When their current occupation is as a working professional. With these in mind, X Education can succeed since they have a very good probability of persuading nearly all prospective customers to change their minds and purchase their courses.

THANK YOU

- Hemanth Kumar Endla
- Tejas Katre
- Priyadarshini swain