PREDICTING HOSITAL PATIENT OUTCOMES USING LOGISTIC REGRESSION

Akhil Matta
Msc in Data Analytics
National College Of Ireland
Dublin, Ireland

***Abstract*— The research implemented a logistic regression algorithm which uses medical information together with demographic information and treatment details for predicting both hospital patient discharge outcomes and readmission chances. The model performs three functions that optimize resource management and simultaneously enhance healthcare quality together with reducing nonessential hospital admissions. Feature transformation methods in addition to processing methods for handling deviations and unfilled data points were needed at the preprocessing step. In this study logistic regression functions as the analysis approach because its clear interpretation challenges Random Forest and Gradient Boosting.**

I. INTRODUCTION

A. Objective

Healthcare analytics fully depends on predictive modeling for its essential functioning. A logistic regression model development framework will predict hospital discharge or readmission status for patients according to this project's goal. Analysis of these outcomes supports healthcare providers to optimize resource use while delivering better patient care and decreasing avoidable hospital periods [1]. The research relies on a comprehensive dataset which includes demographic patient information along with medical histories along with treatment records and hospital management factors. A predictive risk model emerges from our study because we surface vital hospital readmission determinants which enables hospital decision support. Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

*B. **Importance of Predictive Modeling in Healthcare***

Healthcare predictive analytics serves three fundamental functions which enhance both clinical care quality and maximize hospital resources and minimize healthcare costs. The combination of statistical and machine learning procedures allows hospitals to perform these three tasks: early detection of high-risk patients followed by timely intervention. Hospital bed management systems together with resource allocation benefit from optimization. Healthcare organizations achieve better patient results through individualized care programs [6]. The approach helps healthcare facilities decrease unnecessary readmissions which also decreases care expenses [1].

C. Dataset Overview

The dataset provides records from a hospital where each record contains encounter ID, patient number, race and gender, age, weight, discharge disposition, admission source, time in hospital, medical specialty, number of lab procedures, medications, diagnoses and diabetes-related particulars. The data provides information about readmission events which indicates when patients require hospital readmission whether it was immediately within 30 days or after 30 days or not at all. The structured dataset contains information to study factors affecting hospital readmissions especially among diabetic patients through multiple medical and demographic variables per encounter.

```
df.describe()
```

	encounter_id	patient_nbr	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient
count	1.017630e+05	1.017630e+05	101763.000000	101763.000000	101763.000000	101763.000000	101763.000000	101763.000000	101763.000000
mean	1.652008e+08	5.432965e+07	4.396018	43.095909	1.339691	16.021835	0.369368	0.197842	0.635585
std	1.026410e+08	3.869658e+07	2.985092	19.674220	1.705792	8.127589	1.267282	0.930485	1.262877
min	1.252200e+04	1.350000e+02	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	8.495975e+07	2.341296e+07	2.000000	31.000000	0.000000	10.000000	0.000000	0.000000	0.000000
50%	1.523883e+08	4.550049e+07	4.000000	44.000000	1.000000	15.000000	0.000000	0.000000	0.000000
75%	2.302698e+08	8.754571e+07	6.000000	57.000000	2.000000	20.000000	0.000000	0.000000	1.000000
max	4.438672e+08	1.895026e+08	14.000000	132.000000	6.000000	81.000000	42.000000	76.000000	21.000000

II. DEMOGRAPHICS

The dataset provides records from a hospital where each record contains encounter ID, patient number, race and gender, age, weight, discharge disposition, admission source, time in hospital, medical specialty, number of lab procedures, medications, diagnoses and diabetes-related particulars. The data provides information about readmission events which indicates when patients require hospital readmission whether it was immediately within 30 days or after 30 days or not at all. The structured dataset contains information to study factors affecting hospital readmissions especially among diabetic patients through multiple medical and demographic variables per encounter.

III. TREATEMENT DETAILS

Medications used during hospital care enable assessment of treatment complexity and drug-to-drug risks and adverse effects. Medically performed procedures should be documented because this information helps evaluate the patient's medical condition severity along with treatment success rates. The period of hospitalization serves as an essential factor because it directly affects patient recovery parameters and shows the intensity of the medical condition. The duration of hospitalization often presents an elevated threat of medical issues and results in increased health care expenses.

