Lead Score Case Study for X Education

Problem Statement :

X Education sells online courses to industry professionals. 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:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to get converted into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

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.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Problem solving methodology

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Selection.

Feature Scaling and Splitting Train and Test Sets

- Min-Max Scaling of Numeric data
- Splitting data into train and test set.

Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity to evaluate the model.

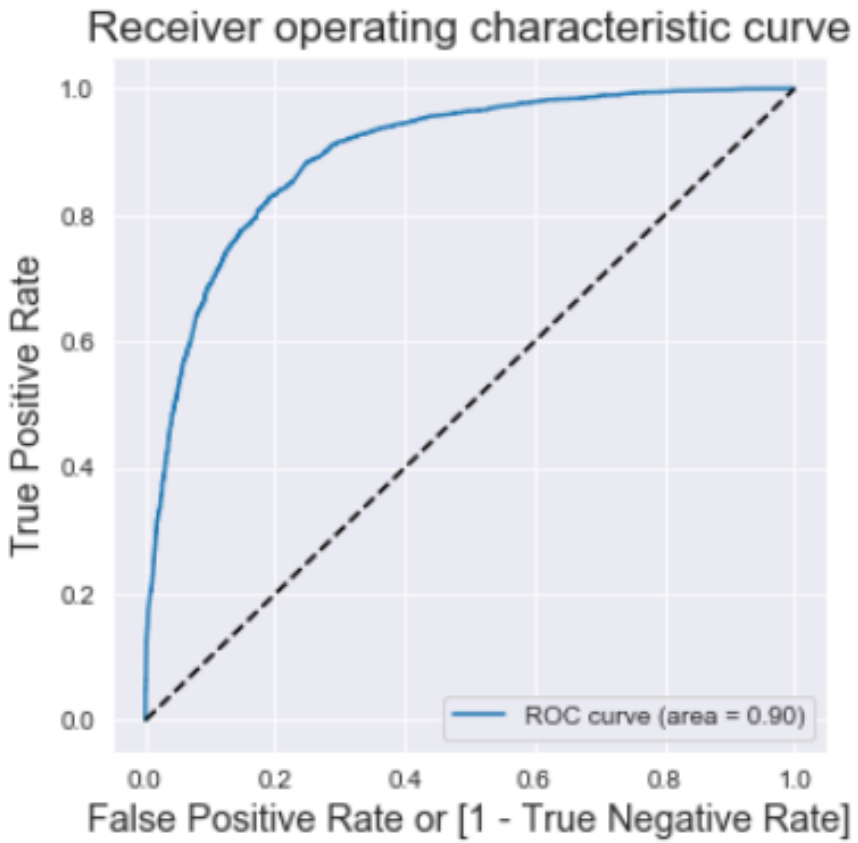
Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Variables Impacting the Conversion Rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Add Form
- Lead Source – Olark Chat
- Last Source – Welingak Website
- Last Activity – Had a Phone Conversation
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – Housewife
- Current Occupation – Working Professional
- Last Notable Activity – Subscribed
- Last Notable Activity – Unreachable

ROC Curve

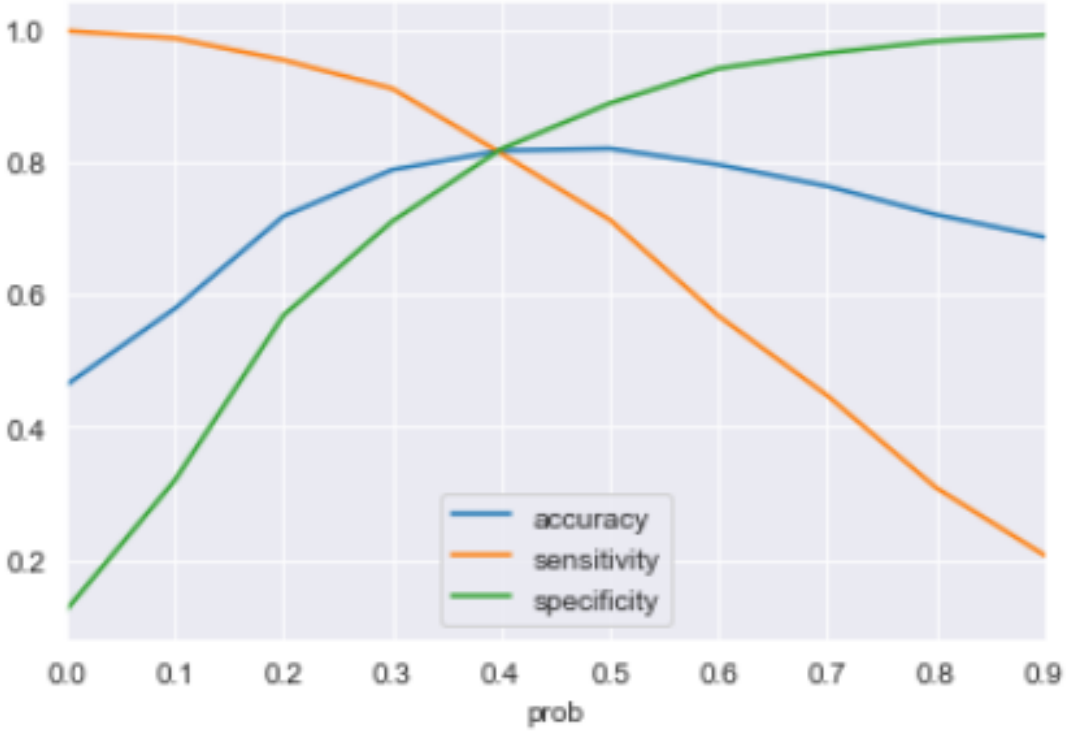


As we can see from the graph on your left, we have plotted a ROC curve. We can see from the graph that the area under the curve is 0.90

So we can say that The accuracy of the model seems to be good.

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.40 based on Accuracy, Sensitivity and Specificity



Confusion Matrix

3206	706
465	2021

- Accuracy - 81%
- Sensitivity - 81 %
- Specificity - 82 %
- False Positive Rate - 18 %
- Positive Predictive Value - 74 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

Confusion Matrix

3206	706
465	2021

- Accuracy - 82 %
- Sensitivity - 80 %
- Specificity - 83 %

Conclusion

- We have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 82%, 80% and 83% which are approximately closer to the respective values
  calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 81% (in train set) and 80% in test set
- The top 3 variables that contribute for lead getting converted in the model are
  - Total time spent on website
  - Lead Add Form from Lead Origin
  - Had a Phone Conversation from Last Activity
- Hence overall this model seems to be good.