LEAD SCORING CASE STUDY SUMMARY

Problem Statement:
X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The typical lead conversion rate is around 30%.

Business Goal:
To 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. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step1: Importing and Reading Data
The various libraries were imported and the data was read.

Step2: Inspecting the Data
The data was analyzed, missing values were checked and values in level 'Select' where converted as null vales. The columns having null values > 30% were dropped, categorical columns were imputed with the mode or new values were created and numerical columns were imputed with mean or median values. The columns having only one value majorly present and columns irrelevant were dropped. Outlier Treatment and EDA was performed.

Step3: Data Preparation
The binary variables were mapped to 1s and 0s. One-Hot encoding or Dummy variables were created for the categorical variables with multiple levels.

Step4: Test-Train Split
The data was split into train and test sets - 70% and 30% respectively.

Step5: Feature Scaling
MinMaxScaler was used to scale the numerical columns.

Step6: Looking at Correlations
A correlation matrix was created to check the correlation among the variables.

Step7: Model Building
A model was built using statsmodels.api. After running the model, we got a statistical view of all the parameters. However, the variables to be dropped cannot be derived as all variables have high values.

Step8: Feature Selection Using RFE
Using Recursive Feature Elimination, we attained the top 15 relevant variables. More variables were dropped manually depending on VIF(<5) values and p-values(<0.05).

Step9: Plotting the ROC curve
Once the model was finalized, the ROC curve was plotted. (ROC curve (area = 0.89)).

Step10: Optimal Cutoff Point

A graph with Accuracy, Sensitivity and Specificity was plotted and 0.37 was considered as optimal cut-off point. The Accuracy score was 81.52%, the Sensitivity score was 80.63%, the Specificity score was 82.07%, the Precision score was 73.50% and the Recall score was 80.63%.

Step11: Making Predictions on the test set

The model was applied on the test data and the various metrics were checked. The Accuracy score was 81.25%, the Sensitivity score was 80.86%, the Specificity score was 81.5%, the Precision score was 73.93% and the Recall score was 80.86%.

Conclusion:

- The variables that mattered the most for variable conversion are Total Time Spent on Website, Lead Origin - Lead Add Form and Last Notable Activity - Had a Phone Conversation.
- The conversion rate on the original dataset was 39%.
- Now, the conversion rate as per the metrics of Accuracy, Sensitivity and Specificity on the test data is around 80% which indicates that our model is well built.