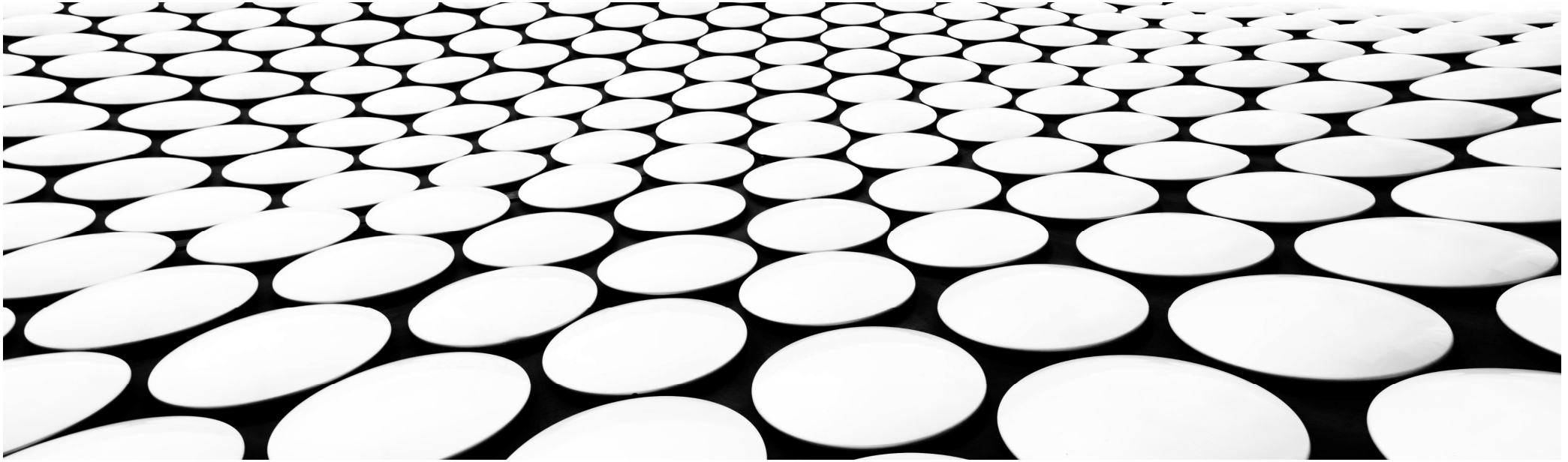
Lead score case study- X Education

SANGEETHA BEHRA

SHUBHAM GUPTA

SHUBHANGI DEOKAR



Problem statement

X Education sells online courses to Industry professionals. The company markets its course on several websites and search engine like Google.

Once buyer visit the website, they either browse the course offered by the website or fill up the form or watch some videos. Any customer filling up the form is considered as a lead along with past referrals.

Based on the leads, the sales team of X Education start making calls or reach out to the customers through various other channels. Through these various modes of conversations, some of the leads get converted and some may not.

As the conversion rate are around 30% , so it pretty worrying for the senior management and looking for option to increase this.

Business Goal

X Education looking for information to go for the leads which are promising enough to have a high probability of conversion that is to into a customer buying the courses.

X education is looking for a model from Data Scientist which can assign a score to a lead implying which lead has higher conversion rate and should be targeted.

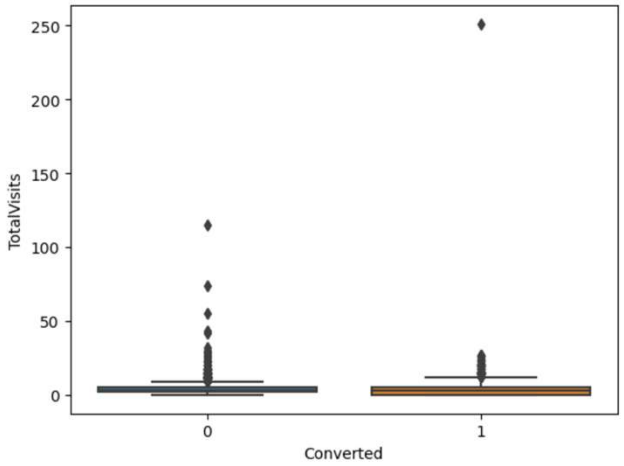
The CEO of the X education has given a ball[ark target of increasing the conversion rate to 80%.

