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.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

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 the future use.

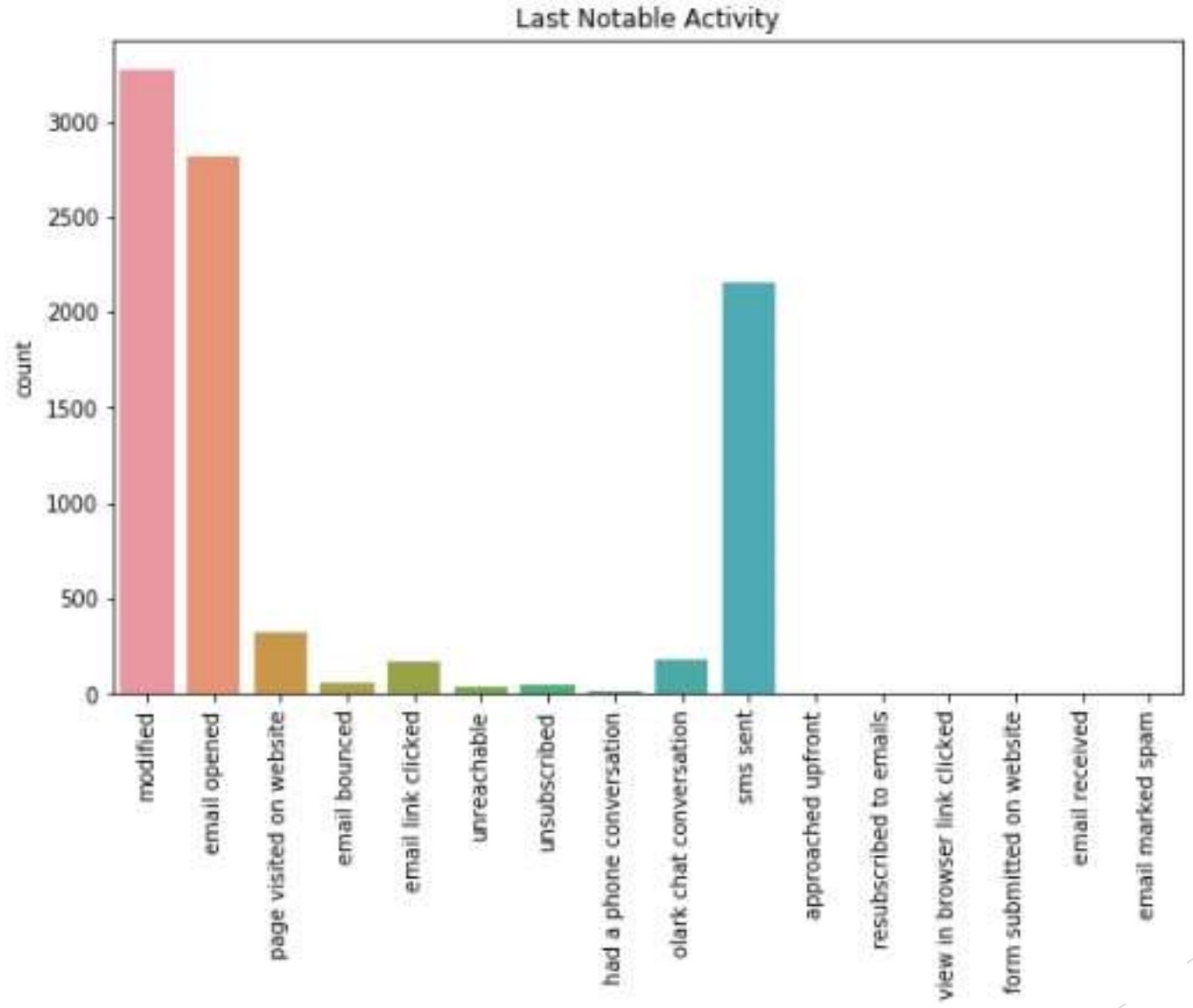
Solution Methodology

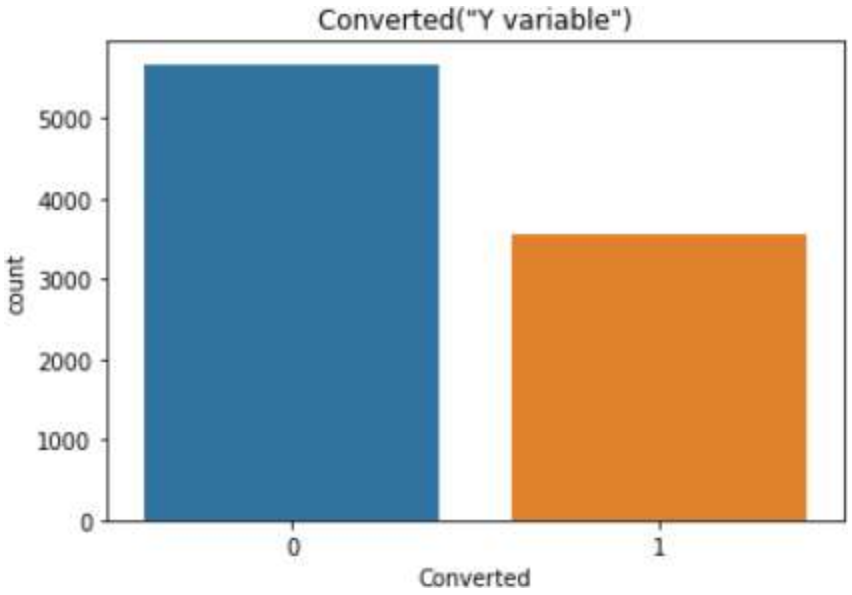
- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers 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.

