Lead Scoring Assignment

- By Ezazaehmad Paliza

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

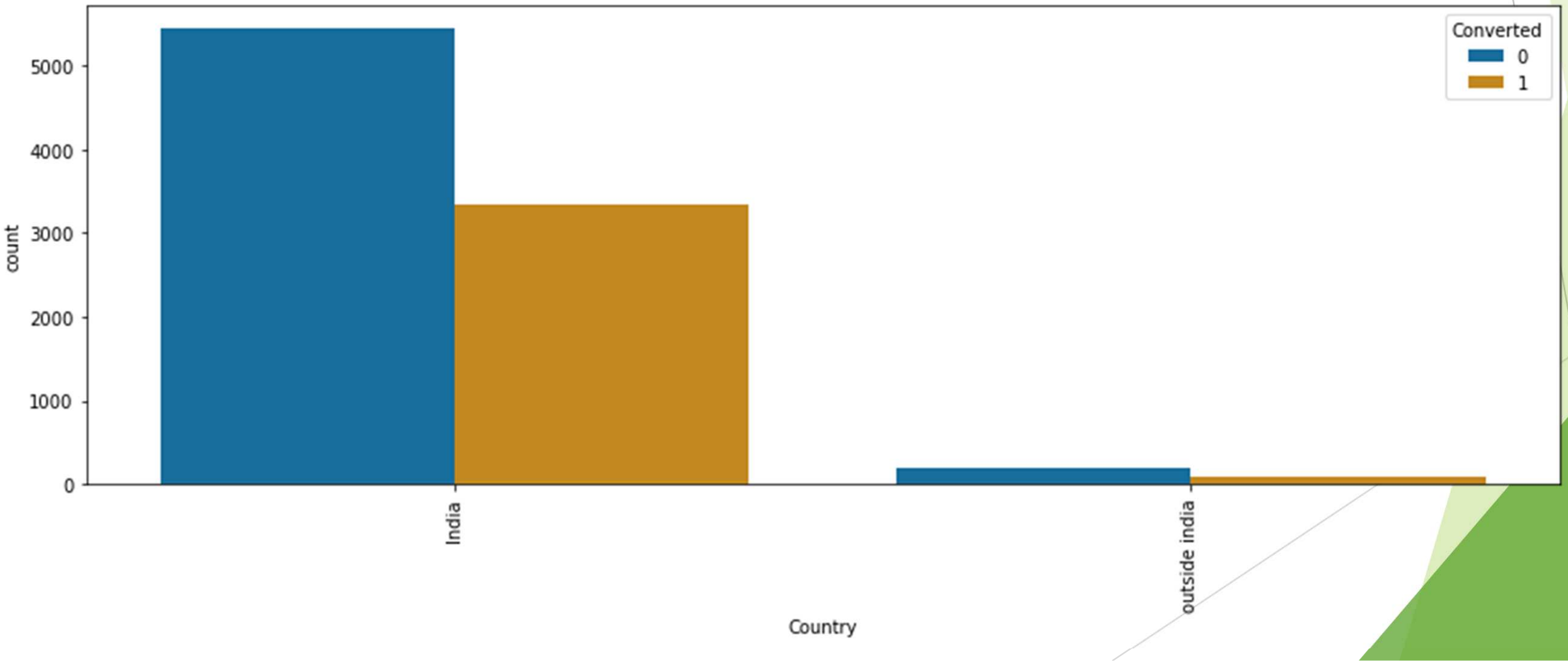
There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step.

