Predictive Modeling of Length of Stay among Heart Failure Patients using MIMIC-III Data

Abstract

This study focuses on developing a predictive model for estimating the length of stay (LOS) in Intensive Care Units (ICUs) for heart failure patients, utilizing the MIMIC-III database. The model aims to aid hospitals in optimizing bed management and resource allocation. The MIMIC-III database provided a diverse range of patient data, from which a cohort of 9,441 patients was selected based on specific criteria. Key features for the model were identified using Recursive Feature Elimination. Various machine learning models, including Logistic Regression, XGBoost, Random Forest, and Deep Neural Networks were developed and compared. The Random Forest model showed the highest accuracy in predicting LOS. The study highlights the significance of certain clinical features like potassium and sodium levels in LOS prediction and discusses the potential implications for patient care and hospital management.

Keywords: ICU Length of Stay, Predictive Modeling, MIMIC-III database, Intensive Care Unit Management, SQL BigQuery

1. Introduction

Effective management of Intensive Care Units (ICUs) stands as a crucial aspect of hospital operations, influencing patient outcomes and operational efficiency. With the nature and prevalence of heart failure as a leading cause of hospitalization in ICUs, the need for predictive tools to manage ICU resources becomes imperative. This project intends to address this need by

developing a predictive model focused on forecasting the length of ICU stay for patients admitted with heart failure, utilizing the MIMIC-III (Medical Information Mart for Intensive Care III) database. The model aims to assist hospitals in making well-informed decisions regarding bed management and resource allocation, thereby optimizing operational efficiency and enhancing patient care.

2. Method

2.1. Data Extraction

The MIMIC-III database is a comprehensive collection of de-identified clinical data from over 40,000 patients who were admitted to critical care units (Johnson et al., 2016). Consisting of 26 tables, the dataset encompasses a wide range of patient and clinical information. The cohort for this study was defined using the following criteria: patients with an ICD-9 code beginning with '428', indicating all types of heart failure; those aged 18 years or older; and patients with a Length of Stay (LOS) of at least one day. The inclusion criteria identified a total of 9,441 patients. Only the first ICU stay for each patient was included in the analysis to avoid high correlation among covariates. These criteria ensured that the sample was more representative of a population more commonly affected by heart failure.

The initial stage of data processing involved utilizing SQL BigQuery to compile the intended final dataset, which includes patient demographics (gender, race, age, weight, height), baseline vital sign measurements (heart rate, blood pressure, body temperature, respiratory rate, oxygen saturation, glucose level), baseline laboratory test results (sodium, potassium, chloride), and treatments (epinephrine, dobutamine, dopamine, norepinephrine). A foundational primary table

was created to include three key elements: LOS, patient identifiers, and ICD-9 codes. Age was calculated by the difference between ICU admission time and date of birth (DOB), with any resulting age of 300 years or more redefined as 90 to correct for potential data entry errors. To further enrich this primary table, a series of SQL JOIN operations were performed. Gender and ethnicity information were obtained from *admissions* and *patient* tables. Height, weight, and other vital sign measurements were sourced from *charevents* and *patients* tables. Lab test results, including average, minimum, and maximum values, were extracted and computed from *labevents* table. These covariates were selected for their significance as key indicators of a patient's cardiovascular and overall health status. Additionally, information on drug administration was gathered from the *inputevents.cv* table due to their direct relevance in heart failure management, providing insights into the severity of the patient's conditions. A crucial aspect of the data extraction process was the identification of specific occurrences in the *chartevents* table using *itemid*. Each entry in the table is linked to vital identifiers such as *subject_id, icustay_id* and *charttime*. This method effectively provides a comprehensive and time-specific record of patient data.