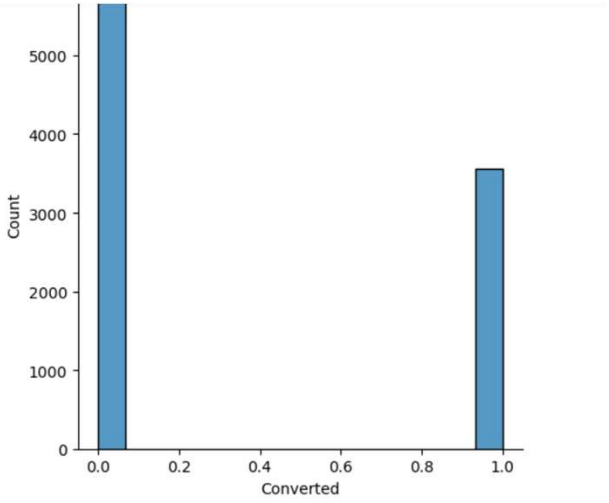
Approach taken

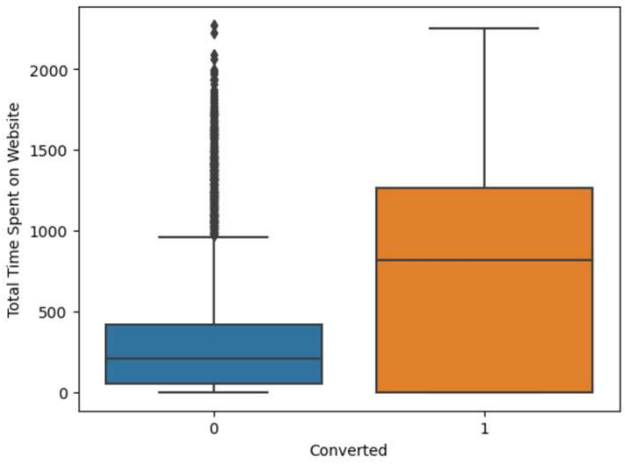
- Analyse the given data
- Apply EDA techniques for cleaning, preparing and analysing
- Spit: Train and Test the data
- Scale the data
- Build logistic regression Model and calculate the lead score
- Evaluate the model using metrics like Specificity, Sensitivity, Precision and Recall
- Do multiple iterations to create different models
- Chose the best model

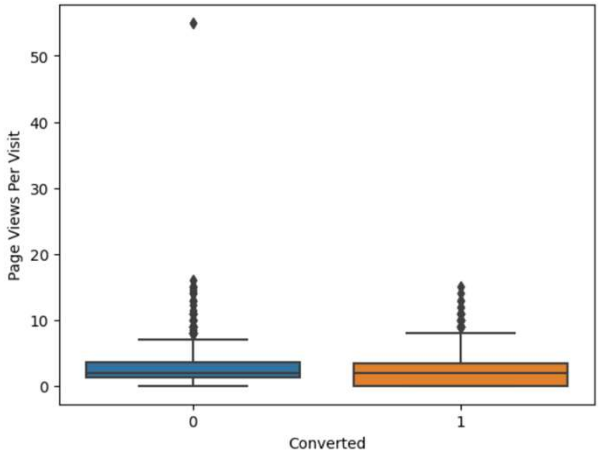
EDA : Exploratory Data Analysis

- Conversion ratio is 38%
- Nothing conclusive from comparing Converted vs Total Visits
- Lead spending more time could result in conversion
- Nothing conclusive from the pages visited per view



