Executive Summary

Goal Statement:

The goal was to build a Logistic Regression based model for X Education company to help in selecting the most promising leads and to figure out the 'Host Leads' of company having a higher conversion chance.

Below are the summaries and steps performed:

1. **Understanding and Cleaning Data:**
   - There were few variables with missing/null values which were needed to be take care by dropping unwanted columns with high percentage null values and some with unique values.
   - Some categorical variables with lots of smaller factors were merged and imputed proportionally to maintain distribution and not introduce bias.
   - Many of them had level called 'Select' which was as good as a null value, therefore, to not lose data they were changed as 'no data' value. Although, later on they were distributed proportionally or were removed while creating dummies.

2. **Exploratory Data Analysis (EDA):**
   - We performed EDA on various categorical and numerical variables to analyse the condition of data. The variables with high data imbalance were removed.
   - With target variable "Converted", we tried to analyse its relationship with all numeric and categorical variables by plotting graphs. No major outliers were found.

3. **Dummy Variables:**
   - Created dummy variables for categorical columns and converted the binary values to numeric elements i.e. 0/1. For rescaling numeric variables, we used MinMaxScaler.

4. **Train-Test Data Split:**
   - The dataset was split into 70% for train and 30% for test sets.

5. **Model Building:**
    - We used RFE to select the top 15 significant variables to perform model building.
    - Later on, the variables with **VIF > 5** and **p-value > 0.05** values were dropped to build model more significant one.

6. **Model Evaluation:**
    - By creating Optimal cutoff (ROC curve) and confusion matrix, the Accuracy, Sensitivity and Specificity was around 79%-80%.

7. **Prediction:**
    - Performing prediction on test dataset using cutoffs 0.35, Accuracy, Sensitivity and Specificity was around 80%.

8. **Precision-Recall:**
    - To recheck the prediction, performed Precision-Recall using cutoff 0.43 on test dataframe, Precision and Recall was around 77% and 73%.

Conclusion:

Some significant variables that mattered the most in identifying the promising "Hot Leads" for company:

1. "TotalVisits"
2. "Total Time Spent on Website"
3. "Page Views Per Visit"
4. "Lead Source"
    a. Google
    b. Olark Chat
    c. Welingak website
    d. Reference
5. "Last Activity"
    a. Olark Chat Conversation
    b. Converted to Lead
    c. Email Bounced
    d. Had a Phone Conversation
6. "Last Notable Activity"
    a. SMS Sent
    b. Unreachable
7. The Leads having Current Occupation as "Working Professional".