  
 BC2407 - Analytics II: Advanced Predictive Techniques

Semester 2, AY2023/2024

**Project HeartHorizon: Resource and Demand Insights for Cardiac Care Optimization**

Instructor: Chew Chee Hua, Neumann

Seminar Group: S01

Group: 4

|  |  |
| --- | --- |
| Name | Matriculation Number |
| BRYNA LEE ZHEN YING | U2210775K |
| JESSALYN AUDREY TANUJAYADI | U2110440H |
| KUAN LI LIN, THADDEA (GUAN LILIN) | U2222673H |
| PIERRE LOW YI WEN | U2210723E |
| SAMUEL QUEK MING YANG | U2110158J |
| ZHANG XINYANG | U2221842C |

Table of Contents

[Executive Summary 3](#_vjizqcij233m)

[1. Background 4](#_7sv7do2a3q70)

[2. Data Acquisition & Cleaning 6](#_ehd8vtejq3qy)

[3. Data Modelling 13](#_oz5l1q6wcryu)

[4. Business Recommendations 18](#_n5i0cj3bz31q)

[5. Limitations and Suggestions 22](#_qbopd5cpxht8)

[6. Conclusion 23](#_amecs6k27wd3)

[7. References 24](#_ofzpbw3evill)

[8. Appendix 26](#_gsfznmkujchh)

# Executive Summary

With the ever-increasing demand for the healthcare sector in Singapore, hospitals face the challenge of providing sufficient beds, staff and resources to patients, especially for those suffering from critical, time-sensitive diseases such as heart disease. Better demand prediction models are needed for a more optimised allocation of resources, thereby streamlining hospitals’ operations and increasing the likelihood of meeting their demand.

ProjectHeartHorizon aims to enhance the ability of hospitals in forecasting demand through 2 main stages: estimating patients’ length of hospital stay, and predicting future admissions, combining which would give us what we call “Predictive Analytics for Resource & Inpatient Care Optimization” (PARICO). To achieve this, sample data from India was used. Thereafter, data cleaning and exploration were performed to prepare the dataset for modelling.

To predict a patient’s estimated length of stay, three machine learning models – Quantile Regression, MARS, and Random Forest – were used to build a prediction model which takes into account various health-related factors of the patient. The best-performing model, Random Forest, was chosen as our final prediction model for the 1st stage of PARICO.

Time-series forecasting through Prophet was used to forecast the number of patients admitted in the future, which constitutes the 2nd stage of PARICO. This forecasting model takes into account any patterns and trends contained in the admission and discharge dates of the dataset.

Integrating the 2 stages, PARICO offers a comprehensive solution that accounts for both individual patient conditions and the broader inpatient admission demand, thereby enhancing efficiency in bed, staffing, and other critical resource allocation.

Limitations of this project were also elaborated in this report, including the limited, non-current data used, as well as potential differences arising from the non-Singapore-based data. Lastly, this report covers suggested improvements that could be done to improve this project in the future, such as the use of more recent data and regular re-training of models to ensure they are updated with the latest public health trends.

# 

# 1. Background

**1.1 Background and Problem Statement**

The primary challenge confronting hospitals in Singapore lies in their struggle to promptly meet the surging demand for wards and beds. According to David Matchar, a professor in Duke-NUS Medical School’s health services and systems research programme, Singapore’s public hospitals are grappling with an unprecedented influx of patients, so patients have to be discharged before new patients can be admitted, contributing to long waiting times for a bed (Lim & Paulo, 2023). This waiting time not only applies to patients waiting within the hospital, but it also extends to those waiting in the ambulances. In late January 2024, Changi General Hospital faced twelve ambulances awaiting the handover of patients, surpassing their capacity of four ambulance bays (Khalik, 2024). Ambulances experiencing delays at a hospital could affect the survival chances of these patients in the precarious race against time. It also means fewer ambulances can attend to emergency requests (Khalik, 2024).

This is an indicator that public hospitals in Singapore could benefit from having a system that could aid with the daily prediction of new patient intake and the demand for beds in the upcoming few days or weeks. To ensure consistency in standard operating procedures across all public hospitals in Singapore, the system could be implemented by MOH. Such a system will address current issues in optimising resource allocation for staff and beds by diminishing unpredictability and allowing time for ward preparation, allowing hospitals to provide timely care to patients, especially those with heart diseases.

**1.2 Case Justification**

Door to Diuretic (D2D) time represents the time taken for the patient suffering from heart failure to receive diuretics upon their arrival at the hospital. A study has shown that the patient group with a D2D time of less than 60 minutes had a 2.3% in-hospital mortality rate compared to 6.0% for the patient group with a D2D time of more than 60 minutes (Abdin et al., 2021). A mismanagement in resource allocation within the cardiology unit can result in delayed access to medication and treatment for heart disease patients and contribute to the in-hospital mortality rate.

**1.3 Objectives**

Our team has recognized a challenge in resource allocation and aims to address this issue by offering demand forecasts. This proactive approach is intended to afford the hospital staff adequate time to optimise resource distribution effectively.

The solution is structured around two main components:

1. **Prediction: Estimating Length of Stay (LOS) for Current Patients**

Leveraging historical data from patients with similar heart disease cases, this segment focuses on estimating the duration of hospital stays based on a mix of qualitative and quantitative factors. This prediction aids in anticipating bed availability more accurately.

1. **Forecast: Projecting Future Patient Admissions**

Using past admission data, this segment aims to forecast the number of new heart disease patients expected over the next 10 days. By predicting incoming patient volumes, the model facilitates timely adjustments to ward staffing and other resources.

Integrating these components, the final model will provide daily forecasts of bed occupancy and patient count for the forthcoming days by combining incoming patient projections with their expected LOS, alongside current patient duration predictions. This unified approach aims to enhance the hospital department’s ability to efficiently manage bed occupancy and staffing allocations, thereby improving resource utilisation and patient care.

# 2. Data Acquisition & Cleaning

**2.1 Data Acquisition & Interpretation**

For our analysis, we chose to use a dataset from Ludhiana, India. It features information on the Hospital Admissions and Discharge Data from Hero DMC Heart Institute from 2017 to 2019. The dataset comprises 15,757 admissions corresponding to 12,244 patients.

**2.2 Rationale**

Our dependent variable to be predicted will be the length of stay or LOS in short. This information can be derived by counting the days from admission to discharge for each patient. By analysing patient data such as platelets, urea, glucose, total lymphocyte count (TLC), creatinine etc. and seeing their relation to the length of hospitalisation, we aim to create a model to predict future patients’ length of stay based on their data.

**2.3 Preparation of Datasets (Cleaning)**

DURATION\_OF\_STAY will be our dependent variable.

**2.3.1 Fixing the Date columns**

The date of admission (D.O.A) and date of discharge (D.O.D) column contained erroneous dates that did not tally with the LOS column given in the dataset. They also contained rows where the day and month were swapped and used a different date format from other rows.

We used the fact that the dataset was organised in chronological entries to weed out erroneous rows that did not fit into the chronological order. The code for the function, parse\_dates\_admission\_discharge, that we used to clean the dates can be found in the Appendix 1.

Thereafter, we found that there were 73 rows where the duration of intensive unit (ICU) stay was longer than the LOS. The LOS is the time between which a patient gets admitted and discharged from the hospital. It is physically impossible for the duration of ICU stay to be longer than the LOS, and as such we decided to drop these erroneous rows.

Our final cleaned columns were named ‘ADMISSION\_DATE’, ‘DISCHARGE\_DATE’ and ‘DURATION\_OF\_STAY’, which represents the length of hospitalisation of a patient.

**2.3.2 Ensuring proper data type**

Next, we ensured that all the categorical data columns were of “factor” data type and all the continuous data columns were of “integer” or “numeric” data type. Those that did not follow this format were converted to the appropriate data type. This ensures that we will not encounter any errors subsequently in our data modelling stage.