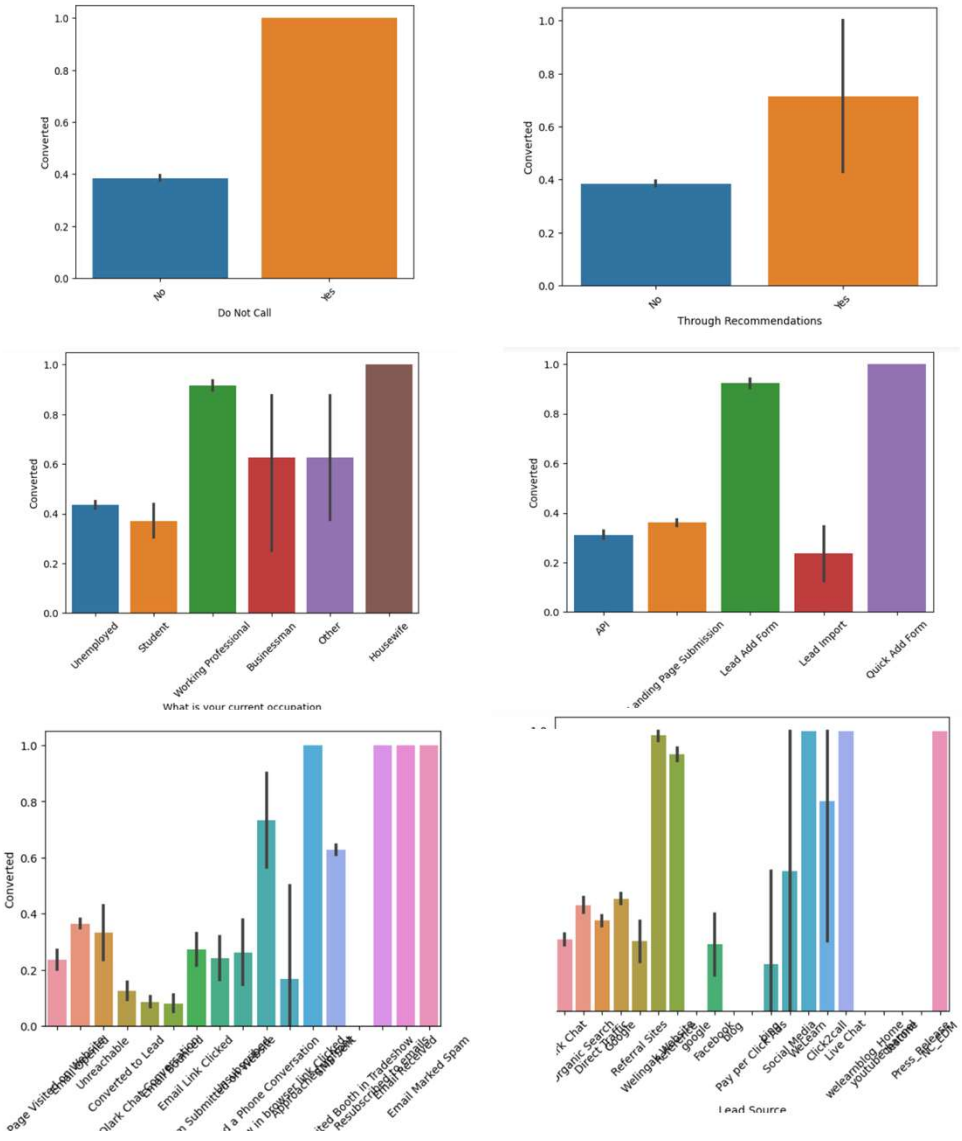






EDA : Numerical Analysis

- Maximum conversion happened on landing page
- Major conversion happened from email sent and call made
- Major conversion in the lead source is Google
- Not much impact through search, digital advertisement and recommendation
- More conversion happened people who are unemployed



