Lead Scoring Case Study Summary

Introduction:

X Education, a provider of online courses for industry professionals, seeks assistance in identifying promising leads, those with the highest likelihood of conversion into paying customers. The goal is to develop a lead scoring model that assigns higher scores to leads more likely to convert, aiming for a target lead conversion rate of approximately 80%.

Logistic Regression Workflow:

- Data Preparation
- Exploratory Data Analysis (EDA)
- Dummy Variable Creation
- Train-Test Split
- Feature Scaling
- Correlation Analysis
- Model Building (using Recursive Feature Elimination - RFE)
- Model Evaluation
- Threshold Tuning
- Precision and Recall Evaluation
- Making Predictions on Test Set

Steps: -

Data Loading:

- Data loaded and analyzed importing essential libraries. Data consists of 9,240 rows, 37 columns (7 numeric, 30 non-numeric). No duplicate rows present.

Data Cleaning:

- Data consists of features having several null values. 'Select' values are also converted to null. Columns with null values more than 40% were dropped.
- Imputing missing values, dropping columns with not much information done. Outlier treatment done for numerical columns.

Exploratory Data Analysis (EDA):

- EDA performed to get more insights from dataset. We found several categorical variables were irrelevant as they contain not much information. We dropped them.
- Conversion rate observed as ~38%.
- All other data insights are noted down.

Data Transformation:

- Binary variables are converted into '0' and '1'.

Dummy Variables:

- Dummy variable created for selected categorical columns.

Train-Test split:

- Splitting of dataset done at 70% and 30% for train and test respectively.

Feature Rescaling:

- Min-Max scaling was performed on the original numerical variables.

Model Building:

- Recursive Feature Elimination (RFE) selected the 20 most important features, refined down to 13 using p-values and VIF.
- We took 0.5 as cutoff for conversion probability and assigned prediction accordingly.
- We calculated Accuracy, Sensitivity and Specificity from Confusion Matrix.

Roc Curve:

- Given our model's high ROC AUC score of 0.90, it demonstrates excellent performance and reliability.

Optimal Cutoff Selection:

- We plotted graphs for Accuracy, Sensitivity and Specificity for different thresholds.
- Intersection threshold came as 0.37, selected as new cutoff for optimal performance.

Evaluation:

- Performance on the train data - accuracy of **81.1%**, sensitivity of **80.5%**, and specificity of **81.5%.**
- Performance on the test data - accuracy of **80.9%**, sensitivity of **79.1%**, and specificity of **81.9%.**

Conclusion:

- The model's performance, with an **accuracy of 80.9%**, demonstrates that it is successfully predicting lead conversions in line with the 80% target. Its **Sensitivity of 79.1%** ensures that a large proportion of "hot leads are accurately identified, helping the sales team focus on the most promising prospects. **Specificity of 81.9%** highlights the model's effectiveness in filtering out leads that are less likely to convert, preventing unnecessary effort on low-potential leads.
- Overall, this balance between sensitivity and specificity ensures that X Education can improve its lead conversion rate efficiently, targeting the right prospects while minimizing wasted resources.