**2.3.3 Handling Missing Values**

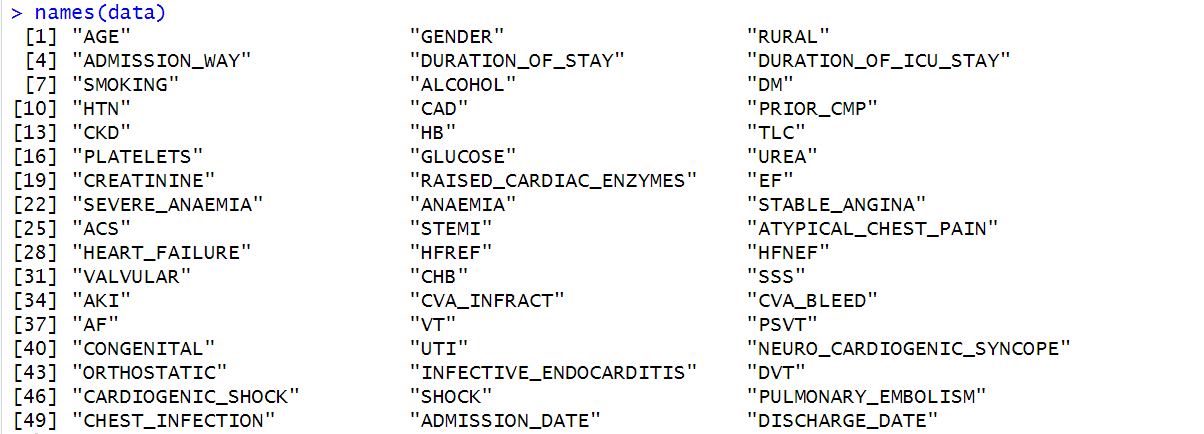
The ‘BNP’ column in the dataset had 9041 missing values, which is a significant number considering the dataset only has 15,757 entries. We thus decided to drop BNP.

For the other missing values in the dataset, we calculated the skewness value of each column and set the skewness threshold to be 1. The missing values in columns with skewness value greater than 1 were replaced with the median, while those in columns with skewness value less than 1 were replaced with the mean.

**2.3.4 Variable Selection**

The column ‘OUTCOME’ in our dataset reflects whether a patient was discharged, expired or discharged against medical advice (DAMA). In a prediction on how long a patient stays in the hospital based on their data at the point of admission, we would not know what happens to the patient on their last day of stay. As such, it would not make sense to include ‘OUTCOME’ in our models, and thus, the column was removed from our dataset.

Our final variables, DURATION\_OF\_STAY included, are as follows:



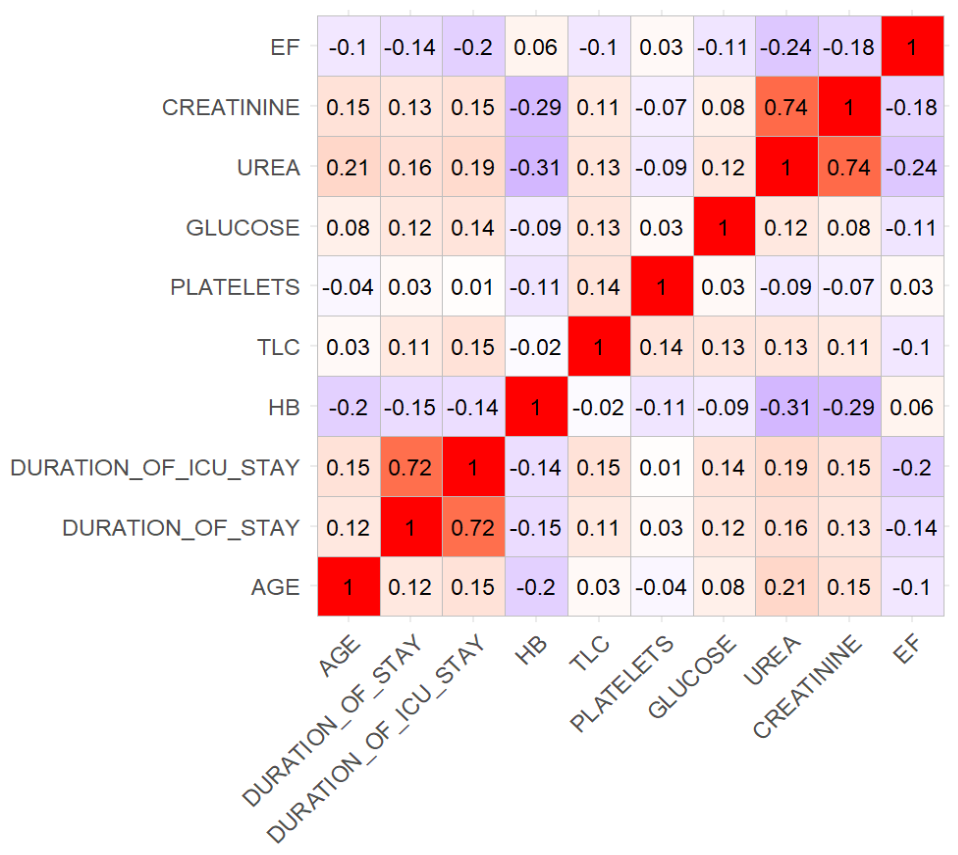
*Figure 1: Final variables after data cleaning*

**2.4 Data Exploration**

Next, we performed some data exploration on the data to examine the correlation of other variables with our key variable, DURATION\_OF\_STAY, the skewness of DURATION\_OF\_STAY and other notable attributes of our data.

**2.4.1 Correlation**

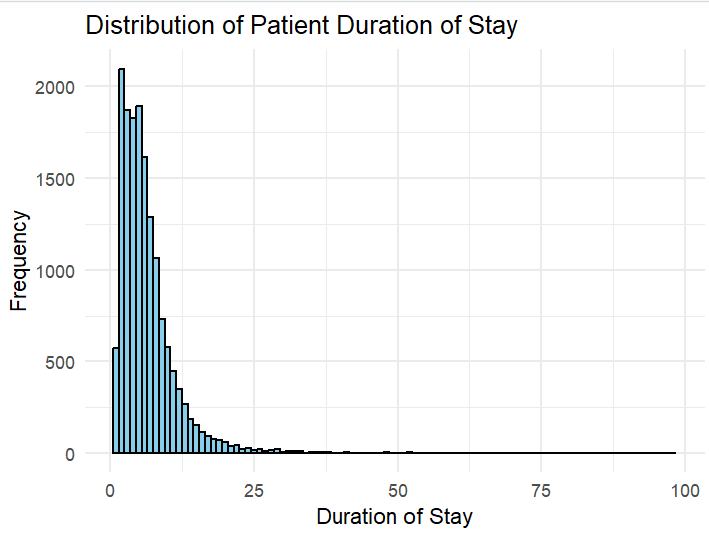
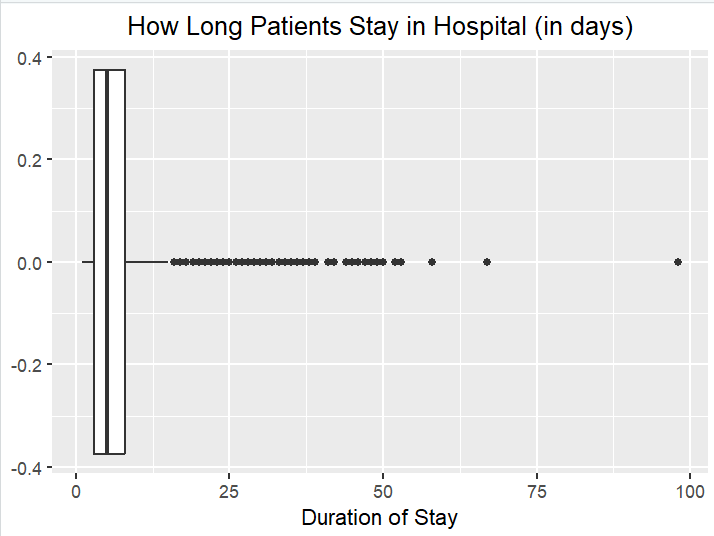
Correlation is a measure of the strength of the linear relationship between two variables (QuestionPro, n.d.). Variables with high correlation to each other may affect the prediction accuracy of our models later on, and as such, should be corrected before we apply machine learning to the dataset. We have set the threshold for correlation to be 0.8, and we can see from the correlation plot below that none of the variables’ correlation is above 0.8, so we do not need to correct anything.



*Figure 2: Correlation Plot between Continuous Variables*

**2.4.2 Data Visualisation**

First, we did a boxplot and histogram plot to see the distribution of data of our key variable, DURATION\_OF\_STAY. We can see that the data is highly positively skewed with some extreme outliers. As such, we will be removing them later on in 2.4.3.

