Summary

X Education is currently facing a challenge with its lead conversion rate, which stands at a modest 30%. The company has set a lofty target of achieving an 80% lead conversion rate, as directed by the CEO. To address this issue, a comprehensive model has been developed to assign lead scores, with the aim of identifying leads with higher conversion potential.

Data Cleaning:

- Columns containing over 40% null values were removed.
- Categorical columns underwent value count assessments to determine appropriate actions, such as dropping, creating a new category, imputing high-frequency values, or eliminating redundant columns.
- Numerical categorical data were imputed with mode, and columns with a single unique response were dropped.
- Various data quality improvements were made, including outlier handling, invalid data correction, consolidation of low-frequency values, and mapping of binary categorical values.

Exploratory Data Analysis (EDA):

- Data imbalance was observed, with only 38.5% of leads converting.
- Extensive univariate and bivariate analyses were conducted for both categorical and numerical variables.
- Key variables, such as 'Lead Origin,' 'Current job,' and 'Lead Source,' provided valuable insights into their impact on the target variable.
- Notably, time spent on the website exhibited a positive influence on lead conversion.

Data Preparation:

- Dummy features were created through one-hot encoding for categorical variables.
- The dataset was split into training and test sets in a 70:30 ratio.
- Feature scaling was performed using standardisation techniques.
- Highly correlated columns were dropped to enhance model efficiency.

Model Building:

- Recursive Feature Elimination (RFE) was employed to reduce the number of variables from 48 to 15, simplifying the dataset.

- A manual feature reduction process was implemented by eliminating variables with p-values exceeding 0.05.
- Three models were developed before arriving at the stable Model 4, characterised by p-values below 0.05 and no signs of multicollinearity (VIF < 5).
- Model 4, consisting of 12 variables, was chosen as the final model for making predictions on both the train and test datasets.

Model Evaluation:

A confusion matrix was constructed, and a cut-off point of 0.345 was selected based on an analysis of accuracy, sensitivity, and specificity.
This cut-off point yielded balanced metrics, with accuracy, specificity, and precision all hovering around 80%.
A focus on sensitivity-specificity view was preferred over precision-recall view, as it aligned with the CEO's objective to boost the conversion rate to 80%.

Making Predictions on Test Data:

Predictions were made on the test data by scaling it and utilising the final model.
Evaluation metrics for both the train and test sets closely approached the 80% target.
Lead scores were assigned to the data, helping prioritise leads with higher conversion potential.

Recommendations:

To enhance conversion rates, allocating a greater budget for Welingak Website advertising and promotion is advisable.
Encouraging customers to provide references that lead to conversions by offering incentives or discounts can be an effective strategy.
Aggressively targeting working professionals is recommended, given their higher conversion rates and potentially stronger financial capacity to afford the services offered by X Education.