LEAD SCORE CASE STUDY

BY:
NIKHIL ALOK &
RAHUL HALLIMANI

Problem Statement:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, but its lead conversion rate is very poor (i.e. 30%). So, if 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 Logistic Regression Model which identifies the hot leads.
- Deployment of the model for the future use.

Approach:

- Data Cleaning and Manipulation:
  1.) Check and handle null/missing values
  2.) Drop columns: if it contains large number of missing values and is insignificant for analysis
  3.) Drop rows: if the column contains less number of missing values and is significant for analysis
  4.) check and handle outliers in the data
- Data Preprocessing:
  1.) Create dummy variables for categorical variables
  2.) Train-Test split
  3.) Feature Scaling
- Modelling:
  Logistic regression is used for making the classification model
- Evaluating the model using ROC curve
- Making prediction of the test set
- Deciding the optimum cut-off of probability using *Sensitivity-Specificity trade off* and *Precision-Recall trade off*.

Data Cleaning:

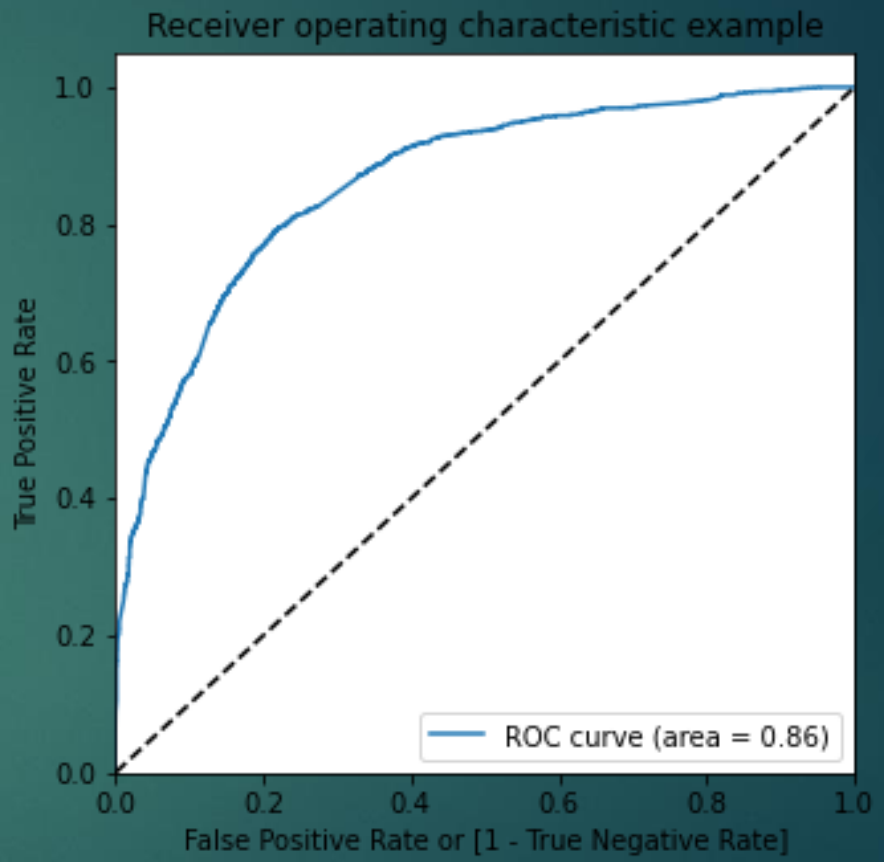
- Firstly we dropped the columns ‘city’ and ‘country’ right off the bat; since it’s not of any significance as X Education is an online platform.
- Then we dropped ‘Lead Profile’ and ‘How did you hear about X Education’ as it had huge number of ‘select’ values in them, which is practically as good as null values.
- After that we dropped 13 other categorical columns ('Do Not Call', 'Search', 'Magazine', 'Newspaper 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’, 'What matters most to you in choosing a course') as they have very little to no variety in them. These variable wouldn’t have contributed in the model.
- We also dropped ‘Prospect ID’ and ‘Lead Number’ as they are useless from analysis point of view.
- Lastly, for columns that have less number of missing values (i.e. ‘Specialization’, ‘Lead Source’, ‘Total Visits’, ‘What is your occupation’), we dropped the rows containing these missing values.
- Finally, after all these steps, we were able to retain ~69% of the records.