Strategy

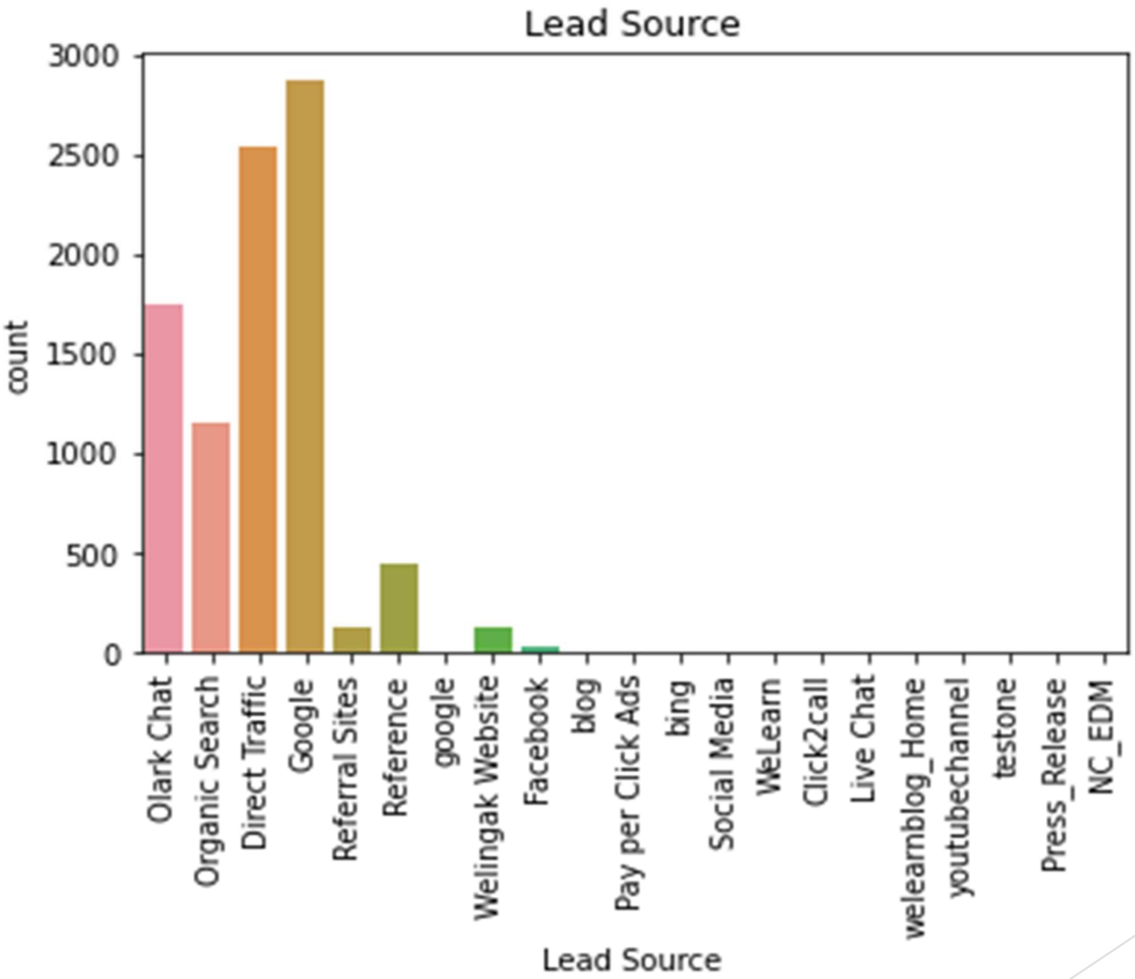
- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Exploratory Data Analysis

