Summary Report

X Education company gets a lot of leads, but its lead conversion rate is only 30%. The company requires us to build a model wherein we need to 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 CEO has given a ballpark of the target lead conversion rate to be around 80%.

This dataset provided to us consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Below are the steps followed:

1. Data Cleaning:
    - Treating 'Select' values: These are basically the columns in which customers have not selected any value from the list and hence we have replaced them with Nan value.
    - Handling the missing/null values: All the columns that had more than 40 % missing values were dropped. All the columns which had single value "NO" and had majorly one value ex: 'No' were dropped as it doesn't add any value to our data analysis.
    - Duplicates Check: There were no duplicates in the dataset.
    - Created a new category "Unknown" for the category columns that are important but has high percentage of missing values.
    - Numerical columns with missing values were replaced with median() after checking the skew
    - Columns that did not add any value for data analysis like "Prospect ID" were dropped.
    - Outlier treatment was done on the numerical columns which had outliers and value standardization was carried out.

2. EDA:
    - Checking Data Imbalance: 39% of people were converted to leads.
    - Univariate analysis and Bivariate analysis were performed.
    - Some important inferences were drawn from the analysis. Customers who spend more time on website and working professionals had high conversion rates.

3. Data Preparation:
    - Binary variables (Yes/No) were converted to 1/0
    - Created Dummy variables for categorical columns.
    - Splitting dataset into Train and Test sets (70:30 ratio)
    - Feature scaling using StandardScaler.
    - Dropping the columns that were highly correlated. "Lead Origin _Lead Add Form" and "Lead Origin_Lead Import" were dropped.

4. Model Building:
    - Feature selection using RFE.
    - Manual feature reduction using P-values by dropping features with P-values>0.05 .
    - VIF values were checked for multicollinearity. All the features had VIF <5
    - Model 2 was selected as the final model with 14 variables.

5. Model Evaluation:
    - Confusion Matrix, Accuracy, Sensitivity and Specificity were checked with the default cutoff Threshold value of 0.5. Then, ROC Curve was plotted.
    - 0.37 was then chosen as the optimal cut off probability where we get balanced sensitivity and specificity. This cutoff gave Accuracy of 80%, Sensitivity as 78% and Specificity value of 82%.
    - Precision and Recall matrix was checked with Threshold as 0.41. This gave precision score and recall score of 75%.
    - Sensitivity-Specificity view was selected for final prediction as this gave better results.
    - Lead Score feature was added to the training data set- A higher lead score means the lead is hot and it is most likely to be converted.

6. Making prediction on the Test Data set:
    - Scaling the test data set and making predictions on the test data set.
    - Drawing ROC curve on the test dataset
    - Evaluation metrics was calculated on test dataset. It gave Accuracy of 80%, Sensitivity as 80% and Specificity value of 81%. The model performed well on Test data as well.
    - Lead Score feature was added to the test data set.
    - Top 3 top features from the model are:
        - Lead Source_Welingak Website
        - What is your current occupation_Working Professional
        - Lead Source_Reference

7. Recommendations:
    - Focus more on the below features:
        - Lead Source_Welingak Website
        - What is your current occupation_Working Professional
        - Lead Source_Reference
        - Last Activity_Had a Phone Conversation
        - Last Activity_SMS Sent
        - Total Time Spent on Website
    - Welingak Website has a very high lead conversion rate. So, more budget can be spent on Welingak Website like spending more budget in terms of advertising etc.
    - Working Professionals can be targeted more as they have a very high conversion rate.
    - The leads who have sent SMS or who had a Phone conversation are more likely to convert. So, the sales team should prioritize contacting them.
    - More focus should be on the customers who have come through Reference as it has a high probability of lead conversion. To encourage more references,

the company can think of providing referral bonus or some discounts to the existing customers if their reference gets converted.

- Targeting customers who are spending more time on the website.