Report: Predicting Hospital Readmission Using Logistic Regression

1. Key Preprocessing Steps

Data Cleaning:

The dataset was examined for missing values, which were either imputed or dropped based on relevance.

Duplicate records were identified and removed to avoid data redundancy.

Feature Engineering:

Encoding Categorical Variables: Categorical features like medical_specialty were converted to numerical form using one-hot encoding.

Outlier Detection: Outliers in numeric columns such as time_in_hospital were handled using the IQR method to ensure robust model training.

Normalization:

All numerical variables were standardized using StandardScaler to ensure that the Logistic Regression model treated each feature equally.

2. Model Choice and Rationale

Logistic Regression:

Why Logistic Regression? Logistic Regression is a suitable choice for binary classification tasks like hospital readmission prediction (readmitted vs. not readmitted). It's simple, interpretable, and performs well when the relationship between independent variables and the target is approximately linear.

Advantages:

It outputs probabilities, which can be interpreted to understand the likelihood of readmission.

It is computationally efficient and easy to implement.

Logistic Regression can handle cases where features are correlated, which is common in medical data.

3. Model Performance Metrics

Accuracy:

Train: 61%

Test: 61%

The accuracy indicates that the model is correctly classifying approximately 61% of patients as readmitted or not.

Precision and Recall:

For class 1 (readmitted): Precision = 0.63, Recall = 0.43

For class 2 (not readmitted): Precision = 0.60, Recall = 0.78

The lower recall for class 1 suggests that the model is missing a significant number of readmitted patients, which is critical in healthcare applications.

F1-score:

F1-scores of 0.50 (class 1) and 0.68 (class 2) indicate a balance between precision and recall, though there's room for improvement in predicting readmissions.

Cohen's Kappa:

Train: 0.60, Test: 0.59

These values suggest moderate agreement between predicted and actual classifications.

4. Theoretical Explanation

Logistic Regression models the probability that an instance belongs to a particular class (readmitted or not) using the logistic function:

$P(y=1|X)=11+e\ -(\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn)$

P(y=1|X)= 1+e −(β0+β1X1+β2X2+⋯+βnXn)

Sigmoid Function: The logistic function outputs a probability between 0 and 1. If this probability exceeds a threshold (usually 0.5), the model predicts the patient will be readmitted.

Coefficients (β): During training, the model learns the coefficients that maximize the likelihood of the observed data. These coefficients represent the relationship between each feature and the probability of readmission.

In Python, this is translated into code using libraries like Scikit-learn, which internally optimizes the coefficients using gradient descent. The sigmoid function is implemented to map the linear combination of features into a probability score, and the final prediction is based on a threshold applied to this probability.

5. Suggested Improvements

Feature Engineering:

Interaction Terms: Introducing interaction terms between variables (e.g., between age and time_in_hospital) could capture more complex relationships and improve model performance.

Regularization:

Adding L2 regularization (Ridge) could help address overfitting by penalizing large coefficients. This is particularly useful when the model has many features, as is common with one-hot encoded categorical variables.

Advanced Models:

Exploring more sophisticated models like Random Forests or Gradient Boosting Machines could improve performance by capturing non-linear relationships that Logistic Regression might miss.

Balancing the Classes:

Readmission cases may be less frequent than non-readmission cases, leading to an imbalanced dataset. Addressing this using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting could improve recall for the readmitted class.

This report summarizes the steps taken to develop a predictive model for hospital readmissions and suggests avenues for future improvement.