Summary Report

Approach

1. CRISP-DM: -
   - Removed variables that have >= 30% missing values
   - Imputed inconsistent values with NULL Value
   - Imputed Null values using Mean and Mode values
   - Outliers handled by capping the values at 99th percentile
   - Engineered features to enable better visual experience for EDA
2. Exploratory Data Analysis (EDA)
   - Created a deep copy of primary dataframe for visualization
   - Univariate, Bivariate & Multivariate analysis for both Numerical and Categorical features
   - Used Vertical and Horizontal Count graphs, Pair plots, Heat Maps and Box plots
3. Data pre-processing for Model building
   - Variable encoding – Yes/No to 1/0
   - One hot encoding for multi-class categorical features
   - Normalization of Test and Train datasets
   - Usage of RFE for automatic variable reduction
4. Model Building
   - Used Statsmodels summary to analyze the significance and weight of features
   - Leveraged VIF calculation to analyze features exhibiting multicollinearity
   - Removing features that proved to be insignificant and multicollinear
5. Model Evaluation & Tuning
   - Predicted probabilities on Train dataset
   - Used a random cut-off value of 0.5 to attain predicted target classes (0 and 1)
   - Evaluated the model through Accuracy, Senstivity / Recall, Specificity and F1 score
   - Evaluated the effectiveness of model through ROC curve.
   - Finding Optimal Cut-off value by leveraging Accuracy, Sensitivity/Recall and Specificity interception curve
   - Utilized optimal Cut-off value to tune the predicted target classes (0 and 1)
   - Evaluated the model again through same metrics to ascertain the stability and robustness of the model
6. Generate Lead scores (Test data)
   - Used the optimal cut-off value to assign a lead score to each of the leads

Learnings

1. Not only did we find out Null values but also inconsistent classes (e.g. Select)
2. There were features that lacked variability (e.g. Only Nos (No Yes) or India as a class with 95% data points)
3. EDA didn't provide any indication of Linear relationship between numerical features
4. However, Heatmap between all the features indicated that multiple features did have multicollinearity
5. Finalized model provided the following insights
   - Sales team should engage Working professionals (2.8292) via Calls (3.5562), Emails (-1.6389) and SMS (1.3515)

- Business should focus on optimizing the following lead sources - Lead Add Form (3.7498), Welingak Website (1.9509), Olark Chat (1.2647)

6. Area Under the Curve (AUC) for ROC curve is 0.88 which proves that the model has a high discriminatory power
7. Accuracy, Sensitivity / Recall and Specificity interception curve provided the Optimal cut-off value of 0.30, which was eventually used for Lead scoring as well
8. Below are the Evaluation metrics and scores obtained on Train Dataset

Metric	**Train @ 0.3 cut-off**	**Test @ 0.3 cut-off**
Accuracy	0.79	0.80
Sensitivity / Recall	0.83	0.83
Specificity	0.77	0.78
F1 Score	0.75	0.77