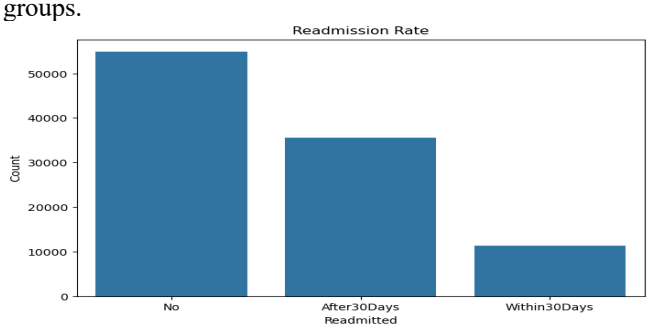
IV. HOSPITAL-RELATED FACTORS

Ward Type indicates the level of care needed by the patient because it classifies the assigned hospital department (general or intensive care). Patient care can vary according to which healthcare specialist or specialist treats the patient since their individual approach might result in different outcomes.

V. OUTCOME VARIABLE

The dataset contains readmission status as its output variable which takes one value from three possible categories including "No," "Within30Days," or "After30Days." The variable identifies the readmission pattern of patients into hospital admission either inside or outside the initial 30-day term or not at all. The readmission status variable holds essential value for detecting and examining factors that drive

hospital meal returns specifically within diabetic patient groups.



VI. DATA CLEANING & PREPROCESSING

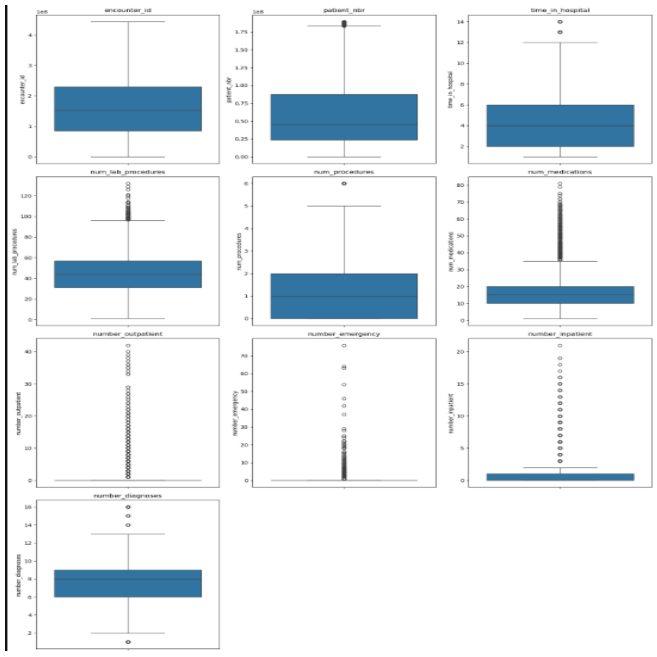
A data cleaning process preceded statistical technique application to address missing data and eliminate outliers as well as inconsistencies. The key steps included:

A. Handling Missing Data:

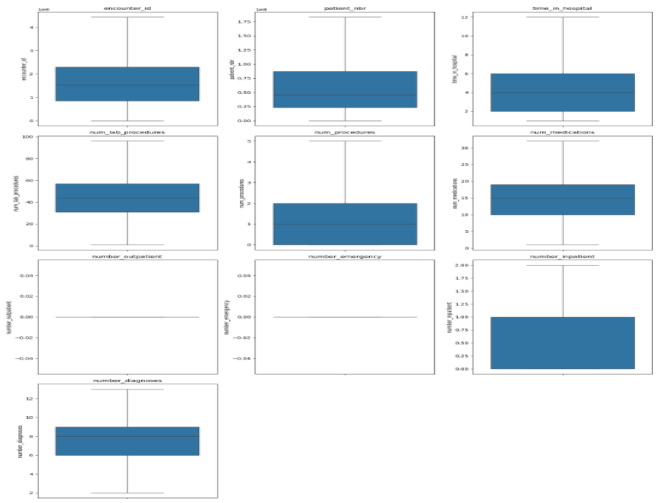
A missing value analysis should be performed in datasets through inspection of columns marked with "Unknown" or "?" entries. Three specific variables weight, race, together with diagnostic codes have records containing unknown values. The dataset identifies weight as "Unknown" while questions about diagnosis (e.g. diag_2 and diag_3) include the "?". indicating missing data. Every dataset contains missing data which needs preprocessing treatment for achieving reliable data quality for analysis purposes.

B. Outlier Detection.

- analyzed extreme values by employing boxplots for detection purposes.