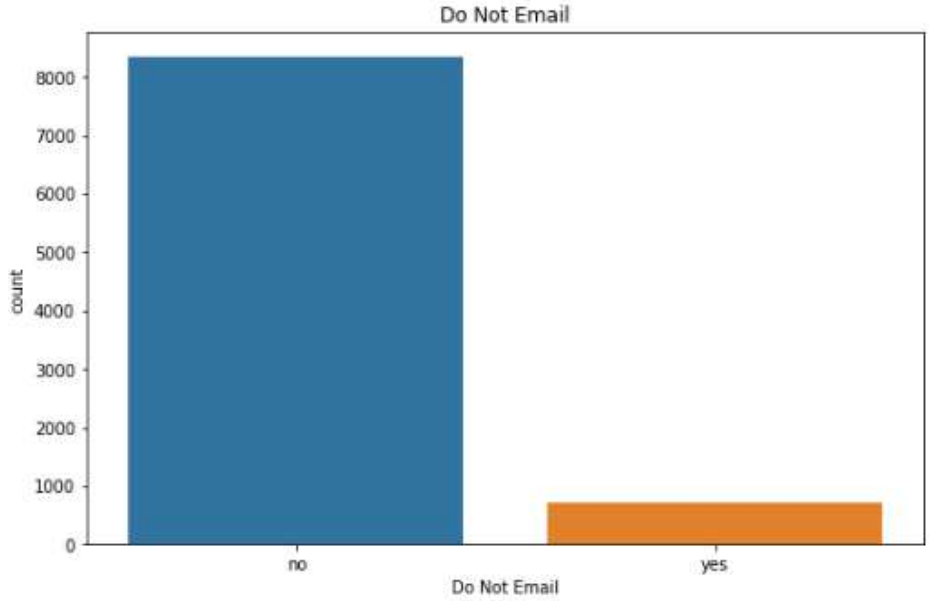
Data Manipulation

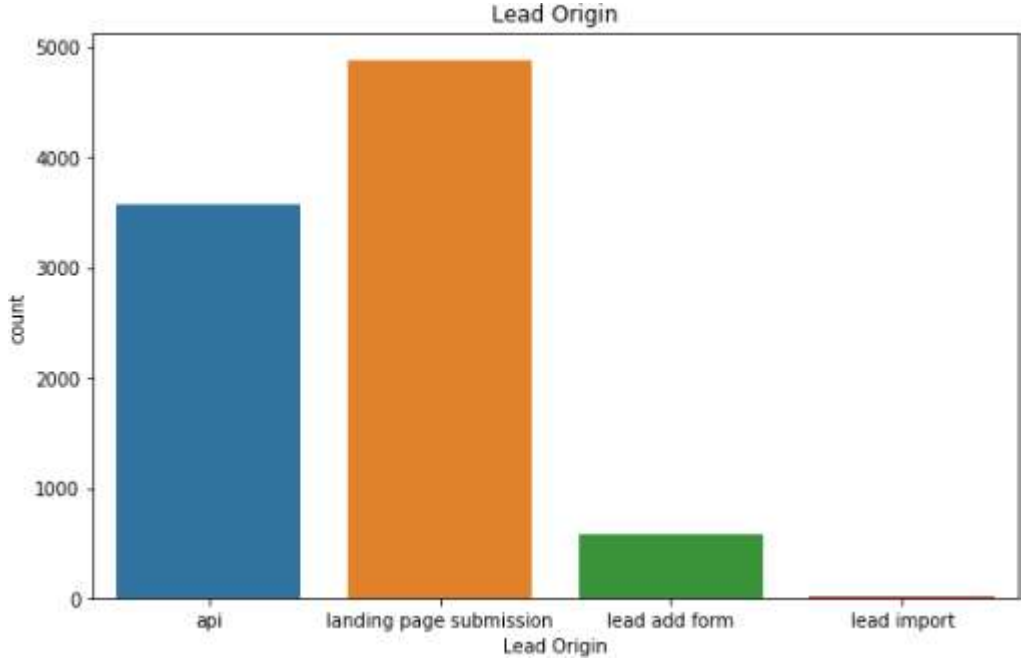
- Total Number of Rows =37, Total Number of Columns =9240.
- Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Magazine", "receive more updates about our courses", "Get updates