2.2. Data Cleaning and Preprocessing

The dataset was then cleaned and preprocessed to ensure its integrity and usability. Given the complexity of representing ethnicity in the original dataset, ethnicity was mapped into broader categories: Black, Asian, Hispanic/Latino, Native American, and Other/Unknown. For covariates with minor missingness (less than 4%), the missing data were replaced by the mean, thereby preserving the overall distribution of the dataset. In cases of high missingness, particularly in treatment drugs data, a shift was made to a binary format focusing on the presence of drug usage

rather than its duration to improve the data's usability. Additionally, LOS was converted from a numerical to a categorical variable, optimizing the model's categorization performance (LOS 1-2 days: '0'; 2-4 days: '1'; 4-10 days: '2'; >10 days: '3'). Only the gender recorded from the first visit was used to account for inconsistencies for the same patient.

2.3. Feature Selection and Predictive Models

Recursive Feature Elimination (RFE) in Python was utilized for feature selection, aiming to identify the most impactful variables for the predictive model. Various machine learning algorithms were employed and assessed for their effectiveness in predicting the length of ICU stay. Initially, Logistic Regression (LR) was used as a baseline model due to its simplicity and interpretability. Then more complex models like XGBoost and Random Forest (RF) were fitted to capture more nuanced patterns in the data. The performance and feature selection of the RF model were further evaluated using SHAP (SHapley Additive exPlanations). Additionally, a Deep Neural Network (DNN) was implemented with Keras. The prediction accuracy was compared across models.

3. Results

3.1. Feature Selection

Our working dataset included 38 explanatory variables, including patients' demographic, vitals, lab results and treatment. The RFE process identified an optimal of 14 features, with the criteria of cross-validation test accuracy. In addition to statistically significant features, we added 6 more clinically significant variables into the features set, incorporating domain knowledge. Therefore, a total of 20 features were used for subsequent modeling, as shown in ***Table 1***.

Table 1. Summary Statistics of 20 Features Used in Modeling

	age	weight_first	heartrate_mean	heartrate_max	sysbp_min	sysbp_mean	tempc_mean	resprate_mean	glucose_min	glucose_mean
count	9431.0	9431.0	9431.0	9431.0	9431.0	9431.0	9431.0	9431.0	9431.0	9431.0
mean	72.0	81.4	85.0	103.5	89.0	117.2	36.8	19.4	107.2	143.5
std	13.7	24.5	15.3	21.1	17.4	16.8	0.6	4.0	35.0	44.1
min	18.0	1.0	32.8	39.0	1.0	29.2	32.0	9.5	0.8	45.0
25%	63.0	65.5	74.4	89.0	79.0	105.3	36.4	16.6	85.0	115.3
50%	74.0	78.5	84.3	102.0	89.0	114.8	36.8	19.0	104.0	134.5
75%	83.0	92.2	94.5	116.0	98.0	126.4	37.1	21.6	124.0	160.3
max	90.0	295.0	147.6	218.0	180.0	194.3	39.9	42.3	592.0	623.0

	spo2_mean	sodium_min	sodium_max	sodium_mean	chloride_min	chloride_max	chloride_mean	potassium_min	potassium_max	potassium_mean
count	9441.0	9441.0	9441.0	9441.0	9441.0	9441.0	9441.0	9441.0	9441.0	9441.0
mean	97.0	135.5	141.8	138.7	100.9	107.8	104.3	3.6	4.9	4.1
std	2.1	5.0	4.9	4.0	6.0	6.1	5.3	0.5	0.9	0.4
min	51.2	3.5	121.0	114.4	66.0	80.0	78.3	0.7	2.5	2.4
25%	96.0	133.0	139.0	136.3	97.0	104.0	101.1	3.2	4.3	3.9
50%	97.3	136.0	141.0	138.7	101.0	108.0	104.4	3.5	4.7	4.1
75%	98.5	139.0	144.0	141.0	105.0	112.0	107.7	3.9	5.3	4.4
max	100.0	161.0	177.0	167.8	133.0	145.0	139.8	6.5	22.9	7.4

3.2. Predictive Models

We built four machine learning models to predict the LOS among heart failure patients in ICU, including LR, XGBoost, RF, and DNN. The RF algorithm achieved the highest test accuracy (60%), followed by DNN and XGBoost with a test accuracy of around 59%. As shown in ***Figure 1***, we observed that the RF classifier is more likely to misclassify nearby classes of LOS, for example, misclassifying class 0 (1-2 days) to class 1 (2-4 days), rather than misclassifying to other classes that are more distant.