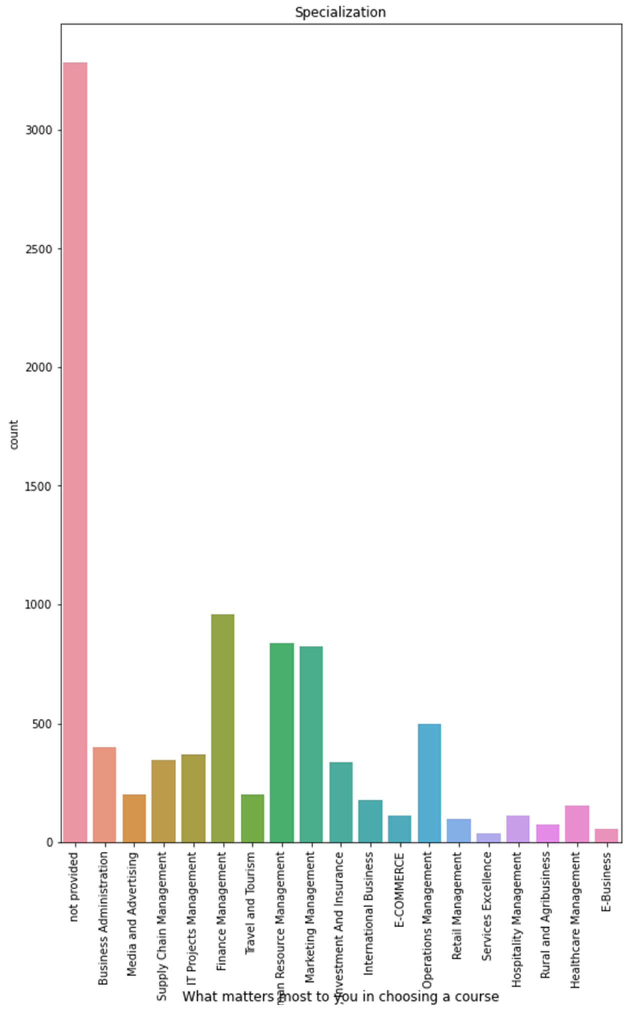
We have 98% leads from india

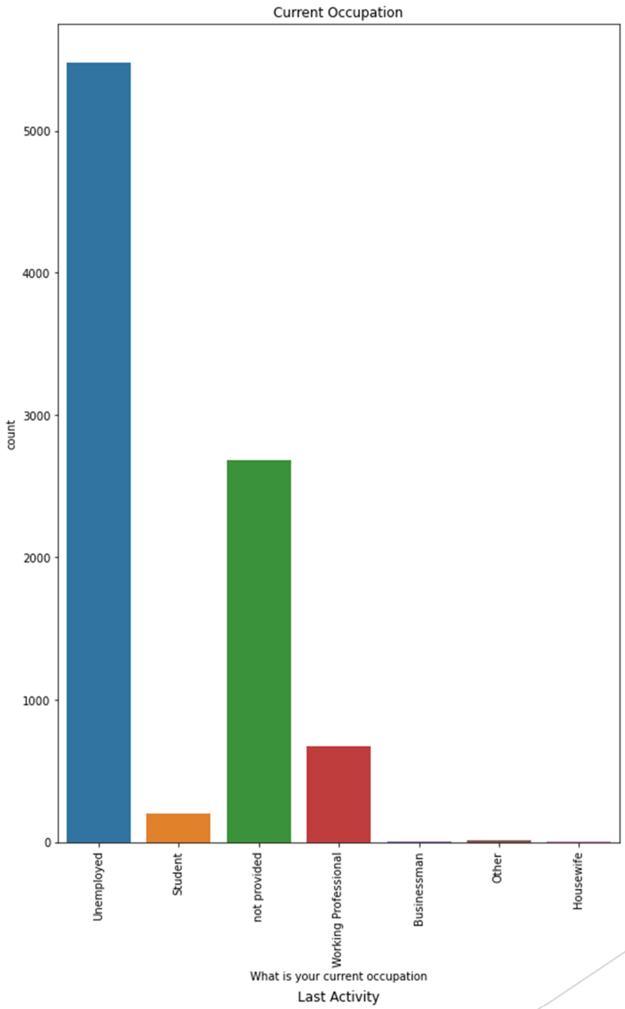


Lead source

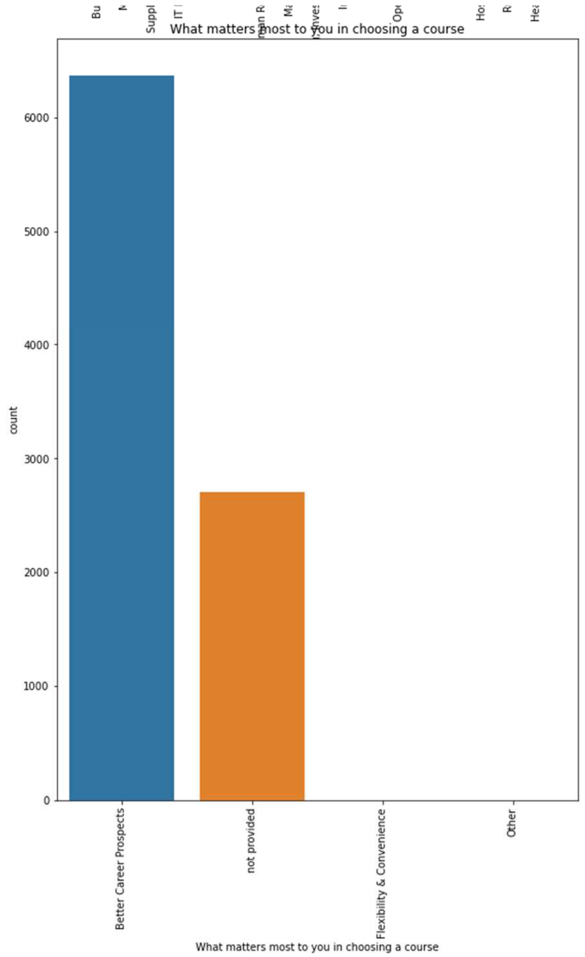

