Summary

X Education receives numerous leads, but its current lead conversion rate is relatively low at about 30%. The company needs us to develop a model that assigns a lead score to each prospect. This scoring should ensure that leads with higher scores are more likely to convert. The CEO aims to increase the lead conversion rate to approximately 80%.

# Data Cleaning:

* Columns with more than 40% missing values were removed. For categorical columns, value distributions were reviewed to determine the best course of action: if imputation led to skewed data, the column was either removed, a new category labeled "others" was created, high-frequency values were used for imputation, or columns that offered no additional value were dropped.
* Numerical categorical data were imputed using the mode, and columns with only a single unique response from customers were discarded.
* Additional preprocessing steps included handling outliers, correcting invalid data entries, consolidating low-frequency values, and mapping binary categorical variables.

# EDA:

* Checked for data imbalance, finding that only 38.5% of leads were converted.
* Conducted both univariate and bivariate analyses for categorical and numerical variables. Variables such as 'Lead Origin', 'Current Occupation', and 'Lead Source' were analyzed for their influence on the target variable.
* The amount of time spent on the website was found to have a positive effect on lead conversion.

# Data Preparation:

* Created dummy features (one-hot encoded) for categorical variables
* Splitting Train & Test Sets: 70:30 ratio
* Feature Scaling using Normalization
* Removed several columns due to their high correlation with one another.

# Model Building:

* Used RFE and VIF to reduce variables from 48 to 15. This will make dataframe more manageable.
* A manual feature selection process was employed to construct the models, which involved excluding variables with a p-value greater than 0.05.
* Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
* Logm3 was selected as final model with 13 variables, we used it for making prediction on train and test set.

# Model Evaluation:

* Confusion matrix was made and cut off point of 0.42 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.
* As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions
* Lead score was assigned to train data using 0.42 as cut off.

# Making Predictions on Test Data:

* Making Predictions on Test: Scaling and predicting using final model.
* Evaluation metrics for train & test are very close to around 80%.
* Lead score was assigned.
* Top 3 features are:
  + Total Time Spent on Website
  + Lead Origin\_Lead Add Form
  + Current\_occupation\_Working Professional

# Recommendations:

* More budget/spend can be done on Website in terms of advertising, etc.
* Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
* Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.