on DM Content", "I agree to pay the amount through cheque", etc.
- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

EDA

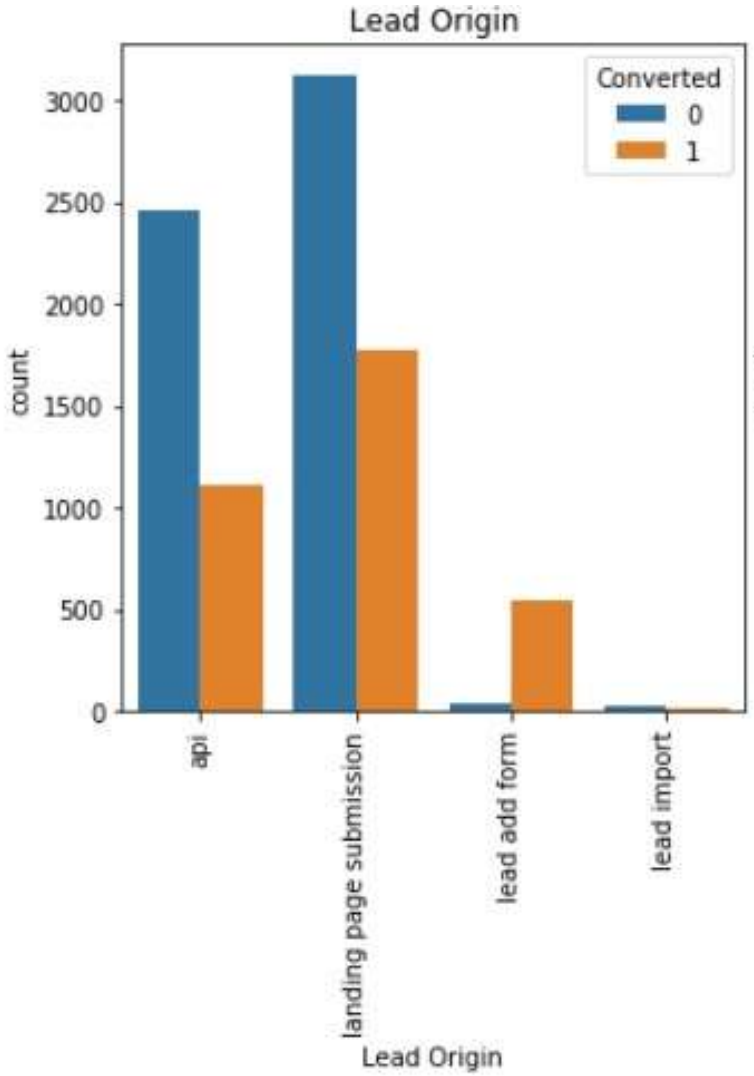

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted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

