BC2407 Analytics II: Advanced Predictive Techniques

Group Report – Project Carity

SEMINAR 03, TEAM 2

1. Au Yew Rong Roydon (U2021424J)
2. Chung Shi Pei Michelle (U2010387K)
3. Paul Solomon Low Si En (U2022421F)
4. Luo Yiying (U2010604L)
5. Timothy Chang Zu'En (U2022114A)

SUBMITTED ON 3 APRIL 2022

Contents Page

1: Executive Summary

This project aims to support HealthMan, a healthcare management consultancy, in solving the problem of inaccurate assessment of Length of Stay (LOS) in South African Hospitals. The problem originates from doctors being incognisant of Administrative Hospital factors and neglecting part of a wide range of data before prescribing LOS. This leads to inaccurate prescriptions of LOS, with patients either receiving insufficient care or staying for unnecessarily long. The implications to hospitals are wide-ranging and include Patient Health Risks, Poor Patient Experience, Higher Healthcare Charges, Operational Delay and Inefficiency, and Increased Financial Burden. All of which fall under the purview of HealthMan's network management and general consultancy services.

We thus devised a two-pronged approach for our solution: To process copious amounts of data and accurately predict a benchmark of LOS for doctors, as well as identify key problem areas and bottlenecks for hospital management to tackle. The end-product of our solution will also promote ease of access to organised data for a Hospital's Management, further attempting to minimise internal inefficiencies that exacerbate inaccurate LOS prescriptions. The Business Opportunities lie in improved Resource Allocation within Hospitals and Patient Satisfaction. Overall, this serves to alleviate the burden on HealthMan in managing its client base of hospitals.

Our team adopted a predictive modelling approach to generate benchmark recommendations of LOS for doctors. We explored Machine Learning Algorithms in Multi-Adaptive Regression Splines (MARS), Classification and Regression Trees (CART), and Random Forest to build the most suitable model to predict LOS. Our evaluation criteria were grounded on key considerations for the Healthcare Sector, such as Predictive Accuracy, Interpretability, and Level of Technology. Ultimately, we adopted CART as the model of choice for HealthMan to integrate in their solution. Linking the Machine Learning approach to the second part of our solution would be Variable Importance.

To address the fringe problem of disorganised data collection and storage rampant in African hospitals, we further explored the use of Interactive Health Management Dashboards to achieve our target outcomes. We developed a Hospital Dashboard for Hospital Management and a Patient Dashboard for Doctors perusal, each containing relevant and summarised data visualisations on key operational aspects of the hospital. This not only promotes the effective management of wards and facilities, but also helps hospital management identify problem areas and bottlenecks in their own organisation.

2: Introduction

2.1: Introduction to Entity

Health Management and Networking Services, also known as HealthMan, is a privately owned healthcare consultancy specialising in the management and administration of hospitals in South Africa. Their services include **healthcare network management**, in providing administration and company secretarial support to hospitals, which aim to facilitate operational excellence for their clients, being healthcare groups and institutions. HealthMan also provides **general consulting services** to hospital management on the topics of practice profiling and costing, as well as technological aspects like data mining, database analysis and coding structures.

2.2: Business Problem

The concept map for our Business Problem and Overall Strategy can be found in **Appendix 2.2** for your reference.

2.2.1: Problem Statement and Background

One of the foremost challenges plaguing HealthMan and the wider African Health System is the **inaccurate assessment of patient Length of Stay (LOS)**. LOS refers to the amount of time a patient spends in the hospital during a single visit and is regarded as one of the most important indicators for a hospital management's efficiency and resource utilisation. **Unbeknownst to many, medical convention suggests that higher LOS does not equate to a better patient experience**, but in fact has negative implications for the patient. (ABOUT Healthcare, 2021b) On the other hand, reduced LOS is associated with decreased risk of infection and medication side effects, improvements in treatment quality, and more efficient bed management leading to increased hospital profit. (Baek, 2018)

2.2.2: Root Cause of the Problem

While doctors are well versed in evaluating their patient's needs accurately, **many neglect the efficient management of hospital resources such as utilities and ward rooms**. They also face difficulty in processing the wide range of both clinical and administrative data needed to predict LOS. Moreover, LOS becomes increasingly difficult to issue at an optimal duration for more complex medical conditions, older patients and for higher LOS.

The problem thus arises when doctors base their judgement almost solely on clinical factors, neglecting broader administrative concerns. Their judgement often deviates depending on the individual doctor's level of experience as well. This leads to some patients receiving insufficient LOS, and others having unnecessarily long LOS.

(1) Insufficient LOS:

Patients may be discharged pre-maturely resulting in further complications such as relapse of medical condition.

(2) Unnecessarily long LOS:

Unnecessarily long LOS results in: (1) Increased healthcare costs from over prescription of medicine and equipment usage. (2) Poor allocation of manpower to care for other patients. (3) Crowding within South African hospitals which prevents others from receiving necessary inpatient treatment. (4) Lower patient satisfaction.

2.3: Case Justification

2.3.1 Implications of inaccurate LOS:

Our group focused on LOS as it is an important measure used in the healthcare industry to ensure the efficiency of hospital management. An inaccurate assessment of LOS will lead to an adverse impact on the 5 key areas as shown below: (ABOUT Healthcare, 2021)

1. **Clinical:** Clinical capabilities such as clinical attention from staff would be impaired due to inaccurate LOS. Resources being used on patients would be directed away from those who need it more.
2. **Financial:** Stagnancy of patient throughput due to inaccurate LOS would result in slower revenue and wasted opportunities in the inflow of new patients.
3. **Operational:** LOS affects many operations such as patient intake being compromised, cleaning and sanitisation procedures delays. Having a lower LOS also frees up more wards for patients to be treated. Hence the operational efficiency of the hospital heavily depends on LOS.
4. **Experiential:** Patients' experiences can be severely compromised if they need to wait in bed longer than clinically necessary. Moreover, the risk of infections also increases.
5. **Health:**

   Impact of Insufficient LOS: Complications from early discharge could be fatal to patients.

   Impact of Unnecessarily long LOS: Spending more time in a hospital can actually hinder a patients' recovery as length of stay is usually tied to higher mortality rates. Longer length of stay also increases the chances of a patient developing healthcare-acquired infections (HAI).

