Summary

The following steps are taken for creating the model:

1. Reading and understanding the given data set with the help of provided data dictionary.

2. **Importing Libraries:**
   Imported necessary libraries and added further libraries as and when required.

3. **Data cleaning**:
   After importing the data it is properly analyzed for various data types, missing values, presence of redundant data, etc, and necessary actions are taken by eliminating or imputing technique. Checked the uniqueness of each column if there is only one category in a column then it is eliminated.

4. **EDA**:
   Various features' behavior wrt the target column 'Converted' is visualized using the plots like countplot, boxplot, heatmap etc. And based on which decisions are taken to deal with certain features. Looking into skewness of some columns they were eliminated and few categorical columns having more options but less counts for each variety are then grouped to a common category. Looking into the correlation matric features are eliminated to avoid multicollinearity in the model.

5. **Creation of Dummy**:
   For the categorical variables like 'Lead Origin', 'Lead Source', 'Specialization', etc dummies are created by dropping first column to convert them into numerical data type for the model. Used pd.getdummies function from Pandas library.

6. **Creating Train_Test data sets**:
   X variable is created by taking all the columns except 'Converted' and y variable includes the target column 'Converted'. From sklearn.model_selection imported train_test_split and train test data are then separated to 70% & 30% respectively. A random state of 42 is used.

7. **Scaling**:
   The numerical columns are then scaled using a standard scaler imported from sklearn.preprocessing.

8. **Feature selection**:
   Using RFE 15 features are selected out of a total of 65 columns which were created after generating dummy columns.

9. **Assessing the model**:
   Added a constant to the train data, data is fit to the logistic regression model. Various results are then checked to determine the wellness of the model.
   By checking the p-value and the VIF value feature are further eliminated

10. **Model evaluation**:
    Metrics – Accuracy, Sensitivity/Recall, Precision, F-1 Score, ROC AUC Score are checked to evaluate the model. A cutoff of 0.5 is first randomly provided and checked for all the metrics then changed to 0.3 cutoff as we can see in the accuracy, sensitivity and specificity intersection point 0.3 is the optimum cutoff for the model.

11. **ROC (Receiver Operating Characteristics Curve):**
    ROC curve of True positive rate vs False positive rate is plotted to check the trend and observed that it follows a normal ROC curve shape.

12. **Lead score column**:
    A lead score column is added to the table for the predicted probability of each customer to help the salesperson determine the priority.

13. **Testing**:
    The model is then tested on the test data using the predict function and it is then observed that test model metrics are having very close values to the train data models, which concludes the wellness of the model.