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 percentage (>40%) of NULL values in them.
    i. Variables like ' 'Tags, Country, 'What matters most to you in choosing a course' etc., were dropped from the dataset
b. Variables like 'Last Ativity', 'What is your current occupation' etc, were imputed
c. 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 - subplots for Lead origin, Lead Source and other categorical variables.
    i. 'Landing Page Submission' was the key for lead origin followed by API,
    ii. Similarly Google and Direct Traffic were the key sources for Lead source.
    iii. Most of the people chooses Finance Management Specialization rather than other Specialization
    iv. The IT Project management have very lees so that most of the People not prefered this Specialization
c. For Bivariate analysis
    i. In Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category
    ii. In Last Activity the number of Hot leads is higher in SMS and in EMAIL cold leads is higher than hot leads.
    iii. In Last Notable Activity it's mostly same as Last Activity.
    iv. In Specialization the most of the leads are comes from Finance management but here Hot leads are lesser than Cold leads.

Step 4 **Creating Dummy Variables**

a. Creating dummy variables for the categorical variables.
b. Also to scale the features 'StandardScalar()' 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(.789), Sensitivity (.629), and Specificity(.888)'
b. Confusion matrix confirms positive predictive rate as 77.68%
c. 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 77.66%
b. 'Accuracy(.7766), Sensitivity (.801), and Specificity(.760)'
c. Where almost all the values doesn't vary much and hence the m0del is finalized.