SUMMARY

X Education receives numerous leads, but its lead conversion rate hovers at approximately 30%. The company tasked us with developing a lead scoring model aimed at assigning higher scores to leads with greater conversion potential. The CEO aims for an 80% lead conversion rate.

Data Cleaning:

- Columns with over 40% null values were removed. Categorical columns underwent value count checks to determine appropriate actions: columns were dropped if imputation skewed data, new categories ('others') were created, or high-frequency values were imputed. Columns adding no value were dropped.
- Numerical categorical data were imputed with the mode, and columns with only one unique customer response were dropped.
- Outliers were treated, invalid data was corrected, low-frequency values were grouped, and binary categorical variables were mapped.

EDA:

- Checked for data imbalance; only 38.5% of leads converted.
- Conducted univariate and bivariate analyses for categorical and numerical variables. Insights from variables such as 'Lead Origin,' 'Current occupation,' and 'Lead Source' were noted.
- Positive impact of time spent on the website on lead conversion was observed.

Model Building:

- Used Recursive Feature Elimination (RFE) to reduce variables from 48 to 15 for better manageability.
- Employed manual feature reduction by dropping variables with p-values > 0.05.
- Built a total of three models, selecting Model 4 as final due to stable performance (p-values < 0.05) and no multicollinearity (VIF < 5).

- Applied logm4 as the final model with 12 variables for predictions on both the train and test sets.

Model Evaluation:

- Constructed a confusion matrix and selected a cutoff point of 0.434 based on accuracy, sensitivity, and specificity plots.
- Achieved approximately 80% accuracy, specificity, and precision with this cutoff. Precision-recall view showed slightly lower metrics at around 72%.
- Chose sensitivity-specificity view for the optimal cutoff to meet the CEO's goal of an 80% conversion rate.

Making Predictions on Test Data:

- Applied scaling and used the final model to make predictions on the test data.
- Evaluation metrics for both train and test data were close to 74% accuracy.
- Assigned lead scores using the chosen cutoff of 0.434.

Top 3 Features:

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional

Recommendations:

- Allocate additional budget for advertising on the Welingak Website.
- Implement incentives or discounts for customers who provide references that convert into leads to encourage more referrals.
- Focus marketing efforts aggressively on working professionals, leveraging their high conversion rates and potentially stronger financial capacity to afford higher fees.