Summary Report

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

Steps:-

- Imported the required libraries and dataset. Inspected the records of the dataset. Checked the data dictionary provided for the column information.
- There were multiple categorical columns(i.e. 'Specialization', 'How did you hear about X Education', 'Lead Profile', 'City') having select as values. Converted these values to null values by using np.nan since these values are default values present from the drop down list (i.e. not selected by the customer).
- Checked for missing records and dropped the columns having high percentage of missing values. Also dropped the columns that were highly skewed.
- Conducted univariate analysis using box-plot to identify for outliers present in the records. Verified the same using inter-quantile ranges. Also used skew method to identify the outliers. Skewness explains the extent to which the data is normally distributed. Ideally, the skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values. Capped the outliers using the using IQR (Inter quantile range) method.
- Conducted binning of categorical variables having many values .Created dummies for categorical values. Split the dataset into test and train dataset. Performed scaling using Standard Scaler for continuous variables.
- Using Recursive Feature Elimination (RFE) for selecting the top 15 features. Build the logistic model and analyzed the stats. Checking the VIF (Variance Inflation factor) to identify the correlation between the predictor variables. Dropped the records which were highly insignificant (i.e. records having high p-value and VIF) and rebuilt the model.
- Created the confusion matrix and calculated accuracy, sensitivity, specificity, recall, precision and F1 score for the model on train data. Plotted the ROC Curve .Obtained the optimum cutoff point to calculate the probability. Made predictions on the test dataset and calculated all the above parameters.

Learnings and Challenges:-

- Learnt about skewed columns as they don’t have much influence on the model.
- Dropped the columns that had only 1 value present in the dataset.
- Major challenge was there were many categorical column with huge number of distinct records. This resulted in generating multiple columns for dummy variables. Had to skew the less percentage of records into a particular category so that less dummies were created. Used RFE to check if any of these columns were selected. If not then had to go back to the code and drop the column itself. Had to go back and forth multiple times.
- Learnt how to cap outliers so that records with high values are not dropped from the dataset. Also learnt skew method the check the outliers in the dataset. Skewness explains the extent to which the data is normally distributed. Ideally, the skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values.

- Learnt about coarse and fine tuning using RFE and manually selection of features respectively.
- Understood the confusion matrix and the significance of the parameters (i.e. how to check the model performance of the data on both train and test dataset).
- Learnt to plot and identify the optimal cut off of value for that dataset for that hot leads can be separated from the cols leads.
- Identified the top 3 attributes/parameters contributing to a lead getting converted (i.e. to 1).