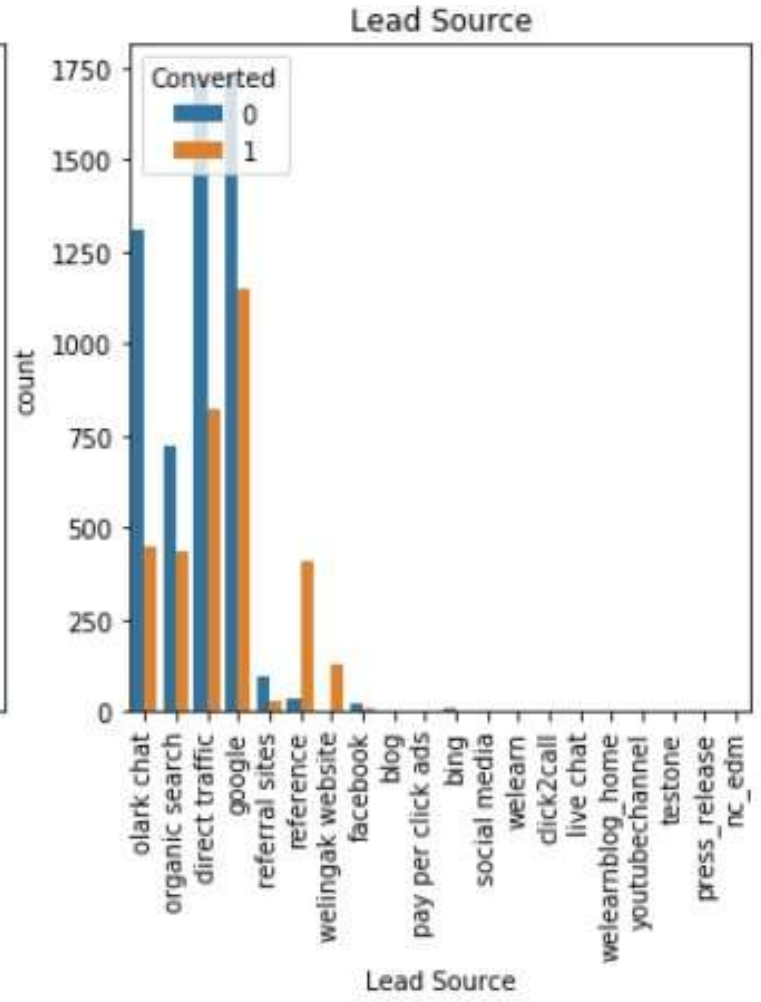

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

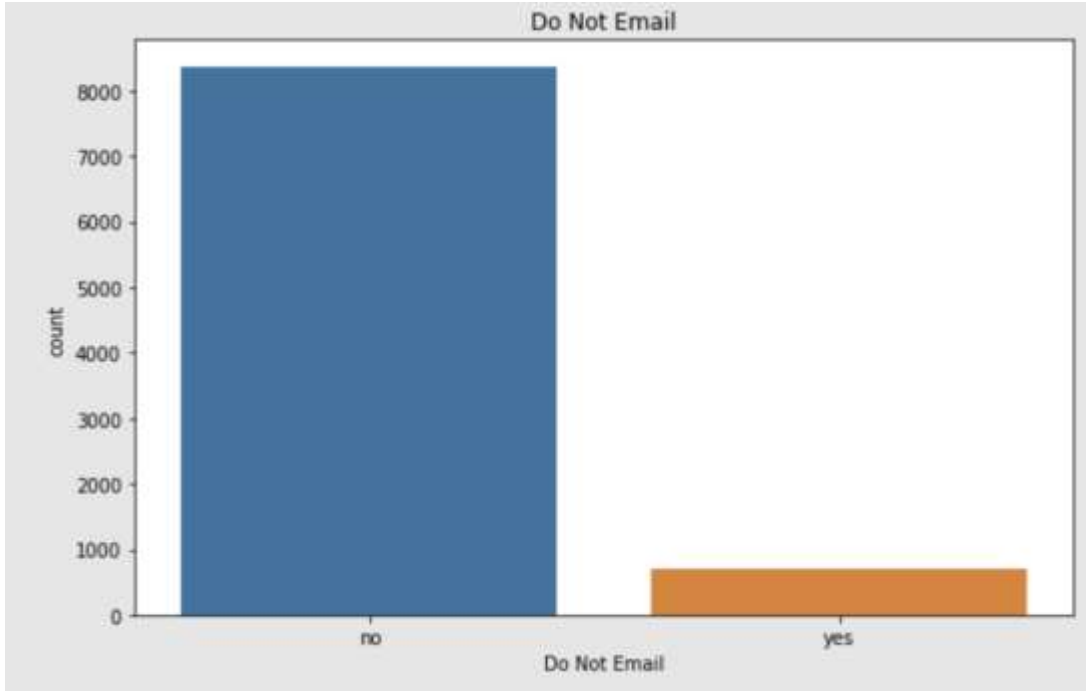

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

