Business Problem:

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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.

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

Data Collection:

All the data was provided in an Excel sheet.

Data Exploration:

Began by exploring the dataset. This includes variables such as browsing behaviour, past interactions, etc. Understand the meaning and significance of each variable in the context of lead scoring.

Data Pre-Processing:

1. Identifying and handling of missing values, outliers, and inconsistencies in the data with the help of graphs and heat maps. We dropped some columns when the values were skewed.
2. Conversion of some features into Category features and created Dummy variables to change categorical data into Discrete values.

Data Split:

1. Separate the Features and Target Variable: Split the dataset into two components: the features or independent variables (X) and the target variable (y). The features are the variables used to predict the target variable.
2. Split the Data into Training and Test Sets: We divided the dataset into a training set and a test set, into 70:30 split using random seed of 42. The training set will be used to train the logistic regression model, while the test set will be used to evaluate its performance.

Feature Selection:

We used RFE feature selection technique to select the most relevant features from a given dataset. We arrived at 20 most significant variables.

Model Selection:

We have used “Logistic Regression “is better choice as we need to understand key features of lead conversion.

Model Training:

We build a model and calculate VIF for each features. Each features should have <5 VIF

Model Evaluation:

When we created data frame for lead having converted probability, our initial assumption is if probability more than 0.5 is 1 else 0.

After deriving Confusion matrix, we calculated Accuracy, Sensitivity and Specificity. For the good model we not only have good Accuracy percentage but also, we should have good percentage for Sensitivity and Specificity.

Model Improvement:

After plotting the ROC Curve, we got decent curve. 88% area is covered.

We plotted probability graph for different values of Accuracy, Specificity and sensitivity. The cut off came as 0.4.

We could also observe we got a new value of the accuracy=81%, sensitivity=68% and specificity=89%

Precision and Recall Matrix:

We found precision and recall matrix, got 79% and 68% respectively.

Based on the Precision and Recall matrix, the cut off was 0.4

Making Predictions on Test Set:

We implemented the learning on test model and calculated probability, we got Accuracy, Sensitivity and Specificity as 81% , 77% and 84%.