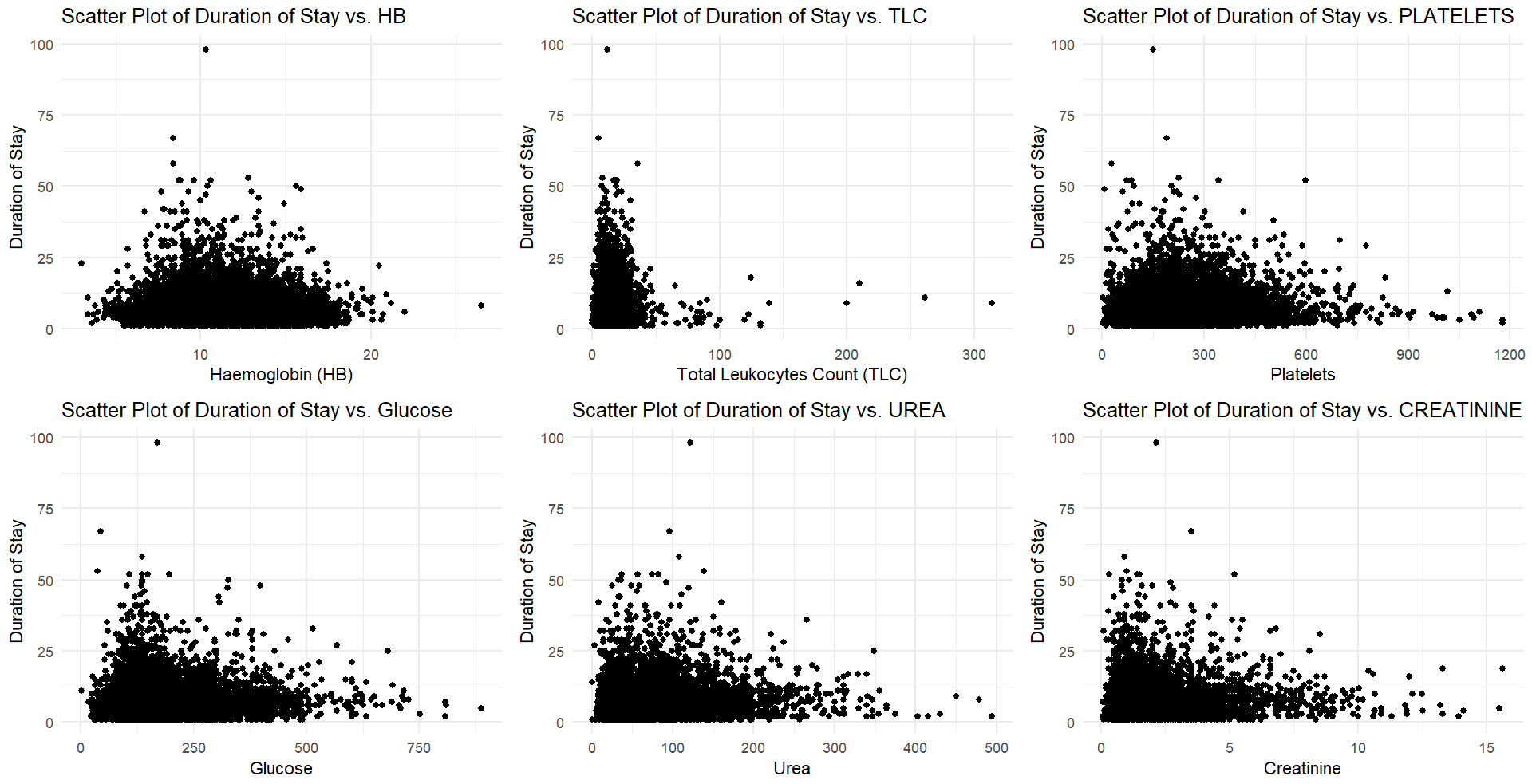


*Figure 3: Distribution of DURATION\_OF\_STAY*

|  |  |
| --- | --- |
| Univariate Data Visualisation | Comments |
|  | The boxplot on the left shows the patients’ age distribution, where we can see the median is around 62 years old, and the ages vary across a wide range. In general, the patients are around 54 to 70 years old. |
|  | The pie chart shows the proportion of Females to Males in the dataset, which is approximately 1:2. |
|  | The pie chart shows the proportion of patients’ area of stay in the dataset, whether rural or urban, which we can see is approximately 1:3. |
|  | The pie chart shows the proportion of patients’ type of admission in the dataset, whether emergency or outpatient, which is approximately 2:1. |

|  |  |
| --- | --- |
| Bivariate Data Visualisation | Comments |
|  | More patients in the dataset are admitted to the hospital through the Emergency services as opposed to the Outpatient route. This can be explained by the hospital specialising in heart issues, which tend to be more critical. In addition, patients admitted via emergency services tend to remain hospitalised for a longer time. |
|  | The scatter plot of DURATION\_OF\_STAY against AGE shows that in general, older people tend to have a longer DURATION\_OF\_STAY at the hospital. This corroborates with the fact that older people tend to have longer recovery times. (Davies, n.d.) |
|  | This plot shows that females and males tend to have around equal lengths of DURATION\_OF\_STAY. |
|  | This plot shows that people staying in urban areas have slightly higher DURATION\_OF\_STAY than those living in rural areas. This could be attributed to lifestyle and diet differences. |

Lastly, we plotted the DURATION\_OF\_STAY against the remaining 6 continuous variables to see the distribution.

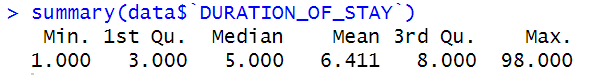


*Figure 4: Relationship between DURATION\_OF\_STAY and the 6 continuous variables*

**2.4.3 Skewness and outliers**

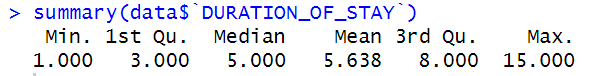
Skewness is the measure of the asymmetry of the data. Highly skewed data may pose problems in our analysis as it may distort the statistical properties of the data, such as the mean, median, standard deviation etc. It may also cause the data to be less representative of the underlying population and more sensitive to outliers and noise (LinkedIn, n.d.). As such, we have created two functions to find and drop outliers outside of the 1.5\*IQR band above and below the upper and lower quartile respectively. These functions will help us to identify and remove data points that are skewed. The exact code can be found in Appendix 2.

These are the statistics of DURATION\_OF\_STAY before removing outliers.



*Figure 5: Statistics of DURATION\_OF\_STAY before removing outliers*

These are the statistics of DURATION\_OF\_STAY after removing outliers.



*Figure 6: Statistics of DURATION\_OF\_STAY after removing outliers*

**2.4.4 Analysis of Historical Trends**

To gain a better idea of the admission numbers, we plotted out several graphs of the number of patients admitted within different periods.

|  |  |
| --- | --- |
| Visualisation | Comments |
|  | From the daily admission numbers, some form of trend can be observed, such as the peak between 2017 and 2018. Overall, there appears to be large swings in patient admissions on a daily basis. |
|  | There is significantly less variance in weekly admission numbers. The peak in early 2018 is much more pronounced. A similar upward trend can be observed towards the end of 2018, but to a much less extent. |
|  | A clear annual trend exists within each year, where patient count peaks at the end of the year. There may also be an overall upward trend between years as numbers of heart disease patients increase every year. However, there are currently insufficient data points to make a definite conclusion. |

Overall, of interest are intra-weekly trends, which can possibly explain the day-to-day variance. Holidays and festive periods may also be relevant and contribute to the noise. Apart from the small-scale variance, large-scale trends (weekly, monthly and yearly) should also be accounted for to derive a comprehensive forecasting model.

# 3. Data Modelling

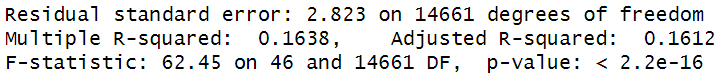
**3.1 Train-Test Split**

A train-test split procedure is used to evaluate the performance of machine learning models when used to make predictions on data not used to train the model (Brownlee, 2020). This is done by dividing the dataset into 2 distinct sets: train set and test set. In this project, we used a 70-30 train-test split ratio for the Quantile Regression and MARS models, as this provides us with sufficient data for model-training purposes, as well as ample data for model-testing purposes, thus minimising the model variance. We did not perform a train-test split when developing the Random Forest model, as random forest has a special feature known as the “Out-Of-Bag” sample, which serves as a natural test set for the data.

**3.2 Predicting Current Patients’ Length of Stay (LOS)**

In building models to predict current patients’ LOS, we excluded the patient’s duration of ICU stay, admission date, and discharge date from the list of independent variables, as these variables were not useful in the prediction of LOS. **3.2.1 Linear Regression**

In order to predict a patient’s LOS from our dataset, one of the simplest methods is to use linear regression, which factors in all the independent variables into a linear equation to predict a dependent variable, in this case the LOS. The results of the linear regression model are shown in Figure 7 below.



*Figure 7: Linear Regression Model*

As shown in Figure 7 above, the adjusted R-squared of the linear regression model (0.1612) is quite low, suggesting that the independent variables used in the model are not able to explain the variability in the dependent variable, making the linear regression model not sufficient for the purposes of predicting the LOS.