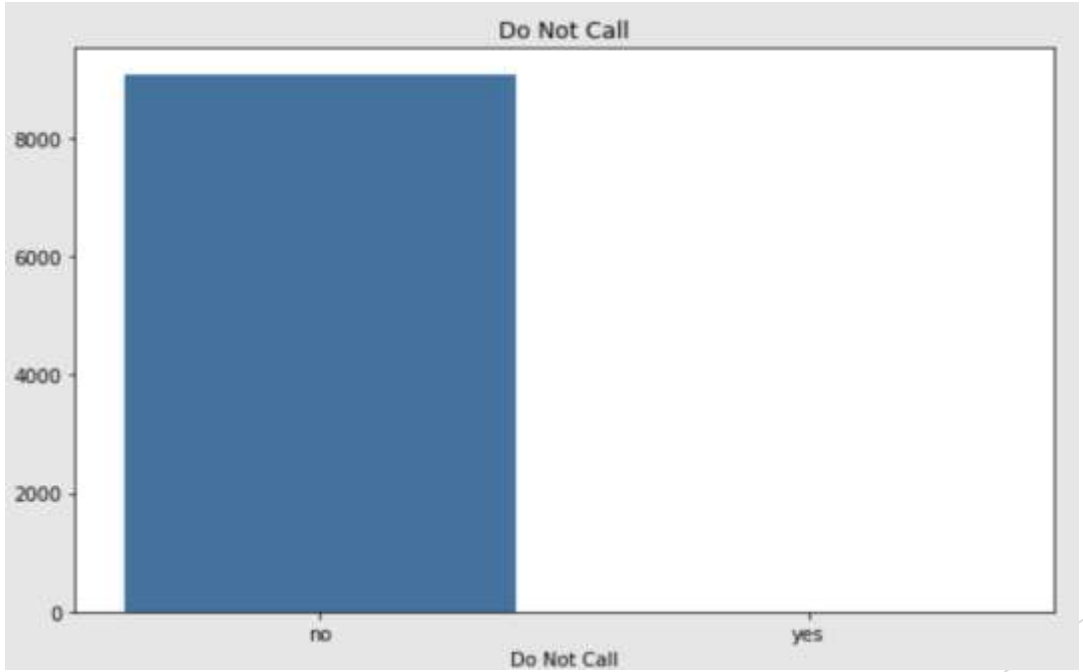

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

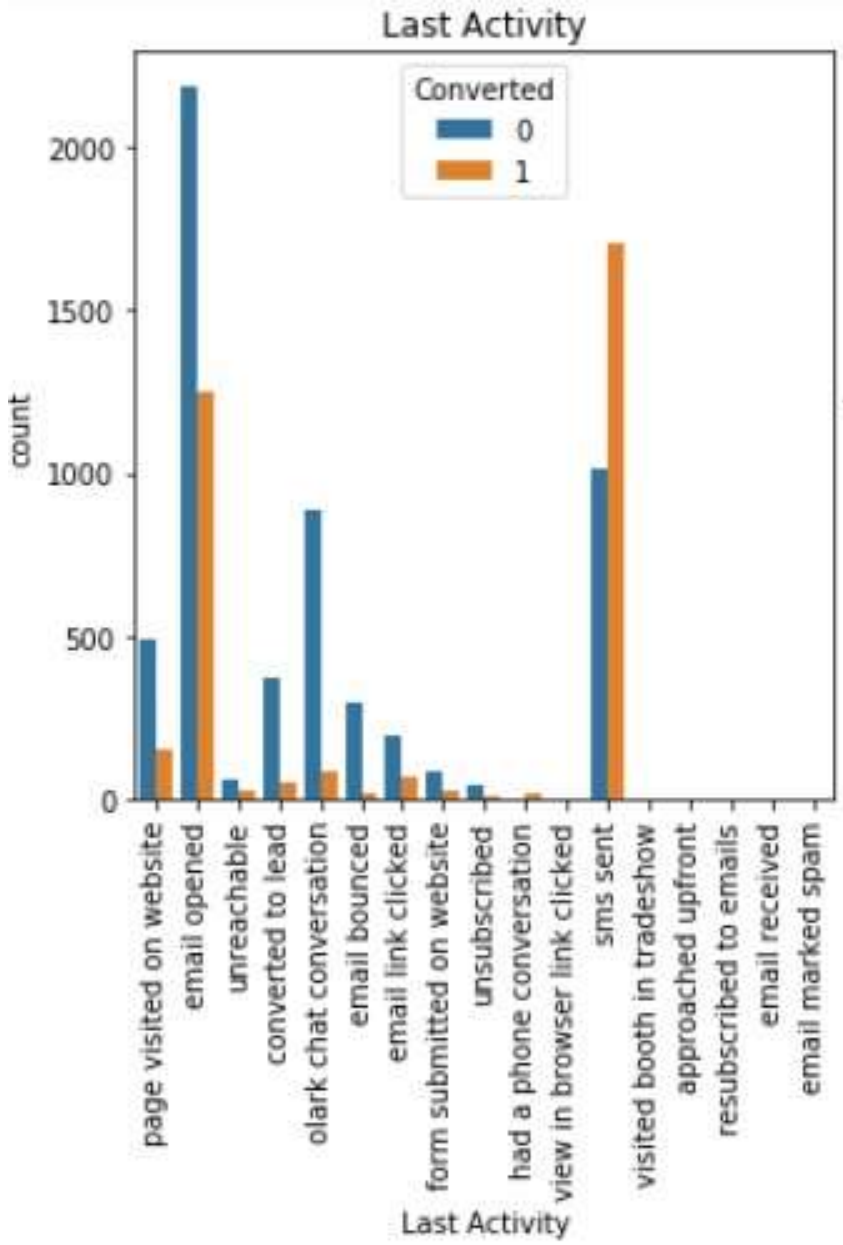
