REPORT

PREDICTING HOSPITAL READMISSION USING LOGISTIC REGRESSION

Introduction

This project aims to predict whether a patient will be readmitted to the hospital within 30 days using historical patient data. The dataset, "Diabetes 130 US hospitals for years 1999-2008" from Kaggle, includes patient demographics, diagnoses, treatments, and medications. Since the target variable, indicating 30-day readmission, is highly imbalanced, data balancing techniques are necessary to improve model performance.

1. Key Preprocessing Steps Taken: The data underwent several preprocessing steps to ensure it was ready for model building:

Dealing with Missing Values: Columns like 'diag_1', 'diag_2', 'diag_3', and 'race' had missing values (represented as '?'). Rows with missing data in these columns were removed.

Feature Engineering:

- The target variable 'readmitted' was turned into a binary class (0 for not readmitted and 1 for readmitted within 30 days).
- Age groups were converted into numeric midpoints to make them more useful for the model.

Dropping Unnecessary Features: Some columns like 'encounter_id', 'patient_nbr', and 'weight' were dropped because they weren't relevant or had too many missing values.

Handling Categorical Data: Columns like 'admission_type_id' and 'discharge_disposition_id' were turned into numerical values using custom encoding for features related to medication.

Outlier Detection and Treatment:

During data visualization, outliers were detected in the dataset. To improve model performance and stability, these outliers were treated using appropriate methods, such as capping or removing extreme values, ensuring the dataset was less skewed.

Data Scaling and Splitting:

The dataset was scaled using StandardScaler to ensure uniform feature ranges, preventing larger scale features from dominating the model. It was then split into training (70%) and testing (30%) sets for evaluation.

Handling Imbalanced Data with SMOTE:

To address the significant class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set. This technique oversamples the minority class by generating synthetic samples, helping the model better identify patterns associated with readmission.

2. Model Choice and Rationale:

The model chosen to predict hospital readmission was Logistic Regression, which was selected for the following reasons:

Easy to Interpret: Logistic regression makes it simple to understand how each feature affects the chances of readmission. The model provides weights for each feature that show their influence.

Efficiency: Logistic regression is fast and works well with large datasets like the one used here.

Binary Classification: Since the goal is to predict either readmitted or not readmitted, logistic regression is well-suited for this type of problem.

3. Performance Metrics of the Model:

The model's performance was evaluated using standard classification metrics:

Accuracy: 65.71%

About 65.71% of the predictions were correct, indicating a reasonable overall performance.

Precision: 17.65%

This low precision indicates that when the model predicts a patient will be readmitted, only about 17.65% of those predictions are correct. This suggests a high number of false positives, meaning the model frequently predicts readmission when it does not occur.

Recall: 56.70%

The recall score indicates that the model correctly identifies about 56.70% of actual readmissions. This shows that while the model misses some readmission cases, it is somewhat effective in detecting those who are likely to be readmitted.

F1-Score: 26.92%

The F1-score, which balances precision and recall, is relatively low at 26.92%. This reflects the trade-off between precision and recall, indicating that it struggles with both false positives and false negatives.

While the model's accuracy is moderate, the low precision indicates it makes quite a few false positive predictions (predicting readmission when it doesn't happen). However, the recall is better, meaning it does well in identifying actual readmission cases, although at the cost of some errors.

4. Theoretical Explanation of the Model:

Logistic Regression is a simple model used for predicting a binary outcome, such as whether a patient will be readmitted or not.

$$p(x) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)}}$$

Where:

- $p(x)$ is the probability of readmission.
- w_0 is the intercept (or bias).
- $w_1, w_2, \ldots, w_n$ are the weights (or coefficients) of the input features $x_1, x_2, \ldots, x_n$.
- e is the exponential function.

The algorithm:

1. Combines the input features linearly.

2. Uses the sigmoid function to map the result to a probability between 0 and 1.

3. Classifies the patient as either readmitted or not readmitted based on a threshold (usually 0.5).

During training, the model uses gradient descent to adjust the weights by minimizing the binary cross-entropy loss. This ensures that the predicted probabilities are as close to the actual outcomes as possible.

5. Suggested Improvements:

Although logistic regression gives decent results, several improvements could enhance the model's performance:

Feature Selection: Some features may be unnecessary or add noise, reducing model performance. Using methods like Recursive Feature Elimination (RFE) can help identify the most important features, improving results.

Regularization (L1/L2): Applying regularization techniques can prevent overfitting by reducing large coefficients, simplifying the model, and helping it generalize better to new data.

Ensemble Models: Using more advanced models like Random Forest or XGBoost can improve performance by capturing more complex relationships in the data that logistic regression, being a linear model, might miss.

Conclusion:

While logistic regression provided an interpretable and efficient model for predicting hospital readmission, addressing the class imbalance and improving feature selection could lead to better results. Ensemble methods and regularization offer potential avenues for enhancing the predictive power of the model.