Lead Score Case Study

Team Members :

Siddharth Sareen

Swapnil Gaikwad

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.

CONTENT

- UNDERSTANDING THE DATA
- DATA CLEANING
- EXPLORATORY DATA ANALYSIS
- SPLITTING THE DATA INTO TRAIN AND TEST DATA
- FITTING THE DATA INTO MODEL
- RESULTS
- CONCLUSION

UNDERSTANDING THE DATA

- We have 37 features in total.
- We have total of 9240 data points.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- We have dropped the following columns because they have null value more than 40% :

  ['How did you hear about X Education','Lead Profile' ; 'Lead Quality' ; 'Asymmetrique Profile Score ' ; 'Asymmetrique Activity Score' ; 'Asymmetrique Activity Index' ; 'Asymmetrique Profile Index']

DATA CLEANING

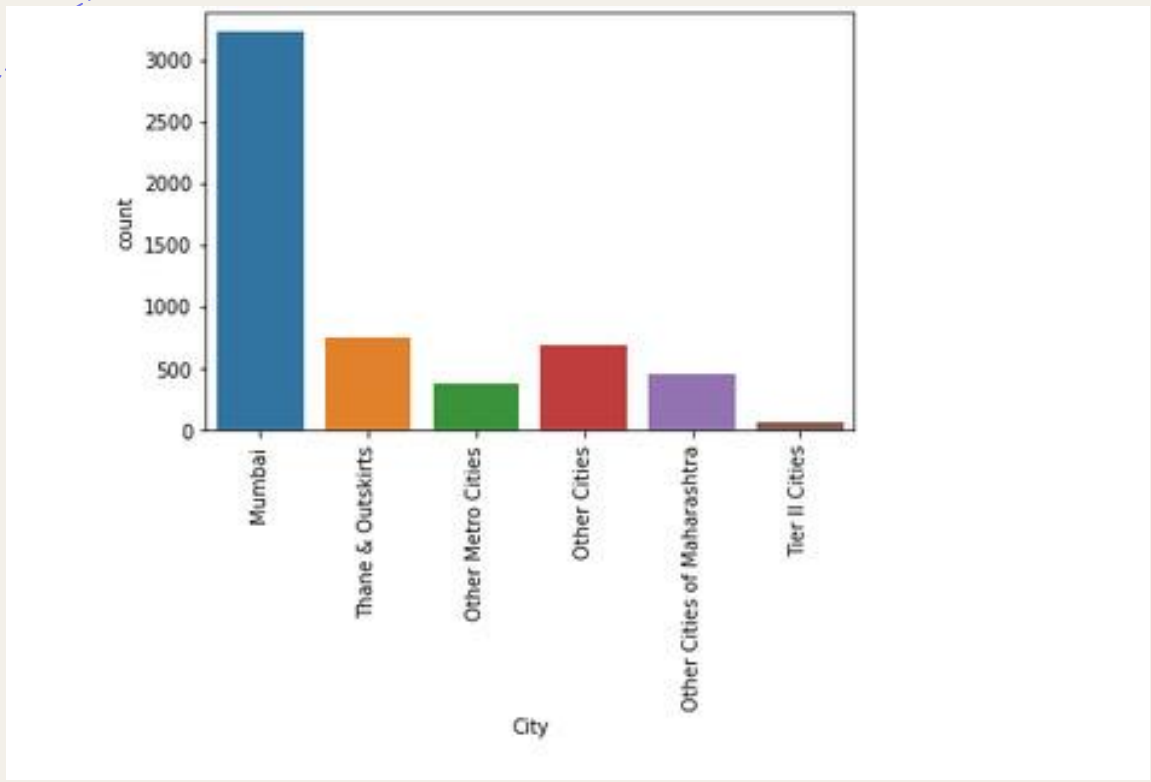
- Below mentioned columns contains null values :

