Summary Report

Lead Scoring Model for X Education

Objective :

The goal of this project was to develop a **Lead Scoring** Model for X Education to maximize conversions and minimize ineffective calls. The process involved **Data preparation, exploratory data analysis (EDA), feature engineering, model building, and evaluation**.

Process Followed:

1. Data Understanding & Cleaning -

- The dataset contained **9,240 leads** and **37 features** (categorical & numerical).
- Features with over 40% missing values (e.g., **Lead Profile, Asymmetrique Scores**) were removed.
- Variables with values **'Select' were replaced with NaN** and missing values in categorical columns were imputed with appropriate values.
- Looking for **duplicate values** and eliminating them if required.

2. Exploratory Data Analysis (EDA) -

- Key categorical variables **(Lead Source, Lead Origin, Last Activity)** and numerical variables **(Total Visits, Time Spent on Website, Page Views per Visit)** were analyzed.
- **Outliers were treated** to improve data quality.

3. Feature Engineering -

- High-cardinality categorical variables were grouped into broader categories.
- **Dummy variables were created, and numerical features were standardized.**
- The dataset was split **(70% training, 30% testing).**
- **Standardization** has been performed on numerical features.

4. Model Building & Selection -

- **Logistic Regression** was chosen for interpretability and efficiency.
- **Recursive Feature Elimination (RFE)** helped select the top 20 features.
- Features with **high p-values and Variance Inflation Factor (VIF)** were iteratively removed.

Key Model Insights -

- **Top predictors**: Total Time Spent on Website, Tags, Lead Source, Last Activity, Specialization.
- **The top 3 contributing variables were:**

    1. **Tags_Closed by Horizzon (5.8192 coefficient)**

    2. **Tags_Lost to EINS (5.0213 coefficient)**

    3. **Tags_Will revert after reading the email (2.9394 coefficient)**

- **Performance Metrics:**
    - Accuracy, Sensitivity, and Specificity: **88%-90%** at an optimum cut-off of **0.37**.
    - A lead score threshold of **30 (probability cutoff = 0.3)** balanced conversion (93.41% sensitivity) and specificity (92.48%).

Learnings:

- Lower thresholds (e.g., 30) improve recall, including more potential leads, while higher thresholds (70-80) maximize specificity to reduce unnecessary calls.

- Tags such as **"Will revert after reading the email"** and **"Lost to EINS"** are critical indicators of conversion potential.

- **Automated engagement** via email and WhatsApp can be more effective than direct calls.

- Data-driven lead scoring enables sales teams to refine outreach strategies for better efficiency.

Conclusion :

This Lead Scoring Model helps optimize conversions by prioritizing key features and fine-tuning lead thresholds. X Education can enhance **efficiency, reduce costs, and improve outreach with data-driven engagement strategies**. Future improvements can include advanced classification models and automation tools.

We have noted that the variables that matter the most in identifying potential buyers are –

- **Total Time Spent on Website**
- **Tags**
- **Lead Source**
- **Last Activity**
- **Specialization.**

These factors significantly influence lead conversion and should be prioritized in future strategies.