Lead Scoring Case Study Assignment Summary

- In this assignment we need to find the hot leads for the education company education so that they can focus on grooming of only the hot leads to get them converted to paying customers.
- **Data Loading:** We started with loading the data set in pandas and checked for data quality by checking on null values in data set, inappropriate values (like 'Select' was present in many variables) and skewness of the data.
- **Checks for Null Values and Imputation:** We then imputed the null values with median/mode depending upon whether the variable was numerical or categorical. Also we removed the invalid values like 'Select' and imputed them with Nan.
- **Univariate/Bivariate Analysis** We then did the univariate analysis and bivariate analysis by plotting the scatter plots of the variables.
- **Outlier Detection:** We also created boxplots of the variables to check for outliers in numerical variables.
- **Binary Conversion:** We then converted the variables containing 'yes' and 'No' values to binary variables (0/1).
- **Dummy Variable Creation:** We finally converted categorical variables to dummy variables such that for n possible values, 'n-1' dummy variables were created. After this the main variables which were converted to dummy were dropped.
- **Train-Test Split:** We then divided the dataset to train and test data with 70-30 ratio respectively and then scaled the numerical variables 'Total Visits' and 'Total Time Spent on Website' so that these variables do not dominate the model when fit.
- **Correlations:** We then checked if there are any correlations in the variables and hence the correlated variables could be removed, but no major correlations were found amongst the variables.
- **Training the model on Train Set:** The model was then trained on the training set using Logistic regression and Recursive Feature Elimination.
- **VIF:** All the variables with Variance Inflation Factor>5 were dropped and initially the 'Predicted' variable was created depending upon whether the converted probability was greater than 0.4, then it was treated as true else false.
- **Accuracy/Precision Statistics:** All the necessary statistics like Accuracy, Precision, Recall and Sensitivity, Specificity were calculated for the training set and the accuracy came out to be ~ 88%.
- **Selection of Optimum Threshold for Converted:** For finding the optimum value, the ROC curve was fitted for Accuracy, Sensitivity and Precision and it was found that the curves intersected at approximately **0.35.**Hence the converted variable was then recalculated considering that all values greater than 0.35 will be true and less than or equal to 0.35 will be false.
- **Making Predictions on the test set:** Once the model is fitted on the training set, the predictions were then made on the test set and the converted variable was created considering 0.35 as the optimum threshold value.
- **Hot Leads Identification:** To discover the hot leads which can turn out to be paying customers, the conversion probability was multiplied by 100 to assign a lead score to every Prospect ID and all the prospects having lead score greater than 75 were considered as hot Leads.

Hence the organization should target all the clients having lead score greater than 75 and call them often and groom them and make them aware of the benefits of their courses so that these leads can be converted to customers.