Summary [Lead Scoring Case Study]

After understanding the Business Problem and Business Objective. We got clear understanding for our goals of the case study.

We performed the following steps:

1. Data Sourcing: Importing the required libraries

2. Data Reading & Understanding:

Reading the dataset “Leads.csv” and understanding it as follows: -

a. Routine Data Check: No of rows, columns, data type of each column, distribution, mean and median for all numerical columns etc.

b. Missing value analysis.

c. Duplicate rows check.

3. Data Cleaning: In this case study, Data cleaning plays a very crucial role. The quality and efficiency of the model depends on the data cleaning step. Hence it must be followed thoroughly.

a. “Select” value is replaced with NAN.

b. Calculation of missing values for each column and dropping Score and Activity variable.

c. Dropping the columns with high percentage of missing values.

d. Checking the unique category for each column.

e. If the columns are highly skewed with one category, such columns will be dropped. Combining different categories of the columns with less percentage values into "Others" category.

f. Imputing the column with least missing values percentage.

g. Finally Checking for the number of rows kept after performing all the above steps.

4. EDA: In EDA, Univariate and Bi-Variate analysis was done on both categorical and numerical variables.

5. Outlier Treatment: We form soft capping of upper range outlier values for Total Visits and Page View Per Visit.

6. Data Preparation: In this step, we performed Data Pre-processing, the dummy variables are created. Performed train test data split and scaled the numerical columns.

7. Data Modelling & Model Evaluation:

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for feature selection.
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.
- Prediction on test data sets.
- Overall accuracy 80%.

8. Conclusion:

Final Observation:

Let us compare the values obtained for Train Data & Test Data:

Train Data:
Accuracy : 79%
Sensitivity : 78%
Specificity : 80%
Precision : 71%
Recall : 78%

Test Data:
Accuracy : 80%
Sensitivity : 78%
Specificity : 80%
Precision : 72%
Recall : 78%

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- When the lead origin is Lead add format.

- When their current occupation is as a working professional.

- When the lead source was:

a. Welingak website

b. Olark Chat

- The total time spent on the Website.

- When the city is Mumbai.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.