2.3.2 Context of South Africa

The problem of LOS is especially significant in Africa, which is known for having one of the worst healthcare systems in the world. African hospitals generally report exorbitantly high LOS of up to 73 hours in the emergency department, way above the acceptable threshold of 12 hours. The top three reasons for this are: Inadequate Human Resources, Poor Resource Allocation, and Poor Maintenance of Healthcare System Infrastructure. (GSMA, 2021). These three areas ultimately present doctors with much raw, unorganised information when prescribing LOS, funnelling into the problem of inaccurate LOS. This limits decision factors to medical-related variables without consideration of administrative variables such as ward quality or availability. Inaccurately prescribed LOS in turn leads to poor staffing and resource allocation, resulting in a spiralling effect. Consequences like increased healthcare costs, morbidity, and mortality rates along with crowding and lower patient satisfaction follow.

Hence, not only is there a significant Business Problem surrounding LOS in Africa, but there is also much potential for the South African Healthcare System to benefit from an analysis model. Our model would enable doctors to make faster and more accurate decisions, improving hospitals' operational efficiency and usage of resources.

2.4: Business Opportunity - Increasing Efficiency of Hospitals:

There lies much potential in solving the problem of inaccurately prescribed LOS. The benefits of doing so extend not only to hospital staff but also patients.

2.4.1: Resource Allocation

Better resource allocation and reduction of wastage of medical resources would improve efficiency and reduce costs, which help to boost profitability. Additionally, improved manpower allocation and

workload distribution serve to reduce the occurrence of overwork among healthcare workers. This positively impacts performance and quality of service.

2.4.2: Inpatient Experience

Inpatient experience can be improved by reducing unnecessary hospital stay as well as providing better quality service. Cutting down on unnecessary hospitalisation also facilitates patient flow, reducing a patient's time on the waitlist and potentially bringing prompt life-saving treatment. Ensuring that patients stay the right duration also minimises the spread of disease internally, reducing morbidity and reduces readmission due to complications.

2.5: Project Outcomes, Measures and Targets

The flowchart of our Project Targets and Solution can be found in **Appendix 2.5** for your reference. To achieve the business opportunity, the key questions to be answered for our project would thus be:

Target 1: How might we develop a system to assist doctors in prescribing LOS consistently and accurately?

The system must accurately predict LOS of a given patient through a wider selection of metrics that accounts for the hospital's administrative constraints as well as the patient's medical conditions, to provide a holistic guideline for doctors to consider when prescribing LOS. This helps prevent insufficient LOS and reduce unnecessary LOS.

Target 2: How might we determine the most significant actionable area for Hospital Management to directly reduce unnecessary LOS?

Besides an accurate prediction model, hospital management would benefit from understanding which predictors in the model are significant. This helps hospital management better allocate resources to remedy target variables that are more important when trying to reduce unnecessary LOS without sacrificing patient's wellbeing. For example, *Administrative Delay* is a predictor for LOS. If it is a significant variable, it could mean that reducing administrative delay would be useful in reducing unnecessary LOS.

Key measurables for the project:

We split our key measures based on the outcomes we hope to achieve:

Prevent insufficient LOS:

(1) Prediction accuracy of the models would be useful in evaluating whether our models could predict LOS accurately and how many predictions are overpredicted vs underpredicted.
(2) Patient re-admission rates can also be tracked periodically to observe re-admission due to premature discharge.

Reduce unnecessary LOS:

(1) Prediction accuracy of the models would be useful. Before looking at variable importance, a high accuracy/low RMSE should be achieved first.
(2) Variable importance of models could be monitored by hospital management to identify significant target areas causing unnecessary LOS.
(3) Periodic tracking of patient inflow; a greater patient inflow would mean a reduction of LOS.

(4) Patient wellbeing surveys would be a good addition to patient inflow since shorter LOS must be achieved without compromising patient wellbeing.

This would therefore improve HealthMan's management of hospitals functions under the African Health System, helping them make better informed decisions with regards to LOS, treatment plans and resource allocation. Hospitals may also benefit from having a means to predict and forecast ward usage, bed turnovers and patient flows.

3: Literature Review

Length of Stay (LOS) has been recognized as a healthcare metric that is immensely beneficial to hospitals and patients when accurately predicted.

Many studies have been done to predict Length of Stay (LOS) in different contexts, with one example being Riascos & Serna, (2017) studying the prediction of LOS for Colombian health care. We examined their results and have found model accuracy metrics that could be used as a benchmark for our models in this report. Their most accurate model is **Linear Ensemble with a Normalized Root Mean Squared Error of 9.524%** (Riascos & Serna, 2017)

4: Data and Approach

4.1: Intended Approach:

Our team split our approach into two parts each attaining our targets as stated above.

Meeting Target 1: Using of Machine Learning to produce accurate benchmarks of LOS, and Dashboards to present doctors with organised patient information and insights

The inaccurate estimation of LOS stems from the difficulty faced by physicians in processing the wide range of data needed to predict LOS. We intend to use machine learning models (MARS, CART, Random Forest) to assist doctors by accurately predicting LOS. The model would serve to guide the decisions of doctors and improve the operational efficiency of the hospital. Common administrative and clinical factors present in hospitals, such as *Age*, *Severity of Illness*, and *Administrative Delay* will be used as predictors.

The dashboard is a collection of visual data which would convey machine learning predictions in an explainable manner to doctors. By summarising all important administrative variables and their effects on LOS, the dashboard provides a concise overview for doctors to not only monitor their patient's hospitalisation, but also prescribe a more accurate LOS through the consideration of both medical and administrative variables.

Meeting Target 2: Using variable importance from machine learning models and dashboards to target important variables to reduce unnecessary LOS

We intend to use the variable importance from machine learning models to decide on significant target areas for the hospital to reduce unnecessary LOS. This helps channel resources to the most significant area to ensure the optimal use of resources. The dashboard further help consolidate information in a concise and explainable manner to help hospitals easily reduce unnecessary LOS.

Fig.1 is our proposed Data Process Pipeline for the Machine Learning Segment of our solution:

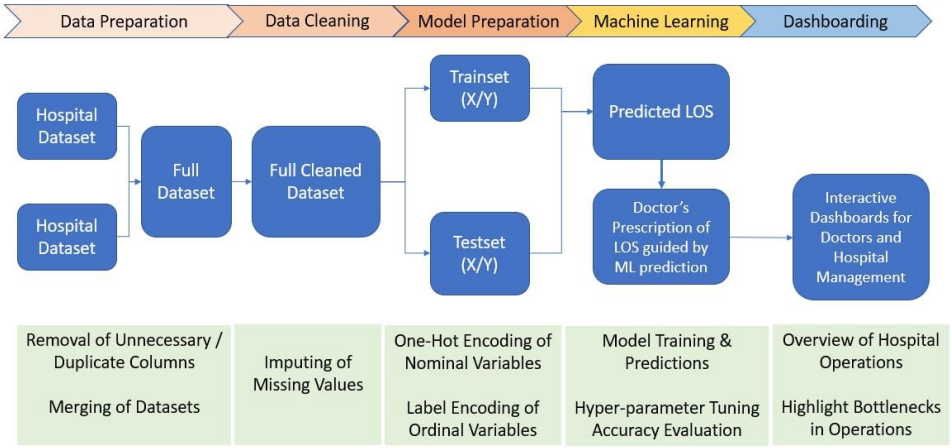


Fig. 1 Data Process Pipeline

4.2: Data Preparation:

Data was sourced from Kaggle.

