Problem Statement:

Help X Education to select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goal:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Summary:

Step1: Reading and Understanding Data:

Read and analyze the data.

Step2: Data Cleaning:

We dropped the variables that had high percentage of NULL values in them i.e., more than 30%. We also figured out if we need to replace null values with median values in case of numerical variables and creation of new classification variables in case of categorical variables.

Step3: Outlier analysis:

The outliers were identified and removed for all numerical variables.

Step4: Data Analysis and visualization:

We did Exploratory Data Analysis (EDA) of the data set with respect to the target variable- conversion to have a better understanding of the relationship.

Step5: Creating Dummy Variables:

We went on with creating dummy data for the categorical variables and binary mapping for variables with 2 categories (Yes/No).

Step6: Train-Test Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values with a random state of 100. Also, this was done by dividing the data set into X and y for further analysis.

Step7: Feature Rescaling with Min-Max Scaling:

We used the Min-Max Scaling method to scale the numerical variables such as 'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit'.

Step8: Model building:

Using the stats model, we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step9: RFE:

Using the Recursive Feature Elimination (RFE), we selected top 20 features. We then repeatedly generate several models and check for high p-values & VIF, and drop those features; creating an efficient model.

Step10: Prediction:

We predict values based on our final model by creating new column 'predicted' with 1 if Converted_Prob > 0.5 else 0.

Step11 - 13: Confusion Matrix, Model evaluation and ROC:

We have evaluated the model based on accuracy, sensitivity and specificity on training data.

Step14 - 16: Finding Cut-off and analyze tradeoff:

We evaluated the train set probability cut-off to be 38%.

Step17 - 19: Analyze the model on test set with the same cut-off.

Conclusion:

- While we have checked Sensitivity-Specificity, Precision and Recall Metrics, we have considered the optimal cut off of 0.38 based on the metrics for calculating the final prediction.

- Accuracy, Sensitivity and Specificity values of test set are around 82%, 80% and 82% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model being around 80%
- Hence overall this model seems to be good as the CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.