LEAD SCORING CASE STUDY

ANURAG AGARWAL
RITHESH PARAMESHWAR
LAKHWINDER SINGH

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%
- Now, although 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

- As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance

GOALS & OBJECTIVES

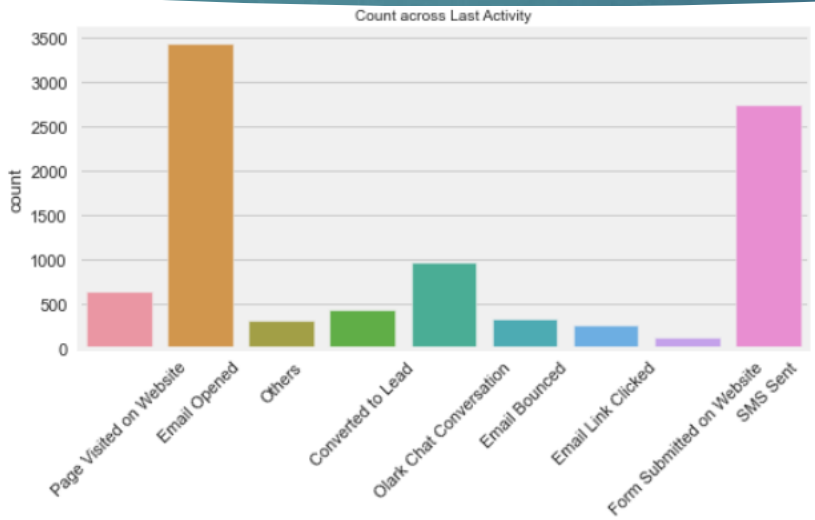
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will 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. Also, make sure you include this in your final PPT where you'll make recommendations

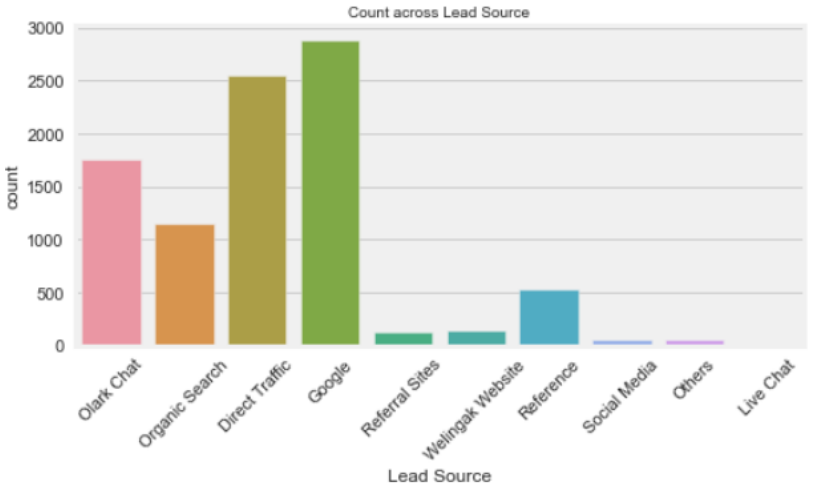
RESULTS EXPECTED

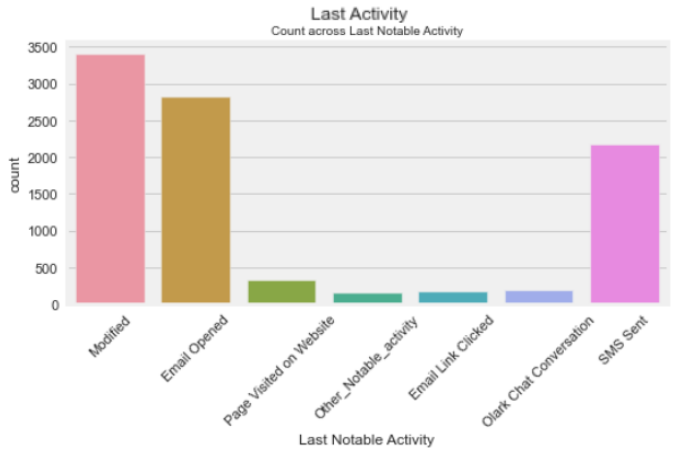
- A well-commented Jupyter notebook with at least the logistic regression model, the conversion predictions and evaluation metrics
- The word document filled with solutions to all the problems
- The overall approach of the analysis in a presentation
  - Mention the problem statement and the analysis approach briefly
  - Explain the results in business terms
  - Include visualizations and summarize the most important results in the presentation
- A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered

