Summary Report

X Education has many leads, its lead conversion rate is very poor at around 30%. Build a model in which 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 and EDA:

- Columns with > 40% nulls were dropped.
- Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
- Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
- Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.
- Data Imbalance Ratio : 1.59 : 1.
- Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Lead Source', etc. provide valuable insight on effect on target variables.
- Time spent on websites shows a positive impact on lead conversion.

Data Preparation and Model Building:

- Created dummy features for categorical variables.
- Splitting Train & Test Sets in 70:30 ratio.
- Feature Scaling using Standardization.
- Dropped a few columns, they were highly correlated with each other.
- Used RFE to reduce variables from 43 to 15. This will make the dataframe more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 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.
- res4 was selected as the final model with 12 variables, we used it for making predictions on the train and test set.

Model Evaluation and Making Predictions on Test Data:

- Confusion matrix was made and the cut off point of 0.36 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 79%.
- As to solve the business problem, the CEO asked to boost conversion rate to 79%, but metrics dropped when we took a 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 0.36 as cut off.

- Making Predictions on Test: Scaling and predicting using the final model.
- Evaluation metrics for train & test are very close to around 80%.
- Lead score was assigned.
- Top 3 features are:
  o Lead Source_Welingak Website
  o Lead Source_Reference
  o Last Activity_SMS Sent

Recommendations:

- Spending more budget can be done on Welingak Website in terms of advertising.
- Discounts for providing references that convert to lead, encourage to provide more references.
- Improve Landing page submission or get to know what is causing the leads to stop using the website.
- Monitoring the Last activity can also help in contacting the users whenever possible.
- Can be more focused on positive coefficients which gives more on marketing strategies.