Predicting Patient Readmission Using Binary Logistic Regression

AARTHI VIJAYARAGAVAN
Data Analytics (MSCDAD_C_JAN25I)
National College of Ireland
Dublin, Ireland
x23438533@student.ncirl.ie

***Abstract*—Hospital patient readmissions are challenging in healthcare, affecting both patient well-being and hospital expenditures. This study investigates the application of Logistic Regression model in predicting the readmission within 30 days or after 30 days following a diabetes-related hospitalization. Binary logistic regression is used here in predicting the readmission of patients. The dataset, adapted from the research of B. Strack et al.(2014) [1], comprises 101,763 patient records with 47 features. The methodology involves preprocessing of dataset, feature selection, evaluation, and performance analysis using multiple metrics. The results highlight key factors influencing readmission rates and the importance of every metric.**

***Index Terms*—Hospital Readmission, Diabetes, Predictive Modeling, Binary Logistic Regression, Machine Learning, Healthcare Analytics**

I. Introduction

Hospital readmission remains an important concern in healthcare, influencing patient safety, hospital operations, and overall financial burden. In the case of certain conditions such as diabetes, heart problems and frequent readmissions often signal gaps in post-discharge care or inadequate disease management. Identifying high risk patients early reduces the fatal rate.

While various machine learning models have been explored for predicting hospital readmissions, Logistic Regression continues to be a widely used approach due to its simpleness and efficiency. This study utilizes Binary Logistic Regression to assess the likelihood of patient readmission within or after 30 days with the given treatment data to build an effective predictive model.

Among the various types of Logistic Regression models, Binary Logistic Regression is a suitable choice for predicting a binary outcome, such as whether a patient will be readmitted or not. Since the dependent variable in this study has two possible outcomes—readmission or no readmission—Binary Logistic Regression effectively models the relationship between patient characteristics, personal medical history, and treatment details to predict readmission.

II. Literature Review

Predicting hospital readmission has been a widely researched topic in healthcare analytics, with numerous studies investigating the factors influencing patient readmissions. Key determinants include patient details, prior hospital visits, primary diagnoses, and treatment history. Various machine learning techniques, such as Logistic Regression (LR), Support Vector Machines (SVMs) and deep learning models such as CNN, ANN have been applied to enhance prediction accuracy.

B. Strack et al. conducted a comprehensive analysis using a large dataset of diabetic patients to identify crucial possibilities of hospital readmission. Their findings concern that factors such as previous hospitalizations, primary diagnosis, and medication history significantly contribute to readmission risks. Building on this work, more recent studies have incorporated more advanced models in machine learning and deep learning, including ensemble learning methods like random forest and deep neural networks, which have demonstrated improved predictive performance. However, these models often suffer from reduced interpretability, making them not suitable for hospital decision-making.

Despite advancements in machine learning, Logistic Regression remains a preferred method for hospital readmission prediction due to its simplicity, transparency, and efficiency. Comparative studies have shown that deep learning models may achieve higher accuracy. As a result, Logistic Regression continues to be a widely used approach in medical related research, particularly when interpretability and ease of implementation are key considerations. Binary Logistic Regression is a suitable choice for predicting a binary outcome, such as whether a patient will be readmitted or not.

III. Data Description

The Hospital dataset used here consists of 101,763 patient records with 47 features, providing detailed information on patient details, medical history, treatment details, and hospitalization outcomes. The primary goal is to analyze these factors to predict readmission cases.

A. Key Attributes

- **Demographics**: Includes age, gender, and race, which are important factors influencing healthcare outcomes.
- **Medical History**: Past hospital visits, and primary/secondary diagnoses, helping identify chronic conditions or previous complications that may increase readmission risk.

- **Treatment Details**: Includes the number of lab tests, prescribed medications, and insulin use, which can indicate the severity of a patient's condition and response to treatment.
- **Hospitalization and Discharge Information**: Factors such as length of stay, discharge disposition and admission type (emergency or elective) help assess post-discharge risks.
- **Readmission Status (Target Variable)**: A binary variable indicating 1 if a patient readmitted within 30 days or after 30 days, and 0 if they were not.

B. Challenges

- **Missing Values**: Some features had missing data, which may cause bias or reduce model accuracy.
- **Class Imbalance**: The dataset had fewer readmission cases compared to non-readmission cases, leading to potential bias in model predictions.
- **Categorical Variables**: Many features, such as diagnoses and medication use, were categorical and needed transformation for machine learning algorithms.
- **Feature Scaling**: Numerical features varied in scale, which may cause bias or reduce model accuracy.

This dataset adheres to ethical guidelines to maintain patient confidentiality and data security, ensuring fair and unbiased predictive modeling.

