Summary Report

Problem statement:

Identify the set of leads of X Education so that the lead conversion rate should go up and the sales team of the company focus more on communication with the potential leads rather than making calls to every customer.

The company needs a model wherein you a lead score is assigned 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.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology:

1. Importing and Inspecting the Dataframe:

1. Importing require libraries for Dataset importing.
2. Dimensions, Missing data and Statistical aspect understanding of Dataset.

2. Data Preparation (Data Cleaning and Treatment)

1. Check for Duplicates
2. To convert the "Select" into the NaN
3. Dropping columns with more than 70% null values
4. Handling of Missing Value
5. Categorical variables encoding
6. Checking Data Imbalance

3. Exploratory Data Analysis:

1. Univariate data analysis: value count, distribution of variable (with respect to target variable and outlier treatment).
2. Bivariate data analysis: correlation coefficient.

Based on EDA various inferences have been drawn and Recommendation also mentioned.

4. Dummy Variables Creation:

1. We created dummy variables for the categorical variables.
2. Removed all the repeated and redundant variables

5. Test Train Split:

Here the dataset divided into train and test set with a proportion of 70:30 ratio.

6. Feature Scaling:

1. Here we used the Min Max Scaling to scale the original numerical variables.
2. Plotted the heatmap to check the correlations among the variables.
3. Dropped the highly correlated dummy variables to reduce the impact of Multicollinearity.

7. Model Building:

1. Classification technique Logistic regression used for the model making and prediction.
2. Using the Recursive Feature Elimination, we selected the 20 top important features.

3. Using the statistics generated, the most significant P- values selected dropped the insignificant values (more then 0.5).
4. The VIF's for these variables were also found to be good (within 5%).

8. Performance metrics for Model:

1. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity, Precision, Recall
2. We then plot the ROC curve for the features and the curve came out be pretty decent. 0.3 is the optimum point to taken it as a cut-off probability.
3. We checked the precision and recall for our final model having

   Accuracy : 91.40%
   Sensitivity : 91.59%
   Specificity : 91.28%

9. Making predictions on the test set:

1. The same model has been applied to test data set after the test data has been scaled.
2. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics found out the

   Accuracy : 90.77%
   Sensitivity : 92.15%
   Specificity : 89.91%

10. Finding the average Lead Score of the predicted converted leads and predicted not converted leads: