Summary

1. Data Cleaning:

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

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

c. Removed columns having more than 30% null values \

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

e. We found for one column is having two identical label names in different format (capital letter and small letter). We fixed this issue by changes the labels names into one format.

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 dataset and scaled 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 counts and compared the model evaluation score like AUC and choose our final model with rfe variables as 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 points and we chose that point for cutoff and predicted our final outcomes.

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

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

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

g. We found the score of accuracy and sensitivity from our final test model is in acceptable range.

h. We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.