Summary Report – Lead Scoring for X Education

Introduction:

X Education, an online course provider for industry professionals, receives numerous leads daily through its website, marketing campaigns, and referrals. Despite a steady influx of leads, the conversion rate stands at about 30%, meaning only 30 out of every 100 leads are converted into customers. To address this, X Education aims to identify "hot leads"—leads with the highest potential to convert—thereby improving efficiency and increasing the conversion rate.

Objective:

Our task is to develop a model that assigns a lead score to each potential customer, indicating the likelihood of conversion. The goal is to achieve a target conversion rate of around 80%.

Data Overview:

We received a dataset containing **9,240 records** and **37 attributes**, detailing customer demographics and online behaviour. After data cleaning, which involved dropping columns with more than **40% missing values**, we were left with **12 relevant attributes and 1 target variable.** These include:

- Lead Source
- Customer Specialization
- Lead Origin
- Preferred Contact Mode (Email/No Email)
- Total Website Visits
- Total Time Spent on Website
- Page Views per Visit
- Last Activity
- Customer Occupation
- Tags based on Response
- Last Notable Activity
- Opt-in for a Free Copy of "Mastering the Interview"
- **Target Variable: Converted (0-No, 1-Yes)**

Exploratory Data Analysis (EDA):

- **Univariate Analysis:** Used boxplots for numerical attributes to identify and handle outliers using the **IQR technique**. Histograms for categorical attributes helped determine which factors contribute most to lead generation.
- **Bivariate Analysis:** Examined relationships between attributes and the target variable to identify trends.

- **Data Imbalance:** The dataset showed a minor imbalance with a ratio of **1.08**, indicating nearly equal representation of converted and non-converted leads.

Data Processing:

- **Dummy Variables:** Created for categorical attributes to convert them into numerical format, essential for logistic regression.
- **Feature Scaling:** Applied to numerical attributes to ensure all variables are on a similar scale, improving model convergence and stability.

Model Building:

- Split the data into a **training set (70%)** and a **test set (30%)**.
- Used **Recursive Feature Elimination** (RFE) to select the **15 most important features**.
- After iterative modelling and checking p-values and VIF, we finalized a model with 12 significant features.

Model Evaluation:

Using the training set with a **default cutoff of 0.5**, we achieved:

- **Accuracy:** 91%
- **Sensitivity:** 90%
- **Specificity:** 93%

The **ROC curve area (AUC)** of **0.97** indicated excellent model performance. The **optimal cutoff point** was determined to be **0.47**, yielding:

- **Accuracy:** 90.6%
- **Sensitivity:** 90.7%
- **Specificity:** 90.6%

For the test set, the model metrics were:

- **Accuracy:** 89%
- **Sensitivity:** 90%
- **Specificity:** 90%
- **Precision:** 87.8%
- **Recall:** 91.4%

These metrics suggest that the model generalizes well to new data and does not overfit

Recommendations:

From the test set, we identified **171 hot leads**. We recommend focusing on leads from the Wellingak website, customers who promised to revert after receiving course information, those who spent more time on the website, preferred email communication, housewives, working professionals, and leads closed by Horizon, as these groups have higher conversion chances.

By targeting these segments, X Education can potentially increase its lead conversion rate and achieve the desired efficiency in its sales process.