Summary

X Education, an online course provider, aims to significantly improve its lead conversion rate from the current 30% to a target of 80%.

Data Cleaning:
- Dropped features with over 35% null values.
- Treated null values in both categorical and numerical columns.
- Removed irrelevant columns.

Exploratory Data Analysis (EDA):
- Checked for data imbalance in the target variable.
- Conducted univariate analysis of both categorical and numerical columns.
- Performed bivariate analysis to understand relationships in the data.

Data Preparation:
- Mapped binary categorical columns to 1/0.
- Created dummy variables for categorical columns with more than two levels.
- Split the data into training and test sets in a 70:30 ratio.
- Applied Min-Max scaling to handle outliers in numerical columns.

Model Building:
- Used Recursive Feature Elimination to select the top 20 features.
- Manually tuned the model to finalize 16 features.
- Iteratively dropped features with p-values higher than 0.05.
- Ensured VIF values were below 5 for all features.
- Finalized Model 6.

Model Evaluation:
- The CEO's goal was to boost lead conversion from 30% to around 80%.
- Chose a probability cutoff of 0.35 based on the sensitivity-specificity curve.
- Achieved a sensitivity of 81%, meeting the CEO's requirement.
- A cutoff of 0.4 yielded a lower sensitivity of 77%, so it was not chosen.
- Obtained an AUC (area under the curve) of 0.88, indicating strong model performance.

Predictions on Test Data:
- Scaled the test data and made predictions using Model 6.
- Achieved a sensitivity of 80% and an AUC of 0.88 on the test set, consistent with the training data results.
- Assigned lead scores to both train and test data.

Recommendations:
- Develop strategies to attract high-quality leads from top-performing sources.
- Increase ads on websites like Welingak to boost total visits and time spent on the site, as these factors have high positive coefficients.
- Analyze features with negative coefficients to identify areas for improvement.