Lead Score Case Study

Group Members

1.Bandaru Girish Kumar

2. Pratik Umesh Pande

3. Riya Gupta

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 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 :

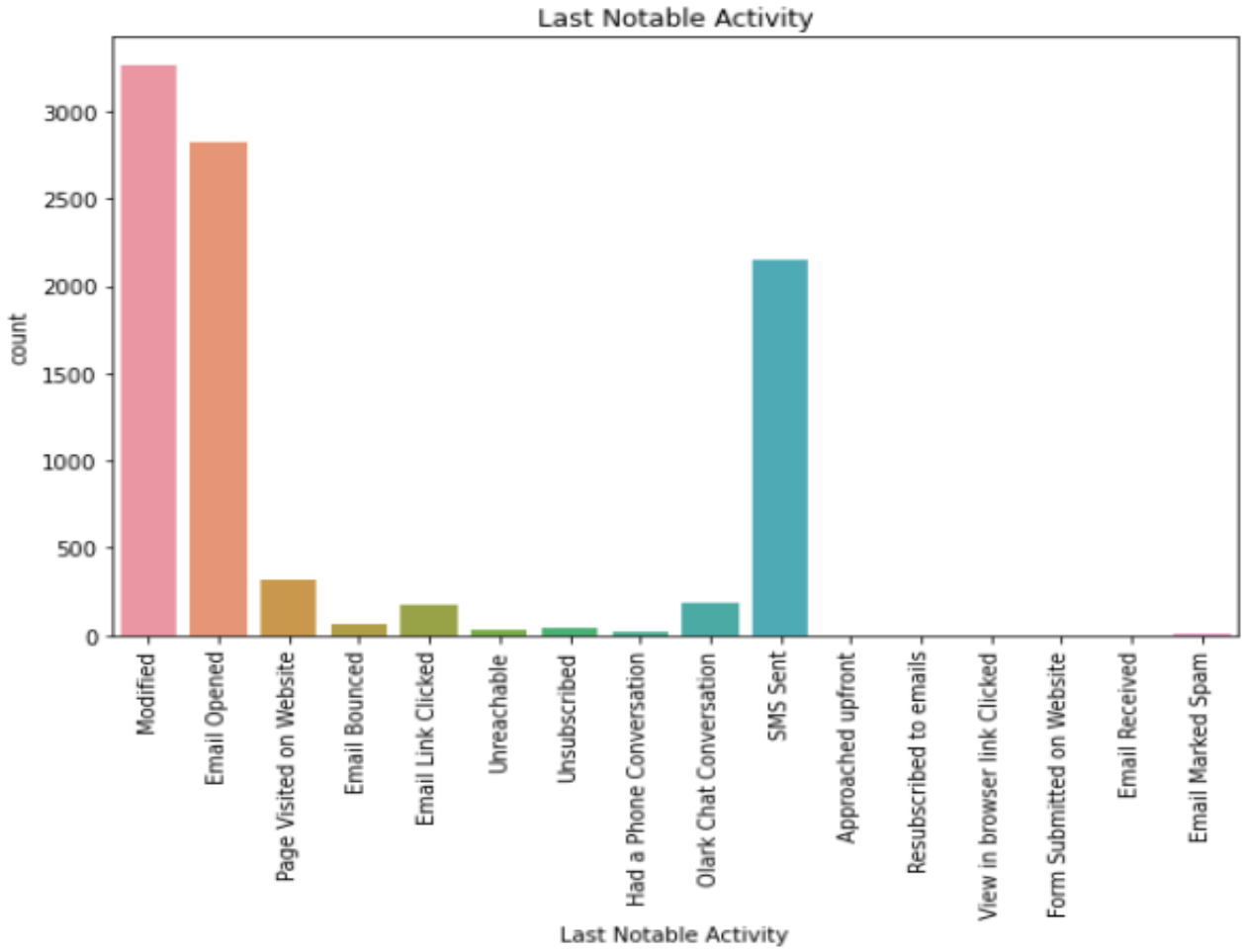
1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