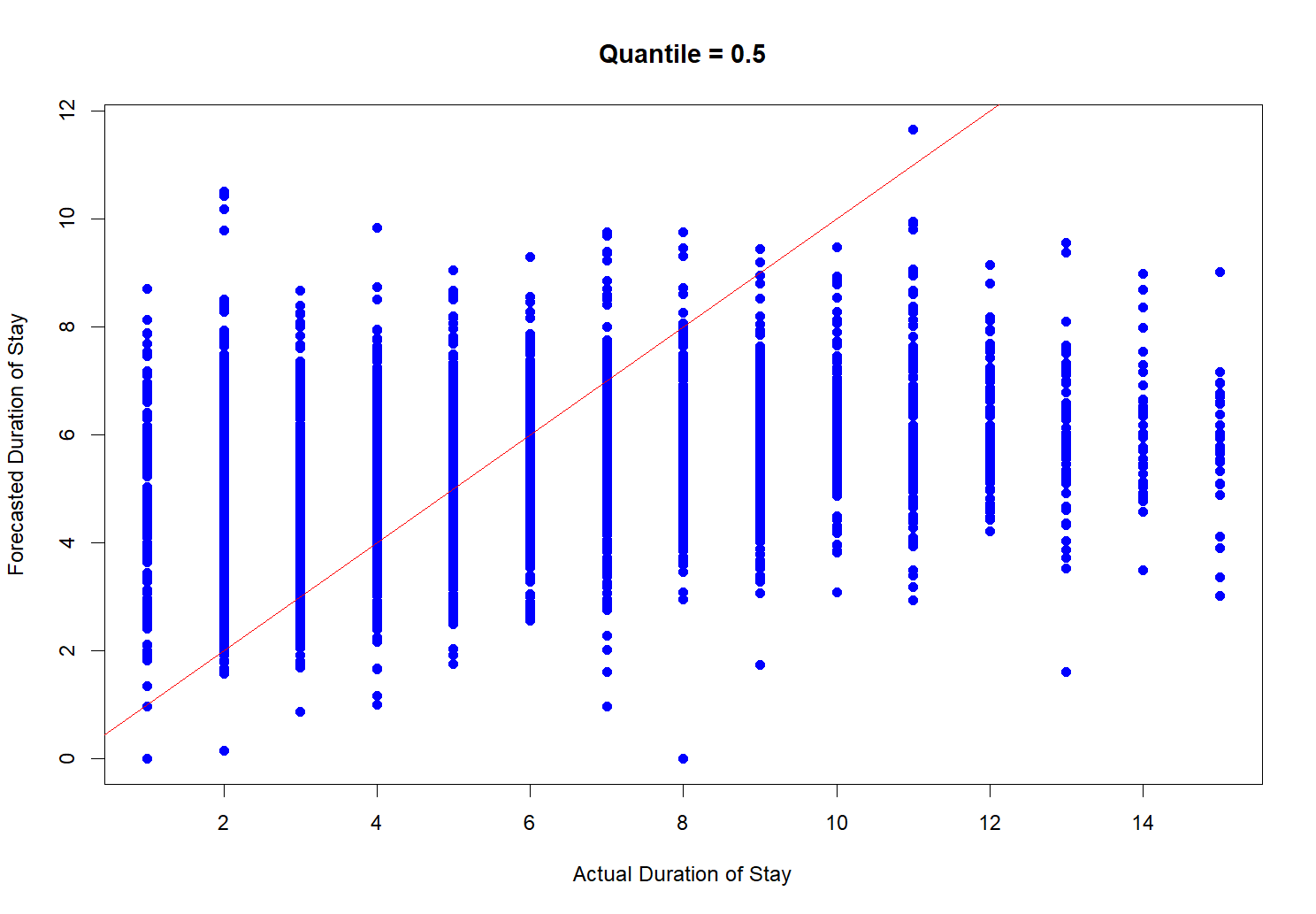
Additionally, the root mean squared error (RMSE) of the linear regression model, which is a measure to estimate the accuracy of a forecasting model, is 2.840641. This RMSE value is quite acceptable, considering the average LOS of patients in the dataset is 5.51 days, but we will try to use other methods which could improve our predictive accuracy.

Thus, in this research we explore 3 alternative machine-learning models below to aid in our prediction of LOS, and we will be comparing their RMSE to identify which model performs best, allowing us to most accurately predict a patient’s LOS.

**3.2.2 Quantile Regression**

Quantile regression is a model used to predict the relationship between a set of independent variables and specific *quantiles* of a dependent variable, which allows more flexibility in modelling data as compared to linear regression. It also doesn’t make assumptions about the distribution of the dependent variable, unlike linear regression which assumes normality and constant variance of residuals.

Due to these benefits, we decided to use quantile regression as one of the models to predict a heart failure patient’'s LOS. For the purpose of this report, we will be using tau = 0.5, or the median quantile, to make our predictions. In relation to our dataset, DURATION\_OF\_STAY is the dependent variable which we will be predicting, and the other variables are the independent variables used to predict the LOS. Upon running the model, we found that the model predicted negative durations of stay for a small number of occurrences, thus we decided to replace those instances with a LOS of 0 days instead, as a negative LOS was logically impossible.



*Figure 8: Quantile Regression Model with Tau = 0.5*

Figure 8 above plots the quantile regression model’s performance when predicting durations of stay from data in the test set. We can see that the majority of the predictions (blue points) were close to the actual values (red line), albeit some predictions being off.

**3.2.3 MARS**

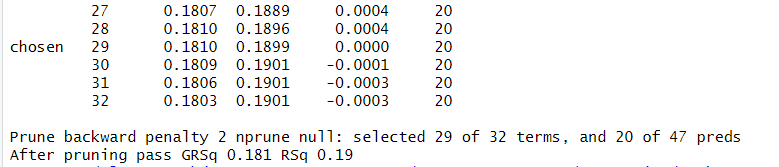
MARS is a modelling technique that addresses some limitations of traditional linear and polynomial regression models, especially in handling non-linear relationships and interactions among variables. Unlike linear regression that assumes a global linear trend, MARS can model complex, non-linear relationships by fitting piecewise linear segments which adapt to changes in the slope of the data. This adaptability is achieved through the use of hinge functions, which are selected and combined through a two-step process involving a forward selection and a backward pruning phase, optimising the model for predictive accuracy while controlling for overfitting.

Due to the ability of the MARS model to adapt to data with varying trends and patterns, resulting in greater flexibility in fitting trends that are local to specific regions of the data, we decided to use MARS to model non-linear relationships between the DURATION\_OF\_STAY and the other independent variables to predict a heart disease patient’s LOS.

Our exploration of the MARS model included comparing models with degrees 1 and 2. The degree of a MARS model dictates the complexity of interactions it can capture: degree 1 allows only additive relationships, while degree 2 enables the model to account for pairwise interactions between variables. Using trace = 3 to view the growing and pruning sequence, we observed the following as shown in Figure 9 and Figure 10:

**Degree 1 Model Analysis**

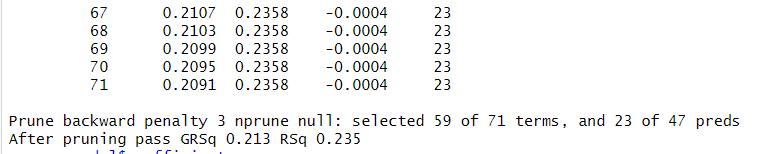
The degree 1 model exhibited a forward pass completion with a final GRSq of 0.181 and RSq of 0.19. The pruning process refined the model to 29 terms out of an initial 32, indicating a focus on the most impactful predictors.



*Figure 9: MARS Model with degree 1*

**Degree 2 Model Analysis**

The degree 2 model, allowing for interactions, completed its forward pass with a higher GRSq of 0.197 and RSq of 0.212. This indicates a better fit and more nuanced understanding of the data complexities. Hence, we went with a degree 2 MARS model, as models with a higher GRSq is generally considered better because it indicates a stronger model performance with a penalty for complexity. We used GRsq as GRSq adjusts for the complexity of the model, penalising the addition of non-significant predictors. This adjustment makes GRSq a more robust measure when comparing models of different complexities, particularly in the context MARS models, where model complexity can vary widely due to the number of hinge functions and interactions included.

