SUMMARY

X Education receives a significant number of leads, but its lead conversion rate is moderate at 38.54%, which is below the CEO's target of 80%. The company requires a model to assign lead scores to better identify high-conversion leads.

Data Cleaning:

- Columns with **>45% nulls** were dropped.
- Categorical column values were assessed for skewness before imputation (high-frequency imputation or new category creation used where appropriate).
- Numerical categorical data were imputed with mode, and columns with only one unique response were dropped.
- Additional processing included **outlier treatment, invalid data correction, low-frequency value grouping, and binary categorical mapping**.

EDA:

- **Lead conversion rate:** 38.54% (not 30%).
- **Univariate and bivariate analysis** were performed to identify variables influencing conversion, such as 'Lead Origin' and 'Lead Source'.
- **Time spent on landing page** showed a positive impact on conversion.

Data Preparation:

- **One-hot encoding** was applied to categorical variables.
- **Train-test split:** 70:30 ratio.
- **Feature Scaling:** Standardization was applied.
- **Highly correlated columns** were dropped.

Model Building:

- **Recursive Feature Elimination (RFE)** reduced variables from 48 to 15.
- **Manual feature reduction** retained only statistically significant variables (p-value < 0.05).
- **Final Model:** logm4, with 12 variables, passed multicollinearity checks (VIF < 5).
- **Predictions made on train and test sets using logm4.**

Model Evaluation:

- Confusion matrix and cutoff selection were performed using **accuracy, sensitivity, and specificity**.
- Initial cutoff of **0.5** resulted in **accuracy, specificity, and precision around 80%**, but **precision-recall metrics dropped to ~75%**.
- Since the CEO's target is an **80% conversion rate**, a trade-off was considered.
- **Final cutoff of 0.32** was selected based on **sensitivity-specificity balance**.

Predictions on Test Data:

- Predictions were made using the final model.
- Train and test evaluation metrics were **close to 80%**.
- **Lead scores assigned to test data.**
- **Top 3 predictive features:**
    1. **Lead Source: Welingak Website**
    2. **Lead Source: Reference**
    3. **Current Occupation: Working Professional**

Recommendations:

- **Increase ad spend** on **Welingak Website**, as it has high conversion potential.
- **Offer incentives** for referrals that convert, to encourage more quality references.
- **Target working professionals aggressively**, as they have a high conversion rate and better financial ability to afford premium courses.