EXECUTIVE SUMMARY

➢ ASSIGNMENT STRATEGY:

To meet the business objective and build logistic regression model, we followed steps as follows

1. Data pre-processing
   - 1.1. Data Understanding, Cleaning
   - 1.2. Address Missing Values
2. Exploratory Data Analysis
   - 2.1. Univariate Analysis for Categorical Variables & Numerical Variables
   - 2.2. Bivariate Analysis for Numerical Variables
3. Build a logistic model
   - 3.1. Convert a categorical variable into dummy variables
   - 3.2. Separate features and the target variable (Converted)
   - 3.3. Split the data into training and testing sets
   - 3.4. Standardize numerical columns
   - 3.5. RFE for Feature Selection
   - 3.6. Build a logistic model using statsmodels
   - 3.7. VIF for Detecting Multicollinearity
   - 3.8. Drop features based on VIF and P-value
   - 3.9. Rebuild the model
4. Evaluate the model
   - 4.1. Confusion matrix
   - 4.2. Plotting ROC curve
   - 4.3. Calculate accuracy, sensitivity and specificity for each probability cut-offs
   - 4.4. Choose Optimal Probability Cut-off
   - 4.5. Making Predictions on the test set
   - 4.6. Compare the metrics between train set and test set
   - 4.7. Assign lead score for each of customers

➢ LEARNINGS:

- Data Cleaning is an important step in building ML model as it can impact the model prediction ability

- Spending time on Exploratory Data Analysis (EDA) is critical to understand data, data distribution, identify and handle missing values & outliers, gain insights and also serve as a base to guess important features that may contribute to the target variable
- Domain Knowledge – It is important to have an understanding on the domain to carry out successful data pre-processing, for example, with domain knowledge, we can decide on the variables which will not contribute to the model and simplify model building process (in this case study, e.g. Lead Quality was an intuition-based variable and hence not significant for model building)
- Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) – To identify most important features, it is important to perform RFE and VIF checks. RFE helps to select 'Top N' most important features for prediction with model and VIF helps to identify and handle multicollinearity which can impact model performance
- Appropriate Model Selection – Logistic Regression model is a good choice for classification problem
- ROC Curve – Using ROC curve to find optimal cut-off value in logistic regression is essential to get optimal model performance
- Model Performance Metrics – We should compare different model performance metrics (beyond Accuracy) to compare model performance on Train and Test data to determine generalization of model on test data. These metrics for Logistic regression includes Confusion Matrix, Specificity, Sensitivity, Precision, Recall. Focusing on which metric improvement depends on the use case