Lead Scoring Case Study Summary

Solution Summary

1. Reading and Understanding Data

We imported data and the necessary libraries

2. Cleaning Data

- Described to find characteristics such as number of rows and columns and other statistics
- Identified Data types
- Replaced null values for numeric variables with mean values of those columns
- For values in columns with Object data type, we replaced the values marked "Select" with null values.
- Categorized columns in categorical, numerical and Target
- Identified necessary columns by eliminating those which were redundant.
- The final variables we were left with are:
    - Last Activity
    - Total Time Spent on Website
    - Page View Per Visit
    - Converted
    - Total Visits
    - Lead Source
    - Last Notable Activity
    - A free copy of Mastering the Interview
    - Do Not Email
    - Lead Origin

3. Exploratory Data Analysis

- Visualisation:
    - We used Pie charts to indicate converted (1) and Non- Converted Lead
    - We used Charts (Line, Dot and Box) for numeric columns
    - We used Bar charts for Categorical columns
- Created a correlation matrix to capture if there are any cases of multicollinearity
- Created Dummy Variables for the categorical variables
- Performed an Outlier Treatment and removed the variables that have +3 std and -3 std
- Normalisation of Continuous Variables such as Total Visits, Total Time Spent on Website, Page Views per Visit.

4. Building the Model

- Divided the data set into train and test sets with a proportion of 70-30.
- Performed Feature Selection to narrow down number of variables:
    - Used RFE to select the top 20 variables from a total of 60 variables

- Then, we checked VIF for all 20 variables and after building 10 models we arrived at 11 most significant variables whose VIF was good.

5. Model Evaluation

- For our final model, we created ROC curve for final model and the coverage area came our to be 84% which proves the legitimacy of the model.
- Based on the table and values found for the confusion matrix, 80% cases are correctly predicted on the basis of the converted column
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set and concluded that the optimum cut-off point is 0.42
- Finally, we made predictions on the test set and calculated the conversion probability based on accuracy, sensitivity and specificity metrics and found the values to be:
  - Accuracy = 76%
  - Sensitivity = 73%
  - Specificity = 81%

6. Conclusion

- The lead score calculated shows the conversion rate of 81% on the final prediction model which clear meets the expectation of CEO as per the given minimum target.
- Good value of sensitivity of our model will help to select the most promising leads.
- Features which contribute the most towards the probability of a lead getting converted are:
  - Total Time Spent on Website
  - Last Activity_Olark Chat Conversation
  - Last Notable Activity_SMS Sent