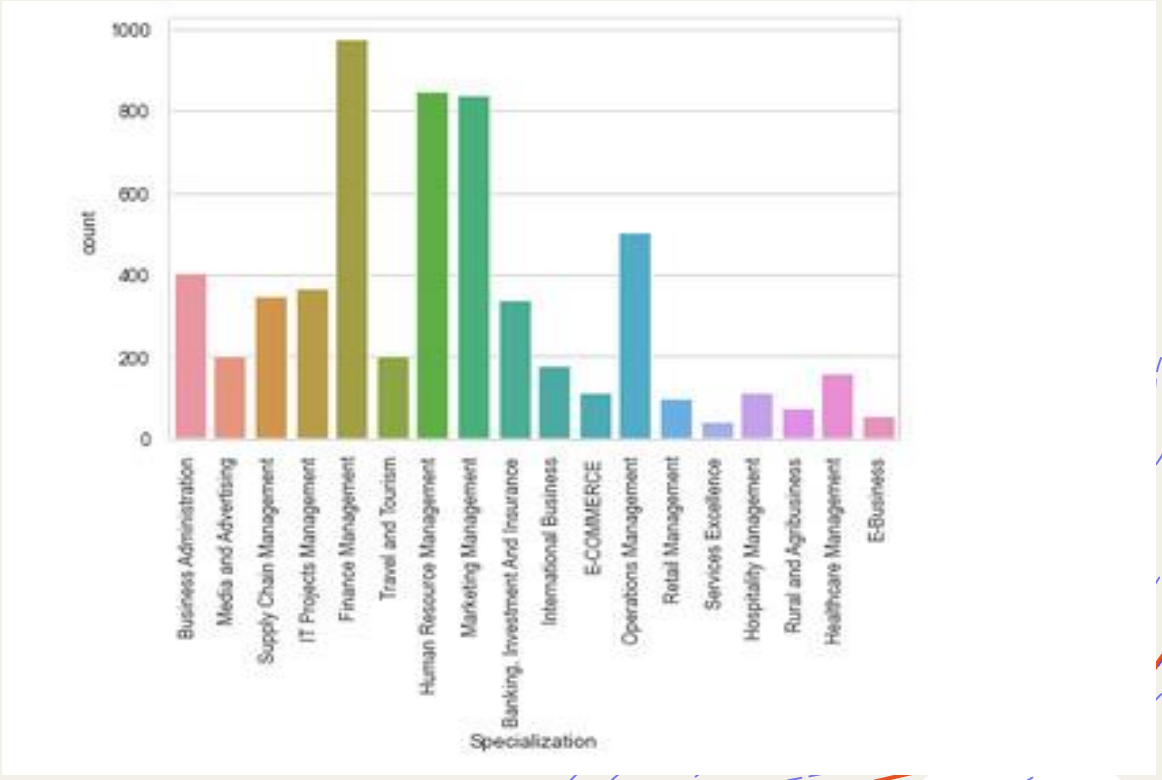
[City ; Specialization ;Tags ; What matters most to you in choosing a course ; What is your current occupation; Country ; Page Views Per ; Visit ; Total Visits ; Last Activity ; Lead Source]

- We have filled null value columns appropriately.
- After analyzing all the remaining columns we have found that some of the features are not important so we have dropped them.
- We have introduced dummy variables for the multilabel features.

Exploratory Data Analysis (Univariate Analysis)

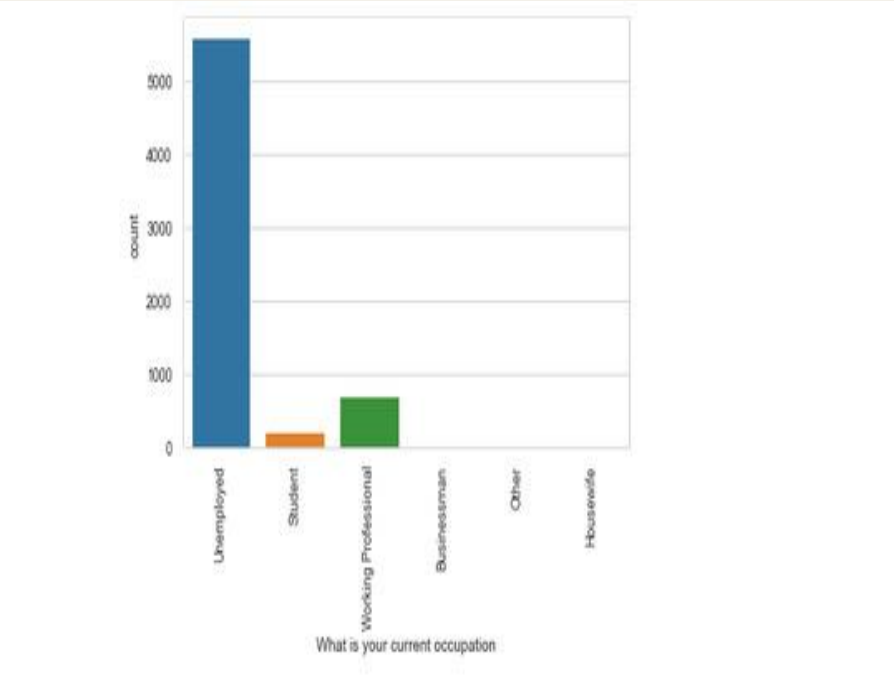
- Analyzing the 'City' and 'Specialization' feature :

