Leads Scoring Case Study

A brief summary report in 500 words explaining how we proceeded with the assignment and the learnings that we gathered post that.
Answer:
Mentioned below are the steps that we followed in our assignment:

1. **Data Cleaning:**
a. Cleaned the dataset by removing redundant and repeated features.
b. After removing redundant columns, we found some columns having label as 'Select' which means the customer had chosen not to answer that 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.
c. Removed columns having more than 30% null values
d. We imputed the remaining missing values with maximum number of occurrences for a column.
e. Renamed 2 labels with same name into one common one

2**. Data Transformation:**
a. Changed multicategory labels into dummy variables and binary variables into '0' and '1'.
b. Checked outliers and created bins for their treatment thereafter.

3. **Data Preparation:**
a. Splitted the dataset into train and test dataset and scaled the dataset.
b. Plotted a heatmap to check the correlations among the variables

4. **Model Building:**

a. We created our model with rfe count 20 and compared the model evaluation score like AUC and chose our final model with rfe 20 variables basis it's stability and accuracy.
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 it for cutoff to predict our final outcomes.
d. We checked precision and recall with accuracy, sensitivity and specificity for our final model and tradeoffs.
e. Prediction was made in test set and predicted value was recorded.
f. We did model evaluation on test set like checking accuracy , recall/sensitivity to find how the model is
g. We found score of accuracy and sensitivity from our final test model being in acceptable range.
h. We gave lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.

5. **Conclusion:**

Learnings gathered are :

- i. Test set is having accuracy, recall/sensitivity in an acceptable range.
- ii. In business terms, our model is having stability and accuracy with adaptive environment skills. In business terms, this model has an ability to adjust with the company's requirements in coming future
- iii. Top features for good conversion rate are:

➢ Last Notable Activity_Had a Phone Conversation

➢ Lead Origin_Lead Add Form