Lead Scoring Case Study

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%. The company has invited us to improve the Lead Conversion Rate.

Analysis Approach

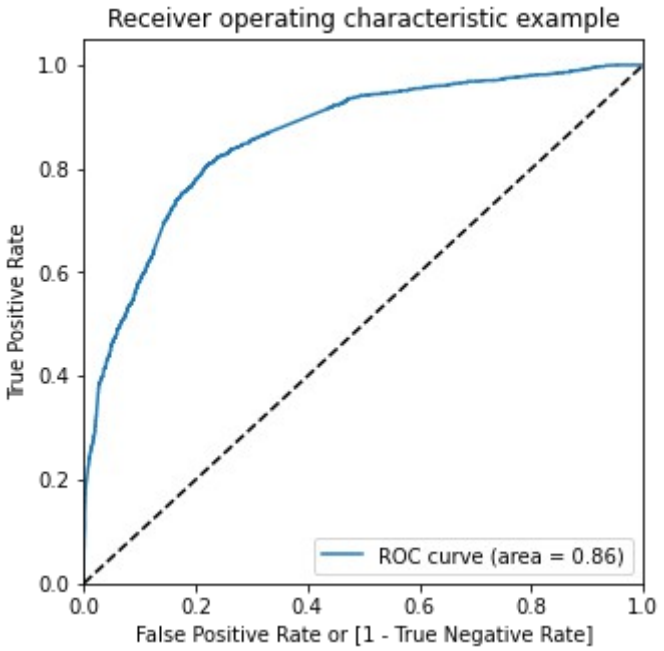
- We inspected the data in the data set and performed data cleaning and preparation.
- This included managing null values and columns which had a large no of null values
- This was done under EDA process
- Next we proceeded to creating a ML model.
- We dealt with the categorical variables by creating dummy variables.

Analysis Approach contd

- Then using the RFE process to eliminate the statistically unimportant variables we executed the Logistic Regression algorithm on the normalised data and built the model.
- Checking the correlation through VIF score and the p values we were left with the most optimum model.

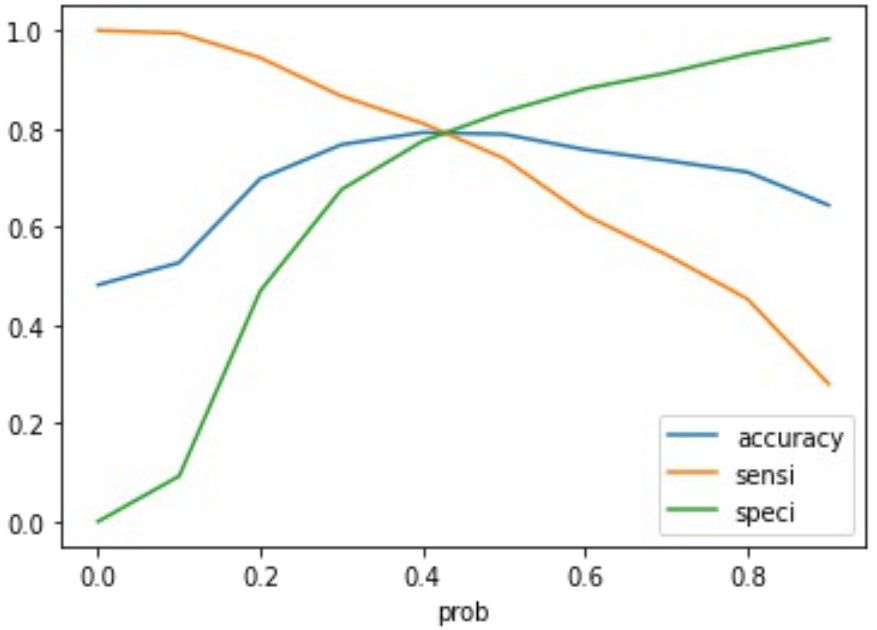
Results and Evaluation

- We found out the ROC AUC value(=0.86) as following which is satisfactory.



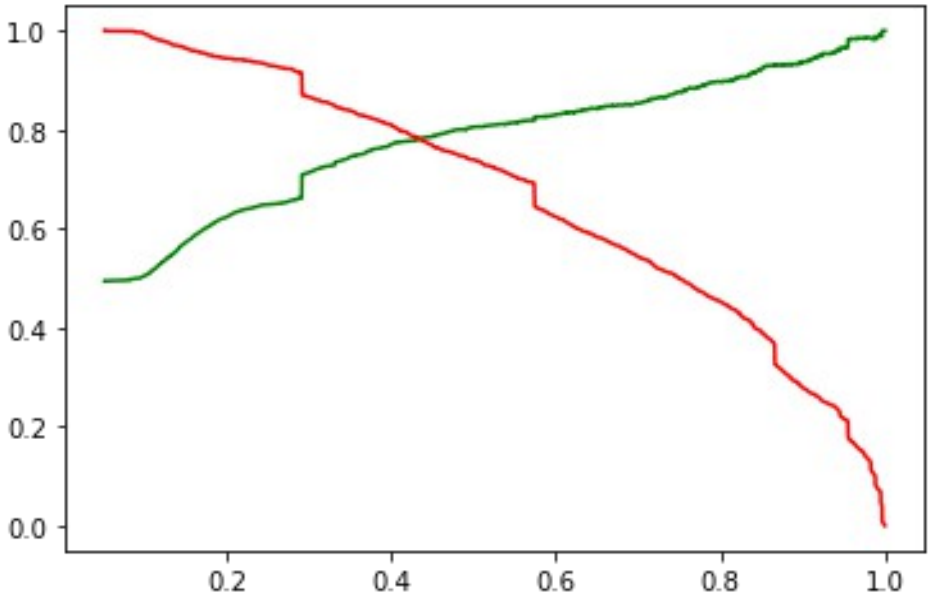
Threshold Probability

- To find out the threshold probability(0.42) which is the optimum probability of Logistic Regression where we achieve optimal values of Specificity, Sensitivity and Accuracy. We achieved this by plotting the Probability and the Specificity, Sensitivity and Accuracy as follows :
- The accuracy of the model calculated here was marginally higher than with Precision Recall view



Precision Recall view

- Next, we used the Precision Recall view to find out the optimum probability which came out to be 0.44
- The accuracy of the model calculated here was marginally lower.



Accuracy, Specificity and Sensitivity

- After using the threshold probability we could calculate the above evaluation metrics of the model by using the model on the test set. They are as follows:
- Accuracy : 78%
- Sensitivity : 77.9%
- Specificity : 78%

Precision Recall View Evaluation Metrics

- By using the model on the test set for prediction, we derived the following evaluation metrics based on precision recall view:
- Precision : 78%
- Recall : 76%
- Accuracy : 78%

Results

- The results of the model are comparable with both values of threshold probability.
- Through the results of the model and the process of RFE we derived the most relevant variables which affect lead conversion.
- These are :

Total Visits, Total Time Spent on the website, Lead Origin_Landing Page Submission , Lead Origin_Lead Add Form , Lead Source_Olark Chat , Last Notable Activity_Had a Phone Conversation, Last Activity_SMS Sent What is your current occupation_Unemployed

- These variables show the features that help to understand whether a lead will be converted or not in the most statistically significant way and linear manner.