CASE STUDY SUMMARY

Steps performed while solving the problem:

1. Understanding the problem statement.
2. Importing the Libraries
3. Importing the dataset
4. Data cleaning and data imputation
5. Checking for outliers
6. Data preparation
    - Train-test-split
    - Feature scaling
    - Checking the correlation
    - Building model on train dataset
    - Using RFE to select best variables
    - Calculating the VIF
    - Predicting the test model
    - ROC curve plotting
    - Finding optimal probability cutoff point
    - Precision and Recall
    - Model Evaluation
    - Lead Score assigning

Understanding the problem statement:

After reading the problem statement there were few points which we made and those were:

- Data would be Large with both Numerical and Categorical values.
- Finding the suitable variable for the business is necessary.
- The model should perform well in the off season as well.
- When the target is achieved way before the time then they should be some parameters on which we have to focus to improve.

Data cleaning and data imputation:

When we started to read the dataset, we found out that the shape of the dataset is (9240,37) which says 9240 rows and 37 columns. When we checked the percentage of null values in those 37 columns, we found out that there were 16 columns which had null values. Out of 16 columns 6 columns had null values above 30%, so we decided to have the null value cutoff as 30% and drop all those columns which has null value more than 30%. Hence, 6 columns were dropped. Now we encountered few columns which were redundant and were making no sense in terms of data, therefore we dropped all those columns (6 columns) as well. Now the shape of the dataset was (9240,25) and we still had 8 columns with null value. We decided to know the unique value of every individual column to understand the type of data and what it actually has. We encountered that the categorical variables had “select” as its data which is redundant value. Therefore, converted all the data mentioned as “select” to null value.

We again verified the null value percentage and dropped the columns having null value percentage more than 30%. We decided to impute the numerical values with mean value of that column and the categorical value with the mode of that column. We used the scikit learn library to impute the values. After imputing the values, we created dummy values for the categorical variables.

Checking for outliers:

For checking the outliers, we used describe method/function of the pandas and found out that there were 3 variables (0'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit') which were likely to have outliers in it. Hence, created boxplot to visualize the distribution of the data. When it was confirmed that there are outliers, we decided not to drop them whereas to create bins and then separate them, so that when we are performing feature scaling it would get normalized.

Data preparation:

Since the data was set right, we divided the data into 2 sets. One for training the model and the other one for testing the model which was termed as X_train, y_train and X_test, y_test. After this we performed feature scaling using StandardScaler, so that the numerical value would be brought under the range of -3 to +3. Once this was done, we created a heatmap to find the correlation between all the independent variables, and we found out that there were 2 variables which had very high correlation. Hence, dropped those columns.

Now we went ahead with creating the model based on train dataset. After creating we saw the p-value of all the independent variables and they were more than the significance level which was 0.005. Hence, used RFE method to choose best 19 variables out of the whole dataset. Again, we created a logistic regression model and then performed the VIF to find the collinearity between all the variables. As, the VIF of all the variable was under 5 which indicated that there was not evident collinearity between independent variables, hence dropped the variables by comparing the p-value which was greater than 0.005. After 2 continued dropping of variables we created a model which had p-value lower than 0.005 and VIF value lesser than 5.

Now we wanted to predict the model on the test dataset and then created the RC curve. The ROC curve was inclined towards Y-axis which emphasized that the model is good. We calculate the **probability, 'Accuracy', 'Sensitivity'** and **'Specificity'** and plotted the graph which showed that the optimal cutoff should be 0.4. We then calculate the **Precision** and **Recall** of the model and it the result was:

Precision:	**0.7512830635609948**
Recall:	**0.7716950527169505**
Accuracy:	**0.8207070707070707**