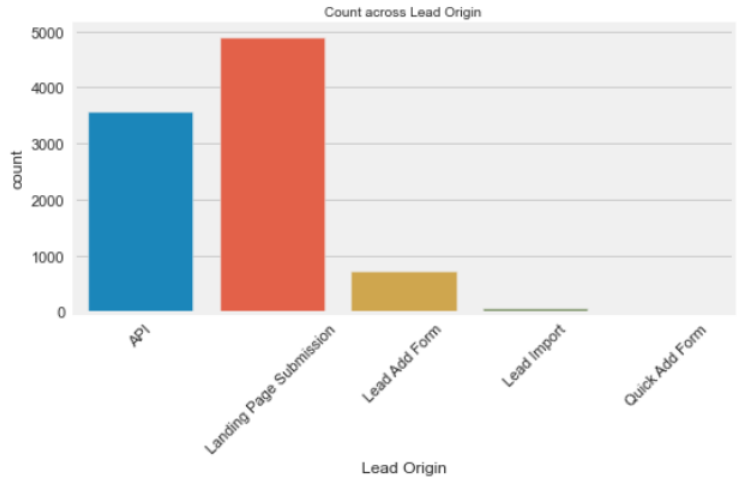
EDA

- It can be seen that Google and Direct Traffic are the two main sources of leads
- Landing page submission and API are the two main lead origins
- Prominent Last notable activity was ‘modified’ and ‘email opened’
- Prominent Last activity was ‘Email opened’ and ‘SMS sent’

