Lead Scoring Case Study

Problem Statement:

An education company named ‘X Education’ sells online courses to industry professionals.

- Need to build a model which assigns 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. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

- The Build model is used by the management to make the process more efficient and understand how exactly the demands vary with different features and accordingly the management can manipulate the business strategy which is used in the conversion of lead to paying customers.

Essentially, the company wants

- To identify the most potential leads, also known as ‘Hot Leads’
- To identify the variables that are significant in predicting the lead customers who are most likely to convert into paying customers.
- To create a Logistic regression Model that quantitatively relates lead customers with variables.
- To know the accuracy of the model, i.e. how well these variables can predict the paying customers.

Steps:

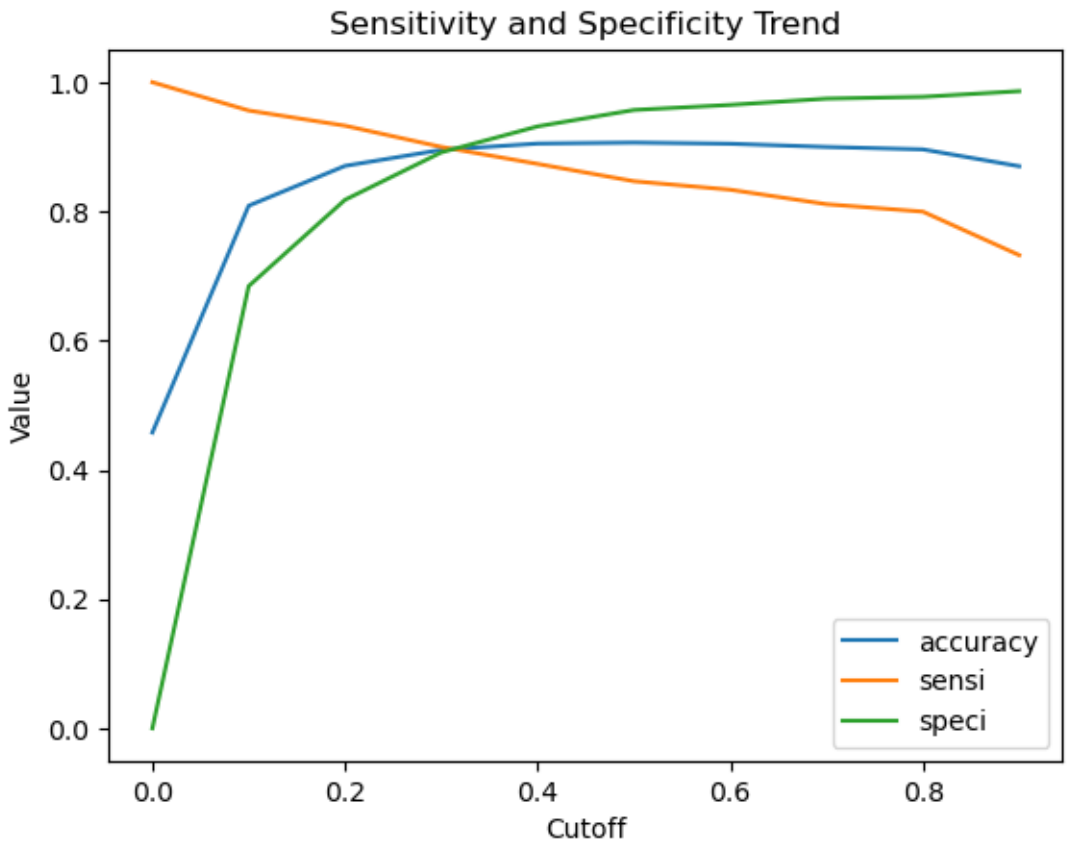
The steps that are performed throughout the model building and model evaluation are:

