SUMMARY

The Objective of the case study was to predict Hot Leads i.e., Potential Leads among all the Leads such that Sales team reach out potential leads first to achieve maximum lead conversion. We were provided data beforehand so that we can analyse & build a predictive model.

Below are the steps that were followed while building the model-

1. **Data Reading & Understanding**
   - Firstly, we imported the data & stored it into a data frame. Then we looked at its shape, size, info, & description to begin with.

2. **Data Cleaning**
   - Data Cleaning is one of the most important steps that should be done first before proceeding with any other step. We observed dataset had columns having null values. Those columns with null value percentages greater than 40% were removed.
   - Categorical columns having only one category were removed, as they won't be helpful for analysis.
   - Columns having highly skewed data were also removed, to counter any possible data-imbalance.
   - Columns having too many sub-category levels were treated so that when dummy variables are created, we don't have too many dummy variables to deal with. Sub-category levels having very low frequency were clubbed into one group.
   - Apart from column wise null value analysis, row wise null value analysis was also done. Rows having greater than 50% null value presence were removed (if any).
   - Finally proper null value imputation technique was used.
   - We found very few columns had outliers, so outlier treatment was done.

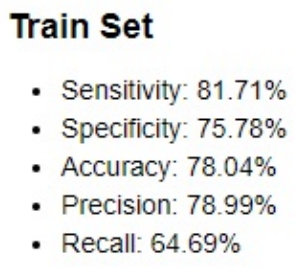
3. **EDA**
   - After data cleaning we proceeded with EDA where we took a closer look at different variables to observe features in detail.
   - We found out sales team specific generated data won't be helpful in analysis as the model being built will be used by sales team before hand to identify Hot leads. So, we removed those columns (if any present till EDA) before moving onto Data Preparation.

4. **Data Preparation**
    - Dummy variables were created for categorical columns having different sub-category levels.
    - For numerical columns scaling was done using Standard Scaler to bring data into one scale.
    - Data was split into Train & Test sets with 70-30 ratio.

5. **Model Building**
    - Balanced approach was used while building the model i.e., Automated (RFE) + Manual Feature Elimination (p-value & VIF).
    - Firstly, RFE was used to attain top 20 features.
    - Then p-value & VIF value were used with backward feature elimination method i.e., one feature was removed at a time until we reached optimum model with all features having p-value < 0.05 and VIF value < 3.

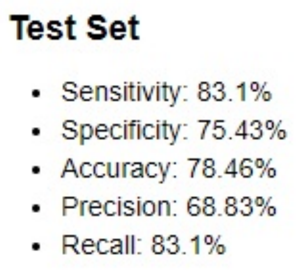
6. **Model Evaluation**
    - To check the model's performance, the model was trained on train set & was then tested on test set.
    - Parameters like Sensitivity, Specificity, Accuracy, Precision & Recall were used for evaluating model performance.

7. **Result**



Train Set

- Sensitivity: 81.71%
- Specificity: 75.78%
- Accuracy: 78.04%
- Precision: 78.99%
- Recall: 64.69%



Test Set

- Sensitivity: 83.1%
- Specificity: 75.43%
- Accuracy: 78.46%
- Precision: 68.83%
- Recall: 83.1%

➢ We faced difficulty while finding the optimal cut-off at which the model's performance was at its peak. So, we run our model a couple of times with different cut-off's ranging from 0.310 to 0.340 & found best output using 0.314 as cut-off.