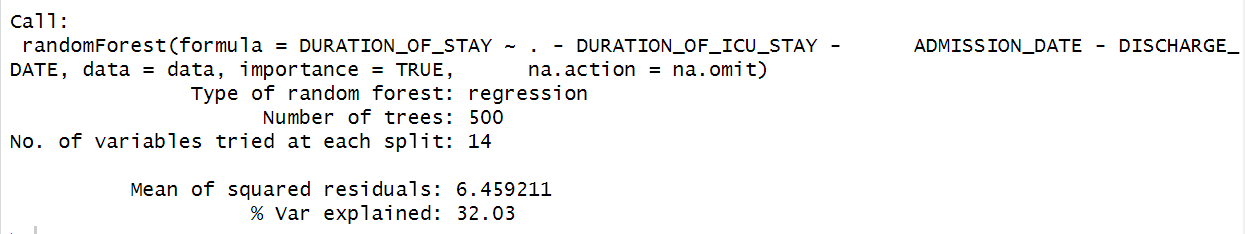


*Figure 10: MARS Model with degree 2*

**3.2.4 Random Forest**

Random Forest is a “machine learning algorithm that creates an ensemble of multiple decision trees to reach a singular, more accurate prediction or result” (Donges, 2024). We choose this model as it offers a reduced risk of overfitting compared to normal decision trees by taking the average of uncorrelated trees to lower the overall variance and prediction error. Random Forest also offers greater accuracy and prediction stability over linear regression (Sahai, 2023). However, it is also a more time-consuming process as compared to the other models since Random Forest algorithms are required to compute the data for each individual decision tree (IBM, n.d.).

Similar to the other models above, we have used DURATION\_OF\_STAY as the dependent variable which we will be predicting, and the other variables are the independent variables used to predict the LOS.



*Figure 11: Random Forest Model*

We can see here that the percent variance explained, which is a measure of how well out-of-bag predictions explain the target variance of the training set (StackExchange, 2023), is not very high. This may mean that our model may not work as well on new data.

**3.2.5 Comparison of Models**

To determine which among the 3 models above perform best in predicting a patient’s LOS, we will be comparing the models’ root mean squared errors (RMSE). RMSE is a widely-used measure to estimate the accuracy of a forecasting model, by calculating the standard deviation of prediction errors from the actual values (Gupta, 2021). Generally, a lower RMSE value indicates that the predictions made are close to the actual values on average, making the model more accurate, and thus more desirable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression (benchmark) | Quantile Regression | MARS | Random Forest |
| RMSE | 2.840641 | 2.885284 | 2.828845 | 2.541498 |

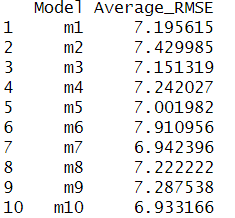
*Figure 12: RMSE of All Models in Predicting LOS*

Figure 12 above summarises the RMSE values for each of the models used in this research. As shown, all the 3 models performed better than linear regression, with the exception of quantile regression. The Random Forest model has the lowest RMSE of 2.54, which means that on average, the Random Forest model’s predicted LOS is 2.54 days away from the actual LOS of a patient. This result is satisfactory, considering the average LOS of patients in the dataset is 5.51 days.

Despite the slightly low percent variance explained of Random Forest, when we compare the RMSE of the 3 predictive models, we can conclude that based on the dataset used in this research, the Random Forest model provides better accuracy than the quantile regression and MARS models, as reflected by its lower RMSE. Therefore, we will proceed to use this model to predict a patient’s length of hospitalisation in order to predict overall hospital beds demand and plan for better resource allocation.

**3.3 Predicting Future Admission Rates using Prophet**

In order to predict the number of patients admitted in the future, time series forecasting will be used. For this section, we will be using Prophet. Prophet is a “modular regression model with interpretable parameters that can be intuitively adjusted by analysts with domain knowledge about the time series” (Taylor & Letham, 2017). Developed by Facebook’s Core Data Science team, Prophet simplifies the process of creating reliable forecasts, even for those with limited statistical or time series forecasting expertise. It is particularly well-suited for forecasting future admission rates in contexts when the data exhibits large swings due to special events or holidays. Furthermore, customised periods can be added as an additional regressor to search for patterns of note.

**3.3.1 Parameter Tuning**

The key parameters that were tested for include ‘daily.seasonality’, ‘weekly.seasonality’, ‘yearly.seasonality’, ‘holidays’ and ‘seasonality.mode’. An additional regressor was also created to test the effects of monthly seasonality.

**3.3.2 Results and Evaluation**

As with all other time-series forecasting models, a generic train-test split such as those used in machine learning models will not work. Cutoffs will have to be used. To evaluate this model, 365 days (from 04/01/2017 to 04/01/2018) were used as base training data. Next, 69 forecasts were made using 69 different cut-off points across 2018-2019, each with a horizon of 10 days. The Root Mean Square Error (RMSE) of each model was then derived by comparing the forecasts against the actual data.  
  
In total, 10 models were created. Based on the RMSE, all models performed somewhat similarly, with minimal difference in RMSE between the models. In particular, model 10, which only included weekly seasonality and used the additive seasonality mode performed the best overall with the lowest RMSE.

# 4. Business Recommendations

**4.1 Current Measures**

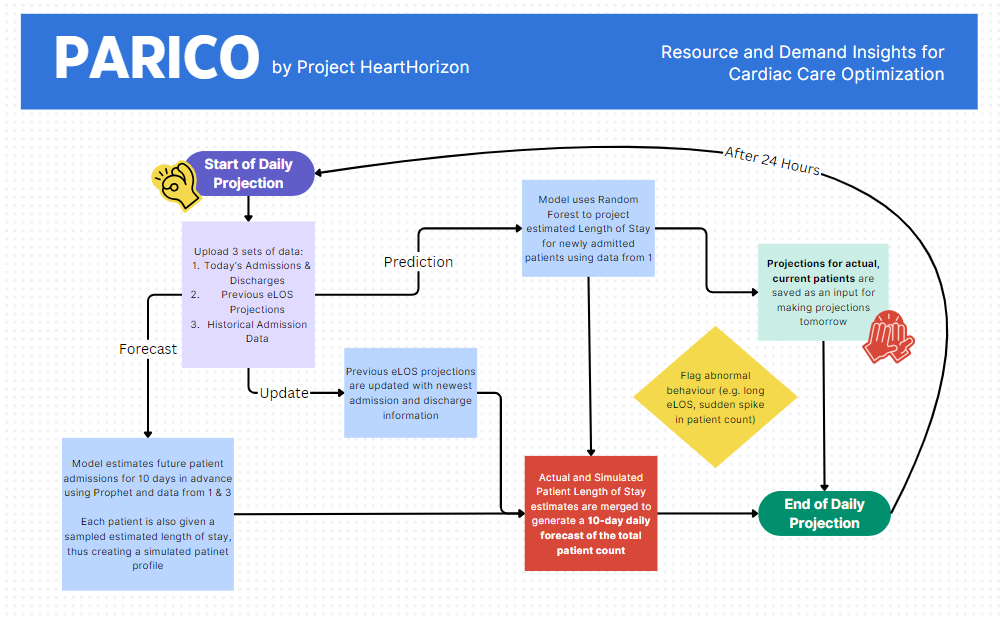
In December 2022, NUHS unveiled the deployment of the ENDEAVOUR AI platform. This platform leverages an AI tool capable of predicting the estimated length of stay (eLOS) for patients admitted to NUHS hospitals, utilising real-time access to patient histories and doctor’s notes to inform its predictions. A notable application of this tool is its ability to flag cases where a patient’s eLOS exceeds two weeks, prompting the medical team to consider management adjustments or early transfers to community hospitals for rehabilitation.

Conversely, a study by Tello et al., published in March 2022, embraced a different strategy for addressing hospital resource allocation issues. Their work focused on employing Machine Learning models to directly forecast inpatient bed demand, forgoing traditional patient-specific focus in favour of a broader view on bed usage and demand predictions.

While NUHS’ method excels in providing a detailed health state overview for individual patients, Tello’s approach shines in its ability to forecast overall bed demand. However, each methodology’s strengths highlight the limitations of the other. NUHS’ ENDEAVOUR AI is unable to predict future admissions and project bed occupancy, while Tello’s models fail to take into account patient status when making their forecasts. To bridge this gap, we propose a novel solution that integrates the detailed predictive insights into individual patient’s eLOS, akin to NUHS’ approach, with a predictive model for future admissions and bed demand, drawing inspiration from Tello’s methodology. This integrated approach aims to offer a comprehensive solution that accounts for both individual patient conditions and the broader implications for hospital resource demand, thereby potentially enhancing efficiency in bed allocation, staffing, and other critical resources.

**4.2 Overview of Solution**

We aim to address the issues in current projection methods by introducing a new platform, Predictive Analytics for Resource & Inpatient Care Optimization (PARICO). PARICO allows hospital employees to rapidly utilise the trained models with most recent patient and admission data, as well as to easily access the predictions made by our platform. For a visual representation of the process flow chart illustrating how hospital personnel will utilise the proposed PARICO platform, refer to Figure 14.