To broaden the scope of our analysis, we sourced additional data on Hospitals in New York to be combined with the original data. This was done to obtain an extensive list of predictors for analysis, which we believe fits well given that New York Hospitals face a similar problem to that of their African counterparts.

Both datasets were merged on *Severity of Illness* as a common column, given that the criteria for severity assessment across hospitals are largely similar. Additionally, we identified *Administrative Delays* as a relevant predictor to be added into the dataset, to better represent the inefficiencies and delays the African Health System is known for. The summary of the variables can be found in a data dictionary in **Appendix 4.2**

4.3: Data Cleaning

We removed rows containing missing values for Length of Stay. Beyond this, no further missing values for continuous and categorical independent variables were reported.

4.4: Train-Test Set Preparation for Machine Learning

4.4.1: Choice of Predictors for Models

Our general selection criterion for possible features would be variables that have actionables in the context of hospital operations. Independent variables that had no applicability for the models such as hospital code were removed.

Some interesting variables worth focusing on include *Available Extra Rooms*, *Admission Deposit*, *Administrative Delays* and *Visitors with Patient*. These variables could be acted upon by hospital management should they prove to be significant in our models. For more information, refer to **Appendix 4.4**. The other variables may be important in the prediction of LOS but more difficult for hospitals to act on in practice.

Independent variables:

Type of Admission	Available Extra Rooms in Hospital	Admission Deposit	Administrative Delays	Severity of Illness	Patient Disposition
Visitors with Patient	Age	Ward Type	APR Risk of Mortality	APR Medical Surgical Description	Payment Typology 1

5: Data Exploration

Exploratory data analysis was conducted to summarise the relationships between variables in the dataset. This provides us with a deeper understanding on how the variables might affect our models and possible trends/relationships between them.

5.1: Data exploration of correlation between continuous variables

We first explore the correlation between continuous variables to view possible relationships as well as identify variables that might have high collinearity, displayed in **Appendix 5.1**. High collinearity would mean the variables provide the same information to the model and would not be as important. It is observed that *Length of Stay* is most correlated with *Administrative Delay* followed by *Age*. This could mean Administrative Delay is more dominant and could affect Length of Stay more than the other continuous variables. This could be further checked by Variable Importance at the machine learning stage of our project. Generally, the other variables do not have a linear relationship with LOS, but it does not mean they do not have any relationship.

Lastly, since all variables generally have a non-linear relationship between one another, there is no collinearity. (Raj, S, 2021) Hence, all continuous variables could be included in models since they generally provide different information to the models.

5.2: Data exploration of Dependent Variable

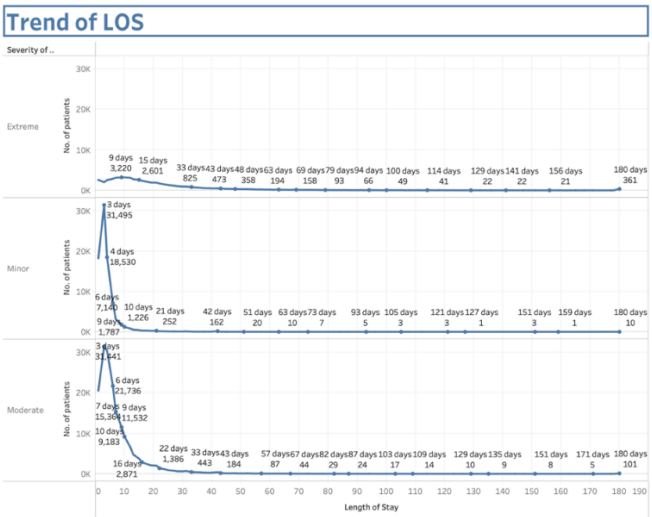


Figure 2.1: Overview of LOS

Figure 2 shows the LOS of all patients in the hospital across the different *Severity of Illness* (Extreme, Minor, Moderate). All three graphs show a similarity of a lower number of patients having high LOS.

This implies that majority of patients only stay in hospitals for less than 10 days except for patients with extreme severity of illness.

It can also be observed that a more serious illness warrants a longer average LOS of patients. Majority of the patients with extreme illness have an average LOS of 9 days as compared to Minor and Moderate conditions where patients stayed around 3 to 4 days.

This could mean that the model trained might be biased towards being able to predict well for lower length of stay but perform poorer in terms of prediction accuracy for higher length of stay. Moreover, the skewness might mean logging of the length of stay might help increase accuracy of prediction.

5.3: Data Exploration of Independent Variable

Since we are unable to view the correlation with regards to categorical variables, we conducted some data exploration to identify possible relationships.

Risk of mortality and LOS vs age:

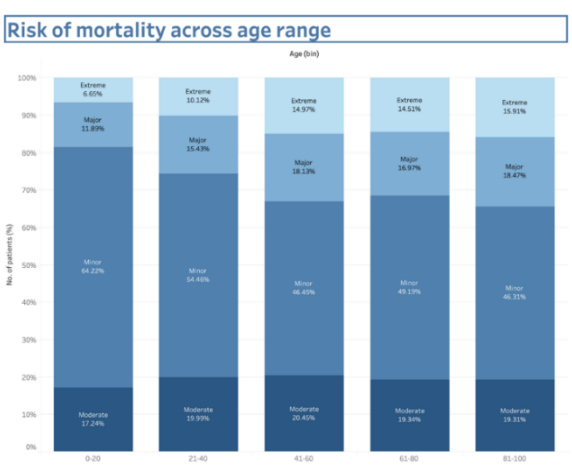
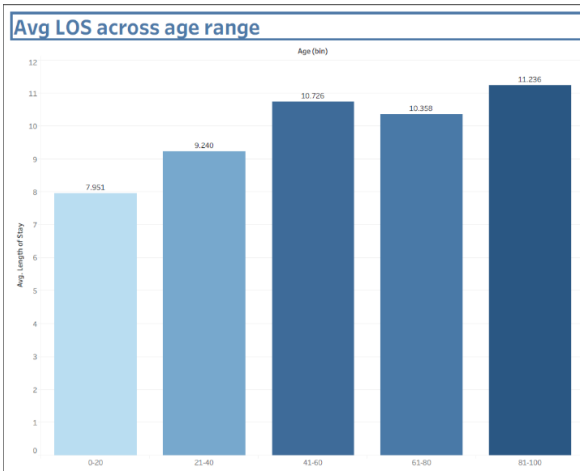


