Summary

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. We are required to build a logistic regression binary classification model and assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The target lead score as mandated by CEO of X Education is around 80%

The following approach was adopted to build the model:-

1. **Data Sourcing:**
   - Leads dataset was imported from the Leads.csv data file provided for this study.
2. **Data Preparation:**
   - First step to clean the dataset we choose to remove the redundant variables/features.
   - The data set was partially clean except for a few null values and the option 'Select' has to replace with a null value since it did not give us much information.
   - Dropped the high percentage of missing values more than 45%.
   - Checked for number of unique Categories for all Categorical columns.
   - From that Identified the Highly skewed columns and dropped them.
   - Treated the missing values by imputing the favourable aggregate function like (Mean, Median, and Mode).
   - Outlier treatment by removing the outliers.
3. **Exploratory Data Analysis:**
   - A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant.
   - Performed Univariate Analysis for both Continuous and Categorical variables.
   - Performed Bivariate Analysis with respect to Target variable.
4. **Dummy Variables:**
   - The dummy variables are created for all the categorical columns.
5. **Scaling:**
   - Used Standard scalar to scale the data for Continuous variables.
6. **Train-Test Split:**
   - The Spit was done at 70% and 30% for train and test the data respectively.
7. **Model Building:**
   - By using RFE with provided 15 variables. It gave top 15 relevant variables. Later the irrelevant features were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value 0.05 were kept).
8. **Model Evaluation:**
   - A confusion matrix was made. Later on the optimum cut-off value by using ROC curve was used to find the accuracy, sensitivity and specificity of 92%.

9. **Prediction:**
    - Prediction was done on the test data frame an optimum cut-off as 0.37 with accuracy, sensitivity and Specificity of 92%.

10. **Precision-Recall:**
    - The method was also used to recheck and a cut-off of 0.42.

11. **Conclusion:**

    Top 3 features that contribute to lead conversion are:

    - Total Time Spent on Website
    - Leads that have come through source as reference
    - Leads that have come through source as Welingak Website

    Top 3 features that need to be improved are:

    - Last Notable Activity_Olark Chat Conversation
    - Last Notable Activity_Modified
    - Tags_Will revert after reading the email