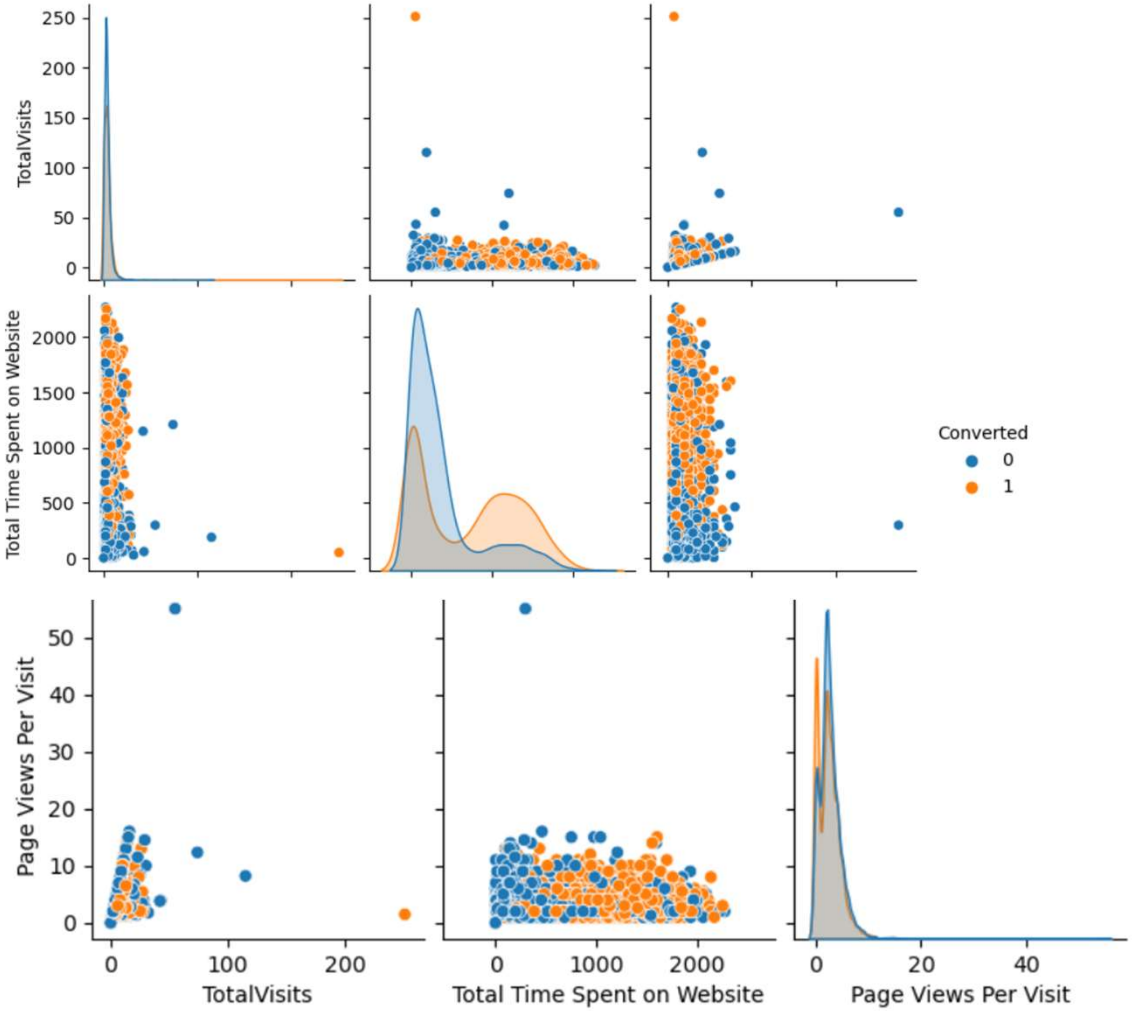
Exploratory Data Analysis continued…

- Checked the columns having null values and dropped following columns have more than 40% null values
  - Lead Quality,
  - Tags, Asymmetrique Activity Index
  - Asymmetrique Profile Index
  - Asymmetrique Activity Score
  - Asymmetrique Profile Score
- As mentioned in the case study dropped columns having maximum number of values as ‘Select’
  - City
  - Lead Profile
  - How did you hear about X Education
- Dropped Country as has only India as most of the values
- Dropping columns having unique values as can’t infer much from it
  - Lead Number
  - Prospect ID

