Summary

X Education receives a number of leads, but only about 30% of those prospects actually become customers. The business wants us to create a model in which we rate each lead individually so that leads with better scores have a higher chance of converting. The CEO aims to convert leads at a rate of about 80%.

Cleaning of Data:

- Numerical categorical data were imputed with the mode, and entries with just one distinct client answer were eliminated.
- Other tasks included handling anomalies, correcting inaccurate data, clustering low-frequency values, and translating binary categorical values.
- Over 40% of invalid column values were removed. Value counts within categories columns were examined to determine the best course of action: if imputation causes skew, the column was dropped; otherwise, a new category (others) was established; high-frequency values were imputation; and value-zero columns were dropped.

Exploratory Data Analysis:

- Time spent on a website has a beneficial effect on converting visitors to leads.
- Checked for data disparity, only 38.5% of leads were transformed.
- Conducted categorical and numerical variable univariate and bivariate analyses. The terms "Lead Origin," "Current Occupation," "Lead Source," and others offer useful information about the impact on the objective variable.

Preparation of Data:

- Scaled the features using standardization
- Sets for the train and test are divided 70:30.
- Created dummy features (one-hot encoded) for category variables
- Removed a few columns because of their high correlation

Building the Model:

- Models were carefully built through feature reduction by removing factors with a p-value higher than 0.05.
- RFE was used to condense 48 factors down to 15. Data frame will be easier to handle as a result.
- We used the final model, logm4, which had 12 factors, to make predictions on both the train and test sets.
- Before arriving at the final Model 4, which was steady with (p-values - 0.05), a total of 3 models were constructed. There is no indication of multi-collinearity for VIF 5.

Evaluating the Model:

- Using a cutoff of 0.345, lead scores were given to the train data.
- As a means of resolving a business issue, the CEO requested an increase in conversion rate to 80%; however, measurements declined when we adopted a precision-recall perspective. As a result, we will decide that the sensitivity-specificity perspective is the best cut-off for final forecasts.
- On the basis of the accuracy, sensitivity, and precision plot, a confusion matrix was created, and a cutoff value of 0.345 was chosen. With a cutoff of 80%, accuracy, sensitivity, and precision were all obtained. The performance measures from the pinpoint recall perspective were only about 75% as good.

Test data predictions:

- Making predictions during testing: Scaling and foreseeing using the end model.
- The train and test evaluation scores are very near to 80%.
- A lead score was assigned.
- Top 3 features are:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current_occupation_Working Professional

Recommendations:

- Working professionals should be actively pursued because they convert well and are more likely to be in a position to pay higher fees.
- Discounts or incentives for supplying references that result in leads, which motivates submitting more references.
- The Welingak website could use more funding for things like ads.