Summary Report

Process Followed

1. **EDA**: Understand the various columns (variables)
    1. Type of data (numerical, categorical – ordinal or nominal)
    - Presence and amount of missing/invalid values. If categorical variables contain Select, this should be considered as missing value and needs to be replaced with NaN
    2. Presence of outliers
2. **Data pre-processing**
    - Impute Missing values:
        - If most values (>45) are missing, the column was dropped
        - If very less values (<2%) are missing, the rows were dropped or these were imputed with mode for categorical variables
        - If the variable made business sense, the missing values were replaced with a new category “UNKNOWN”
        - The numerical values had outliers so the median value was used to impute missing values.
    - Handle Outliers: No specific handling of outliers was done but this information was used in missing value imputation.
    - Create Dummy variables for the categorical variables
    - Test train split: The dataset was split into training (70%) and test(30%) datasets
3. **Build model**
    - Standardise numerical values of the training dataset
    - RFE (Recursive Feature Elimination) was done to automatically retain only 15 important features
    - Logistic Regression model was built
    - The p and the VIF values of the variables were checked. Variables that had a p-value > 0.05 or VIF > 5 were removed (one-by-one) and the model was rebuilt.
    - The above 2 steps were repeated till there were no variables that had a p-value > 0.05 or VIF > 5
4. **Evaluate model on the training dataset**
    - Predict the probability for each data point to get its lead score.
    - For various cut-off thresholds (0.0, 0.1, 0.2….0.8, 0.9) calculate the accuracy, sensitivity and specificity. Plot these to find the point of intersection which is the **optimal cut-off threshold**.
    - Use this optimal cut-off threshold to generate the class
        - if lead score >= optimal threshold Class is 1-Converted
        - if lead score < optimal threshold Class is 0-Not Converted
    - Calculate the accuracy and the sensitivity*

5. **Evaluate the model on the test dataset**
    - Update test dataset to include only the variables that we have identified in the previous step and standardise numerical values
    - Use the optimal cut-off threshold and generate the class
    - Calculate the accuracy and the sensitivity*

If sensitivity > 80%, the model is good and we can proceed with next steps. Else repeat the steps 3, 4 & 5 (with changes in the variables selected) till we get an acceptable sensitivity value.

Learnings

- A good **understanding of the problem domain** helps in feature selection and interpreting the model.
- **Data pre-processing**, though takes the maximum amount of time in the model building, it is an essential step. Bad data will result in a bad model.
- **Feature selection** (variables that give a good model) and the evaluation (how good the model is) steps are most critical to generate and pick the best model.
- Depending on the requirement (whether a good precision or recall is needed), we need to **adjust the cut-off threshold** to generate the classes for the datapoints.