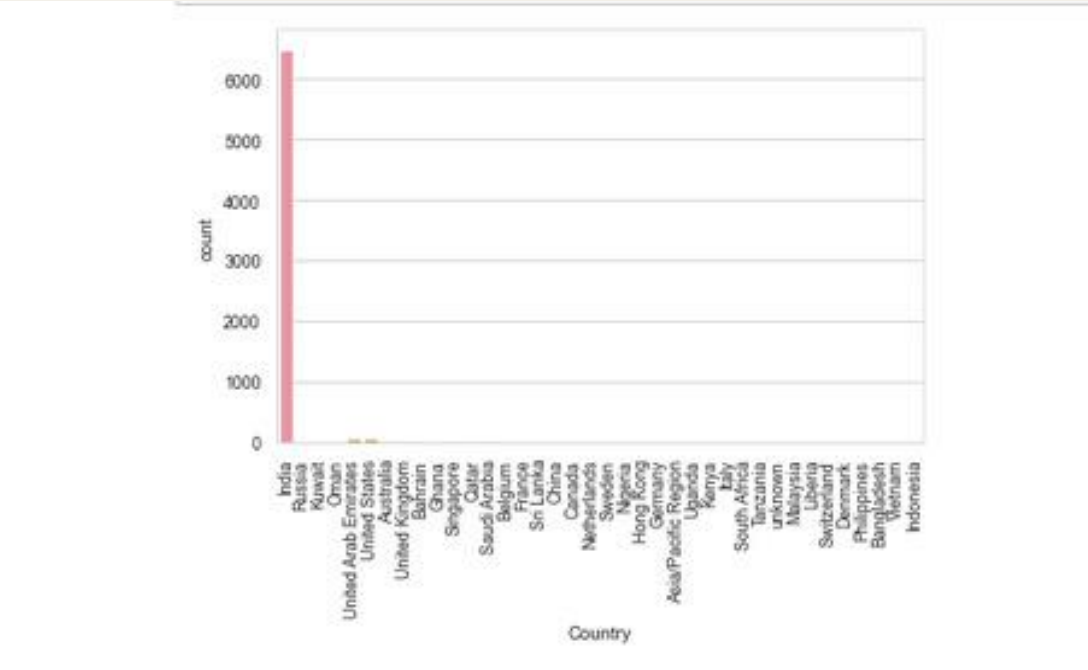




Exploratory Data Analysis (Univariate Analysis)

