Lead Scoring Case Study

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

Data Cleaning:

- Eliminating unnecessary variables and cleaning the data is the first stage.
- Upon removal, we discovered that the columns' labels were "Select," which we must change to null values.
- Columns with more than 35% null values were eliminated.
- We have substituted the maximum number of occurrences of the column for the remaining missing values.
- We have some data with All Capital or All small values by replacing it with their correct format.

Data Transformation:

- Modified the multicategory labels to be binary variables with the values "0" and "1."
- Dummy variables were made for a few variables.
- Checked the outliers and removed some of the numbers using 0.99- 0.1% analysis.

Data Preparation:

- Splitting the dataset into train and test dataset.
- Scaled the dataset using the StandardScaler().
- Plotted heatmap for finding the correlations and dropping them.

Model Building:

- We build our model with the help of RFE with 19 variables.
- Checked the VIF Score for each variables, as all of the variables are having VIF Score < 5.0, we proceed to our next step.
- We then removed the insignificant variables using the P-Value Score.
- For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
- We found one convergent points and we chose that point for cutoff and predicted our final outcomes.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs.
- Prediction made now in test set and predicted value was recoded.
- We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is.
- We found the score of accuracy and sensitivity from our final test model is in acceptable range.
- 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.

Conclusion:

Learning gathered below:

- Test set is having accuracy, recall/sensitivity in an acceptable range.
- In business terms, our model is having stability and accuracy with adaptive environment skills. Means it will adjust with the company’s requirement changes made in coming future.