Summary Report

X Education has a lot of leads, The primary objective of this project is to increase the lead conversion rate from the existing 30% to approximately 80%. This will be achieved by implementing a lead scoring system that evaluates and prioritizes leads based on their conversion potential.

Data Understanding:

- Read the data with the help of pandas library.
- See the shape info and describe the data.
- Look for missing or null values.

Data Cleaning:

- Columns with null values more than 40% were dropped
- Checked the values counts of categorical columns, if imputation causes skew, then column was dropped or created new category (others), grouped low frequency values, dropped columns that don't add any value.
- Numerical categorical data were imputed with mode and dropped columns with only one unique response from customer.
- Mapped binary categorical values.

EDA:

- Data imbalance checked- only 38.53% leads converted.
- Performed Auto EDA.

Data Preparation:

- Created dummy variables for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization
- Saw the correlation and dropped highly correlated columns.

Model Building:

- Used RFE to reduce variables from 48 to 15. To make dataframe more manageable.
- Dropped variables with p – value > 0.05.
- Total 2 models were built before reaching final Model 3 which was stable with (p-values < 0.05) and there was 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 0.347 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 where 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 0.347 as cut off.

Making Predictions on Test Set:

- 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:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - What is your current occupation_Working Professional

Observations:

- Focus on features with positive coefficients for targeted marketing strategies.
- More budget can be spent on Welingak Website in terms of advertising ,marketing etc.
- Working professionals to be can be targeted as they have high conversion rate and will be able to pay high fees too as they are financially able.
- Areas with negative coefficients like 'Lead Origin_Landing Page Submission' should be analysed and put more work on.