Summary

X Education faces a significant challenge with its low lead conversion rate of around 30%. To address this, a lead scoring model is required to prioritize leads with a higher likelihood of conversion. The CEO has set a target of 80% for the lead conversion rate. Here are the steps taken in the project and recommendations to improve the conversion rate further.

Data Cleaning:

- Dropped columns with more than 30% null values.
- Handled categorical columns by creating new categories, imputing high-frequency values, or dropping irrelevant columns and columns with imbalanced data (category > 60% of value).
- Imputed missing values in numerical categorical data using the mode.
- Performed various data cleaning activities like outliers' treatment, fixing invalid data, grouping low-frequency values, and mapping binary categorical values.

EDA:

- Examined data imbalance, with only 38.5% of leads converting.
- Conducted univariate and bivariate analysis to gain insights into the effect of variables such as lead origin, current occupation, and lead source on the target variable.
- Identified that the time spent on the website has a positive impact on lead conversion.

Data Preparation:

- Created dummy variables using one-hot encoding for categorical variables.
- Split the data into train and test sets using a 70:30 ratio.
- Applied feature scaling through standardization.
- Dropped highly correlated columns to reduce redundancy.

Model Building:

- Utilized Recursive Feature Elimination (RFE) to reduce the number of variables, enhancing model manageability.
- Employed a manual feature reduction process by dropping variables with a p-value greater than 0.05.
- Built three models before finalizing Model 4, which exhibited stability and statistical significance (p-values < 0.05) without multicollinearity issues (VIF < 5).
- Selected "logm10" as the final model with 15 variables and used it for predictions on both the train and test sets.

Model Evaluation:

- Created a confusion matrix and selected a cutoff point of 0.34 based on accuracy, sensitivity, and specificity considerations.
- This cutoff yielded balanced performance metrics, with accuracy, specificity, and precision all around 80%.
- While the precision-recall view showed slightly lower metrics at around 75%, the sensitivity-specificity view was chosen as the optimal cutoff for final predictions.
- Assigned lead scores to the train data using the 0.34 cutoff.

Making Predictions on Test Data:

- Applied scaling and used the final model to make predictions on the test set.
- Evaluation metrics for the train and test sets were both close to 80%.
- Assigned lead scores to the test data.

Top 3 Features:

- Lead Source_Welingak Website
- Lead Source_Reference
- Total Time Spent on the Website

Recommendations:

1. Increase budget and advertising efforts on the Welingak Website to attract more leads.
2. Implement incentives or discounts for customers who provide references that convert into leads to encourage more referrals.
3. Develop aggressive targeting strategies for people spending more time on website, considering their higher conversion.

By implementing these recommendations, X Education can improve its lead conversion rate and move closer to achieving the CEO's target of 80%. Continuous monitoring and optimization of the lead scoring model, along with personalized marketing efforts, can further enhance the conversion rate and overall business performance.