SUMMARY

X Education faces a significant challenge with a low lead conversion rate of approximately 30%. Their CEO aims to achieve an ambitious lead conversion rate of around 80%. To address this, a model is constructed to assign lead scores, with higher scores indicating a better chance of conversion.

Data Cleaning:

- Columns with over 30% null values were removed. Categorical columns were assessed, and various actions were taken based on value counts to address nulls, including imputation, category creation, or column dropping.
- Numerical categorical data were imputed with, and columns with only one unique customer response were dropped.
- Several data preparation tasks, such as outlier treatment, data validity fixes, and dealing with low-frequency values, were executed.

Data Preparation:

- Dummy features were created for categorical variables using one-hot encoding.
- The data was split into training and testing sets at a 70:30 ratio.
- Feature scaling was performed using standardization.
- Some highly correlated columns were dropped.

Model Building:

- Recursive Feature Elimination (RFE) reduced variables from 48 to 15, making the dataset more manageable.
- Manual feature reduction was employed by dropping variables with p-values exceeding 0.05.
- Four models were built, with Model 5 emerging as stable with p-values below 0.05 and no signs of multicollinearity (VIF < 5).
- 'logm5' was selected as the final model with 11 variables for predictions on both the training and test sets.

Model Evaluation:

- A confusion matrix was created, and a cut-off point of 0.42 was selected based on accuracy, sensitivity, and specificity plots. This cut-off achieved approximately 79% accuracy, specificity, and precision.
- To meet the CEO's goal of an 80% conversion rate, a sensitivity-specificity view was favored for the optimal cut-off, even though it resulted in slightly lower performance metrics in the precision-recall view.

Lead Score Assignment:

- A lead score was assigned to the training data using the 0.42 cut-off.

Making Predictions on Test Data:

- Scaling and predictions were made on the test data using the final model.
- Evaluation metrics for both the training and test data were consistent, hovering around 80%.
- Lead scores were assigned.

Recommendations:

- Encourage the provision of references by offering incentives or discounts.
- Target working professionals aggressively due to their higher conversion rates and potentially better financial situations for higher fees.