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

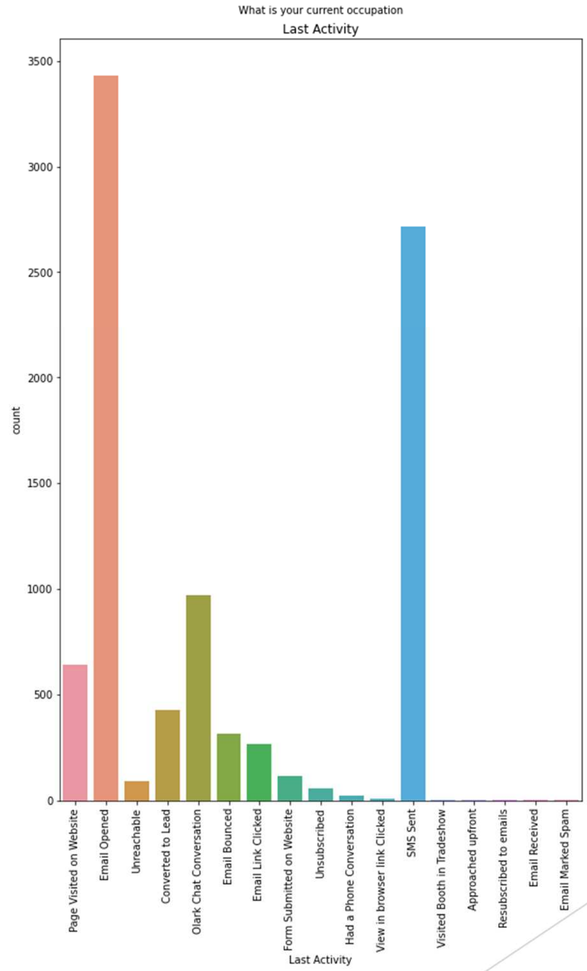
Specialization and current occupation



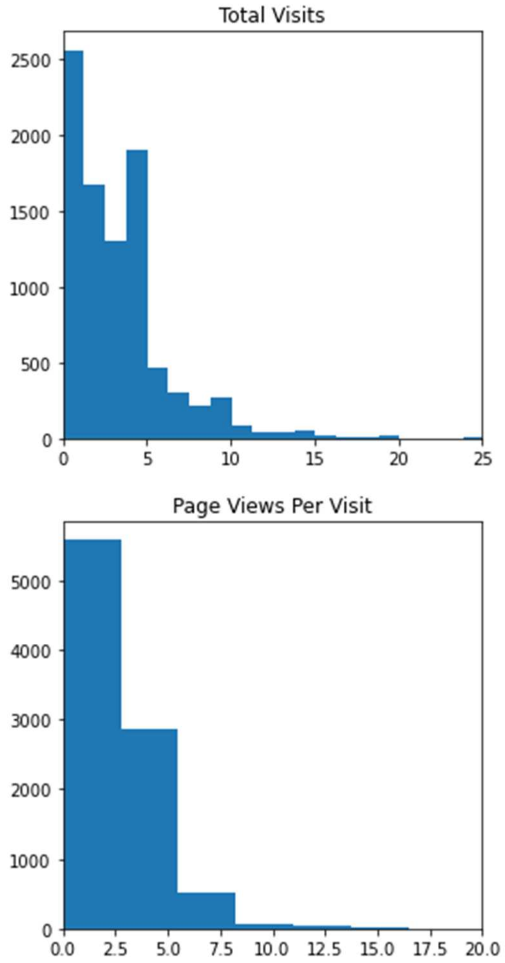


What matters most to choosing course and last activity





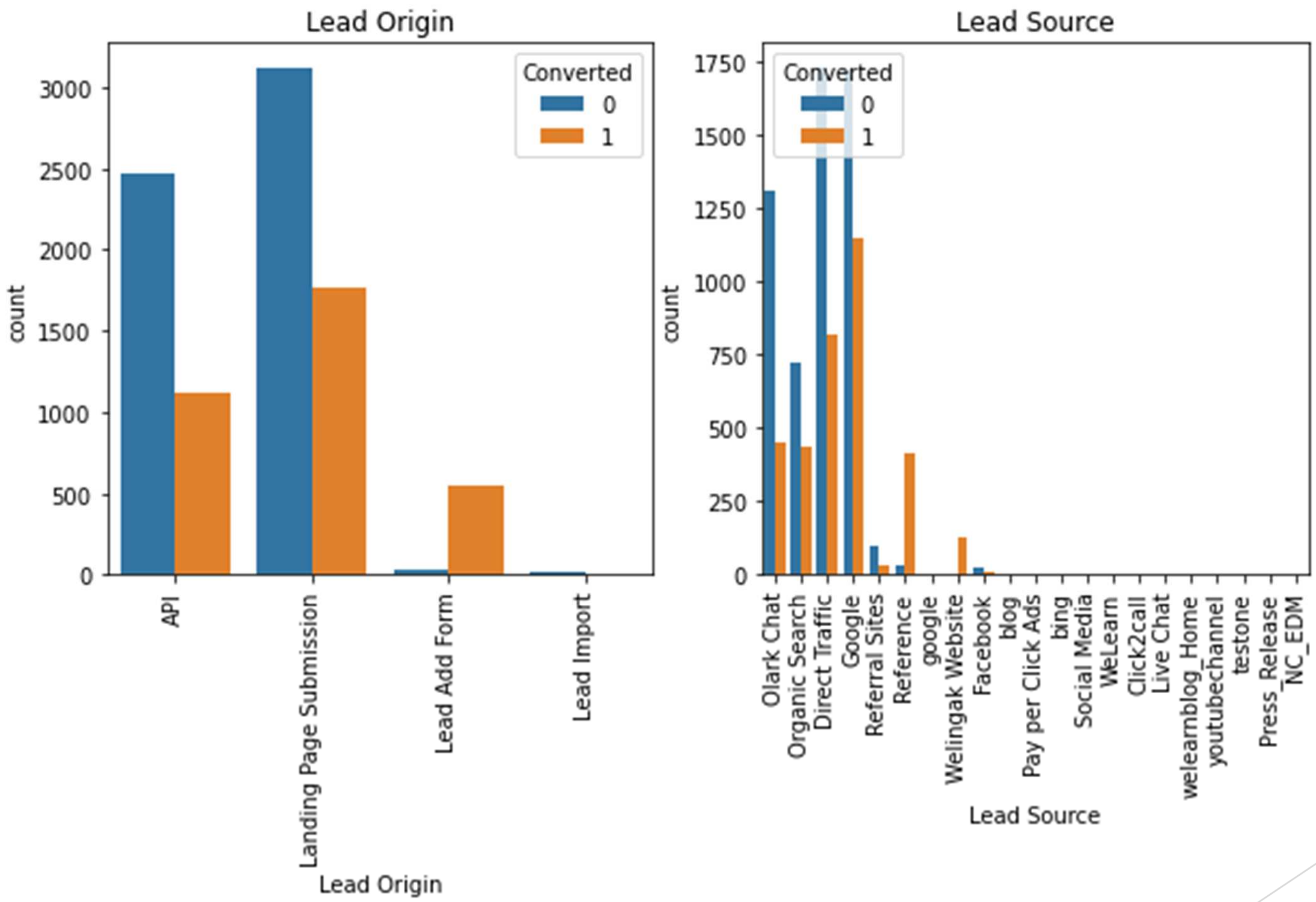
Numerical variables