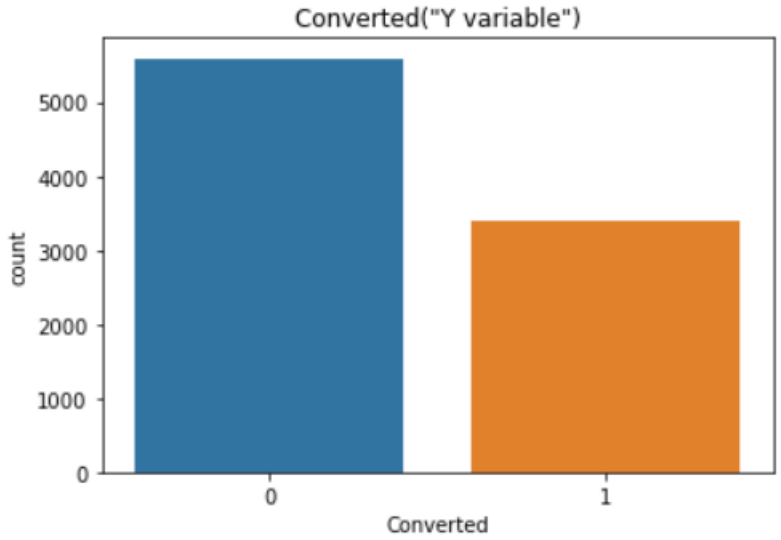
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

