Summary

X Education receives a significant number of leads, but its lead conversion rate is low, hovering around 30%. The company asked us to develop a model to assign a lead score to each prospect, improving the chances of conversion, with a target conversion rate of 80%, as set by the CEO.

Data Cleaning:

- Columns with more than 40% missing data were removed.
- Categorical columns were examined, and where imputation could introduce skew, they were either dropped or merged into an "Others" category. High-frequency values were imputed when appropriate.
- Numerical categorical data was imputed with the mode, and columns with only a single unique value were dropped.
- Other cleaning tasks included addressing outliers, correcting invalid data, grouping low-frequency values, and converting binary categorical variables.

Exploratory Data Analysis (EDA):

- We identified that only 38.5% of leads were converted, highlighting a significant data imbalance.
- Both univariate and bivariate analyses revealed valuable insights, particularly from features such as ‘Lead Origin’, ‘Current Occupation’, and ‘Lead Source’, which were found to significantly influence conversion.
- We found that the time spent on the website had a positive correlation with the likelihood of conversion.

Data Preparation:

- Categorical features were transformed using one-hot encoding to create dummy variables.
- The dataset was split into training and testing sets with a 70:30 ratio.
- Feature scaling was performed using standardization.
- Highly correlated columns were removed to minimize multicollinearity.

Model Building:

- We applied Recursive Feature Elimination (RFE) to reduce the feature set from 48 to 15, making the dataset more manageable.
- A manual feature reduction process was employed, dropping variables with p-values greater than 0.05.

- Three models were developed, with the final Model 4 being chosen due to stable performance and all variables having p-values below 0.05. The Variance Inflation Factor (VIF) analysis confirmed no multicollinearity.
- The final model, *logm4*, with 12 significant variables, was used for predictions on both the training and test sets.

Model Evaluation:

- We used a confusion matrix and selected a cut-off of 0.345 based on accuracy, sensitivity, and specificity analysis. This threshold resulted in approximately 80% accuracy, specificity, and precision.
- However, when evaluating the model through the precision-recall lens, performance dropped to about 75%. Since the business goal was to increase conversion rates, the sensitivity-specificity balance was chosen as the optimal cut-off for final predictions.

Making Predictions on Test Data:

- The final model was applied to the test dataset, with both training and test evaluation metrics remaining consistent at around 80%.
- Lead scores were assigned to the test data.

Key Features:

- Top 3 Features:

    1. Lead Source: Welingkar Website
    2. Lead Source: Reference
    3. Current Occupation: Working Professional

Recommendations:

1. Increase the marketing budget for the Welingkar Website to attract more leads, as it is a top predictor of conversions.
2. Introduce incentives or discounts for leads generated through references, as they show a higher conversion rate.
3. Focus on targeting working professionals, who have a higher likelihood of conversion and can afford higher tuition fees.