Summary Report

This case study is on an education offering platform named X Education which sells online courses to industry professionals. Currently, the conversion rate stands at 30% and the CEO has asked us to reach to a lead conversion rate of around 80%. We have been provided with a dataset with lot of information on customers like the total number of visits they made to the websites, how much time they spent, which country they belong to, which domain they work in, etc.

We performed the following steps to complete this case study:

1. **Importing the necessary libraries:**
   We imported the libraries for data analysis, manipulation, visualization, model building, scaling, evaluation, etc.
2. **Importing the dataset:**
   We imported the dataset and did a thorough inspection post creating the data frame. We also checked the shape, statistical description and data type of each column.
3. **Data cleaning:**
   We performed the below steps under data cleaning:
   i) There were many columns in the dataset which had a level called “Select” which was as good as null value hence we replaced those with null value.
   ii) For variables where there were more than 40% missing values, we dropped those columns
   iii) For the columns where missing values were below 40% we resorted to imputation based on mode.
   iv) We removed the columns like "Prospect ID", "Lead Number" as these were unique identifiers and wouldn’t have helped in the analysis in any way.
   v) For the columns were there we missing values in the range of 1-2% we just removed those rows instead of dropping the entire column.
   vi) We also dropped the columns which were highly skewed.
   vii) We also detected the outliers and performed capping to cure them.
4. **EDA:**
   Next, we performed the exploratory data analysis where we visualized the categorical variables with the target variable and we performed the correlation among the numerical variables to see how correlated they were.
5. **Data preparation:**
   We performed the data preparation steps which are as follows:
   i) We mapped the binary categorical variables.
   ii) We created the dummy variables and dropped the original columns for which we created dummies
6. **Train-Test split:**
   Next, we performed the train-test split where we took 70% of the data as training set and 30% of the data as test set.
7. **Model Building:**
   Under model building we performed the below steps:
   i) We performed the standard scaling on the training data (for the numerical variables).
   ii) We again performed a correlation post dummification and dropped the variables which were highly correlated.

iii) Next we performed RFE and for a favorable list of 15 variables. Then we performed p-value and VIF checks and dropped all the columns where p-value or VIF was high. In our final model, the p-values and VIF look good.

8. **Model Evaluation:**
   Next, we performed model evaluation where we initially took a cut-off of 0.5. Next, we plotted a ROC curve to get an optimal cut-off point where the accuracy, sensitivity and specificity look good. In our case, we concluded 0.3 is the optimal cut-off point.
9. **Prediction on test dataset:**
   We moved ahead to perform prediction on the test data with optimal cut-off point of 0.3 and we got good accuracy, sensitivity and specificity in case of test data too.
   Our **recall** too came above 80% so our model is performing well.
10. Post Model building and evaluation, we came to a conclusion that below are the variables which will help in increasing the conversion rate:
    i) Total Visits
    ii) Total time spent on website
    iii) When the lead origin is “Lead add form”
    iv) When the lead source is “Olark chat”
    v) When the current working occupation is “working professional”