1. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

As a part of the Lead Scoring case study, we have been presented with the details how the company X Education pursues customer leads from various sources and tries to convert them to potential customers.

The current conversion rate is quite low at 30%. So we have been tasked to analyse the data and come up with a model which can make predictions to the order to 90% Lead conversion.

For this, we have proceeded with the basic analysis of the given data set.
- Identifying the columns based on Data Dictionary.
- Elimination invalid / redundant columns
- Removing records with > 40% missing data
- Imputing few columns with missing data
- Identifying the potential data columns which can factor in for accurate prediction
- Identifying the relationship and distribution of column data using graphs
- Removing the outliers in numerical variables
- Plotting heat map to see the correlations

Later, we proceeded with encoding the categorical data into Dummy variables so that we can easily convert them into features which can be fed into a Model used for predictions.

The Others, Unknown values that transformed into columns are dropped from the Dummy columns.

The data is then split into training and test data in ratio of 70:30.

The training data is scaled to avoid any disparities in magnitude of the data values impacting the model prediction.

The training data is fed into a Generalized Linear Model (GLM).

The ineffective variables are eliminated using RFE and VIF. We are then left with 14 variables + 1 constant which has been able to predict the training data set at more than 90% accuracy and precision.

The same model has been applied to test data set after the test data has been scaled. And we have also observed more than 90% accuracy & precision there as well.