Summary Report

X Education faces a significant challenge with its lead conversion rate, currently hovering around 30%. The imperative is to devise a model that assigns a lead score to each prospect, thereby enhancing the likelihood of conversion. The ambitious target set by the CEO is to achieve a lead conversion rate of approximately 80%.

Data Preprocessing:

- Columns with over 40% null values were eliminated. Categorical columns underwent meticulous examination, with actions taken based on value counts, including dropping, creating a new category ('others'), imputing high-frequency values, or removing columns deemed non-contributory.
- Numerical categorical data were imputed using the mode, and columns with only a singular unique response were discarded.
- Various activities, such as outliers' treatment, rectifying invalid data, consolidating low-frequency values, and mapping binary categorical values, were undertaken.

Exploratory Data Analysis (EDA):

- A check for data imbalance revealed that only 38.5% of leads were converted.
- Univariate and bivariate analyses were conducted on both categorical and numerical variables, with a focus on variables such as 'Lead Origin,' 'Current Occupation,' and 'Lead Source,' providing valuable insights into their impact on the target variable.
- Notably, time spent on the website exhibited a positive correlation with lead conversion.

Data Preparation:

- Dummy features were created through one-hot encoding for categorical variables.
- The dataset was split into training and test sets in a 70:30 ratio.
- Standardization was employed for feature scaling.
- Highly correlated columns were identified and subsequently dropped.

Model Building:

- Recursive Feature Elimination (RFE) reduced the variables from 48 to 15, enhancing the manageability of the dataframe.
- A manual feature reduction process, involving dropping variables with a p-value exceeding 0.05, was employed.
- Three models were iteratively built before arriving at Model 4, which exhibited stability with p-values consistently below 0.05. No evidence of multicollinearity was observed, with all Variance Inflation Factors (VIF) below 5.
- 'logm4' emerged as the final model, featuring 12 variables, and was utilized for predictions on both the training and test sets.

Model Evaluation:

- A confusion matrix was constructed, and a cutoff point of 0.345 was selected based on accuracy, sensitivity, and specificity plots. This cutoff yielded balanced performance metrics around 80% for accuracy, specificity, and precision, while the precision-recall view exhibited slightly lower metrics, approximately 75%.
- Given the business objective to boost the conversion rate to 80%, the sensitivity-specificity view was prioritized for determining the optimal cutoff for final predictions.
- Lead scores were assigned to the training data using the selected cutoff.

Predictions on Test Data:

- Scaling and predictions on the test set were performed using the final model.
- Evaluation metrics for both the training and test sets closely aligned around 80%.
- Lead scores were assigned, with the top three influential features identified as 'Lead Source_Welingak Website,' 'Lead Souce_Reference,' and 'Current Occupation_Working Professional

Suggestions for xEducation:

Strategic Approach to Lead Engagement Based on Conversion Probability:

Recommended for Phone Calls (High Conversion Potential):

1. Leads sourced from "**Welingak Websites**" and "**Reference**"
2. "**Working Professionals**"
3. Leads demonstrating extended "**time spent on the website**"
4. Leads generated through "**Olark Chat**"
5. Leads with the last activity recorded as "**SMS Sent**"

Discouraged for Phone Calls (Low Conversion Potential):

1. Leads with the last activity identified as "**Olark Chat Conversation**"
2. Leads originating from "**Landing Page Submission**"
3. Leads with "**Others**" as their specialization
4. Leads who have indicated "**Yes**" to "**Do not Email**" in the specialization section.
