SUMMARY

Problem Statement:

An education company named X Education sells online courses to industry professionals. X Education has appointed you to help them select 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Step1: Importing the libraries & reading the data.

Step2: Data Inspection & Cleaning: Inspection of the data was done. We dropped the variables that had 40% and above NULL values in them. Values of certain variables were imputed & rows with null values & variables with only one category were also dropped. Categorical variables with lower frequency of sub-categories were converted to 'Others'.

Step3: Exploratory Data Analysis: Exploratory data analysis of all the categorical variables with the 'Converted' variable was realised using count plot to find the percentage rate of Converted & Non-converted leads which led to dropping of some variables.

Step4: Dummy variables were created for all the categorical variables.

Step5: Test Train Split: The next step was to divide the data set into test and train sections with a split of 70%train data-30% test data values.

Step6: Feature Rescaling: Standard scaler was used to scale the numerical variables.

Step7: Feature selection using RFE: Using the Recursive Feature Elimination method we selected the 15 top important features. We started assessing & building the model using stats model, based on the P values generated we started dropping the variables with P values >0.05. Next step was to check the VIF's of the variables. Finally, we arrived at the 8 most significant variables. The VIF's for these variables were found to be <5. We then created the data frame having the converted probability values with an initial assumption that a probability value of more than 0.5 means 1 else 0. we derived at the Confusion Metrics and calculated the overall 'Accuracy', 'Sensitivity', 'Specificity', of the model.

Step8: Plotting the ROC Curve: We plotted the ROC curve for the features and the curve came out be with an area coverage of 98%.

Step9: Finding the Optimal Cut-off Point: We plotted the probability graph for the 'Accuracy', 'Sensitivity', 'Specificity', 'False positive rate', 'positive predictive value', 'Negative predictive value' for different probability values. The intersecting point of the graphs was 0.42 which was considered as the optimal probability cut-off point. The values of the 'accuracy=95.98%, 'sensitivity=96.54%', 'specificity=95.51%'.

Step10: Precision and Recall: Precision and Recall metrics values came out to be 95.90% and 95.10% respectively on the train data set. Based on the Precision and Recall trade off, we got a cut off value of approximately 0.5.

Step11: Making Predictions on Test Set: The values of the 'accuracy=95.78%, 'sensitivity=94.43%', 'specificity=96.94%' on the test set.

Step12: Lead score: Lead score was assigned for all the leads with respect to their ProspectID.