EDA: Exploratory Data Analysis continued…

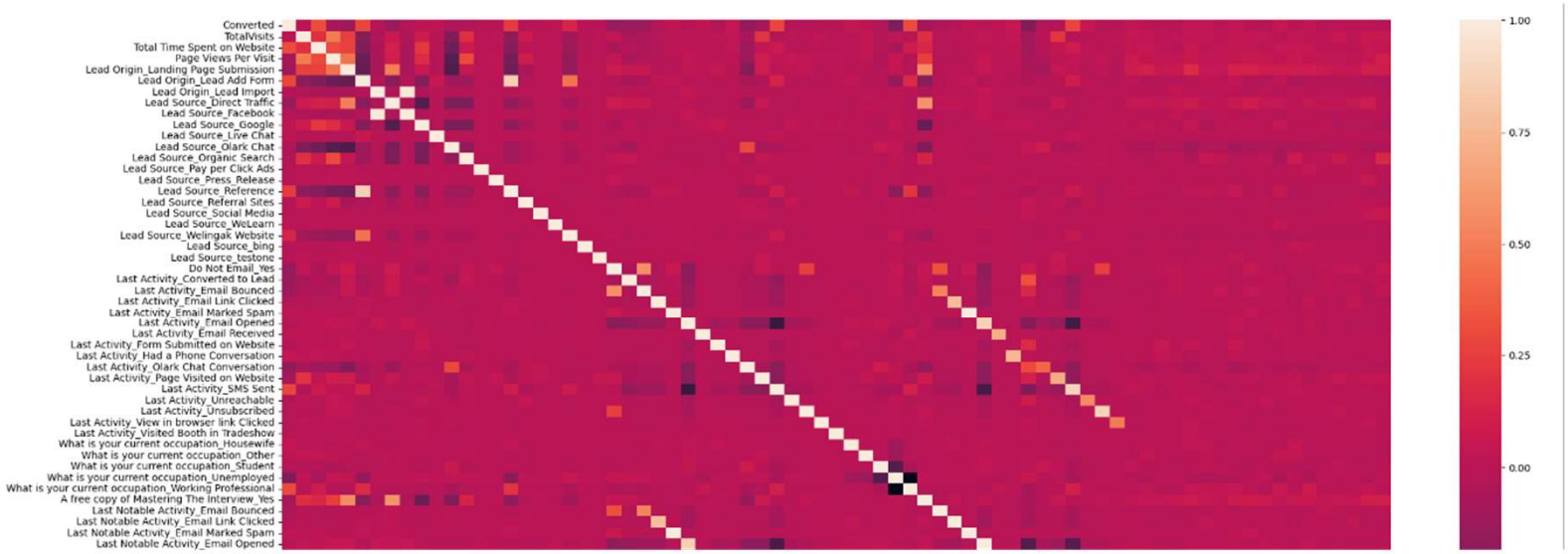
- Dropped columns having values just ‘No’
  - Search
  - Magazine,
  - Article, X Education Forums
  - Newspaper
  - Digital Advertisement
  - Through Recommendations
  - Receive More Updates About Our Courses
  - Update me on Supply Chain Content
  - Get updates on DM Content
  - I agree to pay the amount through cheque
  - Do Not call
- Delete rows of columns having null values not making much sense
  - What is your current occupation
  - TotalVisits’
  - Lead Source
  - Specialization