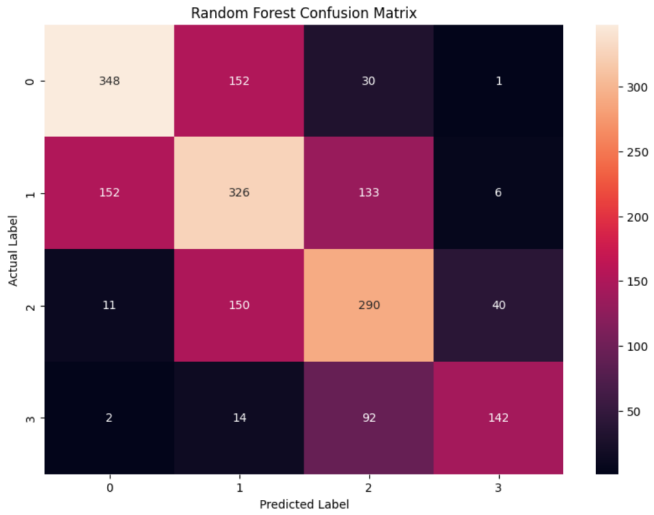

Figure 1. Confusion Matrix of the Random Forest Model with Testing Data

For the RF model, we further assessed the model with SHAP value analysis to dive deeper into the interpretability of the results. The SHAP value graphs are useful in understanding the influence of each feature within the model. As shown in ***Figure 2***, several features significantly contribute to the model's prediction, such as the minimum of potassium level, maximum of sodium level, and minimum of chloride level. Conversely, features like respiration rate, glucose level and weight have relatively low mean SHAP values, suggesting a smaller average impact on the model's predictions. In addition, four SHAP summary plots for each of the four classes were compared to understand model prediction, shown in ***Figure 3***. Several key lab test results, including potassium, chloride and sodium, all have significant impact in predicting LOS in each class. The minimum of systolic blood pressure is an important feature in class 0 (LOS between 1 and 2 days) and class 2 (LOS between 4 and 10 days), but not in other classes. For class 1 (LOS between 2 and 4 days), age contributes to model prediction in this class, but ranks relatively lower in other classes. Our model suggests that minimum and maximum of sodium, potassium and chloride are weighted heavier in model prediction compared to mean values.