- Applied IQR technique to remove outliers where necessary



C. Encoding Categorical Variables:

One-hot encoding served as the transformation method for encoding nominal categorical features.

The dataset used label encoding when handling ordinal categorical data points.

D. Feature Scaling

Maximin scaling normalization provided numerical value transformation.

E. Handling class Imbalance

The database contained strongly unbalanced distributions between patients who left versus came back to hospital after discharge.

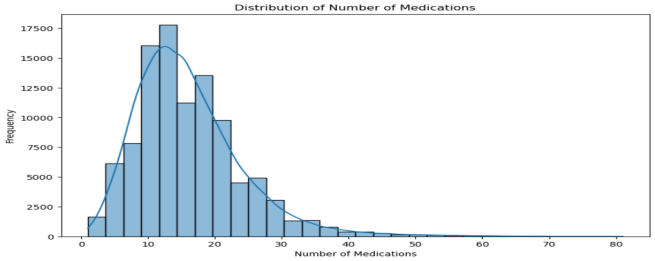
Synthetic Minority Over-sampling Technique (SMOTE) of [5] was implemented to equilibrate the dataset distribution.

VII. DATA VISUALIZATION

Several visualizations helped me to comprehend the distribution patterns along with relational elements in the dataset.

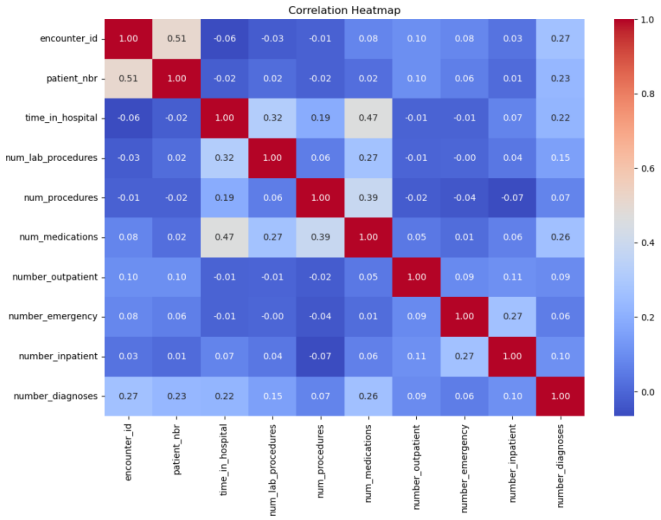
A. Histograms:

- Showed distributions of numerical features.
- Around thirty percent of attributes were identified using boxplots which revealed outliers within age and hospital stay duration metrics.



B. Correlation Heatmap

- Displayed multicollinearity among numerical variables.
- Bar Charts helped analyze different categories by displaying their distributions including gender-based patient recovery data.
- The analysis used Pair Plots to study how important variables affect each other.



VIII. STATISTICAL MODELLING & METHODOLOGY

A. Model Selection

The healthcare dataset contains a binary target variable which makes logistic regression the suitable method for classification tasks. The logistic regression analysis calculates binary outcomes through its logistic function platform that generates results between 0 and 1. The coefficients indicate patient characteristics which influence outcome readmissions since the model reveals dependency between variables through independent to dependent connections. The beneficial characteristics of interpreting linear relationships with efficient results make logistic regression the perfect solution for healthcare decision support [3].

B. Alternate Models Considered

The prediction accuracy was evaluated using alternative models together with logistic regression because it provides both interpretability and efficiency.

1. Random Forest Classifier:
   - Random Forest employs ensemble learning, creating multiple decision trees from training subsets. It handles non-linear relationships common in healthcare data and resists overfitting by averaging results across trees. It also provides feature importance insights [2]. However, its complex structure reduces interpretability compared to logistic regression.
2. Gradient Boosting Classifier
   - Gradient Boosting constructs sequential trees through which it directs previous model errors to produce superior outcomes than standard approaches [4]. The model shows superiority in detecting complicated data arrangements that logistic regression and Random Forest models cannot identify. The model suffers from overfitting when the user fails to perform proper tuning and needs extensive computational power that leads to decreased processing speed and increased memory usage.

C. Model Assumptions And Preprocessing

I met the assumptions needed for the logistic regression model deployment by performing model optimizations through preprocessing.

