Lead Scoring Case Study Summary

Solution Summary:

Step1: Reading and Understanding Data:

Read the data dictionary and inspected the data.

Step2: Data Cleaning:

a. As a First step, we chose to drop the variables having unique values.
b. Then, there were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
c. We dropped the columns having NULL values greater than 52%. Lead Quality is a feature having 52.9% missing values, we chose to keep this feature as it seems important, for imputation we think that null value indicates that the employee may not be sure about option so we imputed with Not Sure.
d. Next, we performed data cleaning and preprocessing. We eliminated variables that were skewed or duplicated. We filled in the missing values with median for numerical variables and created new categories for categorical variables. We detected and removed outliers. We also fixed a column that had inconsistent labels (lowercase and uppercase) by converting them to uppercase.
e. All sales team generated variables were removed to avoid any ambiguity in the final solution.

Step3: Data Transformation:

Changed the binary variables into '0' and '1'

Step4: Dummy Variables Creation:

a. We created dummy variables for the categorical variables.
b. Removed all the repeated and redundant variables.

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.

Step6: Feature Rescaling:

a. We used the Standard Scaling to scale all the variables.
b. Then, we plot the heatmap to check the correlations among the variables.

Step7: Model Building:

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
b. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
c. Finally, we arrived at the 12 most significant variables. The VIF's for these variables were also found to be good.
d. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
e. We then plot the ROC curve for the features and the curve came out to be pretty decent with an area coverage of 95% which further solidified the model.
f. Then, checked if 80% cases are correctly predicted based on the converted column.
g. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on the train set.
h. Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.25.
i. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 90.78%%; Sensitivity= 84.12%; Specificity= 94.58%.

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 84% on the final predicted model which clearly meets the expectation that the CEO has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.
- Features which contribute more towards the probability of a lead getting converted are:
    i. Tags_Lost to EINS
    ii. Tags_Closed by Horizzon
    iii. Tags_Will revert after reading the email