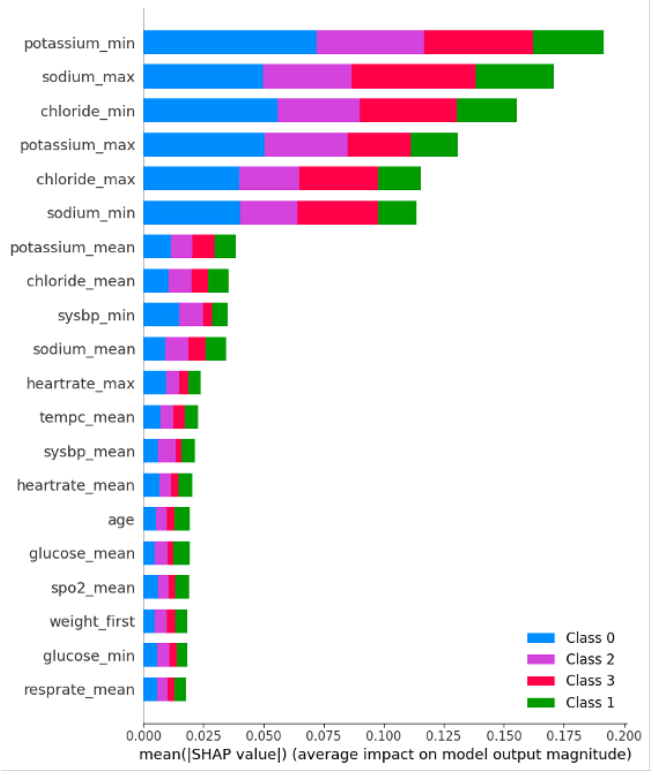

Figure 2. Mean SHAP Value Plot for the Random Forest Model

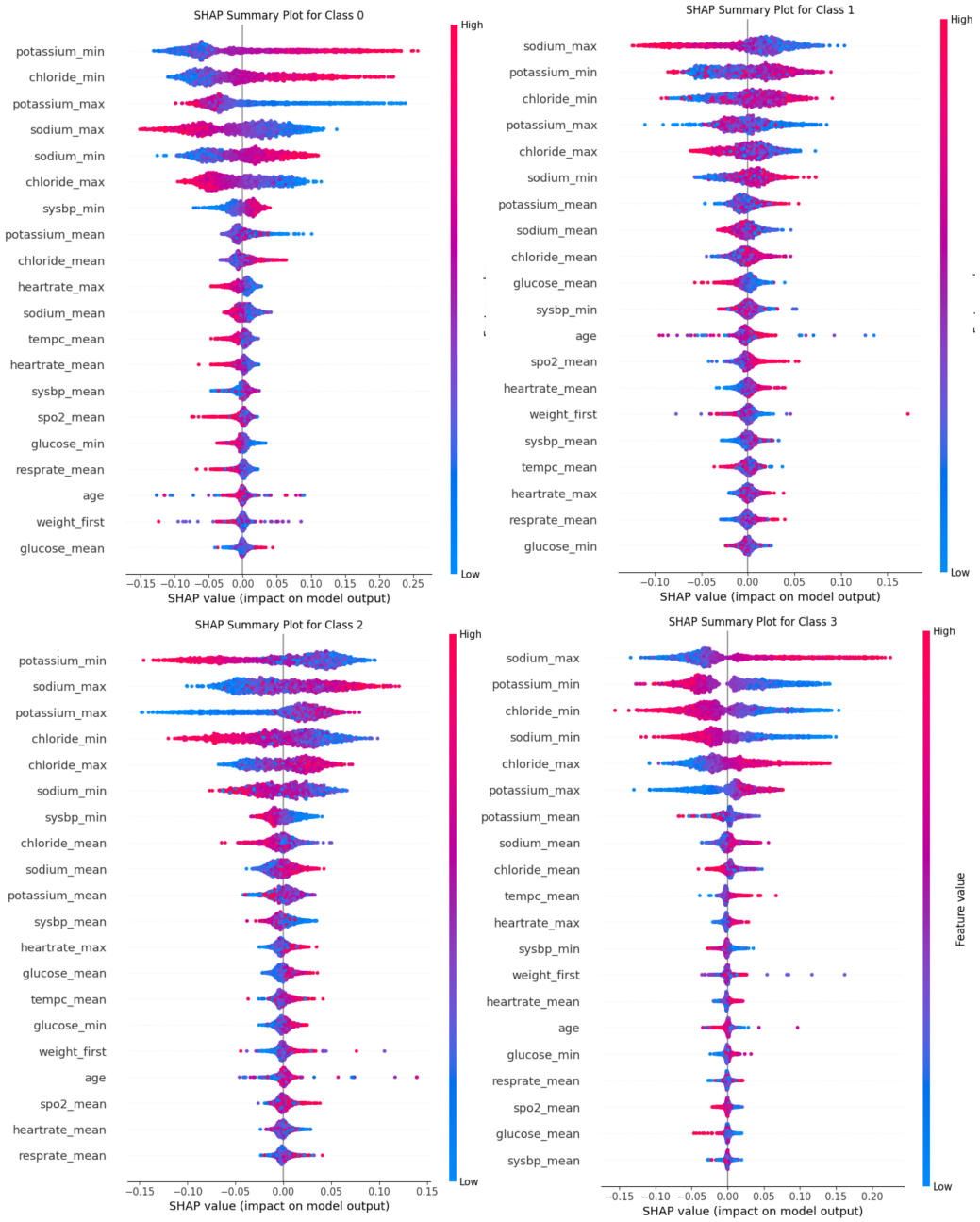

Figure 3. SHAP Value Plots for Each Class of LOS for the Random Forest Model (Class 0: 1-2 days; Class 1: 2-4 days; Class 3: 4-10 days; Class 4: >10 days)

