Lead Scoring Case Study

GROUP MEMBERS:

PRUTHAL AYAR

RAHUL YADAV

JYOTI MALIK

Problem Statement

- X Education sells online courses to industry professionals and professionals who are interested in the courses land on their website and browse for courses.
- When professionals fill up a form providing their email address or phone number, they are classified to be a lead.
- Few leads get converted while most do not and the typical lead conversion rate at X education is around 30%.
- The company wants to identify the Hot Leads and build a model so that the lead conversion rate goes up.
- The target lead conversion rate should be around 80%.

Business Objective

- X Education wants to build a model to identify the Hot Leads.
- The model should be able to adjust and handle the company's future requirements.

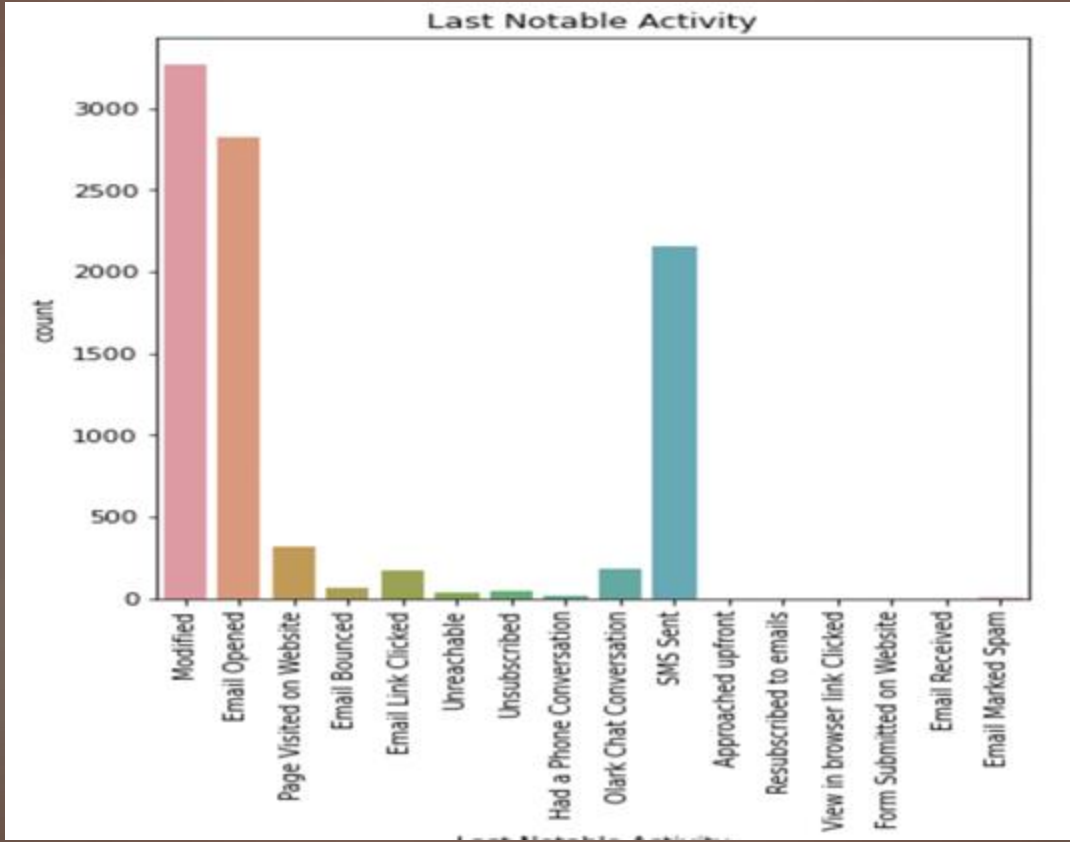
Solution Methodology

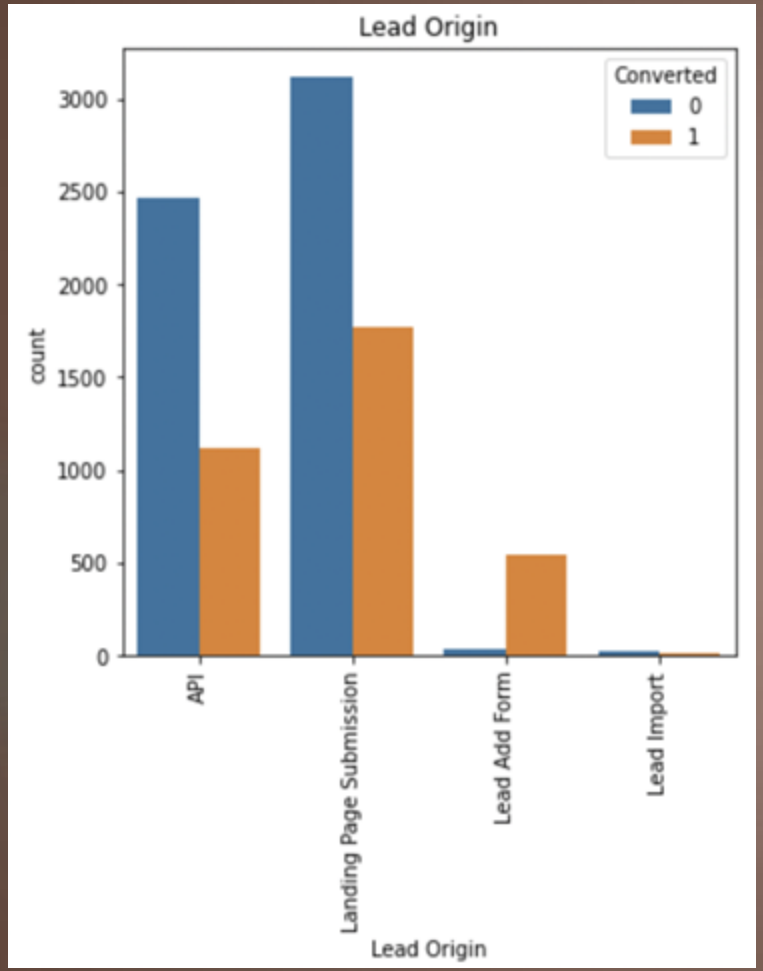
- **Data Cleaning and Data Manipulation:**
  - Checking and handling duplicate data.
  - Checking and holding NA and Missing Values
  - Dropping column if it contains large missing values or not useful for Analysis
  - Imputation of values if necessary
  - Checking and handling Outliers in the data.

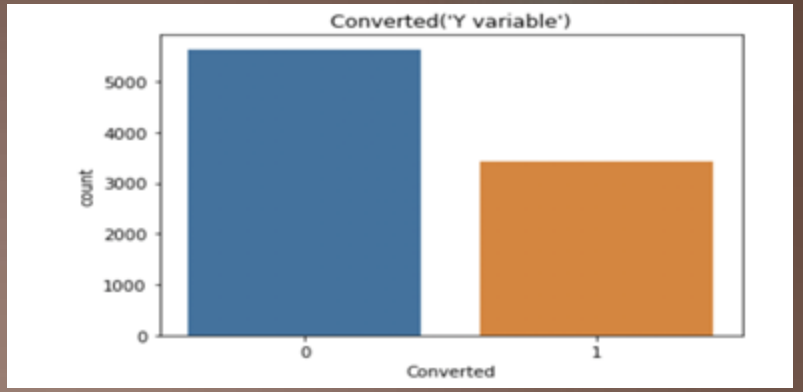
- **EDA:**
  - Univariate Analysis: value count, distribution of variable
  - Bivariate Analysis: correlation coefficients and pattern between variables