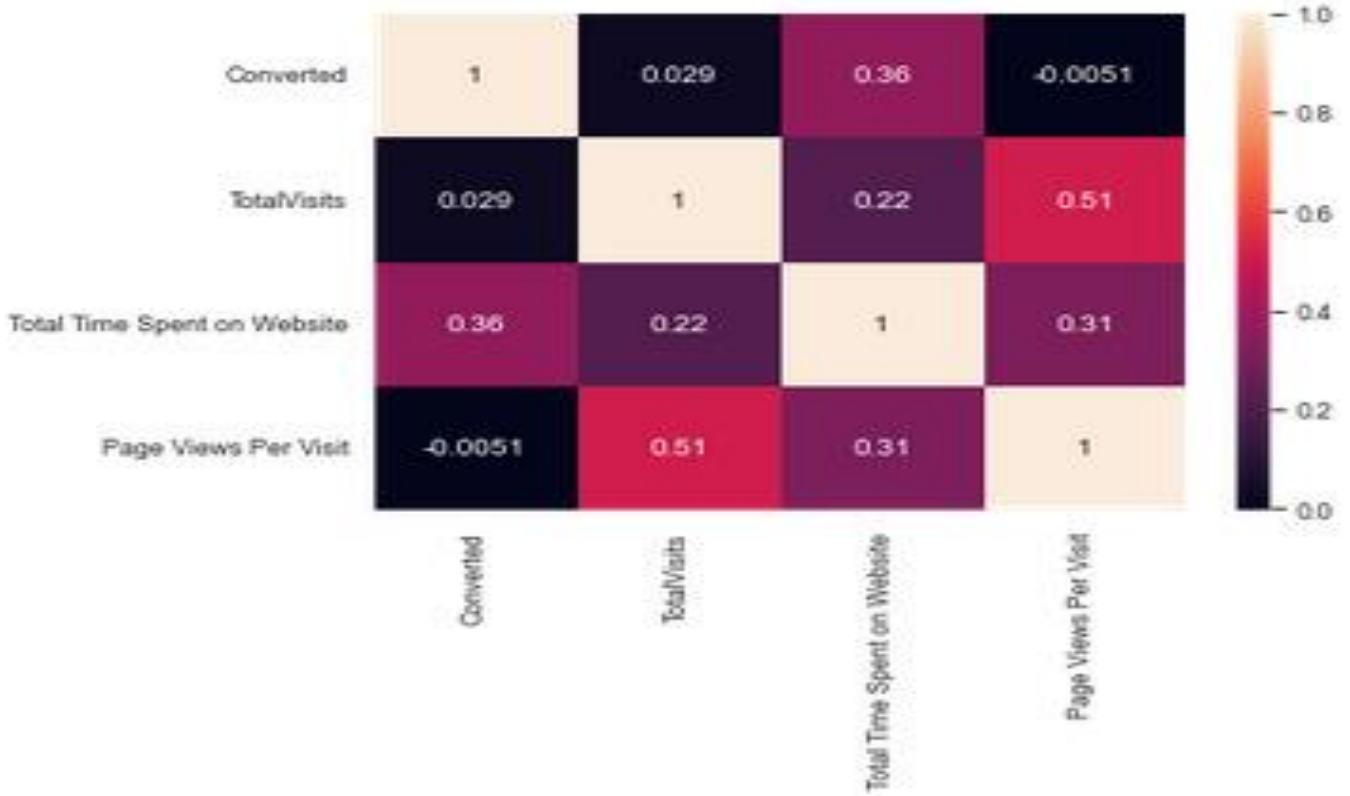
- Analysis of 'What is your current Occupation' and 'Country' Features :





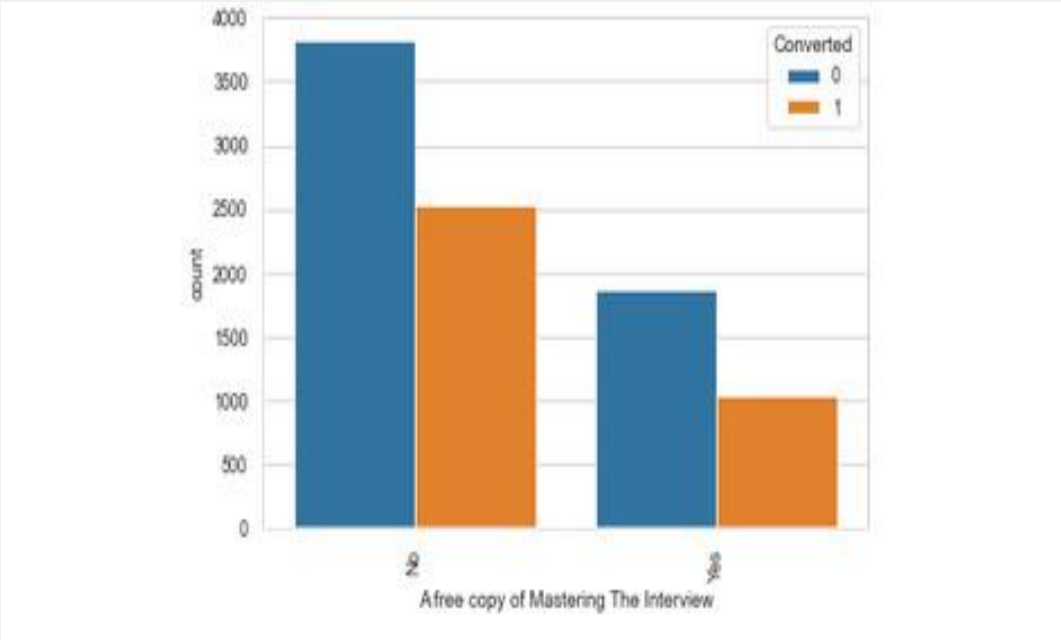
Exploratory Data Analysis (Bivariate Analysis)

