Lead Score Assignment – Summary Report

There are 10 steps to conduct this assignment:

1. Understand the problem statement:
   X Education – an online course provider – would like to conduct a logistic regression model in order to assist them in maximize the lead conversion as well as optimize the efficiency of their sales team.

2. Understand the data and import all required libraries
3. Inspect the dataset:
   - Shape: 37 columns and 9240 rows.
   - 7 numerical and 30 categorical columns.

4. Cleaning the data:
   - Remove all columns having >=40% missing values: 5 columns
   - For those columns having ratio of missing values < 40%, replace missing values with 'unknown'
   - For the columns having "select" response, replace "select" with "unknown" as customers did not interact with them.
   - Remove all imbalanced binary columns where we observed the data is >92% skewed towards a certain response (5 columns).
   - For numerical columns having missing value, replace null with mean score.
   - Numerical columns having outliers outside upper bound, replace those outliers with the upper bound value.

5. Exploratory Data Analysis
   - Conduct univariate analysis for each variable
   - Conduct multivariate analysis for each variable crosstab with "Converted" column to understand the difference between 2 groups "Converted" and "Non Converted".
   - For categorical columns having many granular responses and some response just accounts for small percentage (sample size <30), those responses will be grouped together under 'Others" in order to simplify the list of variables later on.
   - For numerical columns, besides the summary information (mean, min, max, standard deviation, etc.), the data is also binned into different ranges to understand the data better.

From EDA, there are some hypotheses related to what factors might be considered as important and impactful ones in driving for Lead conversion.

6. Feature Engineering and Variable Transformation
    - As all binary columns are removed from above steps due to high missing value ratio or imbalanced data matter, so don't need to conduct binary map.
    - Create dummy variables for all categorical columns.

7. Train_Test_Split
    - Split the dataset into train (70%) and test (30%) sets
    - Scaling the numerical columns by using StandardScaler()

8. Building models
    - Run correlation matrix to have a first glance on the high correlated factors.
    - Building the first logistic regression model
    - Due to high number of variables, RFE method is applied to build the second model with shortlist of the most profound 20 variables.
    - Check p-values and VIF score of the 2nd model and remove those variables having p-value>0.05 or VIF score >5. Repeat this step to run the 3rd and 4th model and see that the 4th model is in good shape to move forward.

9. Evaluating models
    - Check accuracy scores, confusion matrix, sensitivity, specificity, true positive rate, false positive rate with the arbitrary cut off point of converted probability at 0.5
    - Plot ROC curve and trade off chart among accuracy, sensitivity and specificity to define the optimal cut off point at 0.3
    - Rerun the confusion matrix to check balance
    - Recalculate the accuracy, sensitivity and specificity to check the gap among these metrics.
    - Calculate precision and recall score as well as draw the chart to show trade off between recall and precision.
10. Making predictions
    - Apply the 4th model with cut off point at 0.3 on the test dataset.
    - Check on the confusion matrix, accuracy score, specificity, sensitivity and see that these scores are in good shape.
    - Rerun the correlation matrix among the remaining selected columns with "Converted" to define the most important factors impact positively and negatively on the conversion rate.
    - Based the defined classification, run the list of potential leads.