Summary Report on Building a Lead Scoring Model for X Education

In this report, we outline the process of building a lead scoring model for X Education and share the key learnings from the assignment. The objective was to develop a logistic regression model that assigns lead scores to identify potential leads with higher conversion chances.

1. Data Understanding and Preprocessing:

The project started with understanding the dataset provided by X Education. It contained information on various attributes such as lead source, website activity, and last interaction. We observed that several categorical variables had a 'Select' level, which was treated as a null value.

2. Exploratory Data Analysis (EDA):

We visualized the relationships between different features and the target variable, **'Converted**.' Key findings included the significance of the total time spent on the website, the number of total visits, and the impact of certain lead sources on conversion rates. Also we dropped the 'Prospect ID' and 'Lead Number' columns as they were not relevant for modeling. 'Select' values in the dataset were replaced with NaN to handle them as missing data. Columns with more than 45% missing values were dropped to streamline the dataset.

3. Numerical Attribute Analysis:

The analysis of numerical attributes in the dataset reveals the following insights:

Correlation Analysis:

- The correlation heatmap shows no strong linear correlations among the numeric variables, indicating that they are not highly interdependent.

Total Visits:

- The boxplot for 'Total Visits' indicates the presence of outliers.
- The 99th percentile value is 17, and the maximum value is 251, suggesting the presence of extreme outliers.
- The top and bottom 1% of the outlier values have been removed to mitigate the impact of outliers.

Total Time Spent on Website:

- The boxplot for 'Total Time Spent on Website' does not show any significant outliers.
- The distribution suggests that the time spent on the website varies among leads.

Page Views Per Visit:

- The boxplot for 'Page Views Per Visit' indicates the presence of outliers.
- The 99th percentile value is 8, and the maximum value is higher, indicating the presence of extreme outliers.
- Similar to 'TotalVisits,' the top and bottom 1% of the outlier values have been removed.

- In summary, while there are outliers in 'TotalVisits,' 'Total Time Spent on Website' appears to have a positive relationship with conversion.

4. Model Deployment:

Throughout this process, the aim was to improve the model's performance by eliminating less relevant features and maintaining statistical significance. Model 7 represents the final logistic regression model with the most relevant features.

Final learnings and result that we gathered with data analysis:

Comparison of Test and Train Data:

I. **Train Data:**
- Accuracy: 78.37%
- Specificity: 75.76%
- Sensitivity (Recall): 79.98%

II. **Test Data:**
- Accuracy: 78.44%
- Specificity: 76.23%
- Sensitivity (Recall): 79.78%