SUMMARY
Lead Score Case Study

Problem Statement:

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Solution Approach:

1. **Data Cleaning:** The data contained a lot of null values, and 'Select' value in multiple columns. Few columns had Data imbalances as well. Each of these scenarios was analyzed and appropriate handling technique was used:-
   - Columns with high null values (More than 40%) were dropped.
   - For few significant columns, null values were replaced with 'Not Provided'/'Others'.
   - Columns with data imbalances such as Country, was dropped.

2. **EDA:** On the cleaned data, EDA was performed.
   - Outliers observed during EDA were treated using 1.5 IQR Method.
   - Univariate Analysis of Categorical and Numerical variables was performed.
   - Bivariate Analysis of important variables was performed with 'Converted' variable (Target Variable)
   - Based on graphs, less significant categories in few of the columns were clubbed into one.

3. **Data Pre-processing:** The following pre-processing steps were performed.
   - Binary Variables where 'Yes' was encoded as '1' and 'No' as '0'.
   - N-1 Dummy columns were created for given N categories for each categorical column.
   - Data was split into training and test dataset in the ratio of 70:30.
   - Feature Scaling was performed on continuous variables.

4. **Model Building:** Logistic Regression was performed on the training dataset using the following steps.
   - First RFE was done to attain top 15 relevant variables.

- Using these 15 variables, model was built in iterative manner where VIF and p-values were observed for each model.
- Variables with VIF > 5 or p-value > 0.05 were eliminated one by one and the model was rebuilt at every stage.

5. **Model Evaluation:**
   - Predicated values on the training dataset were obtained by using 0.5 as arbitrary cut-off, where in leads with conversion probability < 0.5 were tagged '0' and vice versa.
   - Confusion matrix was created using which accuracy, sensitivity, and specificity(were calculated.
   - ROC curve was plotted and optimal cut off was calculated to be around 0.2.
   - Accuracy, sensitivity, and specificity were re-evaluated and Precision-Recall trade-off observed.

6. **Predictions:** Predictions on test data was made using the following steps.
   - Scaling was performed on continuous variables of test data.
   - Using the model built and cut-off fixed at 0.2, predictions were made on this dataset.
   - Confusion matrix was created using which accuracy, sensitivity, and specificity were calculated.
   - Finally lead conversion score was given to each lead.

Prediction on another Data:

1. As for competition another dataset was given, which was imported in notebook then cleaned dropped and imputed.
2. Dummy variables created and Scaled and prediction was done on test data.

Conclusion:

1. The number of total visits to the website, the total time spent on the website, and the page views per visit. have the highest impact.
2. Time Spent on Website, Olark chat, reference, Emails opened, working professions and hospitality management are major features impacting.