LOGISTIC REGRESSION SUMMARY

X Education aims to significantly elevate its lead conversion rate from the current 30% to an ambitious 80%. To achieve this objective, a comprehensive lead scoring model has been developed, emphasizing the prioritization of leads with higher potential for conversion. Here's a succinct overview of the key steps and insights obtained throughout this initiative:

Data Cleaning:

- Columns with over 40% missing values were eliminated, and appropriate strategies were applied to address skewed value counts in categorical columns.
- Numerical categorical data were imputed using the mode, while columns with only one unique customer response were removed.
- Various data cleaning procedures, including outlier handling, invalid data correction, low-frequency value grouping, and binary categorical value mapping, were executed.

Exploratory Data Analysis (EDA):

- A notable data imbalance was identified, with only 38.5% of leads resulting in conversion.
- Univariate and bivariate analyses were conducted, highlighting the influential role of features such as 'Lead Origin,' 'Current occupation,' and 'Lead Source' on the target variable.
- Time spent on the website emerged as a significant factor positively impacting lead conversion.

Data Preparation:

- Dummy features (one-hot encoding) were generated for categorical variables.
- The dataset was partitioned into training and testing sets with a 70:30 ratio.
- Feature scaling was carried out using standardization, and highly correlated columns were removed.

Model Building:

- Recursive Feature Elimination (RFE) was employed to reduce the variable count from 48 to 15.
- A manual feature reduction process was implemented by excluding variables with p-values exceeding 0.05.

LOGISTIC REGRESSION SUMMARY

- Three preliminary models were constructed before finalizing Model 4, characterized by stable p-values (< 0.05) and absence of multicollinearity (VIF < 5).
- 'logm4' was selected as the ultimate model, comprising 12 variables utilized for predictions on both training and test datasets.

Model Evaluation:

- A confusion matrix guided the selection of a cut-off point of 0.345, ensuring balanced accuracy, sensitivity, and specificity metrics all hovering around 80%.
- To align with the CEO's target of an 80% conversion rate, the sensitivity-specificity view was prioritized despite slightly lower performance in precision-recall metrics (around 75%).
- Lead scores were assigned to the training data utilizing the 0.345 cut-off.

Predictions on Test Data:

- Predictions were made on the test data using the final scaled model.
- Evaluation metrics for both training and test data closely approached 80%.
- Lead scores were assigned based on the predictions, with the top three features being 'Lead Source_Welingak Website,' 'Lead Source_Reference,' and 'Current_occupation_Working Professional.'

Recommendations:

- Increase budget allocation for advertising and promotion on the Welingak Website to enhance lead generation.
- Implement incentives or discounts for customers providing references that successfully convert into leads, thereby stimulating more referrals.
- Intensify marketing efforts targeting working professionals, given their notably high conversion rates and potentially greater financial capacity to engage with the company's offerings.