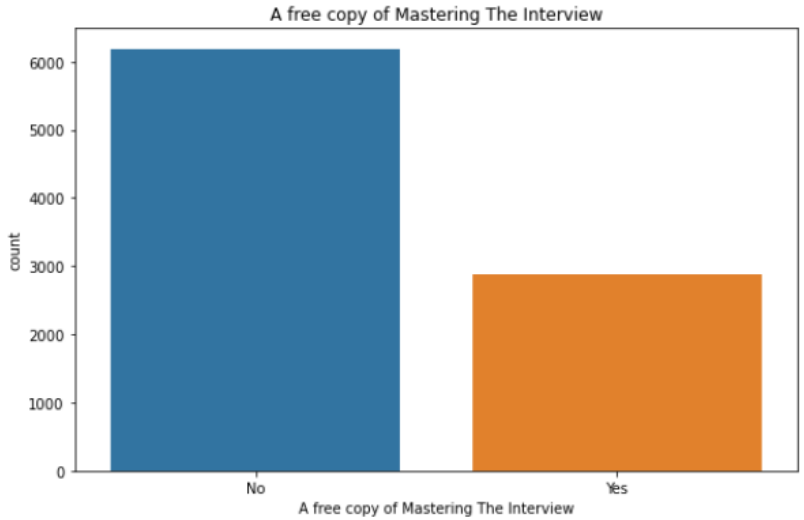
Data Manipulation

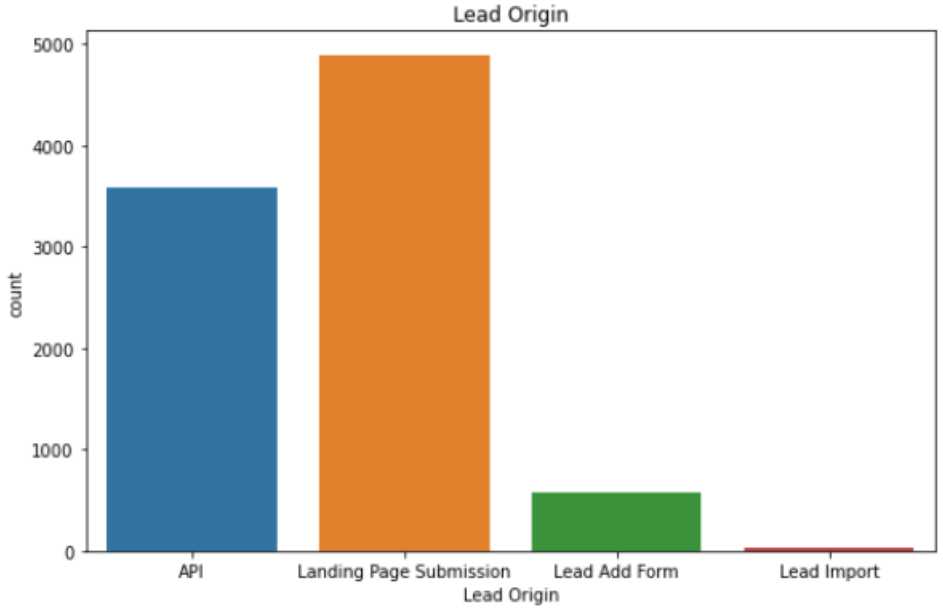
- Total Number of Rows =9240, Total Number of Columns =37.
- Dropped “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- Dropped the columns having more than 45% of missing value such as 'Asymmetrique Profile Score','Asymmetrique Activity Score','Asymmetrique Profile Index’, 'Asymmetrique Activity Index','Lead Profile','Lead Quality', 'How did you hear about X Education’.
- Fillna done on columns, filled with 'not provided’ and the columns are: 'Specialization’, 'City’, 'Tags’, 'What matters most to you in choosing a course’, 'What is your current occupation’, 'Country’.
- In ‘Country’ column all the ‘not provided’ converted to ‘India’. By that in India total leads count is 8953 and outside India 287.
- Regarding 'Country’, 'Tags’ where Number of Values for India are quite high (nearly 97% of the Data), this column can be dropped and we can drop "Tags" ,As tags variable is generated by the sales team after the discussion with students otherwise it will increase the model accuracy .