Figure 2.2a: Risk of mortality across age



2.2b: Avg LOS across age

As proven by an article, the risk of mortality is often associated with the increase in age due to the decline in the rate of damage repair. (Li, Yang, & Anderson, 2014) To explore whether the data used in our report aligns with the article, a bin is created to group the age variable in an interval of 20 (Figure 6.2a). Based on the finding, we can observe a higher risk of mortality starting from the age of 41 by comparing the percentage of patients under the extreme and major category of mortality. The risk of mortality is about 16% higher for patients above 80 as compared to age below 20. Moreover, the length of stay for adults above the age of 40 is generally higher implying that age does play a role in length of stay. Hence it could be important for hospitals to investigate greater care of adults over the age of 40 through better conditions/isolation from patients with severe illnesses to possibly reduce length of stay through better treatment.

Additional findings for categorical variables are found in **Appendix 5.3**

6: Comparison of models

6.1: Overall Strategy on Models and Solutions

We would be running a total of 3 different models and evaluating their accuracies on different dataset or variable splits: (1) Original Dataset, (2) Retaining selected variables that are actionable by Hospital Management. Approach 2 will be used to facilitate analysis of feature importance among the selected actionable predictors, giving hospital management a clearer view of the relative importance and urgency of operational aspects.

To obtain the best possible models, we attempted different parameter values and complexity levels. Through this process of hyper-parameter tuning, we would also experiment with the transformations of variables such as applying logarithmic transformation on skewed variables. All models would have a train-test of split 7:3 with a seed number of 2022 to ensure fair comparison of performance across models.

The evaluation of model accuracy would be done using **Normalised RMSE**. This is to ensure that models trained on different scales of data would have the same basis of comparison.

Regarding Length of Stay, solutions will be generated using insights of the most significant variables that affect it. At the same time, the real-life business context of hospital management and specific circumstances of hospitals in Africa will be considered.

6.2: MARS Model

Multi-Adaptive Regression Splines, or MARS model, was employed to predict Length of Stay. It was selected based on its ability to make predictions for **non-linear regression problems** and **handle large amounts of both continuous and categorical data**.

The *Earth* model from *py-earth* package was trained on both original and logged datasets (to minimise skew) before predicting labels for their corresponding test set features.

To determine the best model in terms of performance, we compare the Generalised Cross Validation Score (GCV) below. GCV measures the model performance and is penalised for more complex models with additional parameters. In general, models with lower GCV for a given dataset are preferred. With reference to Figure 3., MARS with Degree 2 is observed to have higher performance than Degree 1 within both original and log datasets.

Model	GCV
MARS Degree 1 (Original Dataset)	175.6245
MARS Degree 2 (Original Dataset)	148.926
MARS Degree 1 (Log Dataset)	0.3417
MARS Degree 2 (Log Dataset)	0.2800

Figure. 3 Comparison of Generalised Cross Validation Scores

Additionally, **Normalised Root Mean Squared Error (RMSE)** was adopted as a standardised metric for comparing model accuracy, which lends itself to eventually deriving the most accurate predictive technique for our solution. (Figure 4.)

Model	Normalised RMSE
MARS Degree 1 (Original Dataset)	7.484%
MARS Degree 2 (Original Dataset)	6.877%
MARS Degree 1 (Log Dataset)	12.918%
MARS Degree 2 (Log Dataset)	11.657%

Figure. 4 Comparison of Normalised RMSE

The optimal conditions for MARS are thus Degree 2 and using the Original Dataset, as they contribute to the lowest normalised RMSE and in turn highest predictive accuracy among all MARS model trialled. Documentation on the full process of MARS modelling can be found in **Appendix 6.2.**

6.3: Classification and Regression Trees (CART) Model

CART is a decision tree with conditions at each split before the final predicted value is reached. This model was chosen to examine the prediction of Length of Stay (LOS) due to its **simplicity of interpretation and understanding** as well as its ability to interpret both categorical and continuous data.

GridSearchCV was used to optimize parameters to achieve the highest accuracy decision tree without overfitting the data. The optimal depth was found to run the tree before testing it on our test set. To compare the accuracy between varying models, we calculated Normalised RMSE as shown in Figure 5. Screenshots of results can also be found in **Appendix 6.3**.

Model	Normalised RMSE
CART Optimal Max Depth 7 (Original Dataset)	6.776%
CART Optimal Max Depth 9 (Log Dataset)	12.243%

Figure. 5 Comparison of Normalised RMSE for CART

As we can see, the model built on the original dataset has a significantly lower RMSE of 6.776% which makes it the better CART model with the highest accuracy.

6.4: Random Forest Model

Random Forest is a powerful tool that uses the combined predictions of multiple decision trees to give a majority decision. This reduces overfitting leading to a more accurate model. We also expect patient's medical conditions to have high variance, hence the model must be robust to outliers, which is one of the key benefits of Random Forest.

GridSearchCV was used in optimizing parameters to determine a Random Forest with the highest accuracy. Some of the optimal parameters as found from GridSearchCV can be found in **Appendix 6.4.**

The parameters were then used to train and develop a suitable Random Forest model before testing its ability to make predictions. To compare the accuracy between varying models, we calculated Normalised RMSE as shown in (Figure 6.)

Model	Normalised RMSE
Random Forest Optimized Model (Original Dataset)	6.757%
Random Forest Optimized Model (Log Dataset)	11.377%

Figure.6 Comparison of Normalised RMSE for Random Forest

As such, the model built on the original dataset instead of log dataset as it has a significantly lower RMSE of 6.757%, which sets the preferred conditions for our optimal Random Forest model.

7: Model Evaluation

To determine the most suitable model for integration into a hospital's management system, we consider the following criteria:

7.1 Predictive Accuracy

As with most Machine Learning approaches, predictive accuracy is a quintessential evaluation criterion as it determines the quality of predictions, forming the scientific basis for decision-making and policy. Reducing the margin of error for LOS is also important to our context, given that deviations resulting from unnecessary LOS have a compounding effect on resource bottlenecks and patient inflow.

