Leads Scoring Case Study

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

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

1. Data Cleaning:

a. After inspecting our dataset, we checked for the null values for the variables.

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 70% 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 and binary variables, into '0' and '1'.

b. Handled the outliers in the data.

c. Dropped the columns which were not relevant or were not important.

3. Data Preparation:

a. Split the dataset into train and test dataset and scaled the dataset, using StandardScaler.

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

4. Model Building:

a. We created our model with RFE count 15 and compared the model evaluation score as it was more stable and accurate than the other counts.

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

c. We found one convergent points and we chose that point for cut-off and predicted our final outcomes.

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

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 an 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.

5. Conclusion:

Learning gathered are below:

- ✓ Test set is having accuracy, recall/sensitivity in an acceptable range.
- ✓ In business terms, our model is having stability an accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.
- ✓ Top features for good conversion rate:
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Tags_Will revert after reading the email