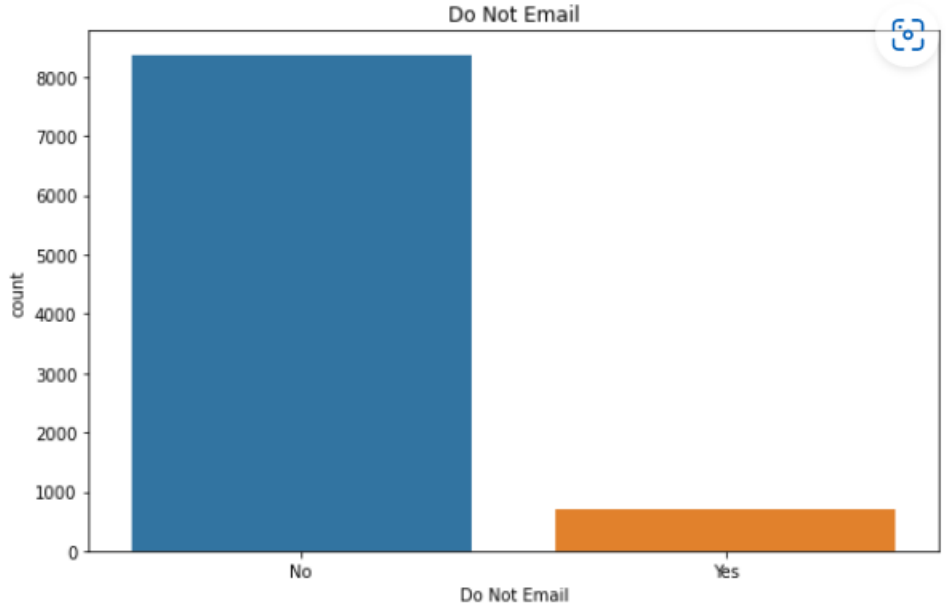
EDA


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
Last Notable Activity

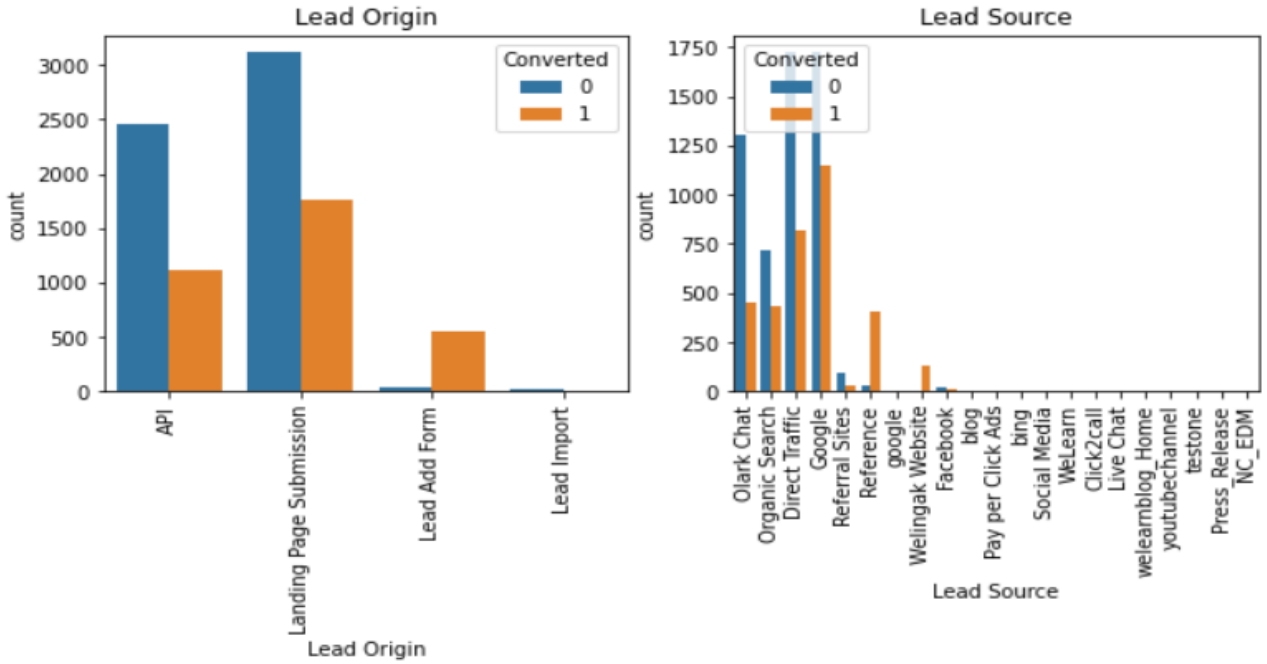

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

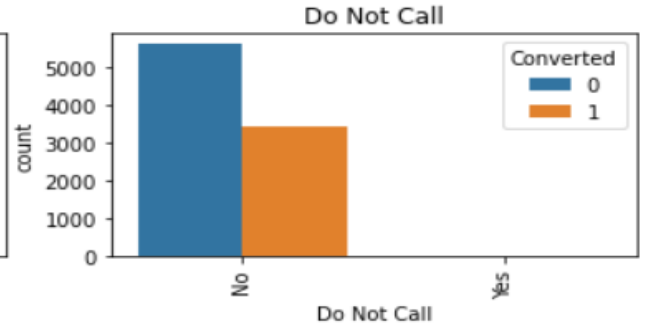

A free copy of Mastering The Interview
count
6000
5000
4000
3000
2000
1000
0
No
Yes
A free copy of Mastering The Interview

