Summary of Lead Score Case Study: -

Steps involved in the case study: -

Step 1: Reading and Cleaning of the data: - We read the data into Leads and then we started cleaning of the data.

1.a Replacing the nan values **with 'select'.**

1.b Cleaning of the data involved removing the columns with null values greater than 60%

Result: 'How did you hear about X Education', 'Lead profile' etc. columns were dropped.

1.c Check for unique values was done in the columns so that 'Select' can be replaced by value that covers more that 50 % of the data.

Result of cleaning: all the data with nan values was treated and no data had unnecessary columns.

Step 2: Exploratory data analysis: - Using EDA we can conclude the following:

2.a We did successfully remove the outliers from the data as outliers effect the data.
2.b EDA also helps to visualize the data i.e. check for the rate that is being converted i.e. which factors affect the most for being converted.

Result: On analysis we found that following are the three variables that affected the most in the rate of being converted:

- Total Time Spent on Website
- Last Notable Activity
- Total Visits

Step 3: Dummy variable creation: - We created dummy variables so that we **can** convert our categorical data into numerical data for better interpretation.
Result: We converted following columns into dummy columns:

a. 'Lead Origin'
b. 'Lead Source'
c. 'Last Activity'
d. 'Specialization'
e. 'What is your current occupation' etc.

Step 4: Splitting the data into training and test set:
- The % of test data is 0.3 and % of train data is 0.7

Step 5: SCALING the data: - Scaling of the data basically helps to normalize the **data** within a particular range.
4.a Here we have used Standard Scaling which helped to normalize the data.

Step 6: Model Building: - We first checked the converted rate which turned out to be 37.86 which was good. Model building Is done on train set.
5.a. Aim: is to identify the hot leads with help of lead score.
5.b. Model used: Logistic Regression because our output variable was categorical.
5.c. Feature Selection: feature selection was done using RFE in which we selected 15 features.
5.d. Feature Elimination: After using RFE we needed to eliminate some features using p value.

Step 7: We checked VIF: On checking we found out that VIF was <5 that means we don't have multicollinearity in our data

Step 8: Plotting of ROC curve: - Done to compare true positive rate and false positive rate.
Result: area under ROC curve turned out to be 0.95 which is pretty good.

Step 9: Finding optimal cut off point: It turned out to be 0.2

Step 10: Finding Accuracy, sensitivity, specificity:

10.a. On train data: On application of fit transform on train set:

Accuracy: 91%
sensitivity :87%
specificity: 94%

10.b. Then we applied transform on test and we concluded:

Accuracy: 90%
sensitivity :87%
specificity: 92%

10.c. Then we did assign lead score on basis of converted probability (predicted through modelling)

Lead score = round (converted probability*100)