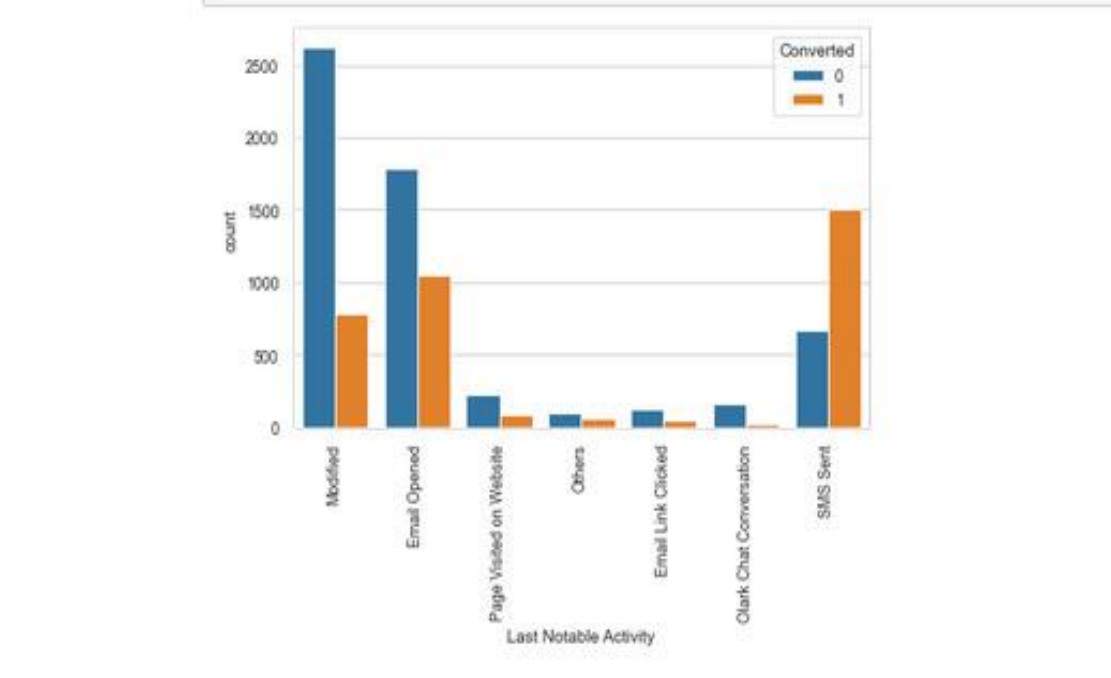
- Correlation Map between Numerical Variables :



Exploratory Data Analysis (Bivariate Analysis)

- Analyzing 'A free copy of Mastering The Interview' and 'Last Notable Activity' features





SPLITTING DATA INTO TRAIN AND TEST DATA

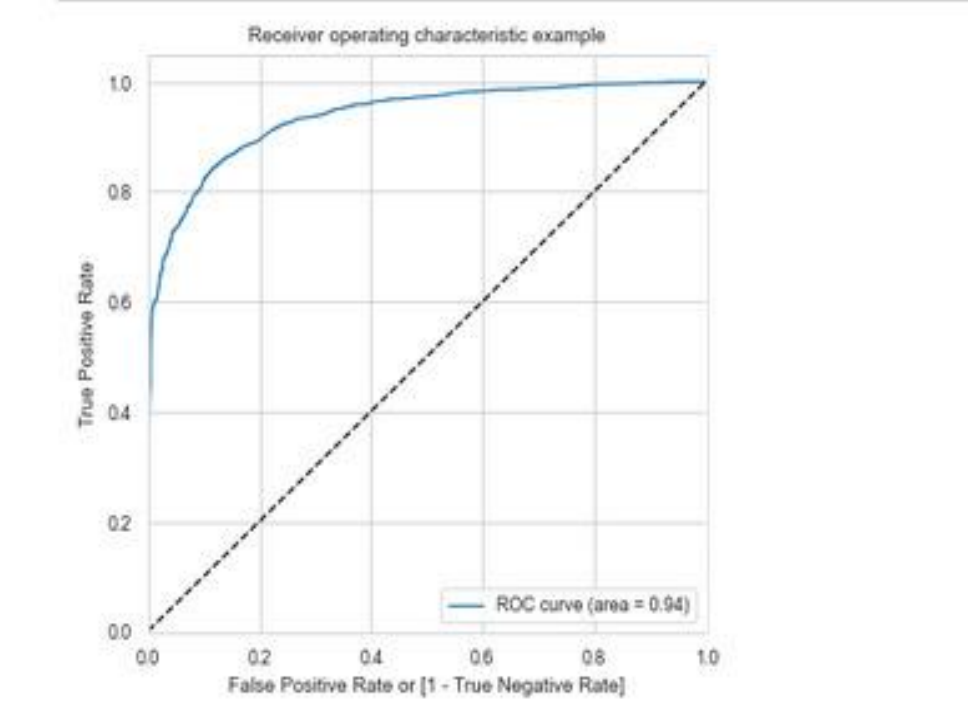
- After data cleaning we have 56 columns and 9240 data points.
- Now we will have to divide this data set into training and testing data set.
- We will divide this data set using train_test_split function of sklearn library
- Our 70% of data will go into training set and remaining 30% into testing set.