Model	Normalised RMSE
MARS, Degree 2 with Original Dataset	6.877%
CART, Max Depth of 7 with Original Dataset	6.776%
Random Forest with Original Dataset	6.757%

Fig. 7 Normalised RMSE Comparison across optimal models

Figure. 7 compares the chosen standardised metric of Normalised RMSE. The error of each model is comparable with Random Forest slightly edging out in accuracy. In general, all our models also perform better than the accuracy benchmark of 9.524% derived from Riascos & Serna, (2017) findings.

Besides RMSE, we further analysed the number of overpredictions as compared to underpredictions. Generally, we would prefer a model which has a lower number of underprediction. Underprediction holds a greater implication than overprediction since underprediction would mean many patients would not be receiving sufficient care resulting in possible complications. We would like to predict Length of Stay accurately without compromising the wellbeing of our patients.

Model	Overprediction (%)	Underprediction %

MARS, Degree 2 with Original Dataset	55.536%	44.464%
CART, Max Depth of 7 with Original Dataset	59.112%	40.887%
Random Forest with Original Dataset	64.375%	35.615%

Fig. 8 Prediction Result of Models

From the results, Random Forest has the highest Overprediction of 64.375% which implies that it yields a result that is more desirable compared to the other 2 models.

7.2 Explainability

Explainability of a model refers to the ease of understanding of the model, such as what parameters the model is considering or if the model contains any bias. Ideally, the model should have high transparency and explaining power, such that hospital management and doctors understand what factors most influence a patient's LOS. Knowing this would allow hospital management to drill down on significant operational inefficiencies and build a pipeline around this to actively reduce any unnecessary LOS.
The hospital staff in Africa have low computer literacy which is a huge consideration when picking our models. (Odekunle et al., 2017) Hence, we would like to propose a model that is easy to understand with a low learning curve for hospital staff to integrate the solution fast.

The explainability of our 3 models would be ranked as such:

CART > MARS > Random Forest

(1) CART presents with the highest relative explainability because users can access the decision tree and its decision split rules. The *feature_importances* function also returns a list of the top variables in descending order of importance with relative ease. It does not take much computer literacy to understand the variable importance in CART since the splitting of the trees can be easily understandable.
(2) MARS (*pyearth*) poses the limitation of not explicitly containing a variable importance measure. It does, however, present various useful metric for model performance such as GCV. However, the understanding of statistical importance may pose a problem for the staff who has little understanding on statistics.
(3) Random Forest is less explainable than CART because it involves the majority prediction of 2500 trees, making it very difficult to interpret the impact of variables and individual trees. The random forest also utilises Random Subset Feature which would mean that not all features are used in the prediction for each tree. Permutation Importance would then have to be used which takes up a considerable amount of time and processing power when the number of predictors is large.
The process is relatively complicated and hospital staff would have to take some time to understand how random forest works.

Additional information of models and their explainability can be found in **Appendix 7.2**.

7.3 Ease of Implementation

One consideration for implementation is the time needed to predict the model using the test set. This would affect the performance of our solution if the time taken to predict LOS for a patient is particularly significant, as it would add to administrative delays.

Model	Prediction Time
MARS, Degree 2	0.03992 s
CART, Max Depth of 7	0.04991 s
Random Forest	35.686 s

Fig. 9 Prediction Time of each Model on Test set

From Fig 9. Random Forest does have a much higher prediction time compared to MARS and CART. However, the performance of Random Forest would ultimately have minimal impact in a clinical situation as only prediction for one patient is needed, compared to 79,610 patients in test set.

7.4 Model of Choice

We decided to recommend CART as our final model despite its higher RMSE to Random Forest since the difference in accuracy is not particularly significant. Moreover, it is more practical to use a CART model as compared to Random Forest model since the processing power required for Random Forest is much higher which hospitals in Africa might be lacking. The high explaining power of CART would also be more useful in providing hospitals with actionable insights on important areas to focus on, providing benefits beyond its predictions. Moreover, the learning curve for CART is less steep as compared to the 2 other models which makes it easy for the hospital management team to understand and utilise it to the full potential.

With a high accuracy of prediction and high explainability power, CART would **help meet target 1: Accurately predicting length of stay**.

8: Variable Importance of Best Model

To meet **Target 2: Reducing unnecessary length of stay**, we first begin by analysing the variable importance of predictors in the model. This helps hospitals in channelling their resources to target areas that are significant in possibly reducing unnecessary length of stay.

8.1: Medical + Administrative Variables

Administrative Delay was found to be the most significant variable which contributed to the prediction of LOS, derived from the best model. Surprisingly, an administrative variable is more significant than the other clinical/medical variables when it comes to predicting LOS.

	Feature_Importance
Administrative_Delay	0.668621
Severity_of_Illness	0.119980
APR_Medical_Surgical_Description_Surgical	0.104354
APR_Risk_of_Mortality	0.041229
APR_Medical_Surgical_Description_Medical	0.016504
Patient_Disposition_Home or Self Care	0.010561
Patient_Disposition_Skilled Nursing Home	0.008066
Payment_Typology_1_Medicare	0.007202
Patient_Disposition_Home w/ Home Health Services	0.006695
Payment_Typology_1_Medicaid	0.006291

Fig. 10: Prediction Time of each Model on Test set

This is further supported by the variable importance shown in Random Forest **(Appendix 8.2.2)** and statistical importance shown in MARS **(Appendix 6.2)**

It was previously noted in 2.3: Case Justification, that the top three problems with Africa's healthcare sector are Inadequate Human Resources, Poor Resource Allocation, and Poor Maintenance of Healthcare System Infrastructure. Most of which also point to the paper-based, analogue data and information collection methods. (GSMA, 2021) These serve as a pivotal cause of Administrative Delay which increases unnecessary LOS.

8.2: Actionable/Administrative variables compared to LOS

We determined that analysing the importance of variables that could not be acted on would provide little practical value to the hospital. We suggest that hospital management could analyse variable importance using a **subset of actionable variables** such as administrative variables instead. A table of such actionable variables can be found in **Appendix 8.2.1**

By re-running CART and Random Forest on these predictors, we observe that *Administrative Delay* is still the highest out of all variables. However, variables such as *Admission Deposit*, *Age*, and *Visitors with Patient* do also play a role in influencing Length of Stay. The RMSE between the model with all predictors (6.776%), and the model with the subset of predictors (7.408%) (Figure 11), is also comparable, hence the results of Feature Importance can still be applicable. To obtain Variable Importance from Random Forest, we can either set *Random Subset Feature* to equal to the number of variables (all predictors would be included) or through *Permutation Importance*. From Figure 12 and 13, Random Forest shows a more balanced variable importance since it is based off multiple CART trees whereas the variable importance of CART is dominated by Administrative Delay. This is due to Administrative Delay being prevalent in Africa Hospitals and most hospitals have poor, inefficient healthcare systems linked to high administrative delay.