Total Visits

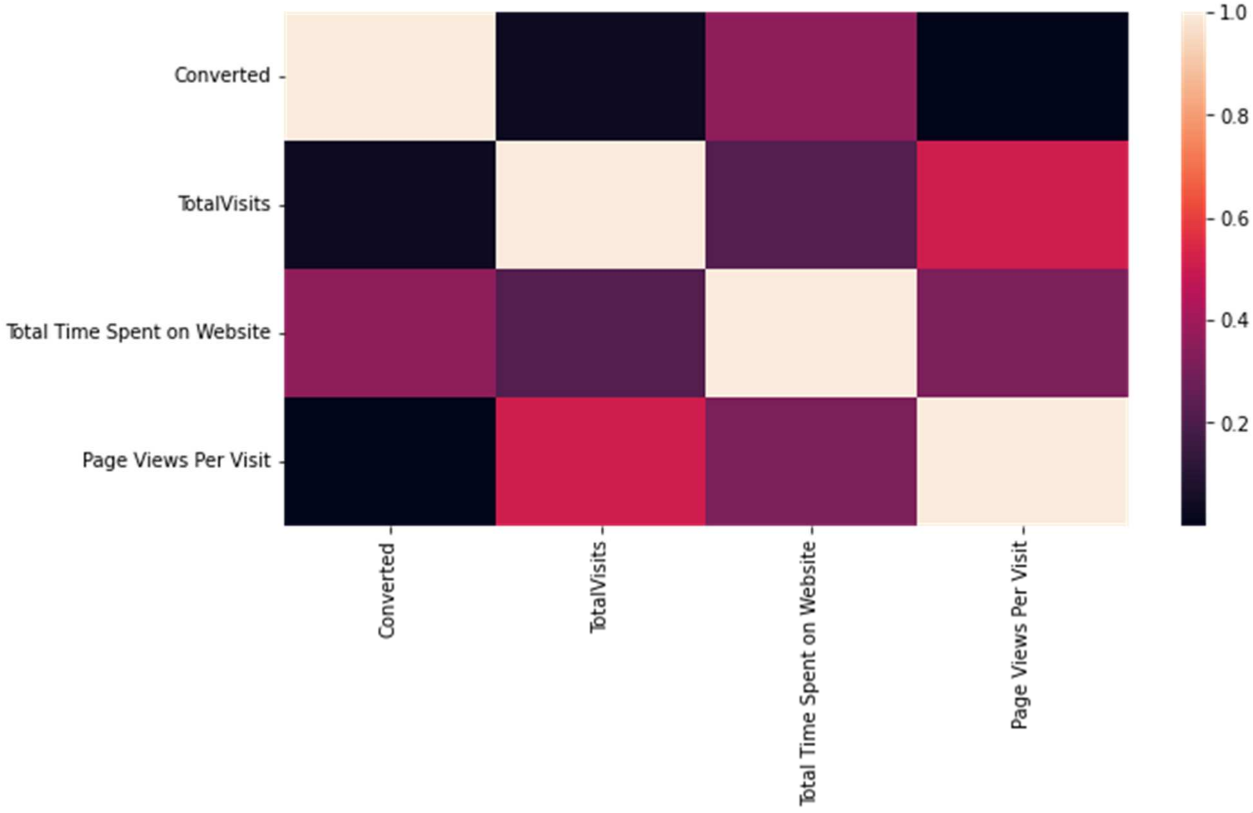
Total Time Spent on Website

Page Views Per Visit

Relating all the categorical variables to Converted

