Lead Scoring Model Overview

1. Introduction X Education faces a challenge with a low lead conversion rate of approximately 30%. The objective is to construct a predictive model for assigning lead scores to enhance conversion chances. The CEO aims to achieve an 80% lead conversion rate.

2. Data Preparation and Cleaning

- Columns with over 40% missing data were eliminated.
- Categorical columns underwent value count analysis for appropriate actions such as dropping, creating an "others" category, or imputing high-frequency values.
- Numerical categorical data were imputed using the mode, and columns with a single unique response were removed.
- Outliers were treated, invalid data was rectified, and low-frequency values were grouped.
- Binary categorical values were mapped.

3. Exploratory Data Analysis (EDA)

- Examined data balance, revealing a 38.5% conversion rate.
- Conducted univariate and bivariate analysis for categorical and numerical variables.
- Key variables like 'Lead Origin', 'Current Occupation', and 'Lead Source' provided valuable insights into their effects on the target variable.
- Positive correlation observed between time spent on the website and lead conversion.

4. Data Preparation

- Engineered dummy features through one-hot encoding for categorical variables.
- Divided the dataset into training (70%) and testing (30%) sets.
- Standardized features using scaling techniques.
- Removed highly correlated columns.

5. Model Development

- Employed Recursive Feature Elimination (RFE) to reduce variables from 48 to 15, enhancing dataset manageability.

- Conducted manual feature reduction based on p-values, dropping variables with p > 0.05.
- Iteratively developed three models before finalizing Model 4, characterized by stable p-values (p < 0.05) and absence of multicollinearity (VIF < 5).
- Selected 'logm4' as the ultimate model, encompassing 12 variables.

6. Model Evaluation

- Constructed a confusion matrix and determined a cut-off point of 0.41, balancing accuracy, sensitivity, and specificity.
- Achieved balanced performance metrics around 80% for accuracy, specificity, and precision using the chosen cut-off.
- Prioritized sensitivity-specificity view over precision-recall for final predictions, aligning with the business objective.
- Applied the cut-off to assign lead scores to the training data.

7. Test Data Prediction

- Scaled and executed predictions on the test dataset using the final model.
- Achieved consistent evaluation metrics, approximating 80%, for both training and test data.
- Assigned lead scores to the test data.

8. Key Insights and Recommendations

- Allocate increased budget for Welingak Website advertising to enhance lead acquisition.
- Introduce incentives or discounts for successful reference-based lead conversions, promoting reference generation.
- Strategically target working professionals, given their higher conversion rates and better financial capacity to afford higher fees.