**Predicting Hospital Readmission: Report**

**DATA SET: Hospital Readmission**

Description: This dataset contains information on hospital readmissions for patients with heart failure, pneumonia, and acute myocardial infarction. It includes demographic information, lab test results, medications, and other clinical factors.

Kaggle Link: <https://www.kaggle.com/datasets/dubradave/hospital-readmissions>

**Preprocess Used in the Dataset**

* **Missing Values**: The dataset was checked for missing values, and it was determined that there were none. This ensured that no imputation or deletion was necessary, preserving the integrity of the dataset.
* **Outlier Treatment**: Outliers were not explicitly treated in this project. This decision was made because the medical records are valid and represent real patient data. Removing outliers could lead to loss of important information about patients' health.
* **Encoding Categorical Variables**: *Label encoding* was used to convert categorical variables into numerical format. This was necessary for the logistic regression model, which requires numerical input for processing.
* **Feature Scaling**: Standardization was applied to the numerical features using Standard Scalar. This scaling ensures that the features contribute equally to the model's performance by transforming them to have a mean of 0 and a standard deviation of 1.

**Model Selection**

The logistic regression model was chosen for this analysis due to its simplicity and effectiveness for binary classification tasks. Logistic regression provides:

* **Interpretability**: The coefficients in the model can be easily interpreted to understand the impact of each feature on the probability of readmission.
* **Efficiency**: It is computationally efficient and can handle a large number of features, making it suitable for this dataset with various patient characteristics.
* **Probabilistic Output**: The model outputs probabilities, allowing us to set a threshold for making predictions, which is important for assessing readmission risk.

**Performance Metrics of the Model**

The model's performance was evaluated using several metrics, which are summarized in the classification report below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **No Readmission** | **Yes Readmission** | **Average** |
| Precision | 0.6 | 0.62 | 0.61 |
| Recall | 0.79 | 0.4 | 0.59 |
| F1-Score | 0.68 | 0.49 | 0.59 |
| Support | 4000 | 3500 | 7500 |
|  |  |  |  |

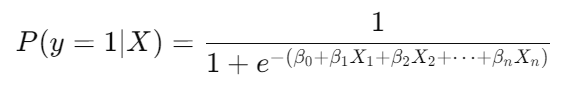
- **Accuracy**: The overall accuracy of the model is 0.61, meaning that 61% of the predictions were correct.

- **Macro Average**: The macro average values indicate an average of the precision, recall, and F1-score across both classes, providing a balanced view of performance.

- **Weighted Average**: The weighted average takes into account the number of instances in each class, providing a more comprehensive performance metric.

**Theoretical Explanation of the Logestic Regression Model**

Logistic regression is based on the logistic function, which models the probability of a binary outcome (in this case, hospital readmission). The logistic regression formula is:



- **P(y=1∣X)**: The probability that the outcome y is 1 (readmitted) given the input features X.

- **β0\beta\_0β0​**: The intercept term.

- **β1​,β2​,…,βn​**: The coefficients for each feature X1,X2,…,XnX\_1, X\_2, \dots, X\_nX1​,X2​,…,Xn​.

The model computes a linear combination of the input features and applies the sigmoid function to transform this linear output into a probability between 0 and 1. The predictions are then made by applying a threshold (commonly 0.5) to these probabilities.

**Future Improvements to the Model**

-**Feature Engineering**: Additional feature engineering could improve the model's performance. For example, creating interaction terms or aggregating features could help capture complex relationships in the data.

-**Advanced Models**: Consider using more complex models such as Random Forest, Gradient Boosting, or Support Vector Machines. These models can handle non-linear relationships and may provide better predictive performance.

-**Hyperparameter Tuning**: Implementing techniques like Grid Search or Random Search for hyperparameter tuning can help optimize model parameters, potentially enhancing accuracy and overall performance.

-**Cross-Validation**: Using k-fold cross-validation will provide a better estimate of the model’s performance by ensuring that the training and testing datasets are representative of the overall data distribution.

- **Handling Imbalanced Classes**: If the readmission rates are imbalanced, consider techniques like SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic examples for the minority class, helping to improve model performance on less represented classes.