IV. Data Exploration and Preprocessing

To ensure model accuracy and precision, several preprocessing steps were performed:

- **Handling Missing Values**: Mode imputation was used for categorical features, while median imputation was applied to numerical features. Features with more than 20% missing values were removed.
- **Encoding categorical variables**: One-hot encoding is useful to nominal variables, and ordinal encoding is useful for ordered categorical variables. Label encoding is a tool used in machine learning models to convert categorical variables (strings) into numerical format as most machine learning models can only operate on numerical data.As shown in Table I, features are encoded for better understanding.
- **Addressing Class Imbalance:** The class weight parameter in Scikit-learn package is a good tool for handling imbalanced data.
  By using `class_weight='balanced'`, one can adjust the weights for each class, which helps to improve the performance.
- **Feature Scaling:** Standardized numerical variables to maintain consistency.
- **Outlier Detection:** Used box plots and Z-score analysis to detect and handle extreme values to reduce their impact without data loss.

Feature	Encoding
race	0: Unknown, 1: African American, 2: Asian, 3: Caucasian, 4: Hispanic, 5: Other
gender	0: Female, 1: Male
age	0: [0-10), 1: [10-20), 2: [20-30), 3: [30-40), 4: [40-50), 5: [50-60), 6: [60-70), 7: [70-80), 8: [80-90), 9: [90-100)
weight	0: Unknown, 1: $>$ 200, 2: [0-25), 3: [100-125), 4: [125-150), 5: [150-175), 6: [175-200), 7: [25-50), 8: [50-75), 9: [75-100)
discharge_disposition_id	0: Home, 1: Other
admission_source_id	0: Other, 1: Emergency, 2: Referral
medical_specialty	0: Other, 1: Cardiology, 2: Emergency/Trauma, 3: Family/GeneralPractice, 4: InternalMedicine
max_glu_serum	0: None, 1: $>$ 200, 2: $>$ 300, 3: Norm
A1Cresult	0: None, 1: $>$ 7, 2: $>$ 8, 3: Norm
metformin	0: No, 1: Down, 2: Steady, 3: Up
repaglinide	0: No, 1: Down, 2: Steady, 3: Up
nateglinide	0: No, 1: Down, 2: Steady, 3: Up
chlorpropamide	0: No, 1: Down, 2: Steady, 3: Up
glimepiride	0: No, 1: Down, 2: Steady, 3: Up
acetohexamide	0: No, 1: Steady
glipizide	0: No, 1: Down, 2: Steady, 3: Up
glyburide	0: No, 1: Down, 2: Steady, 3: Up
tolbutamide	0: No, 1: Steady
pioglitazone	0: No, 1: Down, 2: Steady, 3: Up
rosiglitazone	0: No, 1: Down, 2: Steady, 3: Up
acarbose	0: No, 1: Down, 2: Steady, 3: Up
miglitol	0: No, 1: Down, 2: Steady, 3: Up
troglitazone	0: No, 1: Steady
tolazamide	0: No, 1: Steady, 2: Up
examide	0: No
citoglipton	0: No
insulin	0: No, 1: Down, 2: Steady, 3: Up
glyburide_metformin	0: No, 1: Down, 2: Steady, 3: Up
glipizide_metformin	0: No, 1: Steady
glimepiride_pioglitazone	0: No, 1: Steady
metformin_pioglitazone	0: No, 1: Steady
change	0: No, 1: Yes
diabetesMed	0: No, 1: Yes

TABLE I
CATEGORICAL FEATURES AND LABEL ENCODING

V. Model Selection

Binary Logistic Regression was chosen as the primary model because it is suitable to perform binary classification tasks, like predicting the patient readmission (1) or not (0). This Logistic Regression offers interpretability, efficiency, and ease of implementation, making it ideal for prediction applications.

A. Justification for choosing binary Logistic Regression

While standard Binary Logistic Regression was used in this study, the other types of logistic regression are useful for many applications.

- **Multinomial Logistic Regression**– Used when the target variable has more than two outcomes. This could be

useful if the outcome is multiple cases (No, After30Days, Before30Days).

- **Ordinal Logistic Regression** – Suitable for natural order No<After30Days< Before30Days .
- **Regularized Logistic Regression** – Includes L1 (Lasso) or L2 (Ridge) regularization to stop overfitting. This is useful when dealing with high-dimensional data with plenty of feature's data.

VI. Methodology

A. Data Splitting

The Hospital dataset is divided into two subsets: 80% for training set and 20% for testing set.

1) Feature and Target Selection:

- X consists of all independent variables (features).
- y consists of the dependent target variable (readmitted).

