Summary

To begin with, the task was to enhance the lead conversion rate of X Education, which was currently at approximately 30%. The objective was to build a model that assigns a lead score to each lead, enabling higher-converting leads to receive higher scores. The CEO's target for the lead conversion rate was around 80%.

The process of cleaning the data involved dropping columns that had more than 40% null values, imputing categorical data using appropriate methods, addressing outliers, fixing invalid data, grouping low-frequency values, and mapping binary categorical values. Exploratory data analysis (EDA) included checking for data imbalance, performing univariate and bivariate analysis for both categorical and numerical variables, and identifying variables that significantly influenced the target variable.

Data preparation consisted of creating dummy features for categorical variables, splitting the data into training and test sets, performing feature scaling using standardization, and dropping highly correlated columns. The model building process involved reducing the number of variables using recursive feature elimination (RFE) and manual feature reduction. Three models were constructed before arriving at the final model, which exhibited stability with p-values less than 0.05 and no indication of multicollinearity with VIF less than 5.

The final model, named logm4, utilized 12 variables and was employed to make predictions on both the training and test sets. Model evaluation entailed creating a confusion matrix and selecting a cutoff point of 0.345 based on accuracy, sensitivity, and specificity plots. The lead score was assigned to the training data using the 0.345 cutoff, and the top three features were identified as Lead Source_Welingak Website, Lead Source_Reference, and Current_occupation_Working Professional.

Based on the analysis, the recommendations are to allocate a larger budget for advertising on the Welingak Website, offer incentives or discounts for providing references that lead to conversions, and aggressively target working professionals due to their high conversion rates and better financial situations.

Overall, this project provided a hands-on experience in data cleaning, exploratory data analysis, data preparation, model building, and evaluation. The assignment also highlighted the importance of selecting appropriate evaluation metrics based on the business problem and understanding the tradeoffs between different metrics. Lastly, the analysis offered insights into the factors influencing lead conversion rates and proposed strategies for improving these rates.