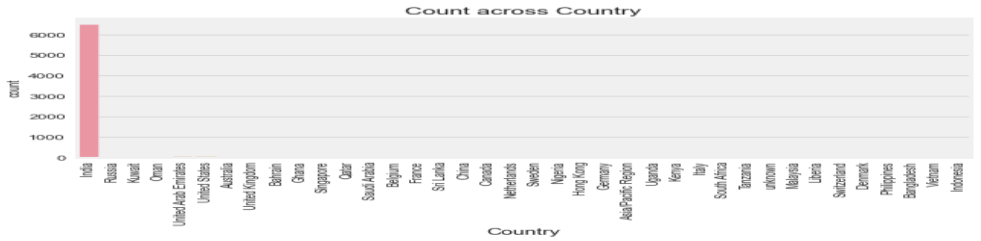


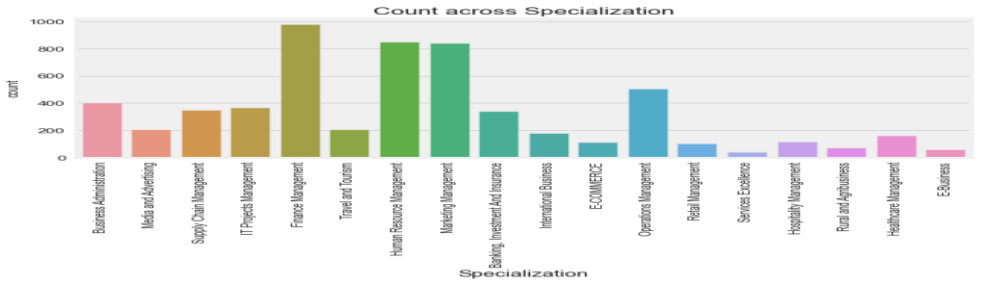


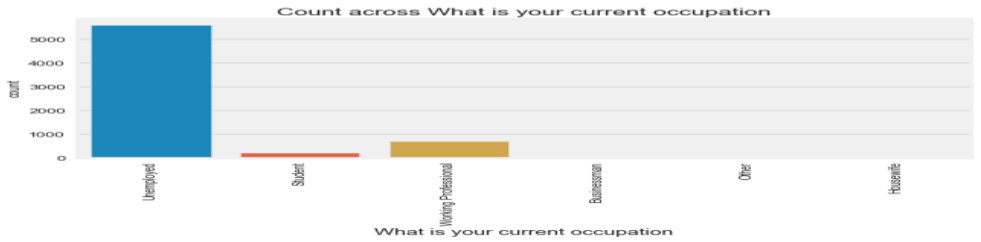


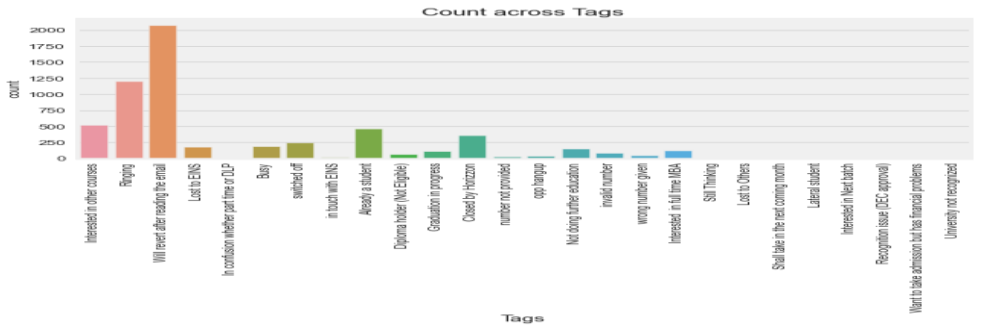


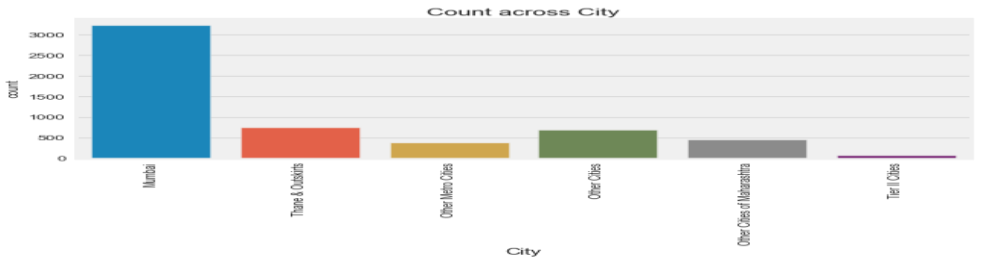
- It can be seen that most of the leads originated from India
- Most of the prospects are interested in Management related courses.
- Most leads generated are from Mumabi