Categorical Variable Relation

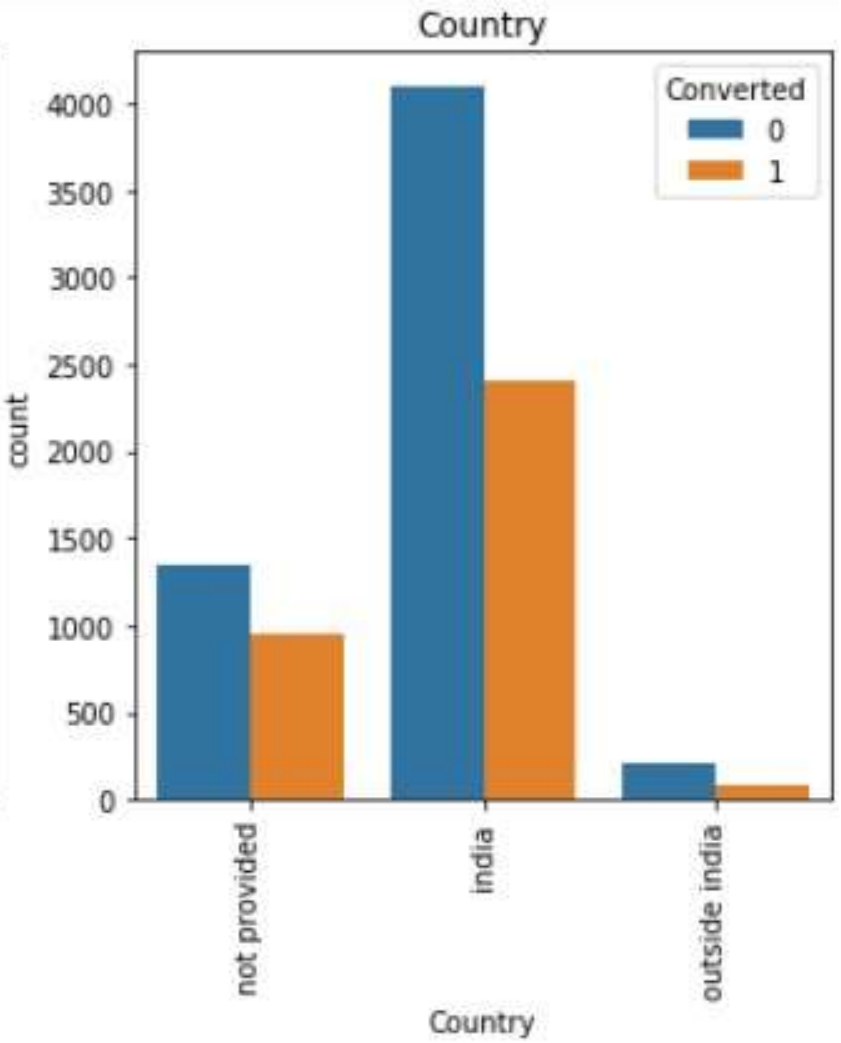





Do Not Email
8000
7000
6000
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email


