Lead Scoring Case Study Summary

Step 1 – Reading and Understanding Data

Step 2 – Understand Data frame, Check Count and Percentage of Unique, Null values.

Step 3 – Data Cleaning

- Drop columns having more than 45% Null values.
- Observe data in Numerical and Categorical Columns.
- Missing values handling.
- Outlier handling with univariate and Bivariate analysis.
- Multivariate analysis with correlation
- Dummy variables creation for categorical variables

Step 4 – Train test split: Define X and Y and split data into 70:30 ratio.

Step 5 – Feature scaling: Scale features

Step 6 – Model Building: Start with first model building.

Step 7 – RFE: select 15 features using RFE.

Step 8 – Drop features having highest P-Value one by one during each model creation iteration.

Step 9 – Drop features with high VIF Values > 0.05 and finalize model.

Step 10 – Evaluated the model accuracy with 81.71% and created confusion matrix.

Step 11 – Plot ROC Curve, which has 89% of area under curve.

Step 12 – Finding Optimal cut-off point, 0.35 is the optimum point to be considered as cutoff probability.

Metrics of the model is given below.

- **Accuracy:** 80.67
- **Specificity:** 81.48
- **Recall / Sensitivity:** 79.36
- **Precision:** 72.54
- **F1 Score:** 75.79

Step 13 – Prediction on test data set is done with below metrics for finalized model.

- **Accuracy:** 79.73
- **Specificity:** 80.8
- **Recall / Sensitivity:** 78.08
- **Precision:** 72.64
- **F1 Score:** 75.26

Step 14 – There are 12 Features finalized, which are important for higher probability of lead conversion rate. 10 of the features if have higher value will impact positively, while 2 of them if value decreases will impact negatively.