Summary

X Education faces a challenge with its lead conversion rate, currently standing at a meager 30%, while the CEO's ambitious target is an 80% conversion rate. To tackle this issue, we embarked on a comprehensive data analysis and modeling journey:

Data Cleaning:

- We began by removing columns with over 40% null values. For categorical columns, we carefully examined value counts and took actions such as dropping skewed columns, creating a new 'Others' category, or imputing the most frequent value.
- We addressed numerical categorical data by imputing them with the mode. Columns with a single unique customer response were removed.
- Further data cleaning tasks involved treating outliers, rectifying invalid data, consolidating low-frequency values, and mapping binary categorical values.

Exploratory Data Analysis (EDA):

- We noted a data imbalance, with only 38.5% of leads converting.
- Univariate and bivariate analyses were conducted for both categorical and numerical variables. Key insights were drawn from features like 'Lead Origin,' 'Current Occupation,' and 'Lead Source.'
- It was observed that the time spent on the website positively influenced lead conversion.

Data Preparation:

- To handle categorical variables, we created dummy features through one-hot encoding.
- The dataset was split into training and test sets in a 70:30 ratio.
- Standardization was applied for feature scaling.
- Some highly correlated columns were dropped to enhance model performance.

Model Building:

- Recursive Feature Elimination (RFE) was employed to reduce the number of variables from 48 to 15, making the dataset more manageable.
- We used manual feature reduction by dropping variables with p-values greater than 0.05.
- Three models were built before finalizing Model 4, which exhibited stability with p-values less than 0.05 and no signs of multicollinearity (VIF < 5).
- 'logm4' was selected as the final model, consisting of 12 variables, and was used for predictions on both the train and test sets.

Model Evaluation:

- We constructed a confusion matrix and determined a cut-off point of 0.345 based on accuracy, sensitivity, and specificity plots. This cut-off yielded around 80% accuracy, specificity, and precision.
- While our primary goal was to boost the conversion rate to 80%, the precision-recall view showed a drop in performance metrics. Therefore, we opted for the sensitivity-specificity view for the optimal cut-off for final predictions.
- We assigned lead scores to the train data using this cut-off.

Making Predictions on Test Data:

- The final model was applied to the test data after scaling.
- Evaluation metrics for both the train and test sets were very close to around 80%.
- Lead scores were assigned to the test data.

Recommendations:

- Allocate a higher budget for advertising and promotion on the Welingak Website, as it appears to be a significant lead source.
- Consider offering incentives or discounts to individuals who provide references that convert into leads to encourage more references.
- Focus marketing efforts on working professionals, as they exhibit a high conversion rate and are likely to have better financial capabilities to afford higher fees.