As a result, the Feature Importance is skewed towards Administrative Delay. Hence more studies must be conducted after reducing Administrative Delay to further understand how the other variables might fare in contributing to LOS. We hence determine that reducing unnecessary Administrative Delays from various operational aspects would be the most useful for Hospital Management to improve LOS. Additional models on Actionable models can be found in **Appendix 8.2.2**

```
Root Mean Squared Error: 13.26116
Normalized Root Mean Squared Error: 7.408%
```

Figure 11: Results of CART using Administrative Variables

	Feature_Importance
Administrative_Delay	0.974970
Admission_Deposit	0.012506
Age	0.007073
Visitors_with_Patient	0.003200
Ward_Type_Q	0.001252
Ward_Type_S	0.000511
Ward_Type_T	0.000282
Ward_Type_R	0.000131
Available_Extra_Rooms_in_Hospital	0.000076
Ward_Type_P	0.000000
Ward_Type_U	0.000000

Fig. 12: CART Feature Importance (Actionable Admin Vars)

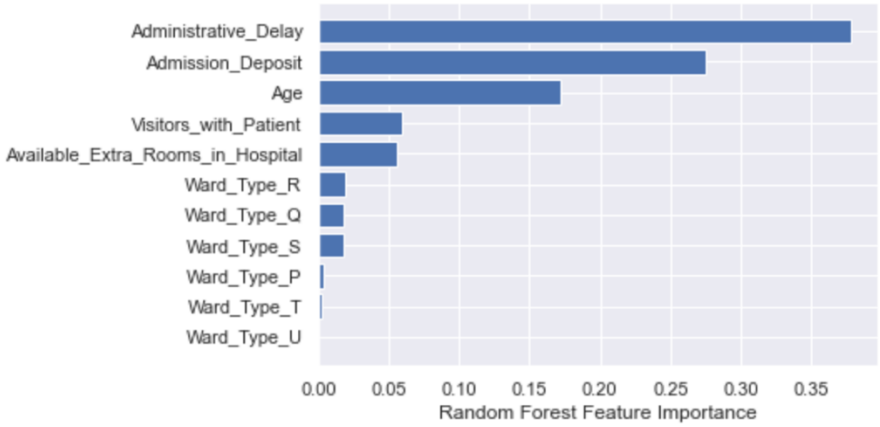


Fig. 13: Random Forest Feature Importance (Actionable Admin Vars)

9: Dashboards

9.1: Tableau Dashboards

Our selected machine learning model provides us with target areas to focus on through Variable Importance. It acts as a guide on which areas should be prioritised first according to importance, thus helping a hospital's management save time and resources in forming strategies to improve accuracy of LOS. Solutions could then be generated to target the most important variables (e.g., *Administrative Delay*). However, the effectiveness of the solution is limited by the scope of the predictors. The target areas/predictors provided by the model could be large which could mean that solutions generated for these areas are very general.

This is where Tableau dashboards would be a good tool to assist the model. Since *Administrative Delay* is of extremely high importance, we demonstrate how Tableau dashboards can serve to identify areas with high Administrative Delay to be improved, thereby supplementing our machine learning approach. Our dashboards assist the models by providing a more in-depth analysis on specific sub areas to work on such as specific ward types with higher Administrative Delay.

Hospitals would receive our end-product in the form of 2 different dashboards, built on 4 design principles. (Sisense, 2022) Using these principles, we hope to create a dashboard that provides sufficient information to the hospital team within mere seconds. Moreover, we utilised the *Inversion Pyramid* format to summarise and organise information in a format that tells a story. In this case, the story would be what affects LOS/Administrative Delay.

The Required Benefits of the dashboards are:

1. Logical – Must present key insights pertaining to the most significant variables affecting LOS, allowing hospital management to monitor the most crucial operational aspects.
2. Organised – Information must be segmented with ease of understanding.
3. Readable - Choosing the right visualisation ensures that hospital stakeholders with no prior technological background can interpret actionable insights.

To encourage minimalism, we split our information across 2 different dashboards which serves different purposes for the hospital staff. **Ultimately, the two dashboards would be used to assist models in achieving targets 1 and 2.** A flowchart of the Dashboard solution can be found in **Appendix 2.2**

9.1.1: Hospital Dashboard

The hospital dashboard as shown in **Appendix 9.1.1** tells a story of how different facilities/general hospital operations might affect hospital performances such as Length of Stay and Administrative Delay. It features the following key metrics: Trend Graphs on Delay and LOS, Department, Mode of Payment, and the Number of Rooms that can be filtered by Hospital and Department. This dashboard will directly aid the hospital administration in tracking the performance of the hospital.

9.1.2: Patient Dashboard

The patient dashboard as shown in **Appendix 9.1.2** tells a story of how patients' conditions might affect the Length of Stay and Administrative Delay. It entails the following key metrics: Trend Graphs and Patients' Conditions that can filter according to Hospital Code, Department, and Ward Type. From the dashboard, the hospital management can identify the summarised view of the patients' condition in each ward type.

9.2: Benefits of Tableau Dashboards

The 2 types of dashboards have many uses for both doctors and the healthcare management team. Our dashboards help aggregate data from multiple sources and provide a concise yet in-depth view of the performance metrics of the whole hospital team. This helps organise important information in an area to eliminate staffing inefficiencies and bridge communication gaps between the medical teams. (Benefits of the Interactive Healthcare Dashboards for Hospitals, 2020) Some opportunities are presented below:

9.2.1: Helping doctors in accurately prescribing LOS with Administrative Variables (Target 1)

It was previously stated that hospitals in Africa have poor healthcare management system with many miscommunications between staff arising from unorganised information. (Oleribe et al., 2019) The unorganised large amount of information also makes it difficult for doctors to make accurate decisions on LOS. Hence this provides our dashboards with an opportunity to assist doctors in making their decisions.

Hospital dashboard: The hospital dashboard provides doctors with important information such as number of available rooms per ward, average Length of Stay per ward etc. This information would be imperative in the decision on prescribing an accurate LOS. For example, doctors would be able view occupancy rate of the ward type and allocate patients accordingly to prevent overcrowding of certain wards.