EDA : Exploratory Data Analysis



Data preparation for the Model

- Prepare data for Modelling
  - Using Fit and Transform functions
  - Dummy Variable creations
- Split the data for training(70%) and testing(30%) the model
- Scale the numeric variables which are at different scales
- From the heat map it was difficult to derive the corrrealation



Model Building

- As there are lot of features, chose RFE to select a small set of variables for building the model
- Using P-Values and VIF, eliminated the variables reducing the accuracy of the models
- Created Logistic Regression model after adding comstant
- Eliminated Lead Source_Reference, Lead Profile_Dual Specialization Student one by one after checking the VIF values
- Dropped What is your current occupation_Housewife, What is your current occupation_Working Professional one by one after checking the P-value

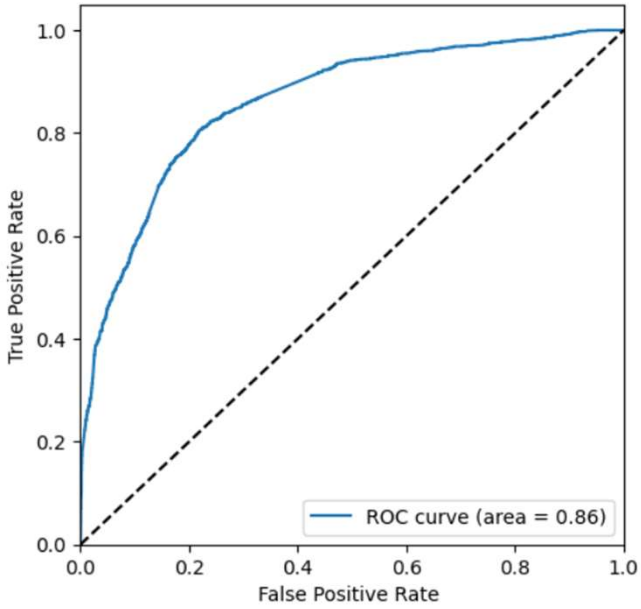
Model Evaluation

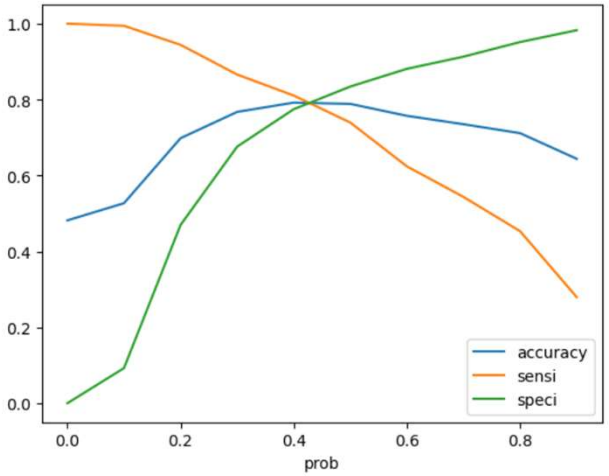
- Prediction was made using the final set of variables/features
  - Total Time Spent on Website
  - TotalVisits
  - Last Activity_SMS Sent
  - Lead Origin_Lead Add Form
  - Lead Source_Olark Chat
  - Lead Source_Welingak Website
  - Do Not Email_Yes
  - What is your current occupation_Student
  - Last Activity_Had a Phone Conversation
- If the probablity is >0.5, then the converted is marked as 1 or else 0
- Created confusion matrix
  - [[1929 383] [560 1589]]
  - Accuracy : 0.7886124187401928
  - Sensitivity: 0.739413680781759
  - Specificity : 0.8343425605536332