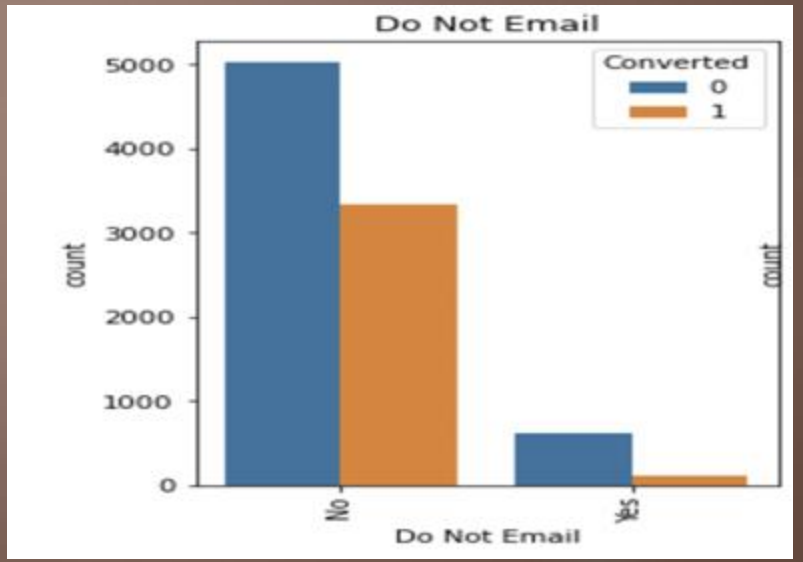
- Featuring Scaling, Dummy Variables and Encoding of data
- Logistic Regression is used for model making and prediction
- Validation of the Model and Model presentation
- Conclusion