2) Train-Test Split (80-20):

- Ensuring an80% training and 20% testing ratio.
- Using stratify=y to maintain the same proportion of readmission cases in both training and test sets.

B. Feature Scaling

- Standardization with StandardScaler Ensured that numerical features are scaled to zero mean and unit variance, preventing any feature from dominating others and Standardization improves model convergence, especially for gradient-based algorithms like Logistic Regression

C. Model Training - Binary Logistic Regression

- Configuration:max_iter=5000 ensures enough iterations for convergence.
- solver='liblinear' is suited for small-to-medium datasets and handles binary classification well.
- class_weight='balanced' adjusts the imbalance of classes by weight adjusting which is inversely proportional to class frequencies

D. Model Evaluation

The model's effectiveness was evaluated using various metrics to provide better understanding.

- **Accuracy Score**: Calculates overall correct predictions.
- **Classification Report**: Gives precision, recall, and F1-score.
- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Curve predicts the ability of the model to distinguish between readmitted and non-readmitted patients. A higher score indicates a better model.
- **Recall (Sensitivity)**: Evaluates the model's effectiveness in correctly identifying actual readmissions.
- **F1-score**: A precision and recall balance, useful for biased datasets.

VII. Results and Graphical Analysis

A. Evaluation Metrics

The evaluation of the logistic regression model is using key performance metrics

1) Accuracy: Accuracy measures the overall correctly predicted cases out of the total admission cases. The model achieved an accuracy of 62.12%, indicating moderate prediction model. However, accuracy doesn't show the system balance and efficiency.

2) Classification Report:

- **Precision**: Precision measures the correct ratio of identified positive cases out of all cases predicted as positive cases. The model had a precision of 0.63 for non-readmitted cases and 0.60 for readmitted cases, indicating a better performance in predicting non-readmitted patients.
- **Recall (Sensitivity)**: The recall was 0.71 for non-readmitted cases and 0.52 for readmitted cases, suggesting that the model is doing better at predicting non-readmitted patients.
- **F1-score**: Mean of precision and recall is F1 score, was 0.67 for non-readmitted patients and 0.56 for readmitted patients, reflecting moderate overall performance. The weighted F1-score of 0.62 indicates that the system does not predict both cases.

3) ROC-AUC Score: Predicts the model's ability to distinguish between readmitted and non-readmitted patients. The model attained an ROC-AUC score of 0.6601, which is more than half but suggests that the model's discriminatory power is poor. The more the score closer to 1.0, the model's ability to differentiate between classes is better.

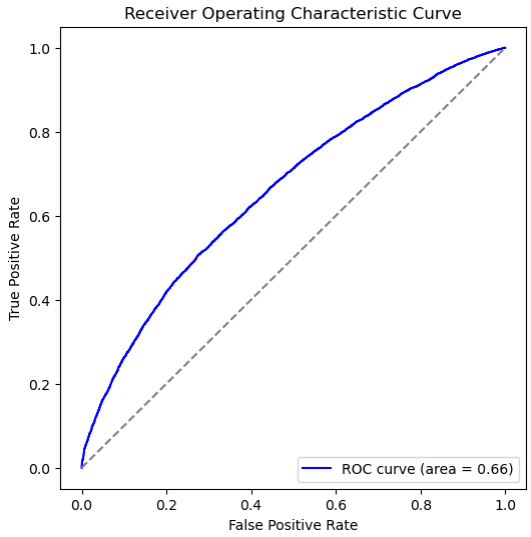

Fig. 1. ROC-AUC Score Curve

4) Confusion Matrix: The matrix gives a detailed analysis of correct and incorrect predictions

- **True Negatives (TN)**: 7,782 cases of correctly predicted no readmission.
- **False Positives (FP)**: 3,190 cases of incorrectly predicted readmission (overestimation).
- **False Negatives (FN)**: 4,519 cases failed to predict readmission (underestimation).

- **True Positives (TP)**: 4,862 cases of correctly predicted readmission.
- This confusion matrix reveals a higher number of false negatives, indicating that the model lacks a significant portion of actual readmissions. But this model predicts an amount of true positives and negatives which indicates the model is bit balanced.

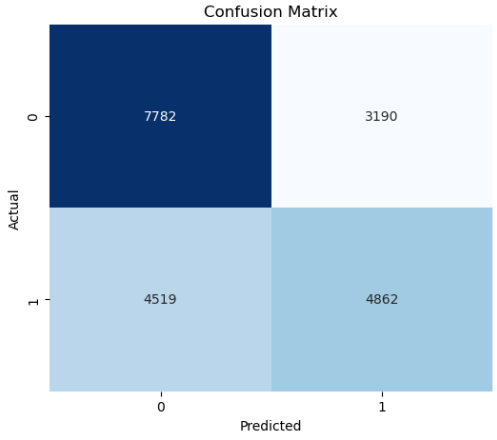

Fig. 2. Confusion Matrix

B. Visualization of Data Results

Several graphs were generated to provide deep analysis into the performance of the model.

