Summary of Analysis for X Education

The objective of this analysis was to identify key factors influencing the conversion of potential leads into industry professionals enrolling in X Education's courses. The provided dataset included information about customer interactions, such as website visits, time spent on the site, lead sources, and conversion rates.

Technical Steps Followed

1. **Data Cleaning**
    - Redundant and irrelevant variables/features were removed.
    - Null values were addressed:
        - Replaced the option "Select" with null since it added no value.
        - Dropped columns with more than 40% null values.
    - Checked the number of unique categories in categorical columns to identify skewed variables.
    - Highly skewed columns were removed.
    - Missing values were imputed using appropriate methods like mean, median, or mode.
    - Detected and handled outliers in numeric variables.
2. **Exploratory Data Analysis (EDA)**
    - Conducted an initial EDA to understand data quality.
    - Identified irrelevant categories in categorical variables.
    - Performed:
        - **Univariate Analysis** for continuous and categorical variables.
        - **Bivariate Analysis** to understand the relationship with the target variable.
3. **Creation of Dummy Variables**
    - Converted all categorical columns into dummy variables for model compatibility.
4. **Scaling**
    - Scaled continuous variables using the Standard Scaler to bring them to a uniform scale.
5. **Train-Test Split**
    - Split the data into training (70%) and testing (30%) sets for model building and evaluation.
6. **Model Building**
    - Used Recursive Feature Elimination (RFE) to select the top 20 important variables.
    - Further refined features manually based on:
        - **VIF values** (retained variables with VIF < 5).
        - **p-values** (kept variables with p-value < 0.05).
7. **Model Evaluation**
    - Evaluated the model using a confusion matrix.
    - Optimized the cut-off point using the **ROC curve**, resulting in an accuracy, sensitivity, and specificity of approximately **80%**.

8. **Prediction**
    - Applied the model to the test data using an optimum cut-off value of **0.37**, achieving an accuracy, sensitivity, and specificity of 80%.
9. **Precision-Recall Check**
    - Re-evaluated the model using Precision-Recall and adjusted the cut-off value to **0.41** for further validation.

Conclusion

The analysis revealed the most significant variables influencing lead conversion:

1. **Total Time Spent on the Website** – Indicates customer engagement.
2. **Total Number of Visits** – Reflects interest and intent to convert.
3. **Lead Source** – Leads originating from **Olark Chat** showed a higher likelihood of conversion.
4. **Last Activity** – Certain activities, such as:
    - **SMS Sent**
    - **Olark Chat Conversation**

    were particularly impactful for conversion.

By focusing on these key variables, X Education can optimize its sales efforts to attract and convert more potential leads into enrolled industry professionals.

Made By: -

A. Anish Gupta
B. Aachal Deshmukh
C. Ninad Kalambe