Exploratory Data Analysis

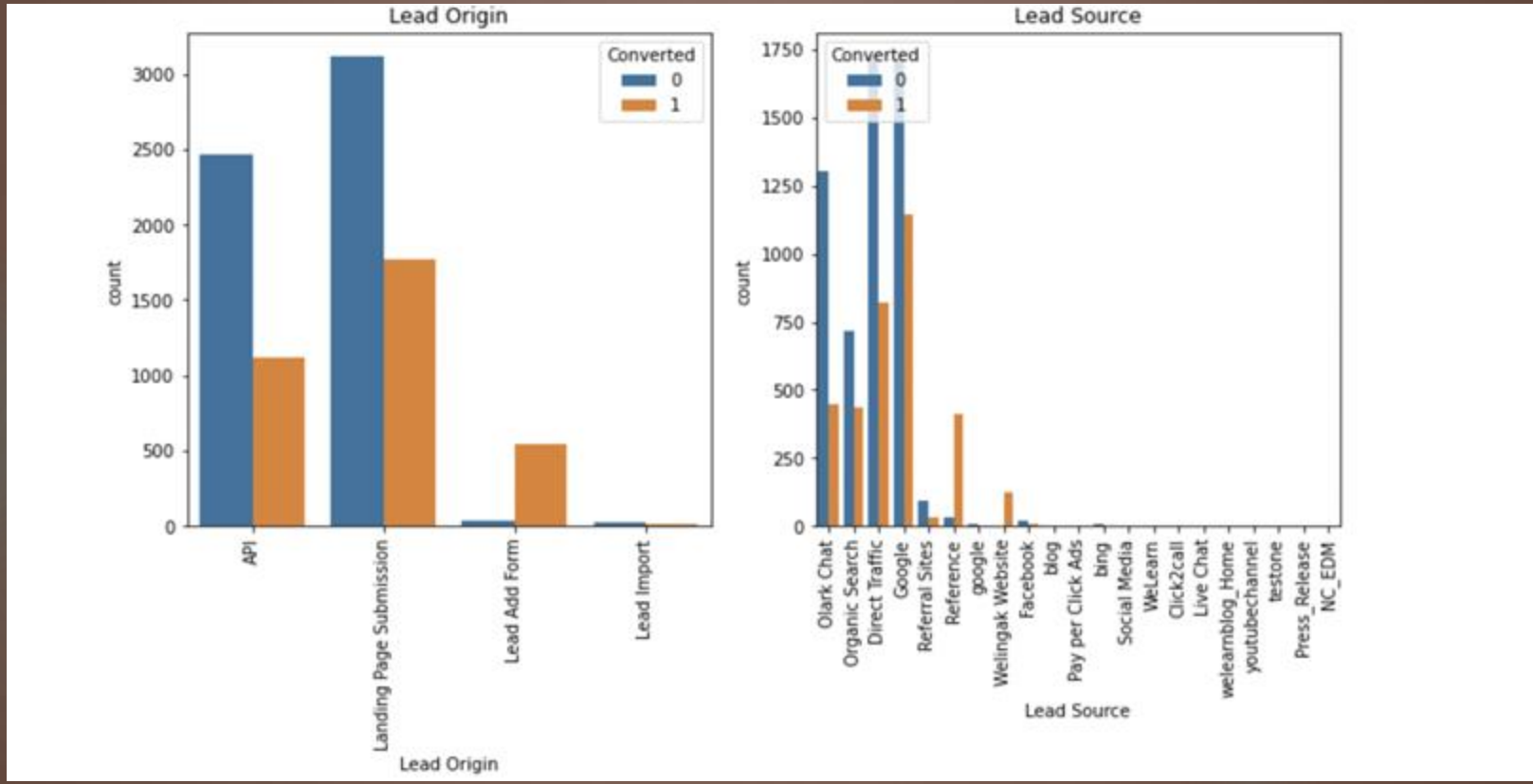


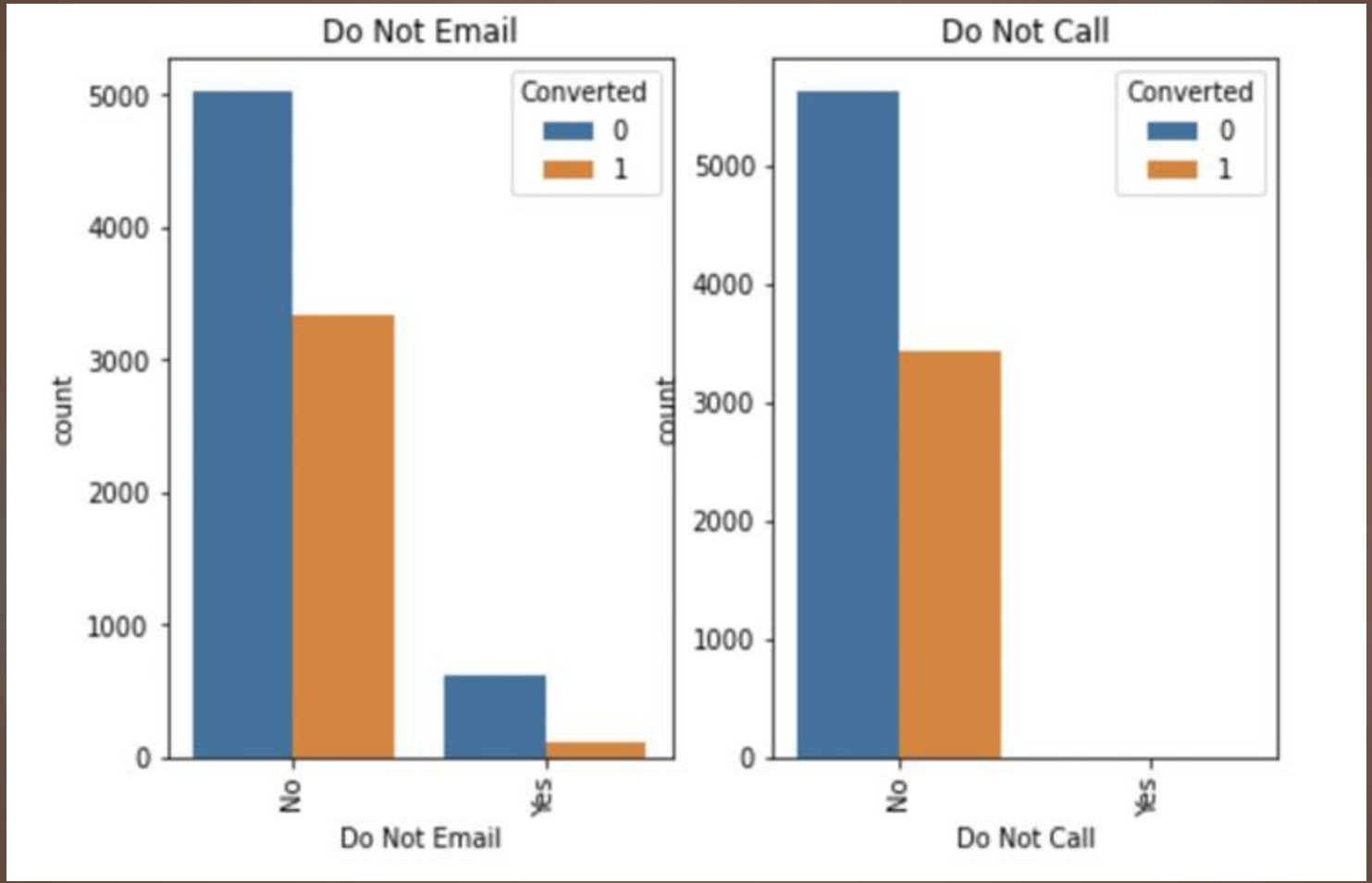

Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

