Lead Scoring Model – Summary

X Education generates many leads, but its conversion rate is low at approximately 30%. The objective is to develop a lead scoring model that assigns a score to each lead, prioritizing those with a higher likelihood of conversion. The CEO aims to increase the conversion rate to 80%.

Data Processing:

- Columns with over 40% missing values were removed. Categorical columns were analyzed, and appropriate actions were taken, such as creating a new category (‘Others’), imputing the most frequent value, or removing non-informative columns.
- Numerical categorical data were imputed with the mode, and columns with only one unique response were dropped.
- Additional data processing steps included handling outliers, fixing invalid entries, grouping low-frequency values, and mapping binary categorical variables.

Exploratory Data Analysis (EDA):

- Data imbalance was observed, with only 38.5% of leads converting.
- Univariate and bivariate analyses were conducted on categorical and numerical variables. Key factors influencing lead conversion include ‘Lead Origin,’ ‘Current Occupation,’ and ‘Lead Source.’
- Time spent on the website showed a positive impact on lead conversion.

Data Preparation:

- Categorical variables were one-hot encoded to create dummy features.
- The dataset was split into training (70%) and testing (30%) sets.
- Standardization was applied for feature scaling.
- Some highly correlated columns were removed to avoid redundancy.

Model Development:

- Recursive Feature Elimination (RFE) reduced the number of variables from 48 to 15, improving model efficiency.
- A manual feature selection process removed variables with p-values greater than 0.05.
- Three models were tested before finalizing **Model 4 (logm4)** with 12 key variables. It showed stability, with all p-values < 0.05 and no signs of multicollinearity (VIF < 5).
- The final model was used to make predictions on both training and test datasets.

Model Evaluation:

- A confusion matrix was used to determine the optimal cut-off point of **0.345**, balancing accuracy, sensitivity, and specificity.

- This cut-off resulted in accuracy, specificity, and precision around 80%, whereas a precision-recall approach yielded lower performance (~75%).
- To meet the business requirement of an 80% conversion rate, the **sensitivity-specificity** approach was chosen for final predictions.
- Lead scores were assigned based on the cut-off of 0.345.

Predictions on Test Data:

- The model was applied to test data, following the same scaling and prediction process.
- Evaluation metrics for both training and test sets were consistent, around 80%.
- Lead scores were assigned to prioritize high-potential leads.

Key Influencing Factors:

1. **Lead Source – Welingak Website**
2. **Lead Source – Reference**
3. **Last Activity_SMS sent**

Recommendations:

- Increase **advertising and budget allocation** for the **Welingak Website**, as it generates high-quality leads.
- Offer **incentives and discounts for referrals** to encourage more lead conversions.
- **Focus on Last Activty**, as they have a higher conversion rate and better financial capability, making them ideal targets for enrollment.

This model helps optimize lead conversion and aligns with the company’s goal of achieving an 80% conversion rate.