# Summary

X Education gets a lot of leads, its lead conversion rate is very poor at around 30%. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have higher conversion chances. The CEO's target for lead conversion rate is around 80%.

## Data Cleaning:

- Columns with more than 40% missing values were removed. Categorical columns were assessed for skewness due to imputation; if imputation caused skew, columns were either dropped or consolidated into an "Others" category, or imputed with the most frequent value. Columns with minimal value added were dropped.
- Numerical categorical data were imputed using the mode, while columns with a single unique customer response were discarded.
- Additional steps included treating outliers, addressing invalid data, grouping low-frequency categories, and mapping binary categorical variables.

## EDA:

- Data imbalance was assessed, with only 38.5% of leads converting.
- Univariate and bivariate analysis were conducted on categorical and numerical variables. Key insights were gained from features like ‘Lead Origin’, ‘Current Occupation’, and ‘Lead Source’, which significantly impact the target variable.
- Time spent on the website showed a positive correlation with lead conversion.

## Data Preparation:

- Created dummy variables (one-hot encoding) for categorical features.
- Split the dataset into training and testing sets using a 70:30 ratio.
- Applied feature scaling through standardization.
- Dropped highly correlated columns to avoid multicollinearity.

## Model Building:

- Applied Recursive Feature Elimination (RFE) to reduce variables from 66 to 15, making the dataframe more manageable.
- Used manual feature reduction by dropping variables with p-values > 0.05.
- Built 2 models before finalizing Model 3, which was stable with p-values < 0.05 and no sign of multicollinearity (VIF < 5).
- Decided lrm3 as the final model with 13 variables, used for predictions on both the training and test sets.

## Model Evaluation:

- Generated the confusion matrix and selected a cutoff point of 0.36 based on accuracy, sensitivity, and specificity plot, achieving around 80% for accuracy, specificity, and precision.
- To meet the business objective of boosting the conversion rate to 80%, metrics dropped when using the precision-recall view. Hence, opted for the sensitivity-specificity view for determining the optimal cutoff for final predictions.
- Assigned lead scores to the training data using the 0.36 cutoff.

## Making Predictions on Test Data:

- Made predictions on the test set after scaling and using the final model.
- Evaluation metrics for both train and test sets are closely aligned at around 80%.
- Assigned lead scores to the test data.
- The top 3 features driving predictions are:
    - Total time spent on the website
    - Lead Source_Welingak Website
    - Current_occupation_Working Professional

## Recommendations:

- Allocate more budget and spend on Welingak Website for advertising and promotion to attract more leads.
- Offer incentives or discounts for customers who refer others that successfully convert into leads, encouraging more referrals.
- Focus on aggressively targeting working professionals, as they have a higher conversion rate and better financial capability to pay higher fees.