Solution for Lead Conversion Optimization

Business Context

X Education is an online course provider that collects leads from multiple sources. The leads are accompanied by metadata, which can help the sales team identify and nurture potential conversions. Prioritizing "hot leads" (those with a higher likelihood of conversion) can maximize efficiency by focusing the team’s efforts where they are most impactful.

To achieve this, a logistic regression model can be developed to predict the likelihood of lead conversion. This model assigns a lead score to each lead, allowing the team to categorize them as "hot" or "cold." Below is the detailed solution framework:

Data Analysis

Handling Missing Values

Columns with a high percentage of missing values (>70%) will be dropped as imputing such values could distort the data.
Columns with missing values below 5% will be imputed:
Categorical Variables: Impute with the mode.
Numerical Variables: Impute with the median (preferred over mean due to robustness).
Default placeholder values like "Select" in categorical variables will be treated as missing and addressed accordingly.

Outlier Management

Although outliers are present, removing them would result in a data loss of approximately 9%. Since the model performs well with outliers included, they will not be removed.

Feature Insights

Exploratory analysis identifies critical levels within categorical variables that significantly impact lead conversion.

Data Encoding

Dummy Variables: Used for categorical features with low to moderate unique levels.
Label Encoding: Applied to variables with a large number of levels to avoid a significant increase in dataset size.

Column Elimination

Features with no variance (constant values) will be dropped, as they provide no meaningful information.

Data Preparation

Scaling

Numerical features will be standardized using a standard scaler to ensure uniform scaling and computational efficiency.
Model Building Techniques

Both Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) will be applied to identify the best approach for feature selection.
PCA is particularly useful for dimensionality reduction and resolving multicollinearity, though interpretability might be reduced.

Custom Functions

Model Creation: Builds and evaluates logistic regression models.
Confusion Metrics: Calculates accuracy, sensitivity, and specificity from confusion matrices.
Cutoff Optimization: Determines the optimal cutoff using precision-recall and sensitivity-specificity trade-offs.

Model Selection Criteria

The selected model will balance accuracy, sensitivity, and specificity while being interpretable.

Model Evaluation

Prediction

The final logistic regression model (using RFE-selected features) will assign probabilities to leads, converting these into lead scores: