Summary of Lead Scoring Case Study

By Harshit Kamani & Gaurav Makhija

The following steps are used:

1. **Data Cleaning / Preparation:**
   a. Binary variables were converted from Yes/No to 1/0
   b. 'SELECT' option in categorical variables was converted to 'NaN'
   c. Unnecessary columns like those having only 1 unique values were dropped.
   d. Missing values were handled – dropping those columns having more than 35% of Null Values and imputing others as per the need.
   e. Options having low representation of categories in Categorical Column were merged together.

2. **EDA:**
   a. All categorical variables impact on target variable 'converted' was observed using charts.
   b. outliers were checked and treated with percentile capping in numerical columns.

3. **Dummy Variable Creation & Correlation Checking:**
   Dummy variables were created for all Categorical Variables and correlation was checked between all the variables, dropping those having high correlation.

4. **Train-Test split & Scaling:**
   The split was done at 70% and 30% for train and test data respectively, with random_state kept at 100. And train set was scaled using fit_transform method.

5. **Model Building:**
   First model was build on train data and RFE was used to select top 15 relevant variables.

6. **Model Evaluation:**
   a. necessary elimination of columns based on p-Value and VIF value was done from 15 columns that were selected by RFE.
   b. ROC curve was plotted and optimal cut off point (0f 0.42) was selected/observed based on the curve.
   c. Accuracy percentage was measured and confusion matrix was made to measure sensitivity & specificity of our model which came out to be 83.37%, 85.03% % 82.33% respectively.

7. **Precision – Recall:**
   This method was also used to recheck and a cut off of 0.42 was found with Precisionaround 75% and recall around 85%.

8. **Predictions on Test Data:**
   Accuracy percentage was measured and confusion matrix was made to measure sensitivity & specificity of our model which came out to be 81.96%, 82.40% % 81.71% respectively.