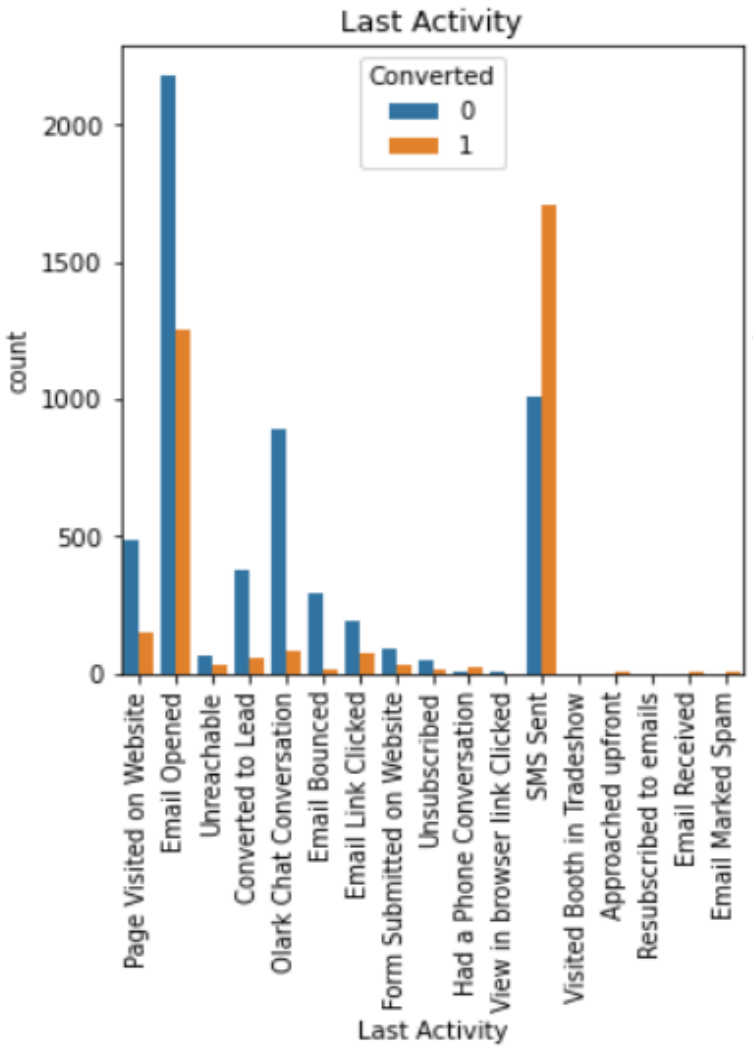

Lead Origin
count
5000
4000
3000
2000
1000
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

