Problem Statement:
X Education company sells online courses to industry professionals.The company markets its courses on several websites and search engines like Google. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30% which is very poor.

X Education needs a model that will help them in identifying the leads that are most likely to convert into paying customers. The model should assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Step 1: Reading and Understanding the Data
Read the input file into a data frame and assessed the shape, columns and info

Step 2: Data Cleaning
identified the columns which has high NULL values and dropped the columns which has >30% null values. Imputed the missing values and outliers with mean/mode or median according to the specific scenarios.

Step 3: Visualising the Data using EDA
Plotted the numerical variables in a box plot and categorical variables as countplot. Dropped the columns which has highly biased data and which do not have any predictive role.

Step 4: Create Dummy Variables
Converted all the categorical variables into dummy variables.

Step 5: Splitting the Data into Training and Testing Sets
Splitted the data frame into Train and Test data sets using 70:30 proportion.

Step 6: Feature Scaling
Performed scaling operation to convert categorical variables into numerical scale.

Step 7: Building a logistic regression model
Build the initial version of the model using all the variables.

Step 8: Feature Selection using RFE
Selected top 20 features using Recursive Feature Elimination technique. Created a logistic regression model using these 20 and observed the P-Values and VIF score.

The model was run in multiple iterations and at the end of each iteration one feature with max P-Value was dropped. Finally after iteration 6 a final model with acceptable P-value and VIF values was determined. There were 15 features left in the final model.

Using the final model, created a new data frame with converted probability values. In first pass all probabilities > 0.5 were considered as converted(1) and remaining as not-converted (0).

Step 9: Plotting the ROC Curve
Plotted the ROC curve to observe the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity. We have observed area as 0.88 which is a good indication of the goodness of the model.

The optimal Cutoff point was determined to be 0.375.

Using the optimal cutoff point of 0.375, calculated the final Predicted value.
Accuracy 80.7%
Sensitivity 79.3%
Specificity 81.6%.
Precision 78%
Recall 69%

Step 10: Making Predictions on the test set using the Final Model
Run the final model against the test data and determined the various metrics.
Accuracy 80.5%
Sensitivity 79.08%
Specificity 81.4%.
Precision 73.4%
Recall 79%