Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses and captures leads from various sources. The goal is to identify "hot" leads—those with a high likelihood of conversion—so the team can focus on nurturing them, improving conversion ratios, and saving time. Logistic regression is used to assign a lead score based on metadata associated with each lead.

Solution Overview:

The approach begins with thorough data analysis to handle missing values. Columns with over 70% missing values are dropped, while columns with fewer missing values are imputed using the mode for categorical data and the median for quantitative data. Default values like "Select" are treated as missing data.

In the data preparation stage, outliers are retained to ensure all leads are considered, though boxplots indicated 9% data loss if removed. Categorical variables are converted to numerical data: low/moderate-level categories use dummy variables, while high-level categories use label encoding. Columns with no variance are dropped. A quick heatmap showed correlations, and VIF (Variance Inflation Factor) was used to handle multicollinearity.

Model Building involved using RFE (Recursive Feature Elimination) and PCA (Principal Component Analysis) to identify the most important variables. Data was scaled using a standard scaler, and several logistic regression models were tested, with Model 6 proving the most effective based on accuracy, sensitivity, specificity, and ROC/AUC scores. The optimal cutoff for predictions was determined by plotting accuracy, sensitivity, and specificity.

Prediction & Model Selection:

Model 6 was used for final predictions, and lead scores were assigned based on predicted probabilities (Lead Score = Predicted Probability * 100). The RFE-based model outperformed the PCA-based one, offering better accuracy and clearer insights.

Key Insights & Learnings:

- Top features influencing lead conversion include Tags, Lead Quality, and Asymmetrique Profile Index.
- The most significant lead origins were "Landing Page Submission" and "Lead Add Form," with "Olark Chat" being the top lead source.
- EDA (Exploratory Data Analysis) is essential for preparing high-quality data and ensuring the model's accuracy. Imputation, scaling, and handling outliers were critical steps.
- RFE is an efficient technique to select important features, while PCA helped with dimensionality reduction.
- The trade-off between sensitivity and specificity must be balanced when selecting the optimal cutoff for predictions. Confusion matrices provided a clear measure of model performance.

In conclusion, logistic regression was an effective method for scoring leads and improving conversion rates.