Lead Scoring Case Study Summary

Step1: Reading and Understanding Data:

Read and inspect the data.

Step2: Data Cleaning:

The variables with unique values were removed as the first stage in cleaning the dataset we selected.
Then, a few columns had the value "Select," indicating that the leads had not chosen any of the available options. These values were modified to be Null values.
We removed the columns with a single value.
Next, we eliminated the duplicate and unbalanced variables. In addition, when necessary, missing values were imputed using median values for numerical variables and new categorization variables were created for categorical data. The outliers were found and eliminated. Also, the same label appeared in one column in several circumstances. This problem was resolved by changing the label's small-capital first letter to an uppercase letter.

Step3: Data Transformation:

Changed the binary variables into ‘0’ and ‘1’

Step4: Dummy Variables Creation:

We created dummy variables for the categorical variables.
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:

We used the Min Max Scaling to scale the original numerical variables.
Then, we plot a heatmap to check the correlations among the variables.
Dropped the highly correlated dummy variables.

Step7: Model Building:

Using the Recursive Feature Elimination, we went ahead and selected the important features.
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.
Finally, we arrived at the 11 most significant variables.
For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
We then plotted the ROC curve for the features and the curve came out to be pretty decent with an area coverage of 85% which further solidified the model.

Then, checked if 80% cases are correctly predicted based on the converted column.
We checked the precision and recall with accuracy, sensitivity and specificity for our final model on the train set.
Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3.
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 80%; Sensitivity= 80%; Specificity= 72%.

Step 8: Conclusion:

The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted model which clearly meets the expectation of 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:
Lead Add Form,
occupation_Working Professional,
Total Time Spent on Website.