Summary

X Education attracts numerous leads, but its lead conversion rate remains low at approximately 30%. The company has tasked us with constructing a model to assign lead scores, aiming to prioritize leads with higher scores, reflecting an increased likelihood of conversion. The CEO has set a target lead conversion rate of approximately 80%.

Data Cleaning:

- We eliminated columns with null values exceeding 40%. To handle categorical columns, we scrutinized value counts to determine the best course of action: if imputation introduced skewness, we either dropped the column, introduced a new category like 'others,' or performed imputation with the high-frequency value. Columns providing minimal value were also removed.
- For numerical categorical data, we imputed using the mode, and columns with only one unique customer response were excluded.
- Various tasks, including addressing outliers, correcting invalid data, consolidating low-frequency values, and mapping binary categorical values, were executed as part of the data preprocessing.

EDA:

- Verified data imbalance, revealing that only 38.5% of leads resulted in conversion.
- Conducted univariate and bivariate analyses for both categorical and numerical variables. Variables such as 'Lead Origin,' 'Current occupation,' and 'Lead Source' yielded valuable insights into their impact on the target variable.
- Notably, the time spent on the website demonstrated a positive correlation with lead conversion.

Data Preparation:

- Generated dummy features through one-hot encoding for categorical variables.
- Split the dataset into training and testing sets in a 70:30 ratio.
- Applied feature scaling using standardization.
- Removed certain columns due to high correlation with each other.

Model Building:

- Employed Recursive Feature Elimination (RFE) to streamline the dataset, reducing variables from 48 to 15 for better manageability.
- Utilized a manual feature reduction process, systematically dropping variables with p-values exceeding 0.05, resulting in the construction of three models. Model 4 emerged as the final choice, demonstrating stability with all p-values < 0.05 and no signs of multicollinearity (VIF < 5).
- Selected 'logm4' as the definitive model, comprising 12 variables, and leveraged it for predictions on both the training and test sets.

Model Evaluation:

- To address the business objective of achieving an 80% conversion rate, a sensitivity-specificity perspective was prioritized over precision-recall, as the latter resulted in a decline in overall metrics.
- Assigned lead scores to the training data using the determined cutoff of 0.345 for final predictions.

Making Predictions on Test Data:

- Executed predictions on the test set by scaling and utilizing the final model.
- Evaluation metrics for both the training and test sets closely approximate 80%.
- Assigned lead scores to the predictions.