SUMMARY

X Education receives a large volume of leads; however, its conversion rate is currently quite low at approximately 30%. To address this, the company has tasked us with developing a model to assign lead scores to each prospect, where higher scores indicate a greater likelihood of conversion. The ultimate goal is to help the company achieve the CEO's ambitious target of an 80% lead conversion rate.

Data Cleaning:

- Columns with more than 40% missing values were removed to ensure data reliability.
- Categorical columns were analyzed for value counts, and appropriate actions were taken:
    - Dropped columns if imputation caused skew.
    - Created a new category labeled "Others" for low-frequency values.
    - High-frequency values were imputed.
    - Irrelevant columns were discarded.
- Numerical categorical variables were imputed using the mode.
- Columns with a single unique customer response were eliminated.
- Outliers were identified and treated to maintain data integrity.
- Invalid data entries were corrected for accuracy.
- Categories with low frequency were grouped together.
- Binary categorical variables were standardized for uniformity in analysis.

EDA:

- Only 38.5% of leads were converted, indicating data imbalance.
- Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
- Time spent on the website positively influences lead conversion rates, indicating higher engagement and interest from potential customers.

Data Preparation:

- One-hot encoded categorical variables to create dummy features.
- Split data into Train and Test sets in a 70:30 ratio.
- Applied feature scaling through standardization.
- Removed highly correlated columns to avoid redundancy.

Model Building:

- Reduced variables from 48 to 15 using RFE for better manageability.
- Applied manual feature reduction by eliminating variables with p-values > 0.05.
- Built four models before finalizing Model 5, which was stable with p-values < 0.05 and no multicollinearity (VIF < 5).
- Selected logm5 as the final model with 11 variables for predictions on train and test sets.

Model Evaluation:

- A confusion matrix was created, and a cut-off point of 0.345 was chosen based on accuracy, sensitivity, and specificity plots. This cut-off provided approximately 80% for accuracy, specificity, and precision, though the precision-recall view showed lower performance metrics around 75%.
- To align with the CEO's target of boosting the conversion rate to 80%, the sensitivity-specificity view was selected as the optimal cut-off for final predictions.
- Lead scores were assigned to the training data using the 0.345 cut-off point.

Making Predictions on Test Data:

- Predictions were made on the test set after scaling and applying the final model.
- Evaluation metrics for both train and test sets were consistent, around 80%.
- Lead scores were assigned based on the predictions.
- Top 3 features are:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current_occupation_Working Professional

Recommendations:

- Increase advertising budget for Welingak's website to boost visibility and lead generation.
- Offer incentives or discounts to encourage referrals that successfully convert into leads, fostering more recommendations.
- Prioritize targeting working professionals, as they exhibit higher conversion rates and are likely to have the financial stability to invest in premium programs.
