Summary

In response to X Education's need for improving lead conversion rates, we embarked on building a predictive model to assign lead scores. The objective was to prioritize leads with higher scores, ensuring a better chance of conversion. Despite the current conversion rate of around 30%, the CEO's ambitious target was an 80% conversion rate.

Data Cleaning:

Our journey began with meticulous data cleaning:

- Columns with excessive null values (>40%) were dropped.
- Categorical columns were carefully treated, with imputation or category creation based on data insights.
- Numerical categorical data were imputed using the mode, and columns with single unique responses were eliminated.
- Outliers, invalid data, and low-frequency values were managed.

Exploratory Data Analysis (EDA): Through thorough EDA:

We addressed data imbalance, where only 38.5% of leads converted.

- Univariate and bivariate analysis illuminated key variables like 'Lead Origin,' 'Current Occupation,' and 'Lead Source.'
- The positive impact of time spent on the website on lead conversion was revealed.

Data Preparation: Our data preparation involved:

- Creating one-hot encoded dummy features for categorical variables.
- Splitting data into a 70:30 train-test ratio.
- Standardizing features for consistent scaling.
- Removing highly correlated columns to enhance model efficiency.

Model Building: Our model development followed a systematic approach:

Recursive Feature Elimination (RFE) streamlined variables from 48 to 15.
Manual Feature Reduction refined the model by eliminating variables with high p-values.

Two models emerged:

Model 1:

Post-RFE, we used stats models to build a promising model, excluding 'Current_occupation_Housewife' due to its insignificant p-value.

Model 2:

This enhanced model, built after further refinement, exhibited substantial coefficients that pinpoint variable influences.

Model Evaluation:

With a focus on achieving the CEO's 80% conversion target, we evaluated our models:

- A cut-off of 0.295 was chosen through an accuracy, sensitivity, and specificity analysis.
- Precision-recall metrics slightly dropped performance, leading us to favor the sensitivity-specificity view.
- By assigning lead scores based on the optimal cut-off, we made predictions on the train data.

Making Predictions on Test Data:

Our final steps included:

- Predicting on the test data using the scaled final model.
- Achieving evaluation metrics of around 85% for both train and test data.
- Assigning lead scores to the test data.

Recommendations:

In line with the model's insights, we propose strategic actions:

- Allocate a greater advertising budget to the Welingak Website to maximize its impact.
- Introduce incentives or discounts for successful referrals, promoting lead generation through references.
- Aggressively target working professionals, leveraging their higher conversion rates and financial capability.

In this journey, we've harnessed data-driven insights to pave the way for remarkable improvements in lead conversion rates for X Education.