Summary

X Education faces a challenge: while they receive many leads, their lead conversion rate is currently only around 30%. To address this, we need to develop a model that assigns a lead score to each lead, indicating their likelihood of conversion. The CEO has set a target conversion rate of 80%.

Data Cleaning:

- Dropped columns with more than 39% null values.
- Checked value counts in categorical columns to decide on appropriate actions: dropped columns if imputation caused skew, created new categories (e.g., "Others"), imputed high-frequency values, and removed columns that added no value.
- Imputed numerical categorical data with mode and dropped columns with only one unique response.
- Addressed outliers, fixed invalid data, grouped low-frequency values, and mapped binary categorical values.

Exploratory Data Analysis (EDA):

- Checked for data imbalance: only 37.40% of leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables. Variables such as 'Lead Origin,' 'Current Occupation,' and 'Lead Source' provided valuable insights into their effect on the target variable.
- Observed that time spent on the website had a positive impact on lead conversion.

Data Preparation:

- Created dummy features (one-hot encoded) for categorical variables.
- Split the data into training and test sets using a 70:30 ratio.
- Scaled features using standardization.
- Dropped highly correlated columns.

Model Building:

- Used Recursive Feature Elimination (RFE) to reduce variables from 48 to 15 for manageability.
- Employed manual feature reduction by dropping variables with p-values > 0.05.
- Built a total of three models before finalizing Model 4, which was stable with p-values < 0.05 and no signs of multicollinearity (VIF < 5).
- Selected **logm2** as the final model with 13 variables for making predictions on the train and test sets.

Model Evaluation:

- Created a confusion matrix and selected a cut-off point of 0.347 based on accuracy, sensitivity, and specificity plots. Accuracy 80.63%
- To meet the CEO's target of an 80% conversion rate, we chose the sensitivity-specificity view for our optimal cut-off for final predictions.
- Assigned lead scores to the training data using a 0.347 cut-off.

Making Predictions on Test Data:

- Scaled and predicted using the final model.
- Evaluation metrics for both train and test data were close, around 80%.
- Assigned lead scores.
- Identified the top 3 features:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current Occupation_Working Professional

Recommendations:

- Increase budget/spending on the Welingak Website for advertising and attracting more leads.
- Offer incentives/discounts for references that convert into leads to encourage more referrals.
- Aggressively target working professionals as they have a high conversion rate and better financial capacity to pay higher fees.

This summary maintains the key points and details while streamlining the language for clarity and conciseness.