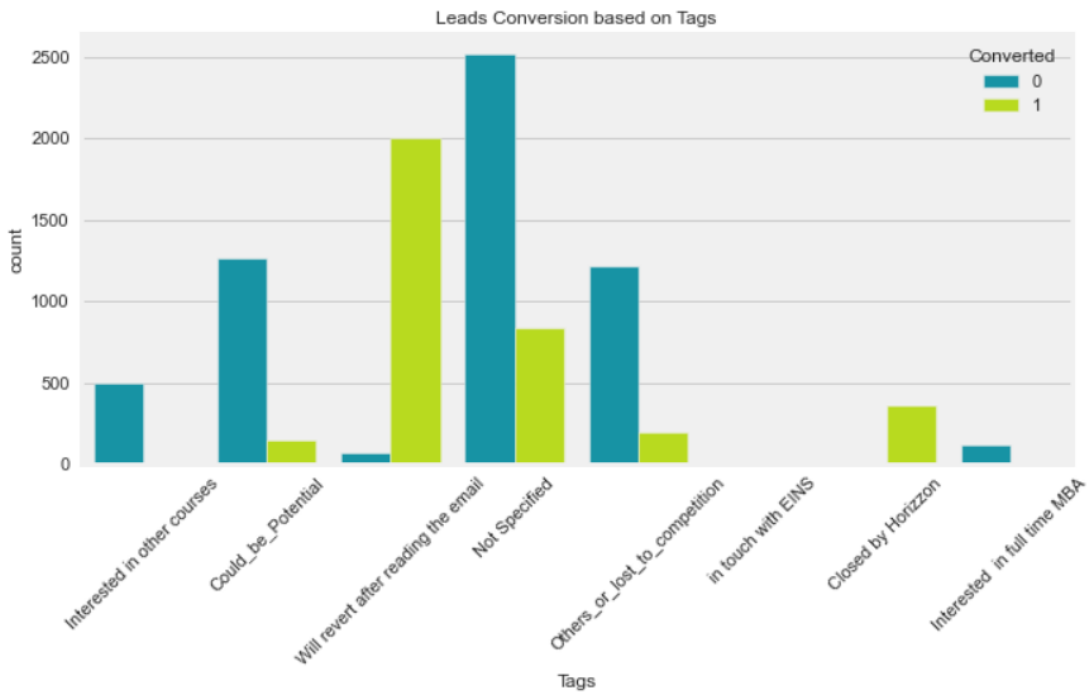


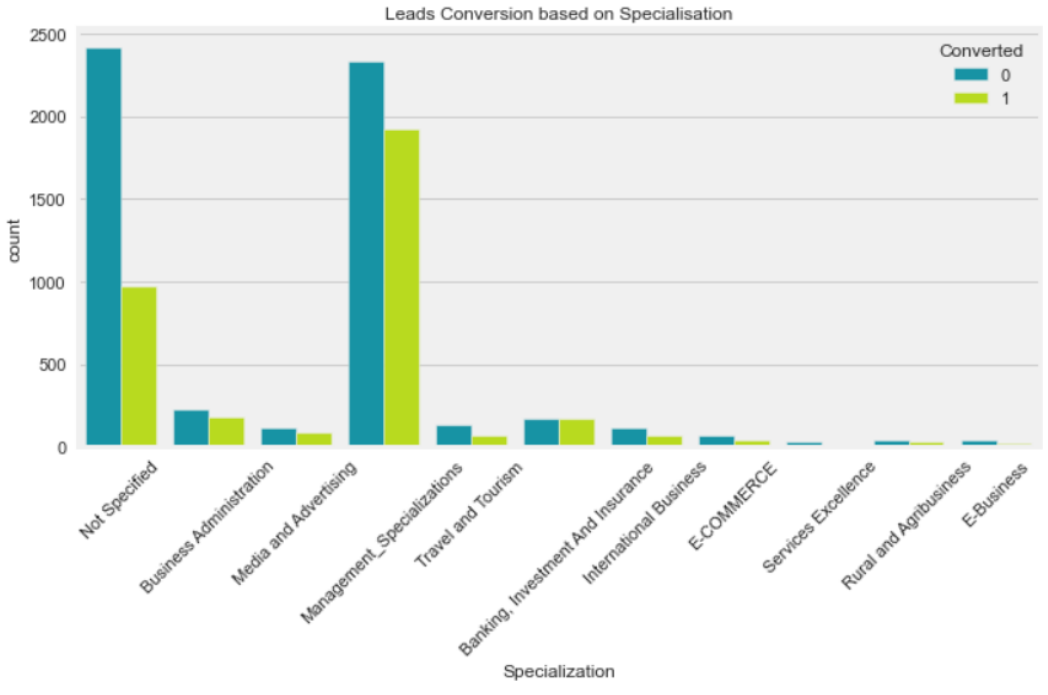


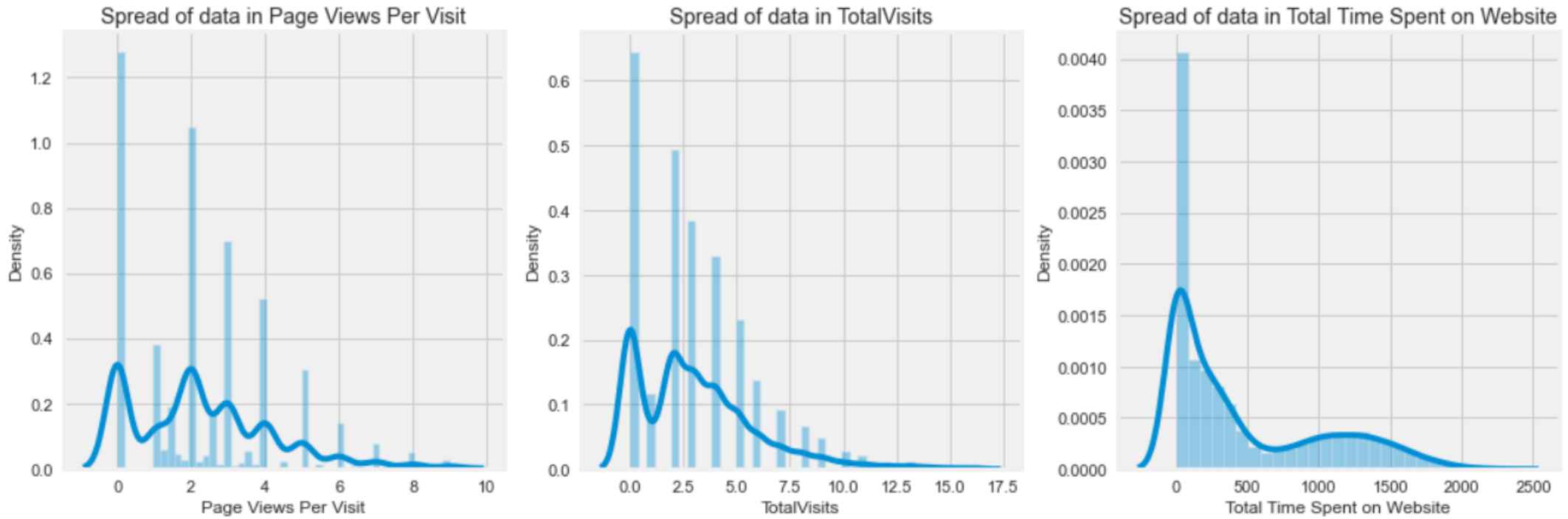




- Most converted leads belonged to management job specializations, whereas significant number of converted leads did not specify their specialization
- Most converted leads reverted after reading the email







