Leads Scoring Case Study Summary

Below are the steps how we have proceeded with our assignments:

1. Data Cleaning:

a. the First step to clean the dataset we chose was to remove the redundant variables/features.

b. After removing the redundant columns, we found that some columns have been labeled as 'Select,' which means the customer has chosen not to answer this question. The ideal value to replace this label would be null as the customer has not opted for any option. Hence, we changed those labels from 'Select' to null values.

c. Removed columns having more than 30% null values

d. For the remaining missing values, we have imputed values with the maximum number of occurrences for a column.

e. We found that one column has two identical label names in different formats (capital letter and small letter). We fixed this issue by changing the names of the labels into one form.

2. Data Transformation:

a. Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.

b. Checked the outliers and created bins for them.

c. Removed all the redundant and repeated columns.

3. Data Preparation:

a. Split the dataset into train and test datasets and scale the dataset.

b. After this, we plot a heatmap to check the correlations among the variables.

c. Found some correlations, and they were dropped.

4. Model Building:

a. We created our model with rfe count 19 and 15, compared the model evaluation score like AUC, and chose our final model with rfe 19 variables as it has more stability and accuracy than the other.

b. For our final model, we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity, and specificity.

c. We found one convergent point and chose that point for cutoff, and predicted our outcomes.

d. We checked the precision and recall with accuracy, sensitivity, and specificity for our final model and the tradeoffs**.**

e. Prediction made now in the test set and predicted value was recoded.

f. We did a model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is

g. We found that our final test model's score of accuracy and sensitivity is in an acceptable range.

h. We have given a lead score to the test dataset to indicate that high lead scores are hot leads and low lead scores are not hot leads.

5. Conclusion:

Learning gathered are below:

- We have a higher recall score than the precision score.
- The Accuracy, Precision, and Recall are in the acceptable range.
- The model can adjust to the company's upcoming requirements.
- The model is in a stable state.
- Important features responsible for a good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

    i) Last Notable Activity_Had a Phone Conversation

    ii) Lead Origin_Lead Add Form and

    iii) What is your current occupation_Working Professional