1) Distribution of Readmission: The bar chart represents the distribution of readmitted patients, where 0 represents non-readmitted and 1 represents readmitted within 30 days or after 30 days. The dataset has more non readmission than the readmission.

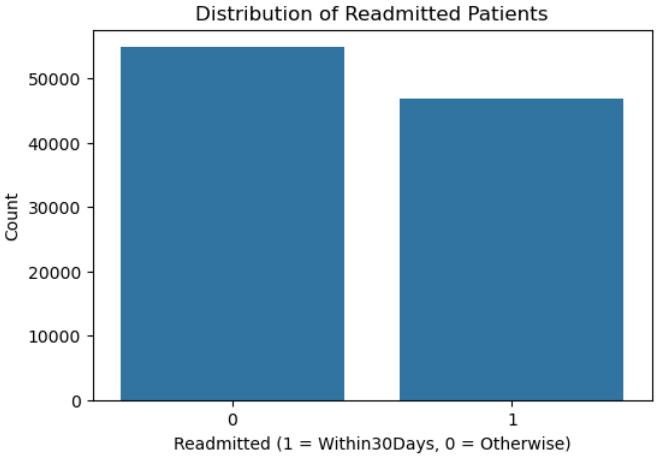

Fig. 3. Distribution of Readmission

2) Feature importance graph: The feature importance graph gives the percentage of important features in influencing the patient readmission.

- Number of emergency visits (10.2%) influences the readmission because it indicates the seriousness of the illness.
- Number of diagnoses (7.8%): Multiple conditions increase the likelihood of readmission. Diabetes medication usage (6.4%): Diabetes medicine intake affects hospital readmission rates.

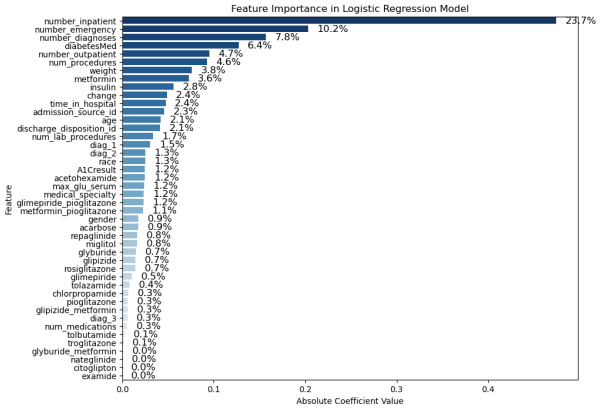

Fig. 4. Feature importance chart

3) Correlation Heatmap: The correlation heatmap visually represents the relationships between variables in the dataset. Key insights include

- Inpatient, emergency, and outpatient visits correlate strongly, indicating frequent hospital interactions increase readmission risk.
- However, A1C and glucose serum levels show weak correlations, suggesting diabetes control plays a role. Race and gender have minimal correlation.

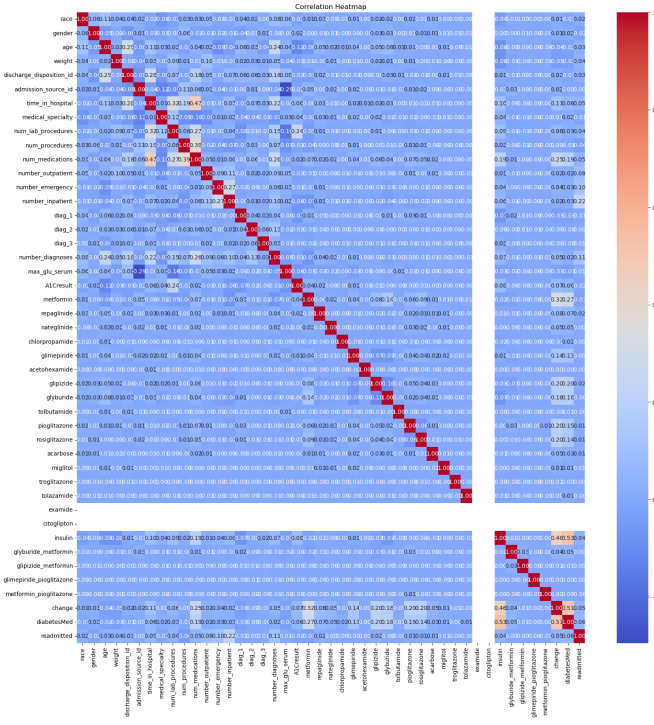

Fig. 5. Correlation Heatmap

4) Comparison of features: The three graphs compare readmission rates across different categorical features. The first graph shows that age and weight influence readmission. The second graph highlights the influence of diabetes-related lab results and medications contribute to higher readmission rates. The third graph examines patient characteristics like race, gender, and medication changes, showing that insulin use and medication adjustments are strong predictors of readmission. Overall, all these graphs suggest the influential factors of patient readmission.