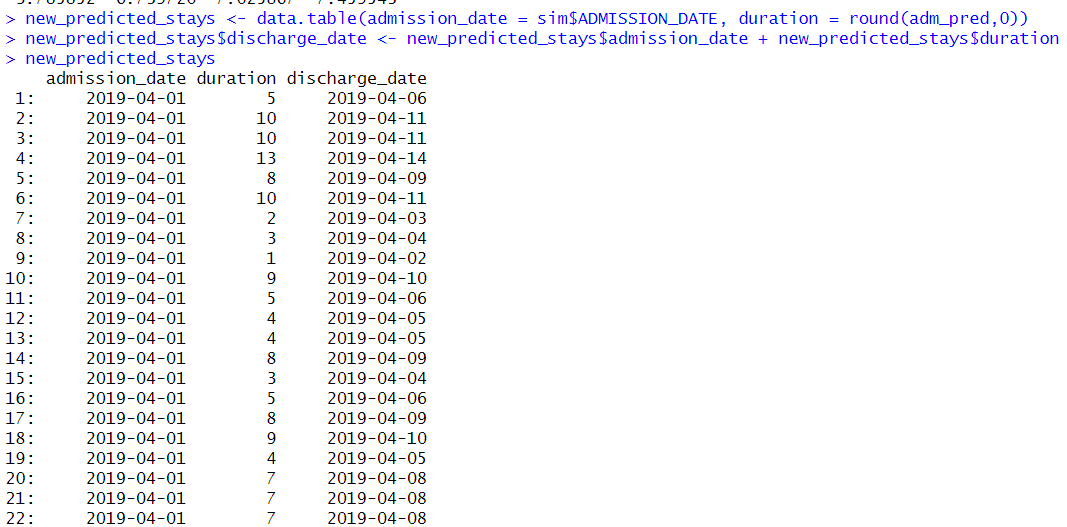
*Figure 14: Flowchart for PARICO platform (For full-sized flowchart refer to Appendix 8.4)*

In the following sections, we will explore the details of our solution and implementation plan. For the purposes of demonstration, let us assume that the current date is 2nd April 2019, at approximately 12am.

**4.3 Prediction: Estimating Length of Stay for Current Patients using PARICO**

Feature 1: Use of Random Forest

We use Random Forest to estimate the length of stay for patients admitted yesterday on 1st April 2019. In the future, predictions will be made for all incoming patients every day.

**

*Figure 15: Estimated LOS for Patients admitted on 1st April 2019*

Feature 2: LOS Projection Record

We retain the projections until the patient is estimated to have been discharged (i.e., estimated length of stay is exceeded). The retention of previous projections reduces computational load as new projections on past admission data do not need to be recalculated.  
  
Note that as it is our first day of using the model, our projections record will need to be initialised based on patients admitted within the past 2 weeks. For subsequent days, this initialisation process will not be necessary, as our current estimates will be carried over.

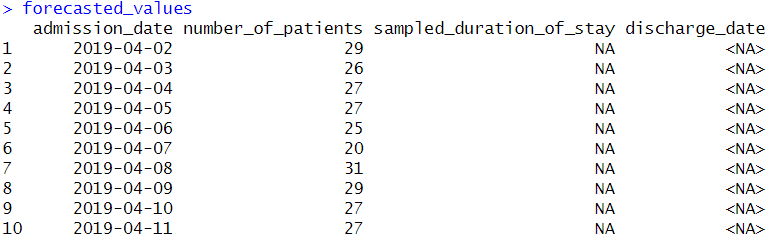
Feature 3: Daily Data Update

To account for errors in predicting, we update the database daily to remove patients who are discharged early. For patients whose current LOS exceeds projections, their estimated LOS will be extended indefinitely for the purposes of projecting patient counts. These cases will also be flagged and brought to the attention of healthcare personnel in order to reassess and potentially revise the care plans for these individuals.

**4.4 Forecast: Projecting Future Patient Admissions**

Feature 1: Use of Prophet to Model

Patient admission count is forecasted 10 days in advance, starting from the upcoming day 2nd April 2019. A new forecast is made every day, hence previous forecasts are not saved.

**

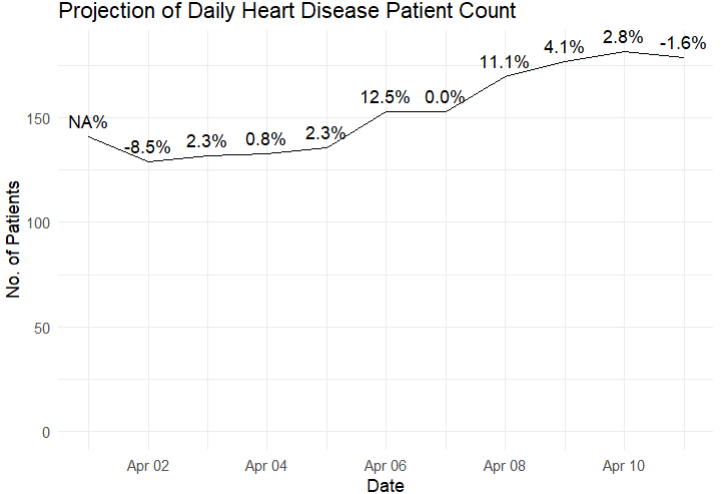
*Figure 16: Forecast of patient admission 10 days in advance*

Feature 2: Sampling LOS

Knowing the number of patients to be admitted is insufficient, as we do not know how long these “future” patients will remain hospitalised. To provide some variance into our projections, we chose to randomly sample from the past records the length of stay of these yet unknown patients instead of using the mean or median LOS. As there are about 30 patients admitted every day, we assume that the actual distribution of their length of stay will be similar to our simulation.

**4.5 Implementation: Integrating Prediction & Forecast**

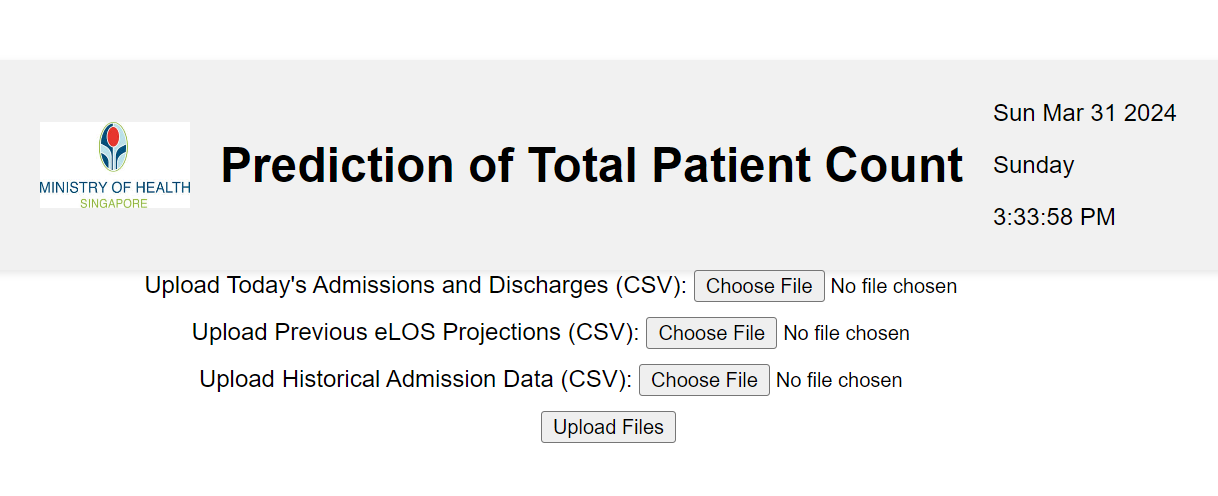
With the estimated Length of Stay of both current and future patients, these predictions are merged together to form our final projection on 2nd April 2019.

**

*Figure 17: Projection of Patient count for the next 10 days*

While the model is more than capable of making predictions further into the future, 10 days was selected as it is short enough to retain a high degree of accuracy in terms of mean RMSE, yet also long enough to contain sufficient information to aid in the decision-making process. Furthermore, this timeframe is selected as it doubles the median duration of patient stays, offering a comprehensive snapshot of expected discharges and admissions that is essential for effective resource allocation and bed management. It ensures a high degree of predictive accuracy, crucial for short-term decision-making, while providing sufficient data to inform actions on staffing, bed allocation, and patient flow management.  
  
**4.6 Application: Real-World Deployment of PARICO**

Based on Figure 17, an initial decrease in patient count over the next 3 days is immediately followed by a sharp rise of patients. When faced with sudden and significant changes to projected patient counts, an alert will be sent to management staff to make the necessary adjustments to resource allocation and staffing schedules.



*Figure 18: Mockup of Interface for Model Integration*

Figure 18 represents a mockup of the interface designed to be used by administrative staff. When the user uploads the necessary CSV files, the model will execute in the backend, generating a graph that displays the projected daily count of patients with heart disease. While the current platform requires manual daily updates, it can be integrated with existing IT ecosystems in hospitals through the following steps.  
  