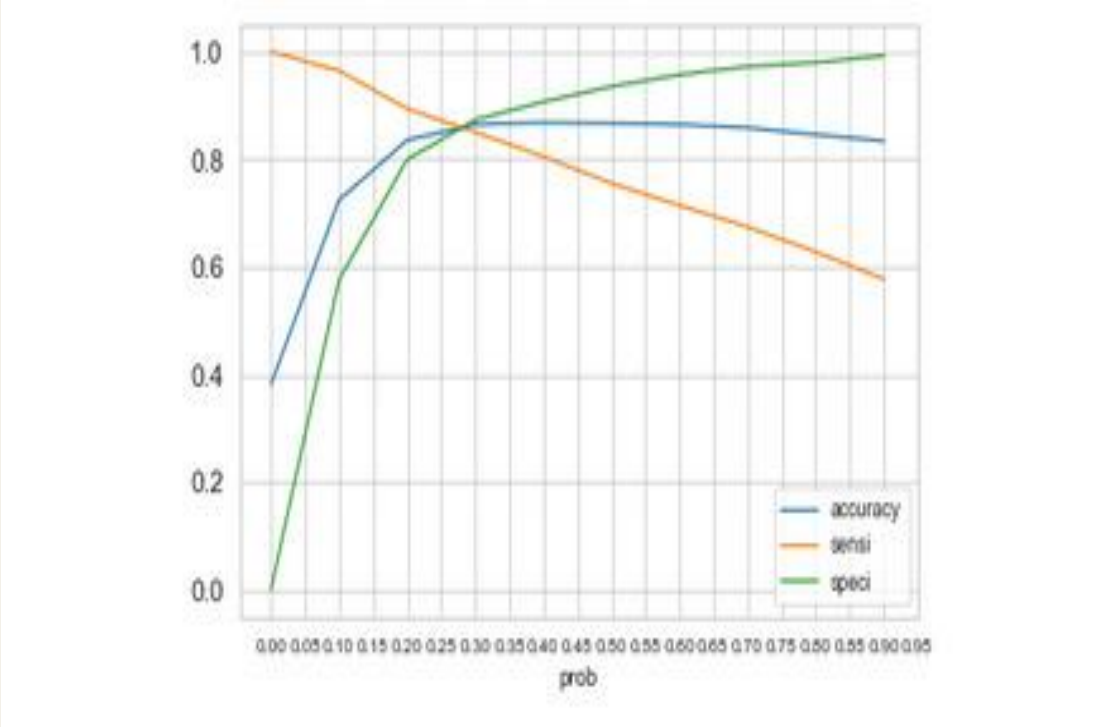
FITTING THE DATA INTO MODEL

- Now we have used logistic regression to create the machine learning model.
- We have used fit the training data to the model using fit function.
- We have used RFE for feature selection.
- We have also used decision tree.
- We have refined our model using GLM.

RESULTS

- Got the final accuracy of 87% on testing data.





CONCLUSION

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

- The total time spend on the Website
- Total number of visits
- When the lead origin is Lead add format.
- With these factors in mind, X Education can thrive since they have a very good probability of convincing practically all potential buyers to alter their minds and purchase their courses.