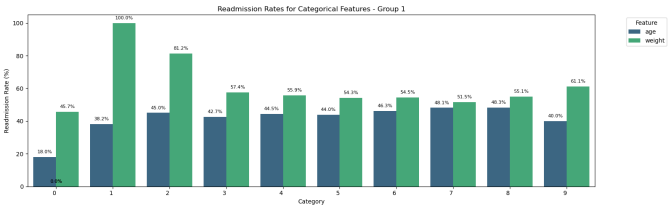

Fig. 6. Group 1

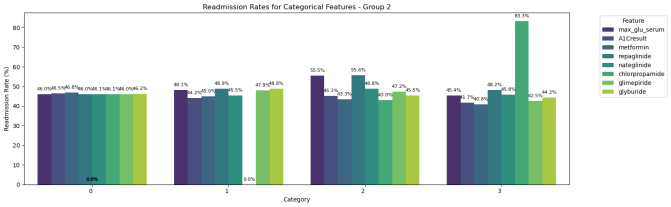

Fig. 7. Group 2

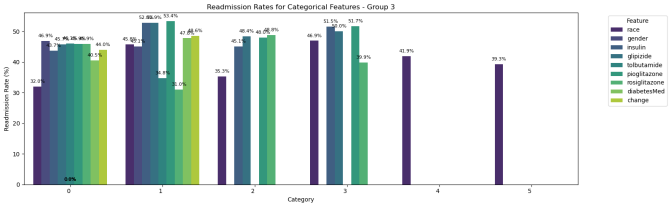

Fig. 8. Group 3

VIII. Conclusion

- This report explored the factors influencing hospital readmission by using Binary Logistic Regression, a model well-suited for binary classification tasks for prediction of readmission of patients.The methodology involved data preprocessing, data analysis (EDA), and model building to predict the readmission.
- Descriptive visualizations, such as bar plots and histograms, were used to understand the distribution of categorical and numerical variables. Readmission rates across various factors were analysed to identify significant influence of key factors.
- A detailed examination of independent variables X which are features, revealed key relationships with the dependent variable Y (readmission). Features such as prior hospital visits, emergency cases, insulin use, and medication for diabetes were found to be significant predictors.
- The model-building process and preprocessing involved feature selection, handling of missing values, and categorical encoders. Several intermediate models were tested, but those models possess weak accuracy and prediction which made the system biased.
- The final model's parameters demonstrated strong balance of system, and its performance was evaluated using various metrics such as accuracy, precision, recall, and AUC-ROC.

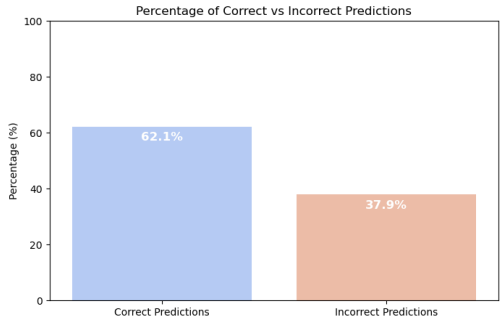

Fig. 9. Prediction

IX. Future Work

- This study demonstrates the effectiveness of Binary Logistic Regression in predicting diabetes-related readmissions of patients.
- The extended research can explore more advanced machine learning models, such as Random Forest, Gradient boosting, SVM, Decision tree or deep learning techniques (ANN, CNN) to improve prediction accuracy of readmission.
- Finally, incorporating real time models integrated into hospital systems could provide detailed analysis about our study.

References

[1] B. Strack et al.,"Impact of Diabetes on Hospital Readmission Rates," *Journal of Biomedical Informatics*, vol. 46, no. 3, pp. 664-672, 2014.

[2] E. W. Lee,"Selecting the Best Prediction Model for Readmission," *Journal of Preventive Medicine & Public Health*, vol. 45, no. 4, pp. 259-266, 2012.

[3] E. Demir,"A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines," *Decision Sciences*, vol. 45, no. 5, pp. 849-880, 2014.