0.7908540685944856

Model Evaluation Continued …

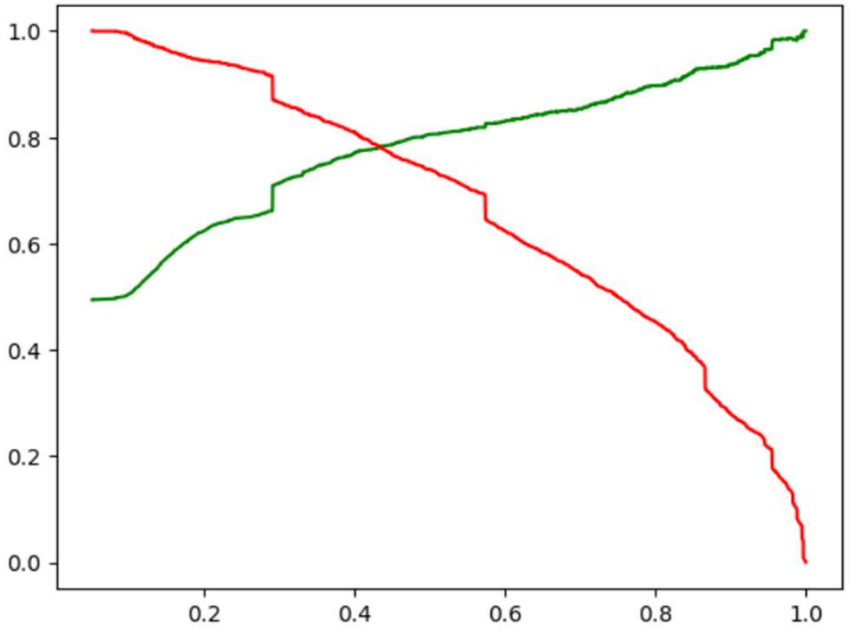
- Initially used 0.5 as the optimal cut off to draw the ROC curve
  - The area under the curve of the ROC was 0.86 which was quite good and seems to be good model
- The optimal value of three metrics came out to be 0.42, which wa considered as Optimal cut off.
- Created confusion matrix with optimal cut off as 0.42 gave
  - [[1823, 489], [444, 1705]]
  - Accuracy : 0.790854068594485
  - Sensitivity: 0.793392275476966
  - Specificity : 0.788494809688581 3





Making Predictions on the Test Set

- Made prediction on the test data using cut off as 0.42
  - Confusion matrix: [[786, 210], [202, 714]]
  - Accuracy : 0.7845188284518828
  - Sensitivity: 0.779475982532751 1
  - Specificity : 0.7891566265060241
- Precision-Recall View
  - Precision : TP / TP + FP = 0.8057809330628803
  - Recall : TP / TP + FN = 0.739413680781759
- Based on the test data the final prediction made are
  - [[801, 195], [213, 703]]
  - Accuracy : 0.7866108786610879
  - Sensitivity: 0.7828507795100222
  - Specificity : 0.7967467248908297



0.7908540685944856

Conclusions

- While we used both Sensitivity and Specificity as well Precision and Recall metrics. For final prediction we have used the optimal cut off based on Sensitivity and Specificity.
- Accuracy, Sensitivity and Specificity for test data has been 78%, 77% and 78% respectively, which is in line to the train data.
- In the final prediction, the lead score calculated, shows the conversion rate is 79% in train set and 78% in the test data set.
- The tope 3 attribute that help in converting the lead are
  - Total visits
  - Time spent on each pages during the visits
  - Last Notable Activity_Had a Phone Conversation
- The above metrics proves that overall model is quite good

0.7908540685944856

Recommendations

- Lots of leads were created in the initial stage, but while flowing through the funnel, it reduced drastically or didn’t get converted
- In order for higher lead conversion, X Education will have continuously/consistently engage the customer and keep on educating them on the various products
- List down all the potential customers from TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit who have high probability of being converted
- Keep these customers/leads engaged with new offers, jobs, services and future programs, studies.
- Carefully plan and deal each lead as per their interest.
- Focus on existing leads for repeat business
- Try to get right information based on QnA, inquiries, appoints to understand their needs and educate them