Summary

X Education grapples with a subpar lead conversion rate of approximately 30%, prompting the need for a model to assign lead scores. The goal is to increase the chances of conversion for leads with higher scores, aligning with the CEO's ambitious target of an 80% conversion rate.

The dataset comprises 9,240 records with 37 features, including columns with lengthy names that require shortening for better handling. Identifiers like "Prospect ID" and "Lead Number" are considered non-contributory and can be dropped. Additionally, five columns contain only one unique value.

Handling null values is imperative, with 17 affected columns. Treatment methods, such as dropping or imputing suitable values, are chosen based on their potential impact on the target variable.

Univariate analysis categorizes the dataset into categorical and numerical columns. For categorical columns, decisions on retention, dropping, or manipulation are influenced by value distribution and their relationship with the target variable. Numerical columns undergo outlier treatment using 99-percentile values.

Multivariate analysis identifies a problematic high correlation between "TotalVisits" and "Page_per_visit," prompting the removal of "Page_per_visit" to address collinearity. Further analysis is required for the correlation between "Lead Source" and "Lead Origin" during model building.

Data processing involves feature creation, converting numerical columns to categorical ones, one-hot encoding, binning of categorical columns, and splitting the dataset into a 70:30 ratio for train and test sets. Numerical features are scaled for uniformity.

Model building utilizes Recursive Feature Elimination (RFE) to reduce variables, resulting in the creation of two models before reaching the stable final Model 3 with p-values < 0.05 and no multicollinearity (VIF < 5). The final model, consisting of 14 variables, is employed to predict both train and test sets.

Model evaluation involves creating a confusion matrix and selecting a cut-off point of 0.34 based on accuracy, sensitivity, and specificity plots. This cut-off achieves an impressive 88% accuracy, specificity, and precision. The precision-recall view yields slightly lower metrics at around 82%. Importantly, these metrics meet the CEO's target of an 80% conversion rate.

Model interpretation reveals key factors affecting lead outcomes. Negative correlation coefficients are associated with leads of uncertain or poor quality. Positive lead outcomes are likely when the "Tag" indicates an intention to revert after reading an email or falls under the category of "Others." Additionally, specific values in features like "Reason for Course Select," "Last Activity," "DN_Email," "Web_Time," "Occupation," "Lead Origin," and "Last Activity" significantly influence lead positivity or negativity.

In conclusion, the comprehensive analysis and modeling efforts successfully address the business problem, providing actionable insights to achieve the CEO's ambitious conversion rate target.