Summary Report - Lead Scoring Assignment

We have been given information on how company X Education seeks customer leads from various sources and tries to convert them into potential customers as part of the Lead Scoring case study.
At 30%, the conversion rate is now relatively low. As a result, we were given the assignment to examine the data and develop a model that could forecast a lead conversion rate of 80% or more.

1. Data Cleaning:

- Treat ‘Select’ as missing values since Select means customers did not choose any options.
- Delete the columns that have more than 40% of missing values. For the columns that have less than 40% of missing values, impute the missing values with mode. For the columns that have less than 5% of missing values, drop the missing values.
- Detect outliers for the continuous variables and replace them with the NULL value. Then drop them since these NULL values are less than 3%.

2. Exploratory Data Analysis:

- Draw different charts to compare Converted customers and Not converted customers and how the variables and features affected them.
- Discover many columns just have ‘No’ values which I could not infer anything.
- People who go to the website and spend more time than normal are promising leads. A lead with more website visits overall is somewhat more likely to be promising.

3. Data Preparation:

- Converting some binary variables (yes/no) to 0/1
- Create dummy features (one-hot encoded) for categorical variables with multiple levels.
- Drop the repeated variables
- Split the data into train set and test set in the ratio of 70:30.
- Feature scaling for continuous variables using a standard scaler.

4. Model Building:

- Build the model and select features using RFE.
- Drop the features with P-values higher than 0.5
- Check the VIFs score for the features and drop them one by one until all the VIFs are below 5.
- Predict the train set.
- Make the predictions on the test set.
- Comparing the values obtained for the Train set & Test set

5. Model Evaluation:

- Draw the ROC curve to see the effectiveness of the model
- Find the Optimal Cutoff Point for sensitivity and specificity
- Decide to choose a cut-off point of 0.3 that has the optimal accuracy, sensitivity, and specificity.
- Check the precision and recall rate.
- Evaluate the test set
- Comparing the values obtained (accuracy, sensitivity, specificity) for the Train set & Test set
- Show the most valuable features and least related features.

6. Conclusion:

- We have succeeded in achieving our objective of estimating the desired lead conversion rate to be around 92%. The sensitivity is 90.8% and the specificity is 94.0%. We should be able to provide the CEO with the confidence to make decisions based on this model.