Summary Report

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Objective: The objective of Lead Score Case Study is to find the hot leads of the X Education Organization with the help of Logistic Regression

The steps followed for the Lead Score Case Study are:

Data Inspection: In this section, we inspected the data to understand the number of rows and columns, statistical information of the data i.e; description and information of the data.

Data Cleaning: In this segment, we tried to clean the data by removing the duplicate rows if there are any and dropped the columns which are not necessary

Data Preparation: In this segment, the data set is treated by removing the outliers, removing the columns which has more than 40% missing values. The learning from this segment is If we have more missing values, it is better to drop it but not to impute by mode or median. So that the model will be better.

Data visualisation: In this segment**,** the data is visualised to know the spread of the different values in each column. So that we can we can either minimise the number of different values in each column or impute the less the less values to one type. In this segment we imputed the less number of different types of each column with other.

Dummy Variables Creation: In this segment, we created the dummy variables for all the categorical variables in the data set. So that for building the Logistic Regression model, we need not lose the data. Once the dummy variables are created we can drop the first column and the original column.

Split the Data into test and train Set: Splitting the entire data set into train set and test set, where we build the Logistic Regression Model with the train set and we test the model with the test set. For this case study we split the Entire data set into Train (70% of the data) and Test (30% of the data) set.

Scaling: In this segment, we scale the entire data set. So that the mean of all the columns is equal to 0. So that the modelling step also becomes easy.

Feature Selection by RFE: In this step, we will pick the required number of attributes or features to build the Model. In our Case study we picked around 15 features.

Logistic Regression Modelling: Before building the Logistic Regression Model, the constant has to be added. Then we can build the Model. Once the Model is built, we have to iterate till the VIF value of all the features is less than 5 and P value for all the features is less than 0.05(5%). If that point reaches then the Model has been ready without Multi Collinearity. We built the model with 11 features. Specificity (94%) and Sensitivity (89%) was calculated. Then the ROC was plotted with 94% of data is under the curve.

Prediction on Test Set: In this step, the test data is been fit with the Logistic Regression Model. The Lead Score is also calculated. Then the Accuracy Score, Confusion Matrix, Sensitivity (87%) and Specificity (91%) of the Test data is calculated. Precision (89%) and Recall (80%) also has been calculated.

Assigning Lead Score to data set: Then the Test Set and Train Data Set has been merged and the Lead Score has been assigned to the data frame. Then determined the features importance.