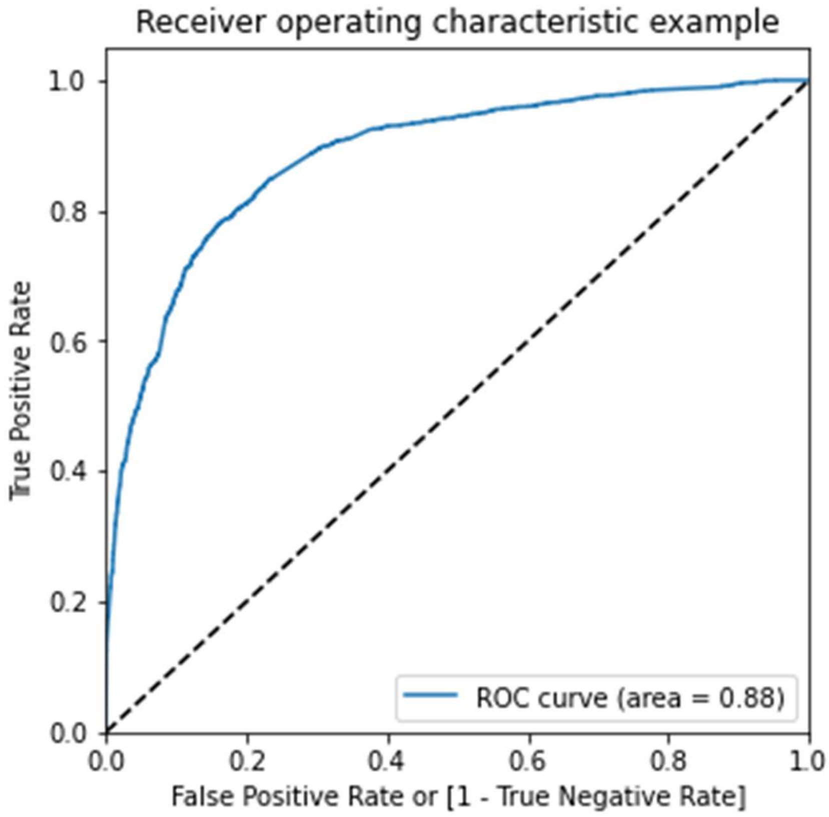


Correlation



ROC Curve

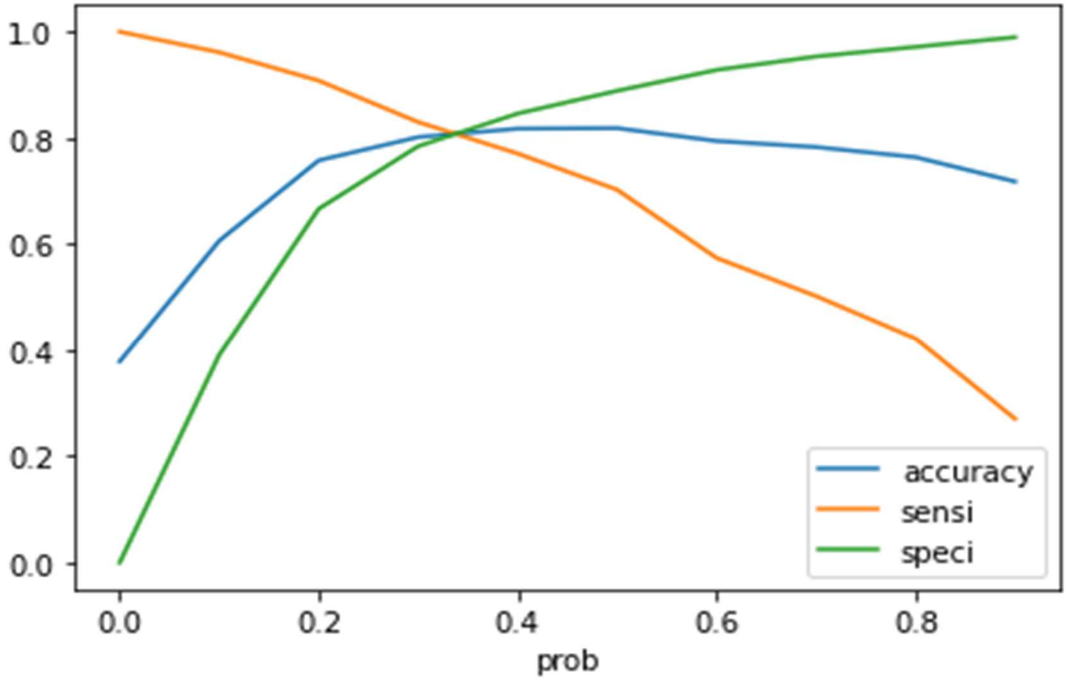
The area of ROC Curve is 0.88 which is very good.



