Summary

The problem statement states to create a Logistic Regression Model with the help of a given data to find out “Hot Leads” out of all the customers. We had to assign a lead score to each of the lead based on the model. If the lead score exceeds 80, the X-Education company would consider it as a hot lead. We started by importing the dataset. Then we observed the distribution of Null Values across the dataset. Imputed the null values with mode, random wherever needed. There are ‘Select’ Strings across the dataset. They are replaced with appropriate values. Somewhere they are replaced with ‘Other’ and so on. Then we started treating outliers for the numerical columns. Now we start preparing the data for the model with splitting the data into train set and test set. 70% of the data was picked for the train set and the rest 30% was moved into the test set. Using a standard scaler, the numeric variables were scaled. The true target percentage was checked. Firstly, a Logical Regression model was created with all the variables. Then RFE(Recursive Feature Elimination) was performed on the columns for 15 variables output. Then a model was build using the features which supported the model. The p-Values and VIF were checked. Then variables with high VIF were dropped to create a more decent model. Gradually, a final model was designed in which all the variables had low p value(less than 0.05) and less VIF(Less than 5). Then Converted probability was compared to the actual value to calculate different metrics such as AUC-ROC, Log-Loss, Mean Squared Error. Then the predictions were made on the test set. Finally, Accuracy score, confusion metrics, sensitivity and specificity were calculated.

This way, a model which tells you the lead score of a lead was developed which works accurately for almost 91% of the total cases.