4. Discussion

4.1. Results Discussion

Since heart failure is a prevalent cause of ICU admissions and places considerable burden on hospital resources, the development of predictive models that estimates LOS among heart failure patients holds substantial clinical significance. Accurate forecasts of LOS can provide profound implications for both patient care and hospital management, such as efficient resource planning and allocation of ICU beds and personnel. Our predictive model outcomes underscore the clinical relevance of this project. With minimum potassium level, maximum sodium level, and minimum chloride level being the most important contributors to LOS prediction, these findings offer valuable insights into the factors that significantly influence the recovery trajectory and care duration of heart failure patients.

To contextualize our findings, we reviewed some previous studies in the relevant field. For example, one of the studies also used patients' vitals and labs to predict LOS but used binary classification to classify LOS as either shorter or longer than the median LOS (Alghatani et al., 2021). They achieved the highest accuracy of 65% in their model using RF. With similar prediction accuracy around 60%, our RF model was able to classify LOS into four different classes, thus offering a more detailed breakdown of predicted LOS and improving the model's practical utility in the clinical setting. In another study, the researchers applied similar ML approaches but analyzed ICU mortality and LOS in separate models, identifying differences in the associations among variables between patients leaving ICU due to recovery and those due to mortality. Results from that study could provide additional information on top of ours, since we did not differentiate LOS scenarios between patient recovery and mortality (Iwase et al., 2022).

4.2. Limitations and Future Directions

Admittedly, this project has some limitations that can be further improved in future studies. First, the MIMIC database used in this project only contains data from a single hospital (Beth Israel Deaconess Medical Center in Boston, MA), so the findings might not be fully representative of the diverse patient populations across different regions in the USA. Future research can expand the sources of data and apply the methods to broader patient populations. Second, our model with the highest prediction accuracy relied on 20 clinical features, but using fewer covariates in future studies might be more feasible and generalizable for applying ML models outside the research setting. Moreover, test accuracy was the only evaluation metric we used for assessing model performance. For multi-class classification, other evaluation metrics such as specificity, precision, recall, and F1 score are also commonly used in previous studies (Abd-Elrazek et al., 2021). Future research can include additional evaluation metrics to comprehensively validate model performance.

5. Conclusion

By leveraging medical data from the MIMIC-III database, this project addressed the practical needs in ICU administration and contributed to predictive modeling advancement in healthcare. We used SQL BigQuery and ML algorithms to predict LOS based on heart failure patients' clinical data. With reasonable performance of the RF model and precise outcomes due to multi-class classification, this project has the potential to optimize hospital resource utilization, enable informed decision making, and eventually improve patient care and outcomes. Future studies could benefit from diversifying data sources and patient populations, reducing covariates quantity or increasing model practicality, and applying additional model evaluation metrics.

References

Abd-Elrazek, M. A., Eltahawi, A. A., Abd Elaziz, M. H., & Abd-Elwhab, M. N. (2021). Predicting length of stay in Hospitals Intensive Care Unit using general admission features. Ain Shams Engineering Journal, 12(4), 3691–3702. https://doi.org/10.1016/j.asej.2021.02.018

Alghatani, K., Ammar, N., Rezgui, A., & Shaban-Nejad, A. (2021). Predicting intensive care unit length of stay and mortality using patient vital signs: Machine Learning Model Development and validation. JMIR Medical Informatics, 9(5). https://doi.org/10.2196/21347

Iwase, S., Nakada, T., Shimada, T., Oami, T., Shimazui, T., Takahashi, N., Yamabe, J., Yamao, Y., & Kawakami, E. (2022). Prediction algorithm for ICU mortality and length of stay using machine learning. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-17091-5

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://doi.org/10.13026/C2XW26.