Summary Report

Data Overview:

Before performing any operations on the data, a general overview of the data was taken which included metadata information pertaining to columns and then metrics on numerical data viz, mean, median, quantiles. This gives a broad picture of the data and helps in further analysis.

Data Preparation:

This step was done in two parts:

- Data Cleaning:
  It was identified that the data contains 4 columns which have 'Select' as values. These values basically signify that the potential lead while entering his data either forgot or deliberately chose not to provide the information.

  It was also identified, that these 4 columns also contain null values. So, a decision on what has to be done for these columns was taken on the basis of the aggregated data which is the percentage sum of null and missing (Select) values.

  Two columns were found to have more than 75% missing values each while the other two were found to have roughly 35% percent missing values. Since these are large numbers and enough to create anomalies in the data, we eventually ended up dropping all of them along with few other columns with high missing values.

- Data Pre-processing:

For the remaining columns we imputed the missing values using medians for numeric and most frequently used for the categorical columns.

We then proceeded with treatment of outliers. Here both outlier removal and capping techniques are used. The outliers which are far away are removed while the ones which are closer are capped.

On observing the data, we found out that a large number of columns are heavily skewed. Such columns give no or little information to the model and are hence removed altogether. For the columns with comparatively lesser skewness and a variety of values, the bottom few values are clubbed together into a separate category called ‘Others’. This reduces the number of dummies created and hence few columns for the model to work upon.

IMPORTANT: Two columns, Country and Current Occupation, though extremely important in business context had to be dropped due to a plethora of missing values and imputation would lead to skewness of data. It is recommended that the relevant data be re-collected for better model building in future.

Train / Test split:

The data set is divided into the train data (data to train the model) and test data (data to evaluate the model) in the ratio of 70% to 30%.

Recursive Feature Elimination (RFE):

We use RFE to bring down the number of columns to 15.

Manual Elimination:

Manually we eliminate few more columns keeping track of the p-values and the Variance Inflation Factor. We keep p-values < 0.05 and VIF < 5.

Building the model:

We build the logistic regression model and make predictions whether the lead will convert or not.

Model Evaluation:

We evaluate the model using various parameters like sensitivity, specificity, precision and recall. We also plot the ROC curve for the model. We first do the evaluation on the test data and then on the complete data set.

Key Findings:

After running the model on the original dataset, we found out the parameters which are the major driving force behind the lead score of a person. Some of these parameters are SMS sent, Mail Subscriptions, Chat Conversations, time spent and the number of visits on the website etc. Any change in these parameters lead to significant changes in the lead score of an individual thereby affecting the probability of a person with regards to opting for the program.

The model is evaluated and we find nearly 80% accuracy in the model. The sensitivity and specificity also stand at near about 80% which means that both the converted leads and non-converted leads are correctly predicted by the model. This also aligns with the ballpark figure presented by the CEO of X Education.

Detailed analysis on how each parameter affects the lead score is also provided by leveraging scatter, bar and line plots to give a better visual understanding of the proportions in which the lead score would be impacted on change of any of these variables.