Lead Score Case Study Summary:

Lead score case study was performed using the Logistic Regression method. As part of the data cleaning, we performed cleaning like checking for null values, grouping of similar values, renaming of columns etc. During the EDA, we removed a few unnecessary variables and created new dummy variables using different variables which converted object type to int data types for the new columns.

Post the EDA, we split the data into train and test datasets in the ratio of 70:30. We included the constant as well. We performed correlation analysis using heatmap. We used RFE to narrow down the top 15 variables which had greater impact on the target variable. We used the generalized linear model (GLM) in the Stats Model to prepare a report providing the p-value. Then we went on to find the Variable Inflation Factor (VIF) of these variables. We eliminated variables one by one by ensuring the p-value was below 0.05 and the VIF was less than 5.

We then tested the model using the test data. We also calculated using the accuracy matrix. Since we were performing Logical Regression, we plotted the sigmoid curve and got the ROC curve area as 0.88.