Problem Statement

X Education seeks to enhance its lead conversion rate, which currently stands at 30%. The goal is to identify 'Hot Leads'—those with a higher likelihood of converting into customers. This allows the sales team to focus on high-potential leads, improving efficiency and overall conversion rates.

- - - - - - - - - - - - - -

Methodology

1. Data Collection:

The dataset was collected, encompassing both the dependent variable (conversion outcome) and independent variables (features). This dataset is crucial for training and testing the logistic regression model.

2. Data Preprocessing:

Handling Missing Values: Any missing values in the dataset were addressed to ensure data integrity.

Categorical Variables: Categorical variables were converted into dummy variables using one-hot encoding to make them suitable for the logistic regression model.

Normalization/Scaling: Numerical features were normalized or scaled if necessary to ensure uniformity in the data.

Data Splitting: The data was divided into training and testing sets to evaluate model performance effectively.

3. Model Building:

Libraries: Essential libraries such as sklearn were imported for model implementation.

Model Initialization: A logistic regression model was initialized to analyze the relationship between features and the likelihood of conversion.

4. Training the Model:

The logistic regression model was trained on the training data. This phase involved fitting the model to learn the relationships between the independent variables and the conversion outcome.

5. Model Evaluation:

Predictions: The model was used to predict probabilities and classifications for the test data.

Performance Metrics: Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate the model's performance.

- - - - - - - - - - - - - -

Observations

1. ROC Curve of Training Data:

AUC Score: An AUC of 0.86 indicates strong performance in distinguishing between positive and negative classes. Although the model performs better than random chance, there is potential for further enhancement.

2. Accuracy, Specificity, and Sensitivity of Training Data:

Sensitivity: At 78.71%, the model correctly identifies approximately 78.71% of positive cases, indicating good performance in detecting positive outcomes but missing 21.29%.

Specificity: At 78.69%, it effectively identifies negative cases, though some false positives remain.

Accuracy: With an accuracy of 78.69%, the model's overall prediction correctness is solid, but accuracy alone may not fully reflect performance, especially in cases of class imbalance.

3. Precision and Recall of Training Data:

Precision: 69.47% indicates that among positive predictions, 69.47% were correct.

Recall: At 78.71%, the model effectively identifies positive cases, showing a trade-off between precision and recall. Improvements in one metric might impact the other.

4. ROC Curve of Test Data:

AUC Score: The test data also has an AUC of 0.86, reflecting good discrimination ability. While this is positive, further improvements are still possible.

5. Accuracy, Specificity, and Sensitivity of Test Data:

Sensitivity: At 79.27%, the model is effective in identifying positive cases in the test data.

Specificity: With a value of 77.04%, it successfully identifies negative cases but has some room for improvement.

Accuracy: The model's accuracy is 77.92%, indicating overall correctness but also highlighting potential limitations in cases of class imbalance.

6. Precision and Recall of Test Data:

Precision: 50.44% reflects that half of the positive predictions are correct.

Recall: 62.31% indicates that the model identifies 62.31% of actual positive cases, showing a trade-off with precision.

- - - - - - - - - - - - - -

Conclusion

The logistic regression model exhibits strong performance with an AUC of 0.86, demonstrating effective class discrimination. However, precision and recall trade-offs suggest areas for optimization. The model shows consistent accuracy, sensitivity, and specificity, but improvements in feature selection, model tuning, or cross-validation could enhance precision and recall. Ongoing refinement and evaluation are essential for balancing these metrics and achieving more reliable predictions, ultimately improving lead conversion rates and sales efficiency.