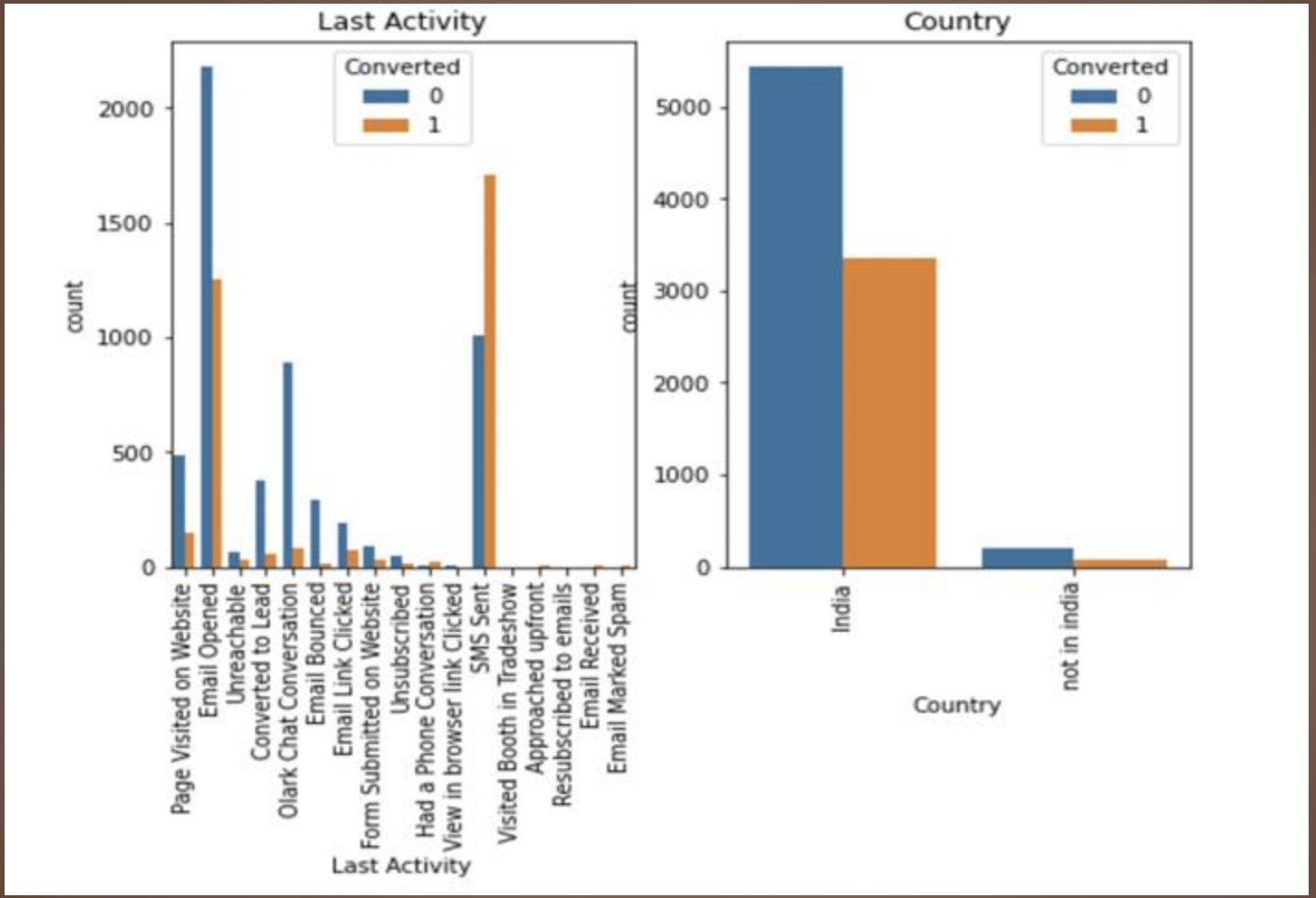

Converted('Y variable')
count
5000
4000
3000
2000
1000
0
0
1
Converted


Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email
count

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
5000
4000
3000
2000
1000
0
India
not in india
Country