1. Importing the dataset
2. Read and understand the data
    - the head of the dataset
    - the dimensions of the data frame
    - the type of each column
    - the statistical aspects of the data frame
3. Data cleaning and preparation
    - Fixing Rows and columns and impute/remove missing values
4. Checking categorical variables
5. Converting categorical variables to Binary variables :
    - converting yes / no to 1 / 0

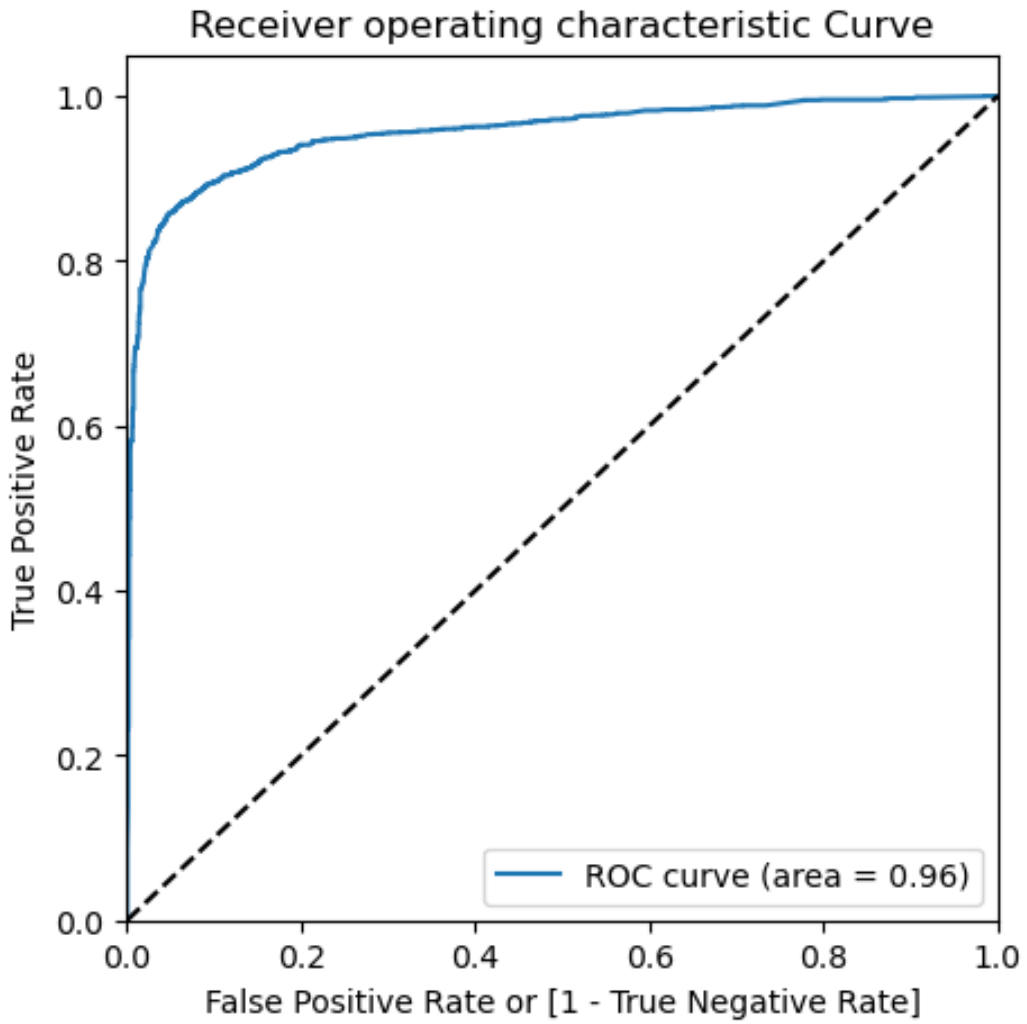
6. Dummy variable creation
    - Creating dummy variables for categorical variables with multiple levels
    - Dropping the repeated columns and the columns which are not required for analysis

