Executive Summary

X Education, with its current lead conversion rate of approximately 30%, has set a target for a significantly higher conversion rate of around 80%. To achieve this goal, a lead scoring model has been developed with the aim of assigning lead scores to prioritize leads with a higher likelihood of conversion. Here is a summary of the key steps and findings in this endeavor:

Data Cleaning:

- Columns with over 40% missing values were dropped, and appropriate actions were taken for categorical columns with skewed value counts.
- Numerical categorical data were imputed with the mode, and columns with only one unique customer response were removed.
- Various data cleaning activities, such as handling outliers, fixing invalid data, grouping low-frequency values, and mapping binary categorical values, were performed.

EDA (Exploratory Data Analysis):

- Data imbalance was observed, with only 38.5% of leads converting.
- Univariate and bivariate analysis were conducted for both categorical and numerical variables, highlighting the influence of features like 'Lead Origin,' 'Current occupation,' and 'Lead Source' on the target variable.
- Time spent on the website was found to have a positive impact on lead conversion.

Data Preparation:

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

Model Building:

- Recursive Feature Elimination (RFE) was used to reduce the number of variables from 48 to 15.
- A manual feature reduction process was employed by dropping 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 signs of multicollinearity (VIF < 5).
- 'logm4' was chosen as the final model, which included 12 variables used for making predictions on both the train and test sets.

Model Evaluation:

- A confusion matrix was created, and a cut-off point of 0.345 was selected based on an accuracy, sensitivity, and specificity plot. This cut-off point achieved accuracy, specificity, and precision all around 80%.

- To align with the CEO's goal of an 80% conversion rate, the sensitivity-specificity view was favored for selecting the optimal cut-off, even though precision-recall metrics showed slightly lower performance (around 75%).
- Lead scores were assigned to the train data using the 0.345 cut-off.

Making Predictions on Test Data:

- Predictions were made on the test data using the final model after scaling.
- Evaluation metrics for both the train and test data were very close to 80%.
- Lead scores were assigned based on the predictions.

Top 3 Features:

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional

Recommendations:

- Allocate more budget for advertising and promotion on the Welingak Website to boost lead generation.
- Implement incentives or discounts for customers who provide references that convert into leads, thereby encouraging more references.
- Focus marketing efforts on working professionals, given their high conversion rate and potentially better financial capacity to afford the company's offerings.