Summary

X Education receives many leads, but its lead conversion rate is low at approximately 30%. The company aims to develop a model that assigns a lead score to each prospect, with the goal of increasing the conversion rate to around 80%.

Data Cleaning:

- Columns with over 40% missing values were removed. Categorical columns were examined to determine the best approach: if imputation would skew the data, the column was either dropped, a new category ("others") was created, or the most frequent value was used for imputation.
- Numerical categorical data were imputed with the mode, and columns with only one unique response were eliminated.
- Additional activities included treating outliers, correcting invalid data, grouping low-frequency values, and mapping binary categorical variables.

Exploratory Data Analysis (EDA):

- Data imbalance was identified, with only 38.5% of leads converting.
- Univariate and bivariate analyses were performed on both categorical and numerical variables. Variables such as ‘Lead Origin’, ‘Current Occupation’, and ‘Lead Source’ provided valuable insights into their impact on the target variable.
- Time spent on the website was found to positively influence lead conversion.

Data Preparation:

- Dummy variables were created through one-hot encoding for categorical features.
- The dataset was split into training and testing sets with a 70:30 ratio.
- Feature scaling was conducted using standardization.
- Several columns were dropped due to high correlation with others.

Model Building:

- Recursive Feature Elimination (RFE) was used to reduce the number of variables from 48 to 15 for easier management.
- A manual feature reduction process was employed to eliminate variables with p-values greater than 0.05.

- Three models were built before selecting the final Model 4, which showed stability with p-values less than 0.05 and no multicollinearity (VIF < 5).
- The final model, logm4, was chosen with 12 variables for predictions on both the training and test sets.

Model Evaluation:

- A confusion matrix was created, and a cutoff point of 0.345 was selected based on the accuracy, sensitivity, and specificity plot. This cutoff achieved accuracy, specificity, and precision around 80%, although the precision-recall metrics showed lower performance at around 75%.
- To address the CEO's goal of boosting the conversion rate to 80%, the sensitivity-specificity view was chosen for the optimal cutoff for final predictions.
- Lead scores were assigned to the training data using the 0.345 cutoff.

Making Predictions on Test Data:

- Predictions were made on the test set after scaling and utilizing the final model.
- Evaluation metrics for both the training and testing sets were closely aligned, around 80%.
- Lead scores were assigned accordingly.
- The top three features identified were:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current Occupation_Working Professional

Recommendations:

- Increase budget allocation for advertising on the Welingak Website to attract more leads.
- Offer incentives or discounts for successful referrals to encourage more references.
- Actively target working professionals, as they have a high conversion rate and are likely to afford higher fees.