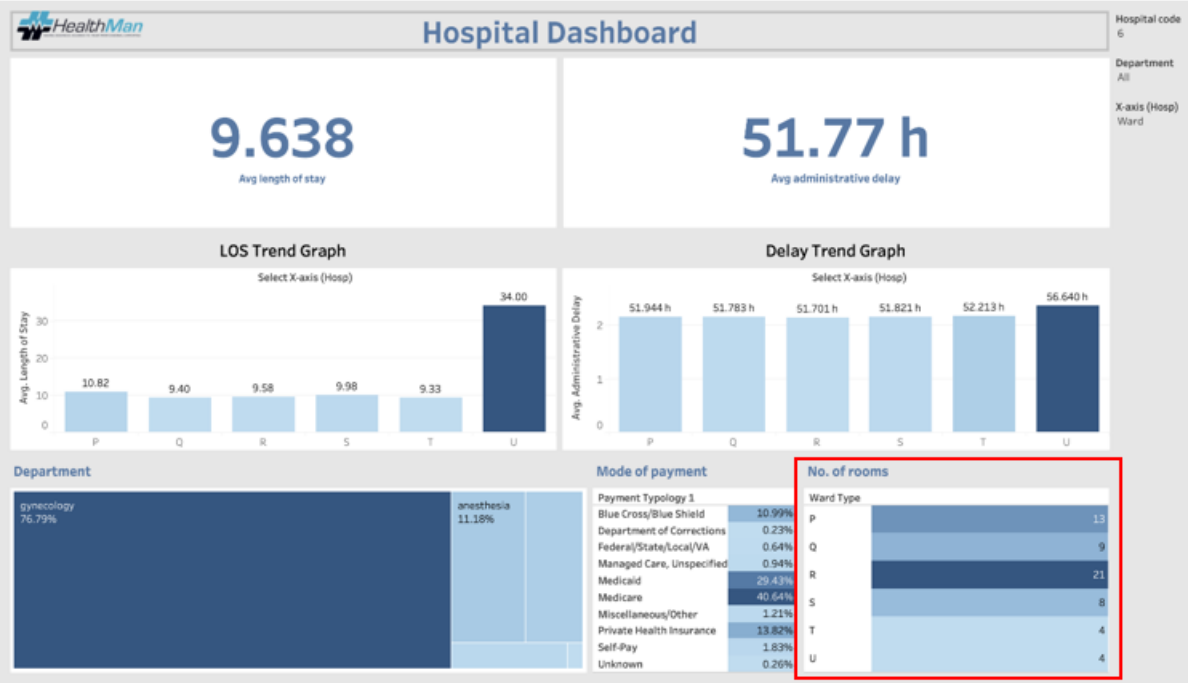


Fig. 14. Hospital Dashboard

Patients Dashboard: The patient dashboard provides doctors with important information pertaining to patients in the hospital such as Severity of Illnesses, Risk of Mortality, Patient Disposition etc. Patient Disposition in this case refers to where the patients would be discharged to after leaving hospitals. Such information is also imperative in the decision on prescribing an accurate LOS. For example, doctors would be able filter and view the severity of patients in different ward types. By understanding the severity spread within each ward, doctors could then better allocate patients accordingly. A patient with a higher risk might need to be placed in a ward that has patients with lower severity/less contagious illnesses. (**Appendix 9.1.2**)

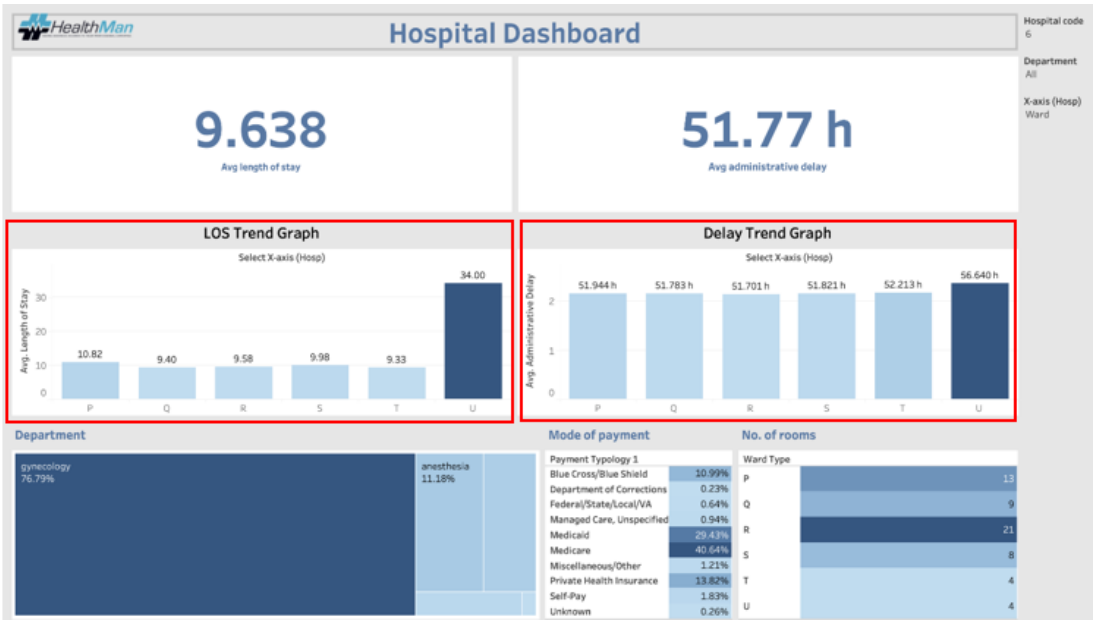
9.2.2: Helping hospital management team identify areas with high administrative delay and LOS (Target 2)

With the organised information on the dashboard, it would make it easier for the hospital management team to track LOS and administrative delays in certain areas and to target areas with unnecessarily high LOS and administrative delays. By understanding which areas have high LOS, hospital management can more easily identify if the high LOS is due to non-medical related reasons that could have possibly increased unnecessary LOS.