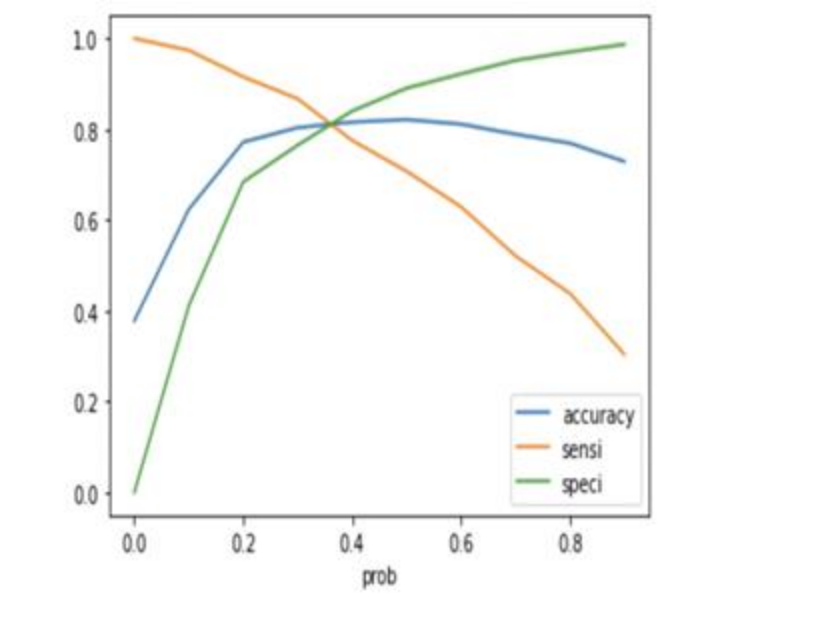
Data Conversion

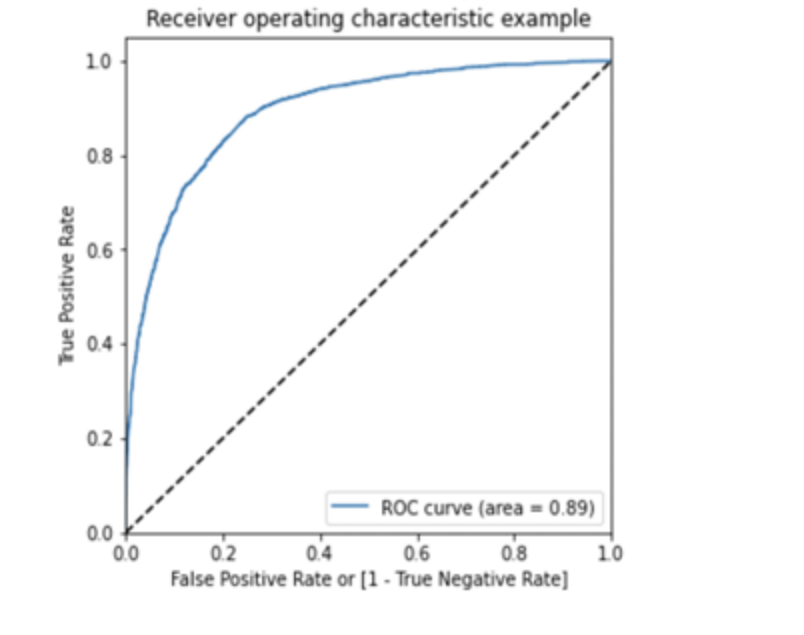
- Numerical Variables are normalized
- Dummy Variables are created for Object type Variables using the 'get_dummies'.

Model Building

- Splitting the dataset into 70 percent and 30 percent for train & test respectively
- Running RFE with 20 variables as Output
- Using RFE for Feature Selection
- Building a model whose p-values are below 0.05
- Creating Prediction on the data set and Overall Accuracy is 82%
- With the current cut off as 0.5 we have around 82% accuracy, sensitivity of around 70% and specificity of around 89%

Model Evaluation-ROC

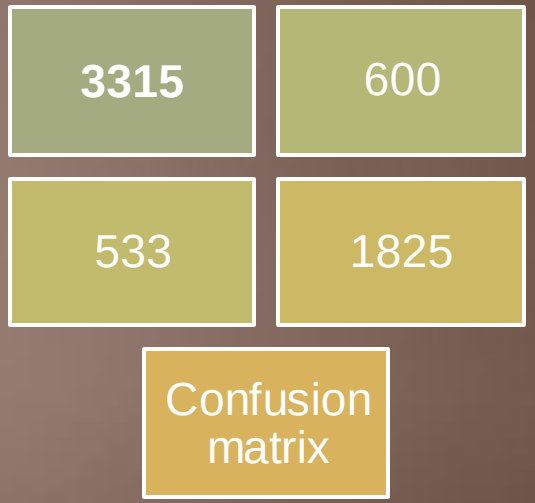




- From the graph it is visible that the optimal cutoff is at 0.35
- The area under ROC curve is 0.89 which is a very good value

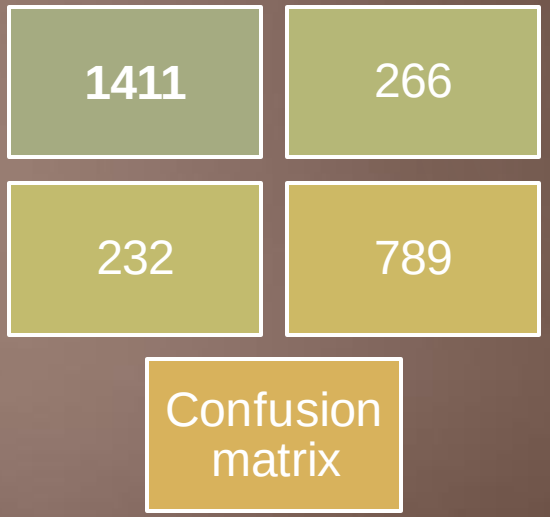
Model Evaluation: Precision & Recall on Train Dataset

- 0.41 is the tradeoff between precision and recall.
- Thus we can safely choose to consider any prospect lead with conversion probability higher than 41 % to be a Hot Lead.
- Precision around 75% and Recall around 77% and accuracy 81%



Model Evaluation: Precision & Recall on Test Dataset

- Precision: 74.78%
- Recall: 77.27%
- Accuracy: 81.54%



CONCLUSION

- It was found that the variables that mattered the most in the potential buyers are (In descending order):
- TotalVisits
- The total time spend on the Website.
- Lead Origin_Lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source Organic Search
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- Do Not Email_Yes
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation

- 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.