SUMMARY

The initial goal was to increase X Education's lead conversion rate, which was already around 30%. The organisation needed to create a model that assigns a lead score to each lead. Customers with a higher lead score are more likely to convert. The CEO targeted a lead conversion rate of approximately 80%.

The data cleaning process included removing columns with over 40% null values, imputing categorical data, treating outliers, repairing invalid data, combining low-frequency values, and mapping categorical values. Exploratory data analysis (EDA) included checking for data imbalance, performing univariate and bivariate analysis for categorical and numerical variables, and identifying the variables that had a significant effect on the target variable.

Data preparation included creating dummy features for categorical variables, splitting into train and test sets, feature scaling using standardization, and removing strongly correlated columns. To develop the model, variables were reduced by both recursive feature elimination (RFE) and manual reduction. Three models were built before reaching the final model, which was stable with p-values < 0.05 and no sign of multicollinearity with VIF < 5.

The logm3 model, with 12 variables, was used to make predictions on both the train and test datasets. To evaluate the model, we created a confusion matrix and chose a cut-off point of 0.345 based on accuracy, sensitivity, and specificity plots. The train data was awarded a lead score with a cut-off of 0.345. The top three features were Lead Source_Welingak Website, Lead Source_Reference, and Current_occupation_Working Professional.

The analysis suggests increasing budget for Welingak website advertising, offering incentives/discounts for lead generation, and targeting working professionals with high conversion rates and better financial situations.

Finally, the analysis provided insights into the factors that affect lead conversion rates and recommended strategies to improve the conversion rates.