Lead Scoring Assignment Summary

Following are the steps involved in the modeling:

Step 1: Reading and Understanding the dataset

- Dataset was loaded, a quick review was done
- The shape of the dataset was (9240, 37)
- 'Converted' was identified as the target variable
- The conversion rate was 38.5%

Step 2: Exploratory Data Analysis

- It was found that the dataset had many null-values, few variables had dominating values, few variables had the default 'Select' value assigned.
- During data cleaning, the variables whose null-value percentage was above 30% were dropped immediately. The categorical variables which cannot be dropped were imputed with the mode. Numerical variables which cannot be dropped were imputed with median.
- Variables with dominating values (presence of a single value in huge quantity) were dropped since they don't provide any value
- Outliers were checked on the numerical variables by plotting box plots and statistical summary
- Graphs were plotted with the feature variables vs target variable and insights were drawn. Many variables which lead the conversion and many variables which doesn't help in conversion were identified. Recognizing both of this feature variables are business critical.

Step 3: Data Preparation for Modeling

- Dummy variables were created for the relevant variables such as 'Lead origin', 'Lead Source', 'Last Activity', 'What is your current occupation', and 'Last Notable Activity'.
- Categorical variables with binary values were encoded with '1' & '0'
- Heatmap were plotted and it was understood that there were 2 groups of 3 variables each which are highly correlated.
- 70% of the dataset were allocated for train set and the rest 30% for the test set.

Step 4: Model Building

- In the initial training model there were many feature variables which were not significant.
- Feature variable selection were made based upon the RFE and 20 variables were selected by the RFE.
- Model building were done one by one and in each case the variables with higher p-values (values > 0.05) were dropped one by one until we reached the final model were all the variables were below 0.05. A total of 6 models were built.

Step 5: Prediction & Model Evaluation based on the final model

- Predictions and Confusion matrix were made based upon the final model and Accuracy, Sensitivity, Specificity and Precision were determined.
- True Positive Rate vs False Positive Rate ROC curve were plotted.
- Plot for accuracy sensitivity and specificity for various probabilities were plotted and the optimal cut-off were calculated to be 0.36.
- Precision – Recall plot were plotted and the optimal cut-off were found out to be 0.42. This cut-off value was disregarded
- Prediction on test set were done and the values of Accuracy, Sensitivity, Specificity and Precision values were calculated.

Step 6: Conclusion

- Accuracy, Sensitivity, Specificity and Precision values of train set are 81.36%, 80.92%, 80.71% and 72.13% respectively
- Accuracy, Sensitivity, Specificity and Precision values of test set are 81.17%, 82.27%, 80.46% and 73.21% respectively
- The values of Accuracy, Sensitivity and Specificity for both train and test are closer and hence we concluded that this is a good model.