Lead Scoring Case Study Summary

Summary:

Step 1 **Reading and Understanding Data**. Read and analyze the data.

a. After performing the basic steps to read the data, we have observed the dataset comprises of 37 columns and 9240 rows.

Step 2 **Data Cleaning**:

a. After checking for missing values, we dropped the variables that had high values(>3000%) of NULL values in them.

   i. Variables like ‘ ‘Tags, Country, city, 'What matters most to you in choosing a course' etc., were dropped from the dataset

b. The outliers were identified and removed.

   i. For ‘TotalVisits’, Page Views Per Visit, Total Time Spent on Website a boxplot was created to check for outliers and the observed outliers imputed with median values.

Step 3 **Data Analysis**

a. Exploratory Data Analysis of the data set to get a feel of how the data is oriented.

b. Performed visualization - pairplots for ‘TotalVisits’, Page Views Per Visit, Total Time Spent on Website

Step 4 **Creating Dummy Variables**

a. Creating dummy variables for the categorical variables.

b. Also to scale the features ‘MinmaxScalar()’ used

Step 5 **Correlation Analysis**

a. The Heatmap provided information that with high levels of correlation can be dropped from the dataset. And the same was performed

Step 6 **Test Train Split**:

a. The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step 7 **Feature Rescaling**

a. We used the StandardScalar() to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step 8 **Feature selection using RFE**:

a. Using the Recursive Feature Elimination we went ahead and selected the 20 top important features.

Step 9 **Logistic Regression**

a. Using the statistics generated, we recursively tried looking at the p-values and VIF Values in order to select the most significant values that should be present and dropped the insignificant values and also the values higher than 5 are dropped one after the other till we obtain significant p-values and VIF values less than 5.

Step 10 **Final Model:**

a. Once we reached the optimal p and VIF Values, we can finalize the model and this leads to test the model.

Step 11 **Model Testing:**

a. Model testing was done using three major attributes ‘Accuracy(78.86), Sensitivity (73.94), and Specificity(83.43)’.

b. The ROC Curve confirms 86% of area under the curve a good sign for model fitment.

Step 12 Lead Scores

a. After finding the lead scores the model accuracy stands at 78.66%
b. ‘Accuracy(78.66), Precision (78.28), and recall(76.74)'.
c. Where almost all the values doesn’t vary much and hence the model is finalized.