Summary

X Education company wants us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. For this we have built a model using the past labelled data, wherein we assigned a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. As per the target set by the CEO, we have achieved a model which can help in converting the 80% of potential leads.

Steps that we have used while building the statistical model for solving this problem are:

1. Data Cleaning:
   - Deleted few columns which have data entered manually by the sales team.
   - Deleted the columns where the proportion of missing values is more than 40%.
   - Then in columns with nulls and select as their members we have applied imputation or other relevant techniques.
   - Plotted data distribution curves for categorical columns and observed that there are many columns with just a single value in them or few columns that have highly skewed data with just 1 member occurring for more than 90% of times. We have deleted such columns as well.
   - Detected and deleted few outliers in the numerical columns using box plots.
2. Exploratory Data Analysis:
   - To gain the insights or analyze trends in the past data, performed univariate, bivariate and multivariate analysis by plotting different charts.
3. Data Preparation:
   - Created the dummy variables for all the categorical columns.
   - Used standard scalar to scale the data for continuous variables.
   - Split the data into train and test data sets with 70% and 30% as proportions respectively.

6. Model Building:
   - Using RFE performed feature elimination and selected 20 most impactful variables.
   - Analysing VIF scores and p-values eliminated features with VIF score is > 5 or p-value is above 0.05.
7. Model Evaluation:
   - Generated confusion matrix and identified optimum cut-off value by using ROC curves which are then used to find the accuracy, sensitivity and specificity which came to be around 80%.
8. Prediction:
   - Prediction was done on the test data set with an optimum cut off as 0.33 and with accuracy, sensitivity, and specificity of 80%.
9. Precision-Recall:
   - This method was used to revalidate the model results. Validated that the cut off is indeed 0.33 with precision score as 70% and recall score as 78% on the test data frame.
10. Conclusion:
    - Below is the list of Features that are most significant in identifying Hot Leads:
    - Top 3 Features:
      - When the lead origin is Lead add format.
      - When the Occupation is Working Professional
      - The total time spent on the Website.
    - Other Important Features:
      - Total number of visits.
      - When the lead source was Google and Olark Chat
      - When Lead Origin is Landing Page Submission