7. Checking Outliers
8. Test-train split and scaling
9. Model Building
    - Build using logistic regression Model
10. Feature selection using RFE
11. Manual feature elimination by using calculating p-values and checking VIFs.
12. Model Evaluation

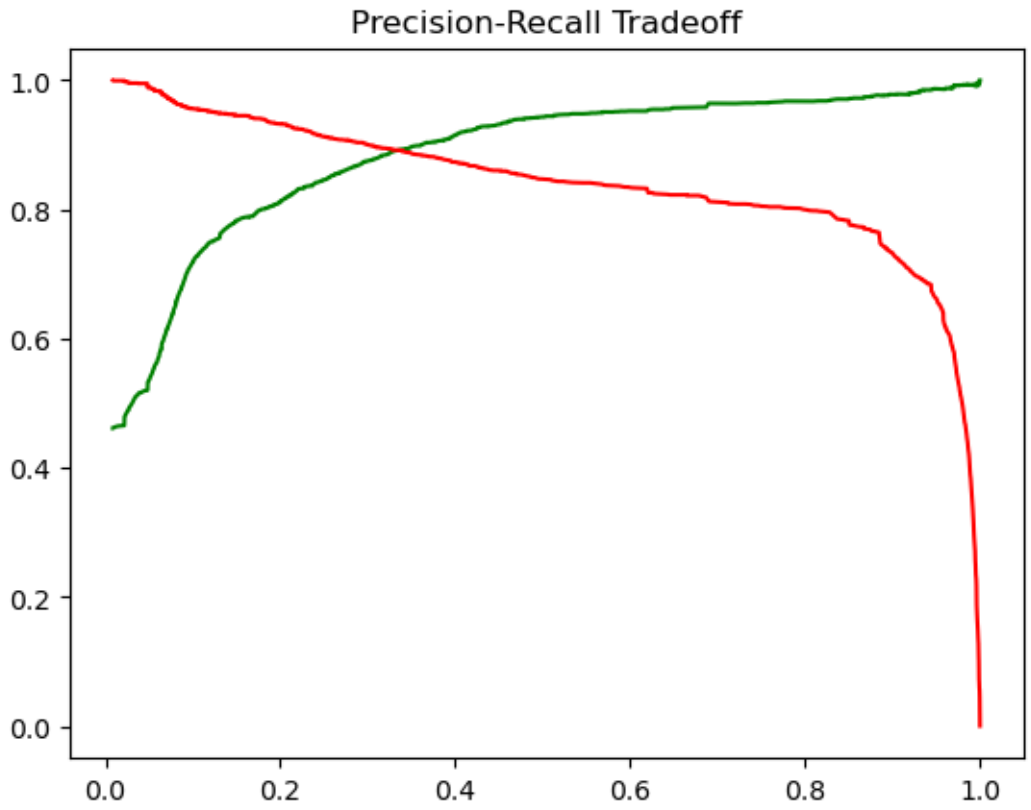
13.Accuracy, Sensitivity and Specificity Curve:



14.Optimal cut-off using ROC curve



15. Precision and Recall Tradeoff



16. Predictions on the test set

The classes were assigned to all the leads in the test data set. For this, a probability cutoff of 0.5 was used. The model thus made, was very accurate (Accuracy = 90%), Sensitivity (84%) and Specificity (95%). The different cutoff was tried out, i.e. 0.3, which resulted in a model with almost same (Accuracy = 89%), but the sensitivity is more (~89%) and less Specificity (89%). Hence, we learnt that we should use 0.3 as the cutoff for probability for this model because the cutoff that equalizes accuracy, sensitivity and specificity.

Recall and the sensitivity are the alternative terms used and that the sensitivity of a model is the proportion of positives correctly predicted by it as positives. Also, the specificity is equal to the proportion of negatives correctly predicted by the model as negatives. The sensitivity increases by changing the cutoff, its specificity goes down.