Lead Scoring Case Study Summary

Summary:

Step 1: **Reading the Data**.

Importing and reading the data.

Step 2: **Data Cleaning**

We dropped the variables that had a high percentage(above 40%) of NULL values in them. We imputed the missing values with the mode in the case of categorical variables. The outliers were identified and capped the outliers with a 99% percentile value.

Step 3: **EDA**

We started with the Data Analysis to get a feel of how the data is oriented. There were around 19 variables that were identified to have unbalanced values, which were dropped.

Step 4: **Creating Dummy Variables**

we created dummy data for the categorical variables.

Step 5: **Train Test Split**:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step 6: Feature Rescaling

We used Standard Scaling to scale the original numerical variables. With the stats model, we created our initial model, which gave us a complete overview of the statistical pattern of the model.

Step 7: **Feature selection using RFE**:

Using the RFE we selected the 15 top essential features. Based on the statistics generated, we recursively eliminated the features with insignificant p values each at a time.

Finally, we arrived at the 13 most significant variables. The VIFs for these variables are good (VIF <5).

We created a data frame with converted probability with an initial assumption that a probability value of more than 0.5

Accd to the above assumption, we derived the Confusion Metrics and accuracy of the model.

We calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

Step 8: ROC Curve

We plotted the ROC curve for the features and the curve was pretty decent with an area coverage of 87% which further solidified the model.

Step 9: Finding the Optimal Cutoff Point

We plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probabilities. The intersecting point of the graphs was considered the optimal probability cutoff point. The cutoff point was found to be 0.34

Based on the new value we could observe that close to 81% values were rightly predicted by the model.

We could also observe the new metrics of the 'accuracy=81%, 'sensitivity=80.08%',

and'specificity=80.53%'.

We calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step 10: Computing the Precision and Recall metrics
we also found out the Precision and Recall metrics values came out to be 72.04% and 80.08% respectively on the train data set.

Based on the Precision and Recall tradeoff, we got a cut-off value of approximately 0.42

Step 11: **Making Predictions on the Test Set**
Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found the accuracy value to be 80.24%; Sensitivity=79.06%; Specificity= 80.91%.