Company X Lead Scoring Case Study Summary

Overview:

Despite receiving a lot of leads, X Education only converts about 30% of those leads into sales. The organization wants us to develop a model where each lead is given a lead score, increasing the likelihood that a customer would convert if their lead score were greater. The CEO wants to convert leads at a rate of about 80%.

Procedure Followed:

1. **Data Cleaning:**

   - The removal of null column values exceeded 40%. Value counts inside categorical columns were examined to ascertain the optimal course of action: delete non-value-adding columns if imputation results in skew, impute high frequency values, eliminate the column, or create a new category (others).

   - The mode was used to generate the numerical categorical data, and columns containing only one unique client answer were removed.

   - Additional procedures included translating binary category data, grouping low frequency values, correcting inaccurate information, and handling outliers.

2. **EDA:**

   - Checked for data imbalance, only 38.5% of leads were converted.

   - Analyzed categorical and numerical data using univariate and bivariate methods. "Lead Origin," "Current occupation," "Lead Source," and so on offer insightful information about the impact on the target variable.

   - Time spend on website shows positive impact on lead conversion.

3. **Data Preparation:**

   - Dummy features (one-hot encoded) were produced for categorical variables.

   - 70:30 ratio for dividing the train and test sets.

   - Standardization-based Feature Scaling.

- Dropped a few columns since they were closely connected to each another.

4. **Model Building:**

- RFE was used to condense 48 variables down to 15. Data frame will be easier to manage as a result.

- By excluding variables with a p-value greater than 0.05, models were constructed manually using feature reduction.

- Before arriving at the final Model 4, which was stable with (p-values 0.05), a total of 3 models were constructed. With VIF 5, there is no indication of multicollinearity.

- We utilised the final model, logm4, which included 12 variables, to make predictions on both the train and test sets.

5. **Model Evaluation:**

- A confusion matrix was constructed, and based on plots of accuracy, sensitivity, and specificity, a cutoff level of 0.345 was selected. At this threshold, the values of accuracy, specificity, and precision were all close to 80%. However, fewer than 75% of performance values were obtained using the accurate recall perspective.

- To address a business difficulty, the CEO asked for a conversion rate increase to 80%; however, if we took a precision-recall approach, metrics decreased. As a result, the sensitivity-specificity view will be our first choice for the final forecast cut-off.

- The cutoff value of 0.345 was used to award the lead score to the train data.

6. **Making Predictions on Test Data:**

- Making predictions while taking a test: Scaling and forecasting using the final model.

- Evaluation metrics for both the train and test phases are very close to 80%

- The score for the lead was assigned.

- The top three features are:

  - Lead Source_Welingak Website.
  - Lead Source_Reference.
  - Current_occupation_Working Professional.

Recommendations:

1. The Welingak website might use more funding for things like advertising.

2. Discounts or incentives for supplying references that result in leads, which motivates submitting more references.

3. Targeting working professionals aggressively is recommended since they convert well and will have greater financial outcomes.

4. Circumstances to pay larger fees as well.