SUMMARY

X Education receives a substantial volume of leads, yet its lead conversion rate remains low at approximately 30%. The company has tasked us with developing a model to assign a lead score to each lead, ensuring that leads with higher scores have a greater likelihood of conversion. The CEO's target for lead conversion rate is approximately 80%.

Data Cleaning:

- Columns containing over 40% null values were removed. Categorical columns underwent value count analysis to determine appropriate actions: columns were dropped if imputation led to skewness, new categories were created for 'others', high-frequency values were imputed, and columns adding no value were dropped.
- Numerical categorical data were imputed with mode, and columns with only one unique response were dropped.
- Outliers were treated, invalid data were corrected, low-frequency values were grouped, and binary categorical values were mapped.

Exploratory Data Analysis (EDA):

- Data imbalance was assessed, revealing that only 38.53% of leads were converted.
- Univariate and bivariate analysis was conducted for both categorical and numerical variables. Insights from variables such as 'Lead Origin', 'Current occupation', and 'Lead Source' provided valuable information on their impact on the target variable.
- Analysis indicated that time spent on the website positively influenced lead conversion rate.

Data Preparation:

- Categorical variables were converted into dummy features through one-hot encoding.
- The dataset was split into training and testing sets using a 70:30 ratio.
- Feature scaling was applied using standardization techniques.
- Several columns were dropped due to high correlation with each other.

Model Building:

- Recursive Feature Elimination (RFE) was employed to reduce the number of variables from 48 to appx. 15, enhancing the manageability of the data frame.
- A manual feature reduction process was implemented, dropping variables with p-values exceeding 0.05.
- Three models were constructed before arriving at the final Model 4, which exhibited stability with p-values below 0.05 and no evidence of multicollinearity with Variance Inflation Factor (VIF) values below 5.
- The final model, logm4, consisting of 12 variables, was selected for making predictions on both the train and test sets.

Model Evaluation:

- A confusion matrix was generated, and a cut-off point of 0.345 was chosen based on the accuracy, sensitivity, and specificity plot. This cut-off yielded accuracy, specificity, and precision all approximately at 80%. However, the precision-recall view showed lower performance metrics, around 75%.
- Despite the CEO's directive to boost the conversion rate to 80%, metrics declined when adopting the precision-recall view. Therefore, we opted for the sensitivity-specificity view to determine the optimal cut-off for final predictions.
- Lead scores were assigned to the training data using the 0.345 cut-off.

Predictions on Test Data:

- Scaling and prediction were performed on the test data using the final model.
- Evaluation metrics for both train and test data were close to 80%.
- Lead scores were assigned.
- The top three features responsible for lead conversion were identified as - Lead Source_Welingak Website, Lead Source_Reference, and Current_occupation_Working Professional.

Final Suggestions:

- Allocation of additional budget towards advertising on the Welingak Website to enhance visibility and engagement will help in lead conversion.
- Implementing incentives or discounts for successful lead referrals, will encourage individuals to provide more references.
- Implementing targeted marketing strategies towards working professionals, will leverage higher conversion rates and potentially stronger financial capacity to pay higher fees.