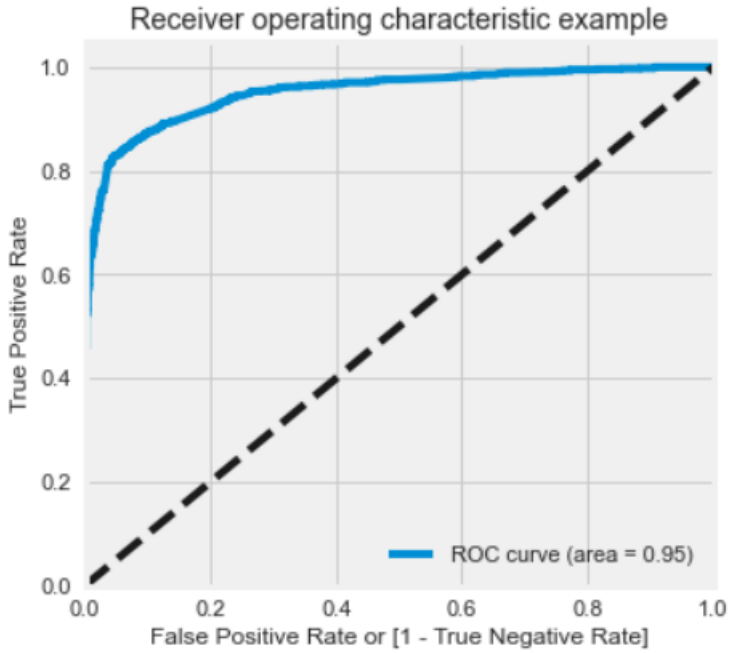
- Continuous variables like ‘Page Views Per Visit’, ‘Total Visits’ and ‘Total Time Spent on Website’ do not follow a normal distribution
- Outliers can be seen in ‘Total Time Spent on Website’ variable

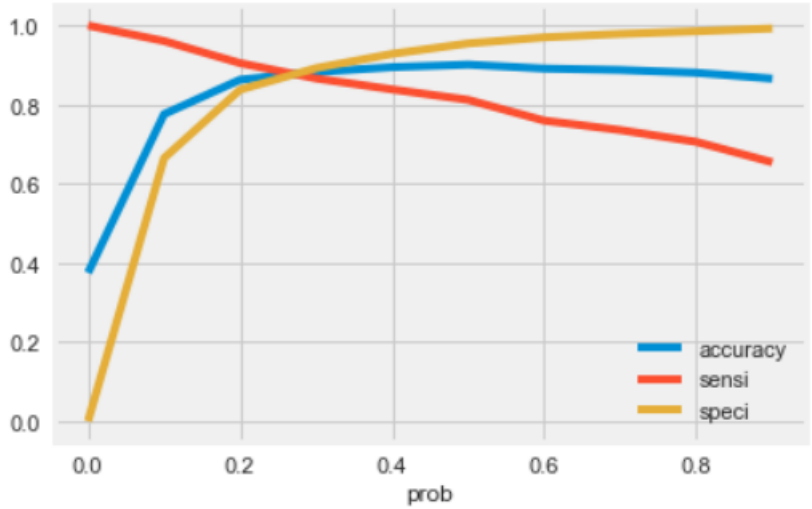
DATA CONVERSION

- Dummy variables are created for categorical features ‘Lead Origin’, ‘Lead Source’, ‘Do not email’, ‘Last Activity’, ‘Specialization’, ‘What is your current occupation’, ‘Tags’, ‘City’, ‘A free copy of Mastering The Interview’
- Variables with more than 40% null values are dropped
- Dataset is split into 70% train set and 30% test sets
- Recursive Feature Elimination (RFE) method is used for feature selection
- Top 15 variables are chosen using RFE
- Model is built by removing variables with p-values greater than 0.05 and RFE values greater than 5
- Predictions are performed on the test data set
- Overall accuracy of the model is found to be 89.41%

ROC CURVE

- ROC curve (receiver operating characteristic curve) is used to find the optimum point of all classification thresholds
- Probability of balanced accuracy, sensitivity and specificity is found to be approximately 0.3





RECOMMENDATIONS

- The good strategy is to focus on below Continuous and Categories or dummy variables as these features are impacting more on potential lead to be converted
  - Lead Source from Welingak website
  - Lead Origin Lead Add form
  - Lead Source with elements Olark Chat
  - Last Activity with elements SMS Sent
  - Working professionals
  - Time spent on website
- The focus should not be on the below features as it is not very relevant on getting a successful lead
  - Individuals who are interested in other courses
  - Individuals who are interested in full time MBA
  - With tag lost to competition