Although X Education receives a large number of leads, its lead conversion rate is quite low, hovering around 30%. The organisation wants us to create a model in which we award a lead score to each lead so that clients with a higher lead score have a better probability of converting. The CEO's goal for lead conversion rate is approximately 80%.

Data Cleaning:

- Over 40% null column values were removed. Value counts within categorical columns were reviewed to determine the best course of action: eliminate the column, create a new category (others), impute high frequency values, and drop columns that don't contribute any value if imputation creates skew.
- Numerical categorical data was calculated with mode, and columns with just one distinct response from the client were eliminated.
- Other operations included the treatment of outliers, the correction of incorrect information, the grouping of low frequency values, and the mapping of binary categorical values.

EDA:

- Checked for data imbalance, only 38.5% of leads were converted.
- 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 spend on website shows positive impact on lead conversion.

Data Preparation:

- Dummy features (one-hot encoded) were produced for categorical variables.
- 70:30 ratio for dividing the train and test sets.
- Standardization-based Feature Scaling.
- Dropped a couple columns because they were very connected with one another.

Model Building:

- RFE was used to condense 48 variables down to 15. Dataframe will be easier to manage as a result.
- By excluding variables with a p-value greater than 0.05, models were constructed manually using feature reduction.
- Before arriving at the final Model 4, which was stable with (p-values 0.05), a total of 3 models were constructed. With VIF 5, there is no indication of multicollinearity.
- We utilised the final model, logm4, which included 12 variables, to make predictions on both the train and test sets.

Model Evaluation:

- Confusion matrix was created, with 0.345 chosen as the cutoff threshold based on accuracy, sensitivity, and specificity plots. Accuracy, specificity, and precision were all around 80% at this cutoff. The precise recall view, however, provided performance values that were less than 75%.
- CEO requested an increase in conversion rate to 80% in order to solve a business challenge, but metrics declined if we adopted a precision-recall perspective. Therefore, sensitivity-specificity view will be our top candidate for the cut-off for final forecasts.
- The cutoff value of 0.345 was used to award the lead score to the train data.

Making Predictions on Test Data:

- Making predictions while taking a test: Scaling and forecasting using the final model.
- Evaluation metrics for both the train and test phases are very close to 80%.
- The score for the lead was assigned.
- The top three features are.
    - Lead Source_Welingak Website.
    - Lead Source_Reference.
    - Current_occupation_Working Professional.

Recommendations:

- The Welingak website might use more funding for things like advertising.
- Discounts or incentives for supplying references that result in leads, which motivates submitting more references.
- Targeting working professionals aggressively is recommended since they convert well and will have greater financial outcomes.
- Circumstances to pay larger fees as well.