Model Evaluation - Sensitivity and Specificity on Train Data Set

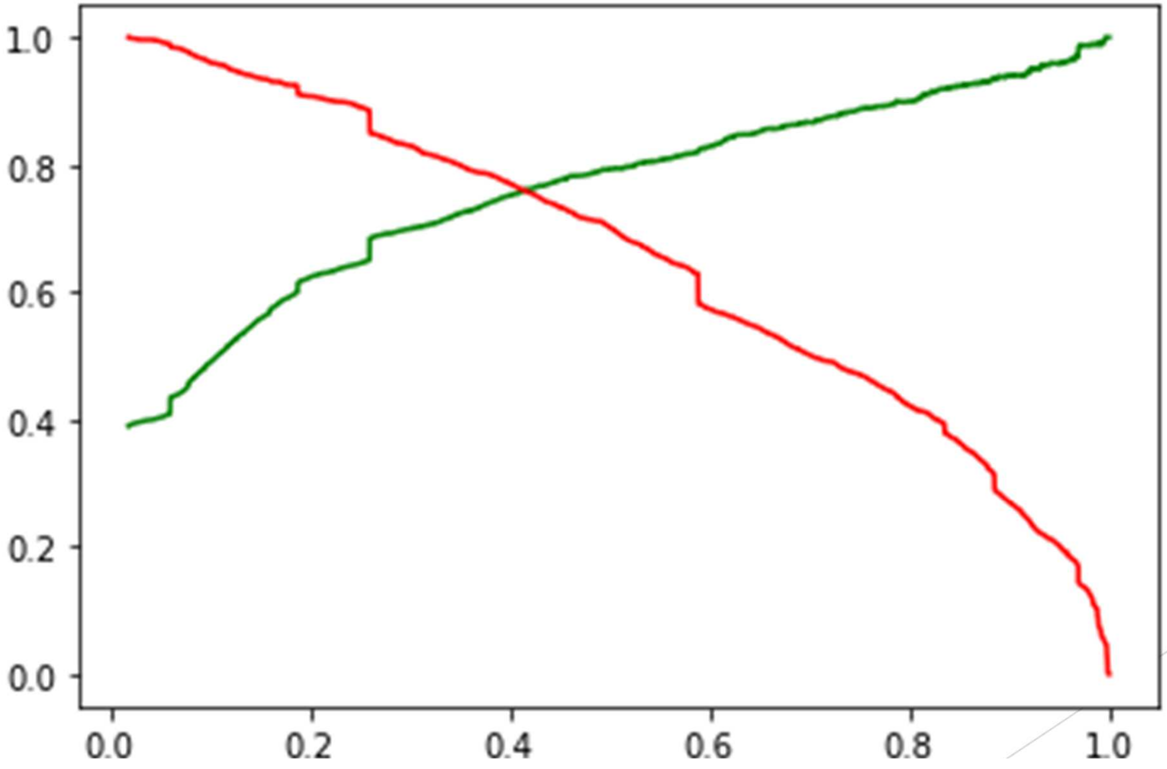
From the graph it is visible that the optimal cut off is at 0.35.

With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 80%



Prediction on Test set

With the current cut off as 0.41 we have Precision around 75% , Recall around 73% and accuracy 80.5%.



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_Welingak Website

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.