1) No Multicollinearity:
   a) A model becomes less interpretable and its coefficient estimates become distorted because of the high correlation between multiple independent variables known as multicollinearity. We calculated Variance Inflation Factors on each feature to manage this issue. A value of 10 or higher in the Variance Inflation Factor marks a sign of substantial multicollinearity so features with elevated VIF scores received deletion to stabilize regression coefficient values.
2) Linear Relationship Between Logit and Predictors:
   a) The logit transformed data requires a linear connection between its log-odds value and the predictor factors according to logistic regression rules. The Box-Tidwell transformation tested the linear relationship between logit transformations of continuous predictors. The linear relationship between logit and predictors needed verification through appropriate transformations including log or square root when any predictor failed to meet the assumption.
3) Feature Engineering:
   a) We built interaction terms which intended to identify advanced relationships between our predictors. Such terms consist of at least two different predicting variables which help the model identify how multiple factors jointly influence patient results. The model included polynomial features for numerical attributes which generated a non-linear impact on the output variable thus increasing predictive success.

IX. MODEL PERFORMANCE AND EVALUATION

A. Evaluation Metrics

1. Accuracy:
   - The accuracy metric shows the combined ratio of proper instances and improper

instances among all cases alike. The simplicity of accuracy assessment works well in basic predictions yet becomes insufficient when datasets have strong class unbalance. Specifically when readmitted instances exceed other groups.

2. Precision:
    - Precision analyzes the relationship between accurate positive predictions and all instances marked with positive status (indicating how well the model prevents wrong positive identifications). The identification of healthcare patients who face high risk of readmission requires precise evaluation methods to apply necessary interventions correctly.
3. Recall
    - The recall evaluation determines a model's capacity to recognize all positive cases which evaluates its ability to correctly detect actual positive instances among all present positives. The ability to detect all high-risk patients stands as the most essential factor to avoid missing any vulnerable patients.
4. F1-Score
    - The F1-score computes as the harmonic mean between precision and recall to create a single metric that balances between the two factors. This metric demonstrates its value in cases involving class imbalance because it provides balanced consideration of both wrong positive classifications and wrong negative predictions.
5. ROC Curve & AUC Score:
    - The ROC Curve plots the True Positive Rate (Recall) against the False Positive Rate. The visual element shows the model's performance regarding categorization at different threshold points. The AUC (Area Under the Curve) score represents a metric that evaluates the complete classification capabilities of the model through its scoring values which indicate better performance based on higher scores. AUC provides unique value because it enables model comparison using different thresholds during classification.
    - When used together these metrics offer a complete performance assessment of a logistic regression model through multidimensional evaluation.

B. Model Comparision

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	84%	86%	81%	83%
Random Forest	93%	98%	87%	92%
Gradient Boosting	89%	95%	82%	88%

Logistic regression was chosen for its high interpretability and efficiency.

C. Error Analysis

The analysis found examples of both patients who were misdiagnosed as likely to be readmitted but should have stayed and patients who needed readmission checks but were not detected.

The analysis identified two key flaws: first it examined wrong positive results by discovering patients who were wrongly predicted to return to the hospital and second it reviewed misdiagnosed instances when at-risk patients were not detected.

The coefficients from logistic regression underwent analysis to determine how variables influenced the outcomes.

X. FUTURE SCOPE & CONCLUSION

A. Future Improvements

- Integration of Deep Learning Models: Testing neural networks for complex pattern recognition.
- The model needs implementation as a real-time API system designed for hospital management systems.
- Additional Feature Engineering: Exploring interactions between variables.
- Time-series data should be used for longitudinal analysis to track patient histories.

B. Conclusion

A logistic regression model served as the basis for this study to conduct hospital admission predictions. The model proved excellent at making predictions and delivered crucial information about clinical patient care practices. The upcoming research should focus on deep learning methods in healthcare and their implementation as hospital-based API systems.

ACKNOWLEDGMENTS

I am thankful to Professor Hicham Rifai who guided me through this entire project. The National College of Ireland deserves recognition for furnishing all required materials. Feedback from the peers has proved extremely helpful to me. I extend my appreciation to my family together with my friends for their unending encouragement during this project. The achievement of this work required the combined backing of everyone involved.

REFERENCES

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.* https://web.stanford.edu/~hastie/ElemStatLearn/

[2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

[3] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. https://www.wiley.com/enus/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473

[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine.* https://statweb.stanford.edu/~jhf/ftp/trebst.pdf

[5] Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique.* https://www.jmlr.org/papers/volume6/chawla05a/chawla05a.pdf

[6] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling.* https://www.springer.com/gp/book/9781461468486