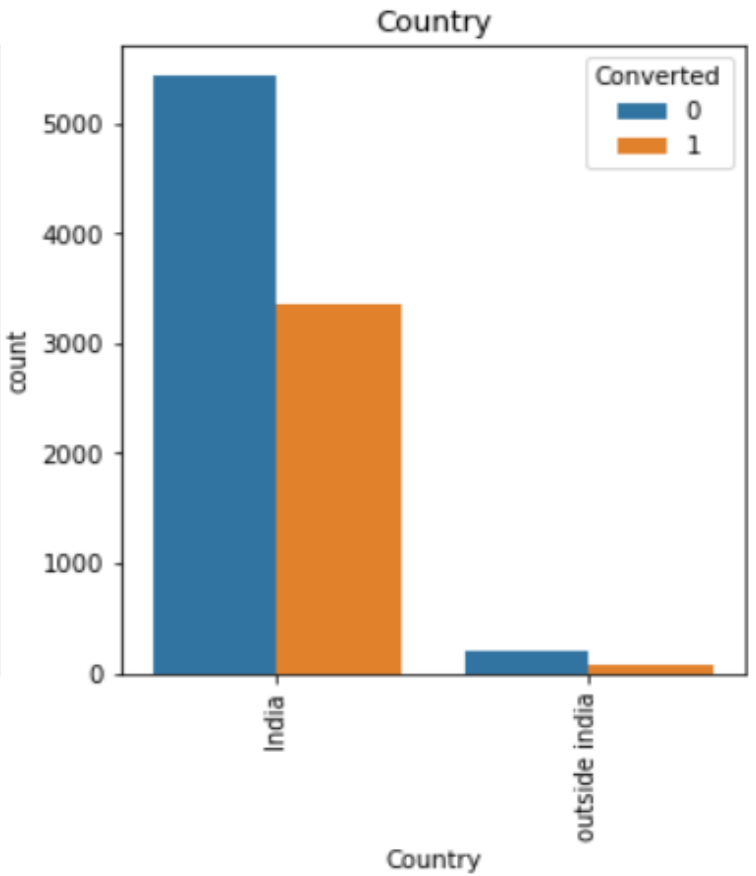

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

Categorical Variable Relation






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
outside india
Country

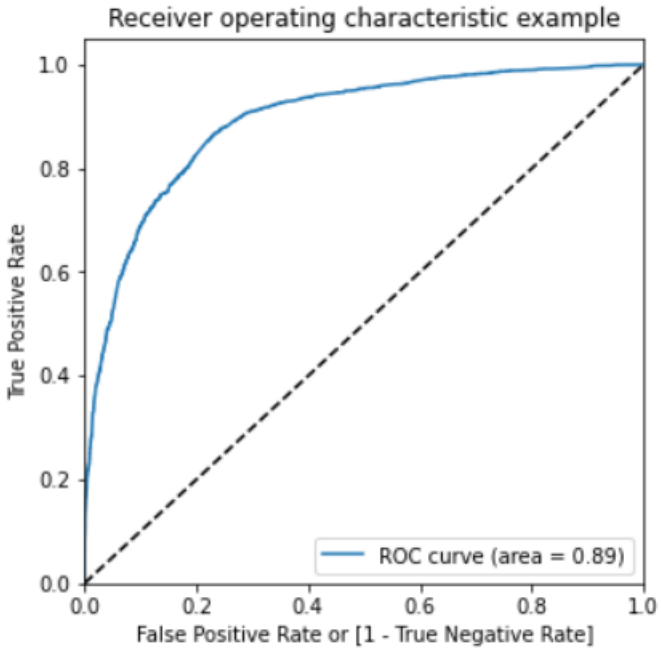
Data Conversion

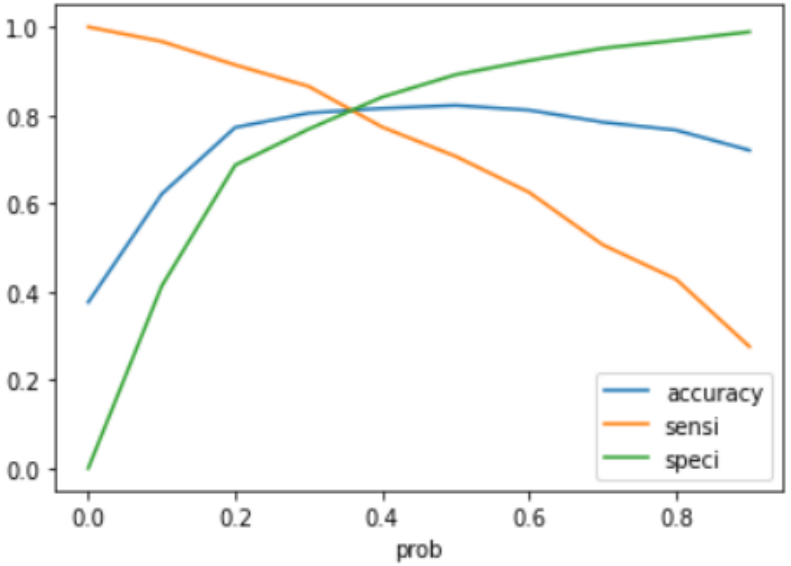
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- We dropped cloumns: 'City','What is your current occupation_not provided','Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call','Last Activity', 'Specialization', 'Specialization_not provided','What is your current occupation','What matters most to you in choosing a course', 'Search','Newspaper Article', 'X Education Forums', 'Newspaper','Digital Advertisement', 'Through Recommendations','A free copy of Mastering The Interview', 'Last Notable Activity’
- Total Rows: 8991
- Total Columns: 21

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 20 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 82%

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

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

- Total number of visits.
- The total time spend on the Website.
- When the lead source was:

a. Google

b. b. Direct traffic

c. c. Organic search

d. d. Welingak website

- When the last activity was:

a. SMS

b. b. Olark chat conversation

- When the lead origin is Lead add Form.
- 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