Executive Summary

Post analysis on Lead Scoring Dataset

The following are the steps used in this case study

- Cleaning Data
    a. The data was partially cleaned at first step by removing some of the null values and then some of the null values were generalised by replacing it with “Not Provided”
- EDA
    a. Using categorical analysis very few categorical values were actually related to lead conversion and was relevant but rest were irrelevant. Numerical analysis had few cases with outliers but overall seems good.
- Dummy Variables
    a. Dummy variables were created for categorical values
    b. Scaling was done using standard scaler function
- Train test Split
    a. Train Test split was done 70-30 % ratio
- Model Building
    a. RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-Values.
- Model Evaluation
    a. A confusion matrix was made. Later, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 79%,67%,88.7% respectively.
- Prediction
    a. Prediction was done on the test data frame and with an optimum cut off as 0.3 with accuracy, sensitivity and specificity of 78.9%,83%,75.8% respectively.
- Assign Lead score to find Hot leads
    a. Finally lead score was assigned to test data to find hot leads from the data set.

It was found that

- There are total of 551 Hot leads which can be converted from all the data set.
- Leads coming from leads who are working professional are Good Leads
- Leads coming from phone conversation are good leads
- leads where emails have bounced should not be considered