Summary

X Education faces challenges with lead conversion, currently standing at approximately 30%. The task involved building a lead scoring model to assign scores to leads, prioritizing higher scores for customers with a better chance of conversion. The CEO's target for the lead conversion rate is 80%.

Data Preparation:

Initially, columns with over 40% null values were dropped, and categorical columns were handled by either dropping, creating new categories, or imputing based on frequency distribution. Numerical categorical data underwent mode imputation, and columns with single unique responses were removed. Further steps involved outlier treatment, data validation, grouping low-frequency values, and mapping binary categorical values.

EDA:

A check for data imbalance revealed that only 38.5% of leads converted. Univariate and bivariate analyses for categorical and numerical variables showcased influential factors like 'Lead Origin,' 'Current Occupation,' and 'Lead Source.' Notably, time spent on the website displayed a positive impact on lead conversion.

Data Preparation:

Categorical variables were converted into dummy features (one-hot encoded), and the dataset was split into 70:30 ratios for train and test sets. Feature scaling was applied using standardization. Correlated columns were dropped to reduce multicollinearity.

Model Building:

The Recursive Feature Elimination (RFE) technique was employed, reducing variables from 48 to 15. Manual feature reduction was executed by eliminating variables with p-values > 0.05. The final stable model, logm4, consisting of 12 variables, showcased good stability (p-values < 0.05) and no signs of multicollinearity (VIF < 5).

Model Evaluation:

Evaluation involved creating a confusion matrix and selecting a cutoff point of 0.345 based on accuracy, sensitivity, and specificity plots, yielding metrics around 80%. Sensitivity-specificity view was chosen over precision-recall for the final predictions to align with the CEO's goal.

Making Predictions:

Predictions were made on the test data using the finalized model, achieving performance metrics close to 80%. Lead scores were assigned to the dataset based on the chosen cutoff.

Top Features and Recommendations:

The top three influential features for lead conversion were identified as 'Lead Source_Welingak Website,' 'Lead Source_Reference,' and 'Current_occupation_Working Professional.' Recommendations included allocating more budget to the Welingak website, incentivizing references, and targeting working professionals more aggressively due to their higher conversion rates and potentially better financial situations.