SUMMARY

Following process flow was used for the business problem of lead scoring model building:

Data cleaning:

1. Firstly, I started with good understanding of business problem followed by importing all important libraries in python notebook.
2. Loading the data in the python notebook and doing basic data checks such as shape of data, data types of the columns, null values check, percentage of null values in each column etc.
3. After that I started with data cleaning. But before that I changed all 'Select' options in each column to 'NaN' as suggested in the business problem.
4. I then removed all the columns that had null values > 40%.

EDA:

1. I started with univariate analysis for numerical features that included removing outliers as they could skew the results.
2. Post univariate analysis I then went to bivariate analysis with the target variable 'Converted' using correlation in order to find strong positive correlation to know the feature that impacted the target variables the most.
3. After numerical columns, I then went to categorical variables where I removed various columns that had data imbalance.
4. Null values in the categorical feature columns were imputed with either new category such as 'Others' or 'Not specified' or with highest occurring category feature in the respective column. These imputations were done keeping various factors in mind.
5. Following this, many categories of the categorical features that had very low count were merged with others or a new category was created in order to replace all these new categories into 1.
   REASON: This was done because, firstly they were less impactful due to very negligible count and secondly, they would **have led to creatin of lot of additional columns when creating dummy variables.**

Dummy variables, Train test split and feature scaling:

1. I then created dummy variables for all categorical columns.
2. I then split the data in to train data that was 70% of the data and test data that comprised of 30% of the data.
3. For the training data, I then used StandardScaler (standard scaler) for feature scaling of numerical features/columns.

Model Building:

1. I then built the 1st Logistic regression model and evaluated it based on p-value. I then deleted columns that had higher p-values and re-built the model.
2. Once all p-values were under a specified range of hypothesis, I then checked VIF or variation inflation factor. VIF number if greater than 5 was noted for any column then that column was dropped and 1st and 2nd step was repeated.
3. 2nd step is repeated till VIF is all under a specified range.

Prediction:

1. Prediction was the done with training data and confusion matrix and other metrics such as accuracy of model, sensitivity and specificity was noted.
2. I then optimum cut off value using ROC curve was used.

Precision Recall:

1. This method was also used for training and test data. Using the model the accuracy on the test data and metrics was calculated.