**4.7 Future Development Roadmap**

Firstly, PARICO will be given access to Singapore’s National Electronic Health Records (NEHR) systems’ application programming interfaces (APIs) to fetch patient data directly. The use of APIs will allow PARICO to access, analyse, and return predictions without disrupting the NEHR's core functions. Furthermore, the volume (< 1 MB) and frequency (daily) of data requested is expected to be minimal and is not likely to overload the system’s server.

Secondly, deploying PARICO on cloud platforms ensures scalability and accessibility, making it easier to update and maintain. Additionally, employing middleware solutions can help bridge the gap between PARICO and various hospital management systems, ensuring data consistency and integrity across platforms.

Finally, training sessions and technical support for hospital staff are crucial to ensure smooth adoption and effective use of PARICO, enhancing its acceptance and integration into daily hospital operations. Dedicated technical staff should also be trained to aid in the maintenance of the platform and to further refine the models used.

**4.8 Value Proposition**

The proposed Predictive Analytics for Resource & Inpatient Care Optimization (PARICO) platform stands at the forefront of enhancing hospital resource management and patient care through advanced predictive analytics. PARICO uniquely combines the detailed insights of individual patient estimated lengths of stay (eLOS) with comprehensive forecasts of future patient admissions, filling a critical gap in current healthcare resource planning methodologies. This dual approach not only optimises bed allocation and staffing requirements, but also facilitates proactive patient care planning.

# 5. Limitations and Suggestions

While our chosen models can predict the foreseeable influx of patients in public hospitals and the estimated LOS in the hospital, nonetheless there are some areas for improvement,

**5.1 Obsolescence: Time Frame**

The model utilises data collected from 2017 up till 2019. To ensure its relevance and accuracy, it is crucial to retrain the model with updated patient data from the hospital, accounting for the intervening five years between the original data collection and the present day.

**5.2 Irrelevance: Singapore Context**

As our team was unable to retrieve patient data from Singapore’s hospitals, we built all of our models on data collected from India. As such, the profile of their population of heart disease patients may differ from those of Singapore, and thus affect the importance of different variables, and therefore the applicability of the current trained model to Singapore. In addition, India may have different seasonality trends from Singapore. These combined factors can result in a different distribution for the LOS and trends in the number of patients expected for the next day.

**5.3 Inaccuracy: Limited data points**

The chosen dataset does not include daily updates on the changes in patient’s health. As such, our prediction model is purely based on the initial state of the patients when they were first admitted to the hospital and is not dynamic. Hence, users of our model may accidentally overlook significant events or developments that could affect the accuracy of predictions regarding the patient’s LOS in the hospital.

**5.4 Possible improvements and suggestions**

To resolve the issue of obsolescence as well as irrelevance, recent data (up to 5 years ago) from Singaporean hospitals should be used to retrain our models before they can be used in local hospitals.

To further improve the accuracy of our model, the random forest model should also undergo regular retraining weekly to account for recent trends. The Cardiology department could then rely on the most updated information to conduct resource planning.

Additionally, newer, different models can be trained for other departments within the hospital due to the similar problems of resource allocation faced by those departments. This broader application will enhance the potential to optimise resource allocation and regulation of bed occupancy across various areas, benefiting a wider range of patients within the hospital.

These improvements can be made during the implementation of the models by MOH, as the organisation and public hospitals in Singapore would have access to a bulk of the patient data within the hospitals.

Given that public hospitals in Singapore possess access to comprehensive daily patient data, including detailed records of ICU stays, our model could be expanded to predict ICU stay durations. This extension is particularly significant as ICU stays demand substantial resources, as highlighted in the literature (Katz & Becker, 2019).

# 6. Conclusion

All in all, our project aims to help MOH, and therefore all the hospitals in Singapore, be able to predict future patient demand better and as such make the necessary resource and staffing changes to their equipment and roster. Through the development of predictive models and the introduction of the innovative Predictive Analytics for Resource & Inpatient Care Optimization (PARICO) platform, we are confident that hospitals will be better able to meet patient needs on busy days and cut inefficient resources on slower ones. PARICO’s easy to use interface allows for widespread adoption of PARICO with minimal staffing training needed to operate. Should our project be extended to other departments apart from cardiology, retraining of data with newer models to predict LOS would produce more optimal results. Despite the presence of some limitations, we believe that these are minor hurdles that can be solved by implementing our proposed improvements. Overall, this project offers a significant leap forward in managing hospital resources, ensuring that the cardiac healthcare system remains responsive, resilient, and ready to deliver high-quality care to every patient who needs it.

# 

# 7. References

Abdin, A., Anker, S. D., Butler, J., Coats, A. J. S., Kindermann, I., Lainscak, M., Lund, L. H., Metra, M., Mullens, W., Rosano, G., Slawik, J., Wintrich, J., & Böhm, M. (2021, October 16). “time is prognosis” in heart failure: Time-to-treatment initiation as a modifiable risk factor. ESC heart failure. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8712849/#:~:text=HF%20is%20an%20urgent%20clinical,severity%20or%20duration%20of%20HF>.

Brownlee, J. (2020, August 26). *Train-Test Split for Evaluating Machine Learning Algorithms*. Machine Learning Mastery. <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>

Davies, E. (n.d.). *Why do you take longer to heal as you age?* BBC Science Focus. <https://www.sciencefocus.com/the-human-body/why-do-you-take-longer-to-heal-as-you-age>

Donges, N. (2024, Mar 8). *Random Forest: A Complete Guide for Machine Learning*. Built In. <https://builtin.com/data-science/random-forest-algorithm>

Gupta, S. (2021, April 26). *RMSE: What does it mean?*. Medium. <https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e>

IBM. (n.d.). *What is random forest?*. <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems>

Katz, J. N., & Becker, R. C. (2018, November 7). Evolution of the coronary care unit: Past, present, and future. Cardiac Intensive Care (Third Edition). <https://www.sciencedirect.com/science/article/abs/pii/B9780323529938000011>

Khalik, S. (2024, February 25). *Ambulances delayed as emergency departments struggle under high patient load*. The Straits Times. <https://www.straitstimes.com/singapore/ambulances-delayed-as-emergency-departments-struggle-under-high-patient-load>

LinkedIn. (n.d.). *All Engineering Data Science How can you handle data skewness in preprocessing?*. <https://www.linkedin.com/advice/0/how-can-you-handle-data-skewness-preprocessing-skills-data-science-yyrnf>

Ministry of Health. (2023, August 3). *Principal causes of death*. <https://www.moh.gov.sg/resources-statistics/singapore-health-facts/principal-causes-of-death>

NUHS. (2022, December). *NUHS deploys the ENDEAVOUR AI platform which hosts AI tools that predict hospital stay durations and waiting times.* National University Health System. Retrieved March 3, 2024, from <https://www.nuhs.edu.sg/sites/nuhs/NUHS%20Assets/News%20Documents/NUHS%20Corp/Media%20Releases/2022/Media-Release-NUHS-deploys-ENDEAVOUR-AI-platform.pdf>.

Paulo, D. A., & Lim, K. (2023, November 26). *Singapore’s Public Hospital Bed Crunch: Are Radical Solutions needed?* CNA. <https://www.channelnewsasia.com/cna-insider/singapore-public-hospital-bed-crunch-waiting-time-radical-solutions-3943586>

QuestionPro. (n.d.). *Using correlation analysis to identify linear relationships between two variables*. <https://www.questionpro.com/features/correlation-analysis.html#:~:text=Correlation%20analysis%20in%20research%20is,the%20change%20in%20the%20other>

Sahai, N. (2023, September 6). *Random Forest Regression – How it Helps in Predictive Analytics?*. AnalytixLabs. <https://www.analytixlabs.co.in/blog/random-forest-regression/>

StackExchange. (2023, January 17). *Interpreting percent variance explained in Random Forest output*. <https://stats.stackexchange.com/questions/161709/interpreting-percent-variance-explained-in-random-forest-output>

