Summary

1. Data Cleaning and Preparation

-Dropped columns that do not add value to the analysis like Lead Number, Prospect ID.

- It was observed that a lot of columns have “Select” as values, so replaced it with “np.nan”.

-Checked the Null value percentage of all the columns and dropped columns with Null value percentage>45%.

-Imputed Null values of the remaining columns with the “mode” of the column.

-Checked for columns with outliers by plotting a boxplot. Removed top 3% and bottom 1% of outliers from the data.

-Created Dummy variables for all the multiple categorical columns.

-Converted columns with binary category as “Yes” and “No” to numerical “1” and “0”.

-Dropped the redundant columns from the dataset.

2. EDA

- Plotted boxplot to identify columns with outliers like “TotalVisits”, “Total Time Spent on Website”, “Page Views Per Visit”.

- Plotted heatmap to find variables with high correlation.

-By plotting a countplot of “Last Activity” column, it was analysed that conversion is higher for folks that have last activity as SMS Sent, Email Opened. By plotting a countplot of “Lead Score”, it was analyzed that conversion rate is higher for folks that come through reference, Google.

3. Splitting the Train and Test data

-Assigned column “Converted” as “y” since it is the target variable and the rest of the columns as “X”.

-Split the dataset into Train and Test data.

-Scaled numerical variables in the dataset using Standardization. Fitted and transformed the training dataset.

4. Building the model and evaluating metrics

- First, Recursive Feature elimination (RFE) was performed to select the top 15 columns.

- Built a model (LRModel1) using the top 15 columns. Dropped the columns with high p-value(p-value>0.05) and built the model again (LRModel2).

- Calculated Variance Inflation Factor (VIF) of the columns. Dropped columns with high VIF(VIF>2) and built the model again.

-Since the p-value and VIF of all the variables were fine, finalized the model (LRModel4).

-Predicted "y_train" using the final model as "y_train_pred" and built the confusion matrix using actual and predicted values.

- To determine the optimal cut-off point, accuracy, sensitivity, and specificity were plotted and the optimal cut-off point was found to be "0.3".

- The train data had a high accuracy score of 0.92, sensitivity of 0.91, and specificity of 0.92.

-Plotted the ROC curve and got an AUC value of 0.97, which is indicative of a good model.

- For the test dataset, scaled the test data by fitting the test data.

- Predicted "y" in the test data using the final model.

5. Conclusion

-Calculated the variable "Lead score" as Lead Score = y_pred*100, where the value of Lead Score is between 0 and 100, where a higher lead score indicates that the lead is hot and most likely to get converted and vice-versa.

- Test data has a high accuracy score of 0.92, high sensitivity of 0.91, and high specificity of 0.92. Conclusively, the model seems to predict the conversion rate quite well.