Do Not Call
8000
6000
4000
2000
0
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

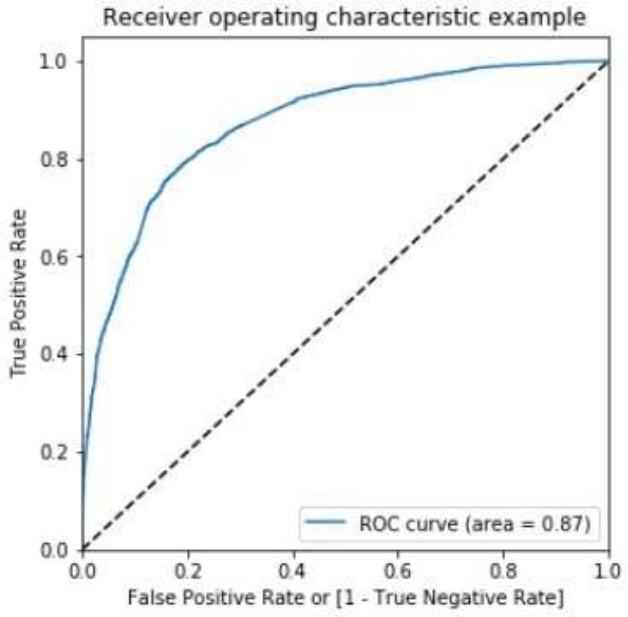
Data Conversion

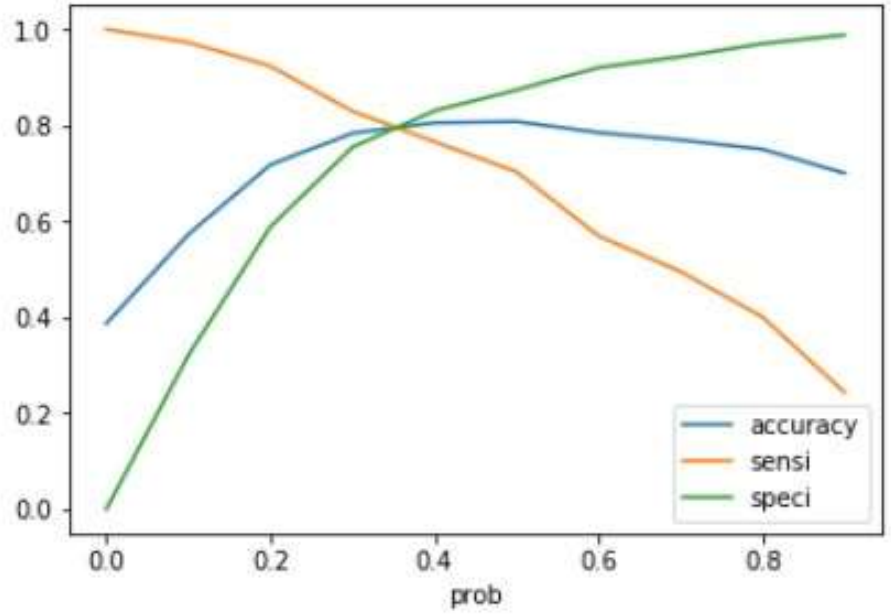
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

Model Building

- Splitting the Data into Training and Testing Sets
- 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.

Making Predictions

- With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 80%.
- With the current cut off as 0.35 we have Precision around 79% and Recall around 70%
- With the current cut off as 0.41 we have Precision around 75% and Recall around 76%
- With the current cut off as 0.41 we have Precision around 73% and Recall around 77%

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :
1)The total time spent on the website.
2)Total number of visits on the website.
3)Lead source:
a. Google
b. Direct traffic
c. Organic search
d. Welingak website

4)Last activity:
a. SMS
b. Olark chat conversation

5)Lead origin is Lead add format.
6)Current occupation is Working professional.