Lead Score Case Study Summary Report

Introduction:

We were provided with the leads dataset with shape 9341*37. Our goal is to make a logistic regression model which will help the sales team to find out the most potential leads or the hot leads on the basis of the lead score so that they can focus on these leads with full resources and effort.

Approach:

Initially, we imported all the basic libraries required for Exploratory data analysis. After that we read/import our data's .csv file. Then we followed the following steps:

- **Data Cleaning and Processing:**

  First, we checked for duplicates in prospect id and lead number, as a result, we did not find any duplicate data in these two columns so we dropped these columns as they contain the unique data which is not required for the modeling prediction. Then we handled the 'select' level that was present in many categorical variables by assigning it to 'NaN'. After this, we found the percentage of the missing values in the data and dropped the columns with more than 45% of missing values. Then we started with 'the categorical attributes analysis' here we use imputation techniques like mean, mode, and median. we also clubbed some values like in specialization we combined all the low-frequency categories of management into one single category called 'management'. After this, we looked forward to finding some imbalanced variables that can be dropped. In 'the Numerical attributes analysis' we checked the correlation of numeric values using heat maps and did analysis for all the variables in order to find and remove outliners. We left with the data of 6640*14 shape. We retained about 71% of the data.

- **Data Preparation:**

  We started with the creation of dummy variables for all the categorical columns which resulted in the creation of 51 more columns. We converted the columns with values 'YES'/'NO' with binary '1'/'0'. With this step our whole data becomes numeric. We perform the test and train split with a ratio of 30:70 and random_state=100. Once the data got split into test and train we performed scaling to the train set using a standard scaler.

- **Modelling:**
  Once all the processing of the data is done we started with the model-building process. Once the model is made based on summary statistics, we inferred that many variables might be insignificant and hence we need to do more feature elimination. We started off with an automated feature selection technique i.e. RFE because the number of features was large. After completing the RFE process we were left with 15 features. By looking at the summary we realized that some features have very large p-values. So we started the manual elimination of features, which has high p-values and VIF values. After two or three rounds of manual elimination by re-running the model, we finally got the features with zero p-values and low VIF values. Once the model is made we start with the prediction on our train data set with a probability cut-off of 0.7. after this, we checked the performance of the model using the confusion matrix and find out the overall accuracy, sensitivity, and specificity. Finding the optimal cut-off was important so that the model is balanced we used the ROC curve. Once we get the optimal cut-off for the model we start putting this model on our test data set. We checked the model performance and compared it with the test data set. Also as required by the business we also calculated the lead score for our leads.