Data Preparation:

- We created dummy variables for all the categorical variables and dropped the original column
- For variable ‘Specialization’ we took special care; we didn’t just use drop_first to drop one of the dummy variable, but rather we specifically dropped Specialization_Select column (since ‘Specialization’ still had ‘select’ level in it).
- Then we did the Train-Test split (70-30) using random state as 42.
- After that we normalized the numerical variables using MinMaxScalar.

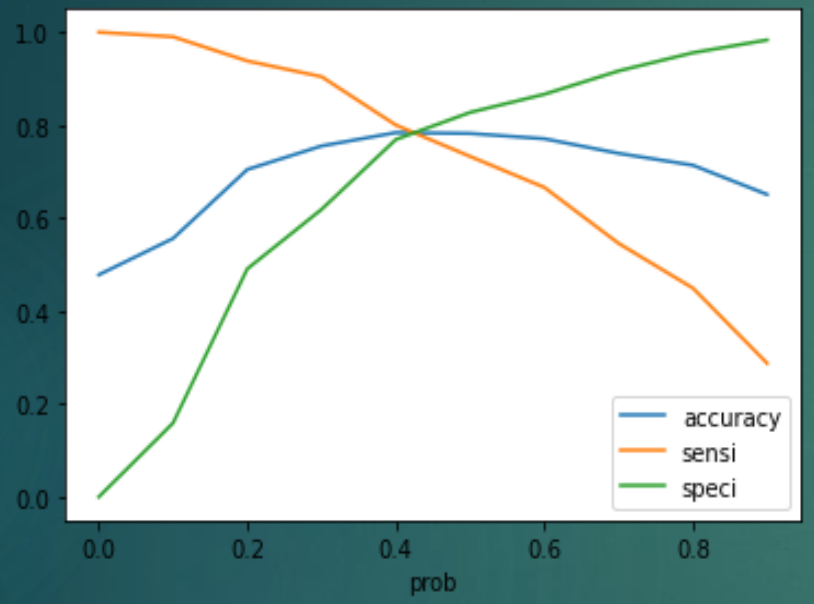
Model building:

- Firstly we did Feature Selection using RFE utility present in SciKit Learn (selecting top 15 feature coarsely).
- After that we used Statsmodel the model and manually dropped feature that had p-values > 0.5 and VIF > 5
- Then we evaluated the model using ROC curve an checking the AUC which came out to be 0.86



Finding the optimal cut-off:

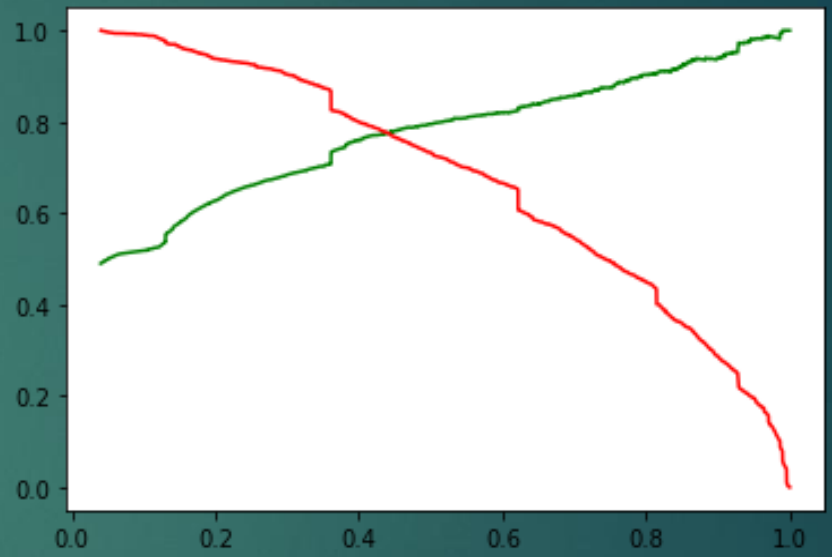
- Using Sensitivity-Specificity trade-off:



Using the cut-off of 0.42, we found:

Accuracy, Sensitivity and Specificity of ~80% each on the test set.

- Using Precision-Recall trade-off:



Using the cut-off of 0.44, we found:

Accuracy of 79%, Precision of 80% and Recall of 77% on the test set.

Conclusion:

It was found that the variables that mattered the most in deciding the potential buyers are:

- Lead Source_Welingak Website
- Total time spent on website
- Total Visits
- Lead Source_Reference
- What is your current occupation_Working Professional

Keeping the above factors in mind, the sales team can now approach these hot leads selectively instead of going after every lead that is generated.