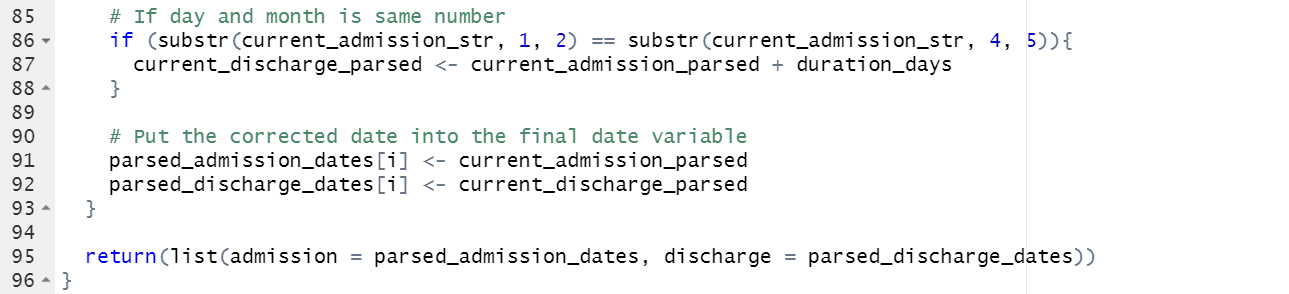
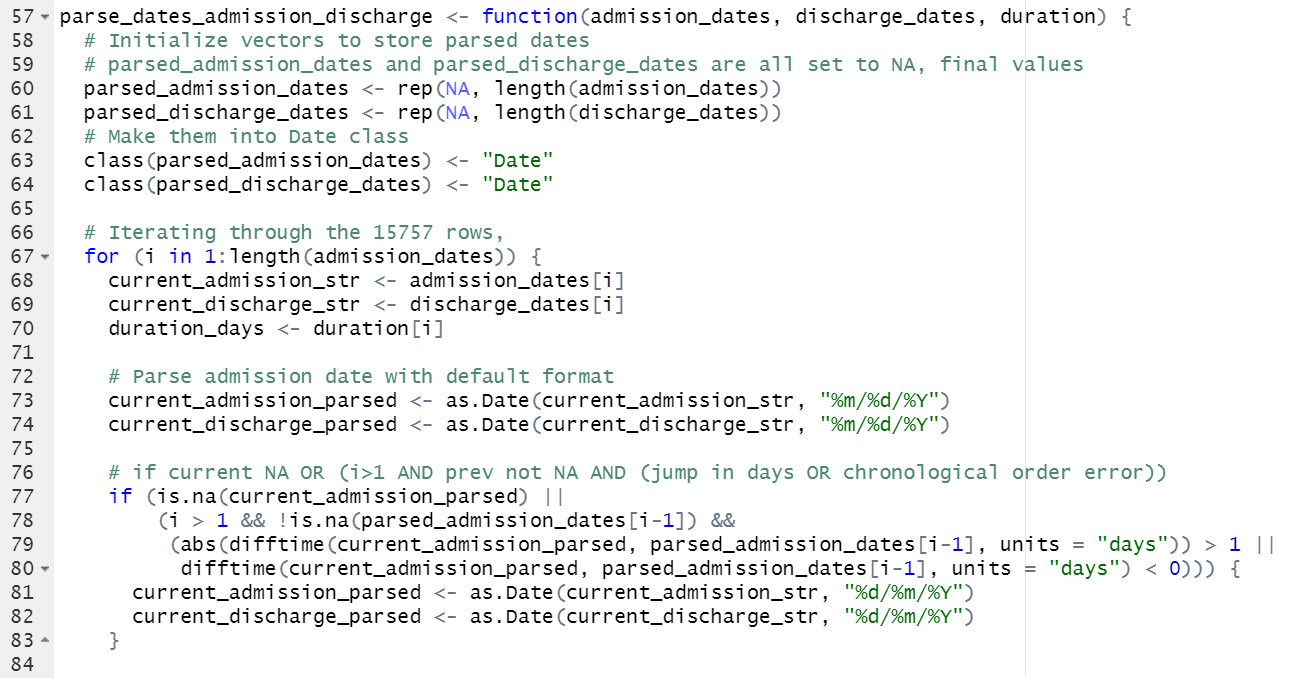
Taylor, S. J., & Letham, B. (2017). Forecasting at scale. *PeerJ Preprints*. <https://doi.org/10.7287/peerj.preprints.3190v1>

Tello, M., Reich, E. S., Puckey, J., Maff, R., Garcia-Arce, A., Bhattacharya, B. S., & Feijoo, F. (2022). Machine learning based forecast for the prediction of inpatient bed demand. *BMC Medical Informatics and Decision Making*, *22*(1). <https://doi.org/10.1186/s12911-022-01787-9>

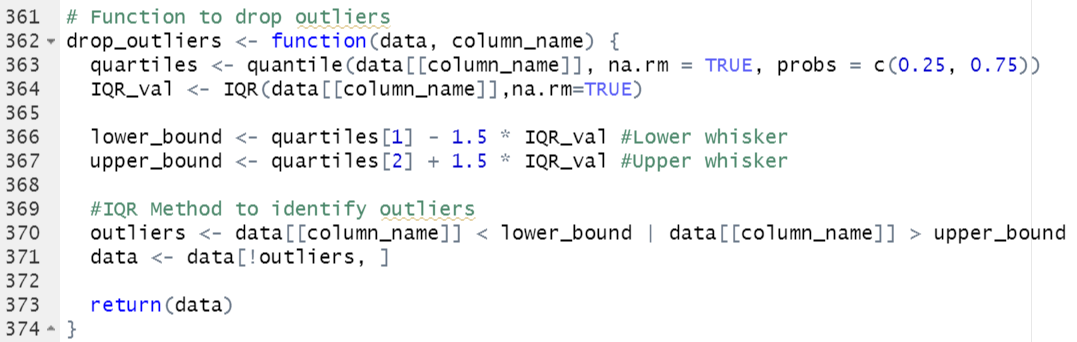
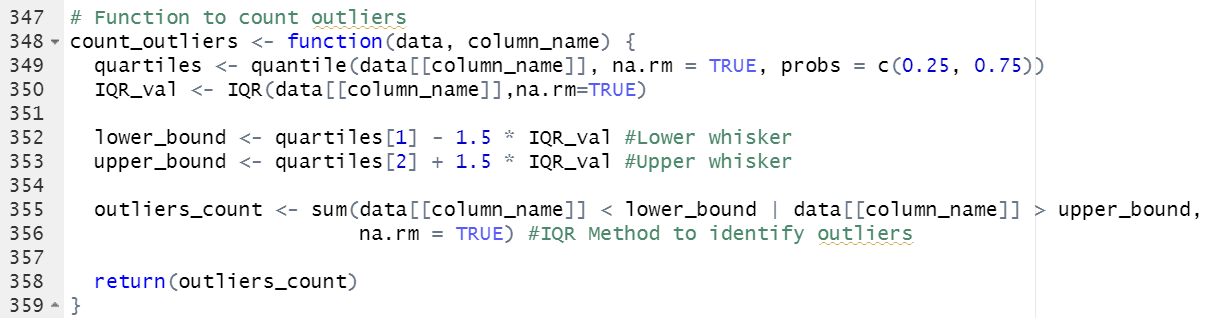
# 

# 8. Appendix

**8.1 Appendix 1**



**8.2 Appendix 2**



**8.3 Appendix 3: A Brief Report on Investigation into Hospital Readmission Rates**  
  
Apart from data from Hero DMC Heart Institute, we also explored similar data from Zigong Hospital in Sichuan, China. This dataset contained more information on readmission dates such as readmission with 1 month, 1-3 months and 6 months. The dataset was also more balanced overall with a significant proportion of patients readmitted. Overall, there were 2008 rows and 167 columns of data. In comparison to HDHI, Zigong’s dataset consists of fewer data points, but has more variables to explore.  
  
The overall data cleaning process was similar to that performed for HDHI data. To handle missing values, rows and columns with a significant amount of missing information were removed. Variables were classified as continuous or categorical depending on the number of unique values present in the column. Details on the cleaning process can be found in Readmission\_pipeline.R.  
  
Univariate Analysis was performed on the data. In general, continuous variables followed skewed normal distributions, while categorical variables had a tendency to be highly unbalanced. As such, several categorical variables were combined. When performing Bivariate Analysis, several variables were found to be correlated. These variables were removed as well.  
  
Despite our best cleaning efforts, modelling proved to be difficult. Although the outcome was changed to a binary output of whether the patient is likely to be readmitted, there were significant classification errors, with an accuracy of around 60-65% for all models attempted.   
  
The reason for the poor performance for all models tried (Random Forest, xGBoost, Neural Networks etc.) could possibly be attributed to a possible poor relationship between hospital data and readmission rates. And it makes sense, the patient’s actions after discharge is likely to be a significant determinant of their readmission rates. If the patient remains disciplined and maintains a healthy lifestyle while obeying doctor’s orders, it is likely that they will not be readmitted within 6 months, and vice versa.  
  
Furthermore, readmission did not account for a significant proportion of patients in the HDHI dataset. As such, readmitted patients are not expected to be a significant proportion of all admissions.   
  
Hence, as readmitted patients are not likely to occupy many beds, and the poor performance of the predictive models, the decision was made to drop readmission modelling and focus on patients as a whole. More information can be found in our submitted code, “Readmission\_pipeline.R” for our modelling process.

**8.4 Appendix 4: Full scale Flowchart**

