Summary

X Education generates a large number of leads; nevertheless, its lead conversion rate is only 30%. To address this, the company has assigned us the responsibility of creating a model that allocates a lead score to each possible consumer. The goal is to boost conversion chances for clients with higher lead scores, with a target lead conversion rate of around 80%.

Data Cleaning:

- Eliminated columns with over 40% missing data. For categorical columns, we evaluated value distributions to determine actions: columns with skewed imputations were dropped, new categories ('others') were created, high-frequency values were imputed, and columns with negligible value were removed.
- Applied mode imputation to numerical categorical data and dropped columns with a solitary unique customer response.
- Conducted various tasks like handling outliers, rectifying invalid data, grouping infrequent values, and mapping binary categorical values.

EDA:

- Verified data imbalance, noting that only 38.5% of leads converted.
- Conducted univariate and bivariate analyses for both categorical and numerical variables. Notable variables such as 'Lead Origin,' 'Current occupation,' and 'Lead Source' yielded valuable insights into their effects on the target variable.
- Discovered a positive correlation between time spent on the website and lead conversion.

Data Preparation:

- Introduced dummy features (one-hot encoded) to represent categorical variables.
- Segregated the dataset into Train and Test sets using a 70:30 ratio.
- Employed feature scaling via standardization.
- Removed certain columns with high intercorrelation.

Model Building:

- Utilized Recursive Feature Elimination (RFE) to streamline the variables from 48 to 15, enhancing manageability.
- Employed manual feature reduction by dropping variables with p-values exceeding 0.05.
- Constructed three models before finalizing Model 4, which demonstrated stability with p-values below 0.05. Multicollinearity was absent, with Variance Inflation Factor (VIF) below 5.
- Selected 'logm4' as the definitive model, featuring 12 variables, for prediction on both training and test sets.

Model Evaluation:

- Utilized Recursive Feature Elimination (RFE) to streamline the variables from 48 to 15, enhancing manageability.
- Employed manual feature reduction by dropping variables with p-values exceeding 0.05.
- Constructed three models before finalizing Model 4, which demonstrated stability with p-values below 0.05. Multicollinearity was absent, with Variance Inflation Factor (VIF) below 5.
- Selected 'logm4' as the definitive model, featuring 12 variables, for prediction on both training and test sets.

Making Predictions on Test Data:

- Constructed a confusion matrix and identified a threshold of 0.345 by evaluating accuracy, sensitivity, and specificity plots. This threshold achieved balanced metrics around 80% for accuracy, specificity, and precision. However, metrics from the precision-recall perspective yielded slightly lower performance, around 75%.
- Considering the business goal of achieving an 80% conversion rate, a trade-off was observed in precision-recall metrics. Hence, sensitivity-specificity perspective was adopted to determine the optimal cut-off for final predictions.
- Utilized the 0.345 cut-off to assign lead scores to the training data.

Recommendations:

- Applied scaling and made predictions on the test data using the final model.
- Evaluated metrics for both training and test data, yielding results close to 80%.
- Assigned lead scores to the test data.
- Key contributing features include 'Lead Source_Welingak Website,' 'Lead Source_Reference,' and 'Current_occupation_Working Professional.'