# Summary report

Although X Education receives many leads, it has a low lead conversion rate of approximately 30%. The company has instructed us to create a model that involves assigning a lead score to each lead, which will help identify customers with a greater likelihood of conversion. The CEO aims for the lead conversion rate to be approximately 80%.

**Data Understanding and Data Cleaning –**

- Columns that had more than 40% null values were removed. To determine the appropriate course of action, the value counts in categorical columns were examined. If imputing values leads to an imbalance, the column would be eliminated, an others category would be created, the most frequently occurring value would be imputed, or columns without value would be discarded.
- Various other tasks were conducted, such as addressing outliers, rectifying any invalid data, categorizing infrequent values, and converting binary categorical values.

**Exploratory Data Analysis (EDA)** –

- I analyzed categorical and numerical variables separately and together using univariate and bivariate analysis methods. Words such as 'Lead Origin', 'Current occupation', and 'Lead Source' are examples. Offering significant input regarding the impact on the desired outcome.
- The amount of time spent on a website has a beneficial effect on the conversion of leads.

**Data Preparation** –

- We created fabricated features (encoded as binary) to represent categorical data.
- Dividing the data into two sets, with 70% for training and 30% for testing.
- Standardization is a technique for scaling features.

- A few columns were removed as they were closely related to one another.

**Model Building** –

- RFE was employed to decrease the number of variables from 48 to 15. This will help to handle the data frame more easily.
- The process of manually reducing features was used to create models by eliminating variables that had a p-value greater than 0.05.
- Before arriving at the final stable Model 4 with (p-values < 0.05), a total of three different models were constructed. There is no evidence of multicollinearity as the VIF is less than 5.
- The logm4 model, which comprised of 12 variables, was chosen as the ultimate model and subsequently employed to predict outcomes on both the train and test sets.

**Prediction on Test Data** –

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

**Recommendation** –

- On the Welingak website, additional spending can be made on things like advertising.
- Offers of rewards or discounts in exchange for references that result in leads motivate people to offer more references.
- Working professionals should be aggressively targeted because they convert well and are more likely to have the money to pay higher fees.