Lead Scoring Case Study Summary

Below are the steps we followed to build the model and achieve a target lead conversion rate of 80%.

Step1: Reading and Understanding Data:

Read and inspected the data like, size of the data frame, the data type of each feature, statistical information and got a basic understanding the features included.

Step2: Data Cleaning

a. Few categorical variables have a level called 'Select' which means the leads did not choose any given option. So, replaced the 'Select' values in the data set with null.
b. We dropped the columns having null values greater than 40%.
c. Next, we inspected the columns having comparatively higher null values and imputed them with Mode value or using 'Other' value for categorical features.

Step3: Exploratory Data Analysis

a. Performed univariate, bivariate analysis and multivariate analysis for categorical and numerical variables.
b. Columns with huge imbalance are dropped. For instance, in 'Newspaper Article' feature all customers choose 'No' option creating imbalance in that feature.
c. In case of numeric features, outliers were identified, and we imputed the outliers to 5% or 95% value for analysis.
d. In case of categorical features, created new classification variables to group all the low frequency values. For example, in 'Last Activity' feature, all the values with low count were categorised as 'Others'.
e. Dropped few features which we felt not relevant for model building like 'Lead Number', 'Tags', 'Country'.

Step4: Data Preparation

a. Changed the binary variables into '0' and '1'
b. We created dummy variables for the categorical variables.
c. Splitting the data into train and test set in the ratio 70-30% values.
d. Feature Scaling using Standardization.

Step5: Model Building:

f. Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.

g. Using the statistics generated, we recursively created models and tried looking at the P-values to select the most significant values that should be present and dropped the insignificant values.
h. Finally, we arrived at the 13 most significant variables. The VIF's for these variables were also found to be good.
i. Then we made predictions on the train data set.
j. We then plotted the ROC curve for the features and the curve came out be decent with an area coverage of 89% which further solidified final model.
k. We then calculated the Lead score for each lead using the calculated probability prediction values.
l. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity, and the optimal cut-off is 0.34
m. Next, based on the Precision and Recall trade-off, we got a cut off value of approximately 0.4.
n. Since we got a higher recall for the cut-off value of 0.34, we decided to go with that value.
o. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.39%; Sensitivity= 80.49%; Specificity= 80.33%.