Summary

To begin with, the problem was to improve the lead conversion rate of X Education, which was currently at around 30%. The company required building a model to assign a lead score to each lead so that customers with a higher lead score have a higher conversion chance. The CEO's target for the lead conversion rate was around 80%.

The data cleaning process involved dropping columns with more than 40% null values, imputing categorical data using appropriate methods, treating outliers, fixing invalid data, grouping low- frequency values, and mapping binary categorical values. The exploratory data analysis (EDA) involved checking 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.

The data preparation involved creating dummy features for categorical variables, splitting the data into train and test sets, 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 built before reaching the final model, which was stable with p-values < 0.05 and no sign of multicollinearity with VIF < 5.

The final model, logm4, had 12 variables, and it was used to make predictions on the train and test sets. The model evaluation involved creating a confusion matrix and selecting a cut-off point of 0.347 based on accuracy, sensitivity, and specificity plot. The lead score was assigned to the train data using the 0.347 cut-off, and the top three features were Lead Source_Welingak Website, Lead Source_Reference, and Current_occupation_Working Professional.

Based on the analysis, the recommendations were to allocate more budget/spend on Welingak Website advertising, provide incentives/discounts for providing reference that converts to lead, and aggressively target working professionals as they have a high conversion rate and better financial situations.

Overall, the assignment provided a hands-on experience in data cleaning, exploratory data analysis, data preparation, model building, and evaluation. The project also demonstrated the importance of selecting appropriate evaluation metrics based on the business problem and understanding the trade- offs between different metrics. Finally, the analysis provided insights into the factors that affect lead conversion rates and recommended strategies to improve the conversion rates.