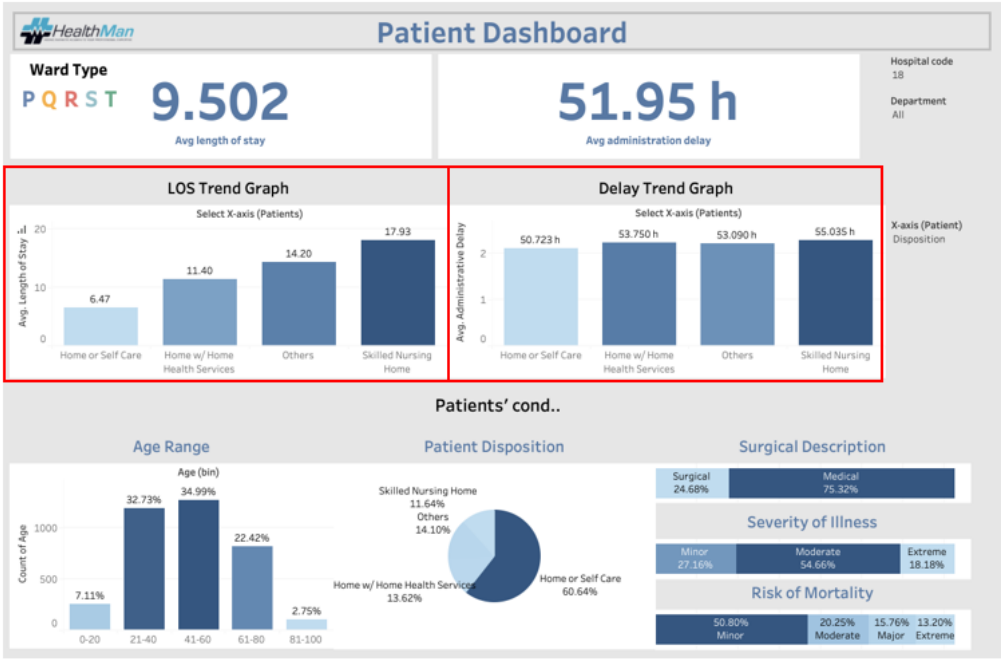
Hospital dashboard: The hospital dashboard helps the management team track facilities/processes that may have high administrative delay/ LOS.

For example, the hospital management team can easily filter the trend graphs to observe the relationship between ward types and LOS/administrative delay. The team could then identify the ward type with the highest amount of administrative delay to investigate into the causal reasons for the prolonged delay. In Figure 14, it can also be observed that ward U has the highest LOS and administrative delay out of all other wards. Hence it could be useful for the hospital team to investigate as to the reasons behind the high administrative delay and LOS. Another example would be how the hospital management team can review modes of payment by understanding the delays behind each mode of payment. It would then be useful for hospital management teams to try to reduce payment delays through a review of their

payment policies such as ways to make payment easier and faster for certain types of payment. (**Appendix 9.1.1**)



Patient dashboard: The patient dashboard helps the management team track processes with regards to patients and their respective administration delay and LOS. For example, the hospital management team could observe the administrative delay and LOS with regards to patient disposition. Hospital management could then identify transfer processes that could have the highest administration delay as well as review their policies with regards to making transfers faster.



9.2.3: Helping hospital management identify potential bottlenecks: (Target 2)

The dashboards would also provide information to the hospital management team with regards to potential bottlenecks which could arise from shortage of equipment or lack of number of available rooms etc.

Hospital dashboard: The hospital dashboard provides real time information on potential bottlenecks with regards to facilities. For example, the combination of administrative delay, length of stay and the number of rooms per ward type doubles as a gauge for the hospital administration in times of high patient inflow. It offers an estimate of the capacity of each ward so that the management can make contingency plans to mitigate the foreseeable bottleneck. In the case that a certain ward type is running out of available rooms, the dashboards would provide an early warning to the hospital management team to ensure that they would have sufficient time to prepare spare rooms for patients if needed. Without the help of the dashboard, administration delays would possibly further increase especially since Africa hospitals have a lack of healthcare workers and may need time to reallocate healthcare workers accordingly.

Patient dashboard: The patient dashboard also provides real time information on potential bottlenecks with regards to the attending of patients. For example, the severity of illness and risk of mortality spread in different ward types provide hospital management a gauge on the number of healthcare workers required for each ward. Wards with patients that have higher severity would require more healthcare workers to monitor and care for them. This would hence provide hospital management with better planning capabilities with regards to manpower allocation to curb the little amount of manpower that the hospital has. The information provided also provides hospital management with better resource allocating abilities since they would understand the number of equipment or resources needed for different patient types in different wards.

10: Possible Recommendations from Dashboard Findings

Besides the assistance that dashboards provide to the models to attain **target 1** and **target 2**, the dashboard did provide some interesting findings that hospitals could act on.

10.1: Spread of patient within wards

It was observed that some hospitals in Africa do not split the patients of different severities when it comes to allocating them to a particular ward. Research has shown that an important reason as to prolonged LOS is due to patients of different severity of illness being allocated to the same ward. (Refer to **Appendix 10.1.1** and **Appendix 10.1.2** (Organizational Factors Affecting Length of Stay in the Emergency Department: Initial Observational Study | Israel Journal of Health Policy Research | Full Text, n.d.)

We hence propose that hospitals can look into better segmentation of their patients so as to prevent prolonged length of stay. Although this restructuring process would be extremely difficult since it requires a lot of time and cost, it would be beneficial in the long run.

For instance, wards are typically categorized into ICU wards and normal wards where patients with a higher Risk of Mortality and the Extreme Severity of Illness will be administrated to the ICU ward. This is because ICU wards have a higher staff-to-patient ratio and more advanced medical resources. Without proper patient allocation, the management process becomes messy since different equipment would be required for different levels of severity/risk of mortality. This makes it difficult for the hospital management team to plan for resource and manpower allocation hence further increasing administrative delay. The mixing of patients with different severities could also result in medical complications especially if the disease is contagious.

10.2: Inefficiencies in transfer processes

Using the patient dashboard, we observed that there are high delays in discharging of patients. This provides a case for Hospital Management to explore improvements in their patient discharge and transfer processes. A possible recommendation from our team would be the implementation of a Shared Discharge Plan, which embeds discharge plans across various wards into the hospital's Electronic Medical Records (EMR). Modelled after UK Hospitals, this enhances communication among caregivers to coordinate a smoother outflow of patients. Moreover, transfer processes are inherently poor in Africa due to its extremely fragile and understaffed social care sector. Hospitals that identify categories of Patient Disposition as having high Administrative Delay should thus look to improve communications and procedures with external care facilities like hospices, nursing homes, and specialist clinics.

11: Summary of Business Opportunity

In summary, the solution provided would solve our Key Business Questions (**targets**) in the following manner:

1) **LOS Prediction Model** (Used by doctors and management team): The CART model would be provided to the Hospital Management Team and Doctors. **The Hospital Management Team** would oversee the **retraining of the models** provided that **new variables/data are made available**. In the process of retraining, the management team could also utilise the Variable Importance function provided by the model to further identify new areas to work on. The trained model would then be provided to doctors to predict Length of Stay of their patients during clinical sessions. The predictions would then serve as a benchmark to guide doctors in their decision-making process.
2) **Tableau Dashboards** (Used by doctors and management team): Tableau dashboards are provided to both doctors and the management team but differ in their use and purpose. For Hospital management, the dashboards would be provided to them as a tool to provide greater explaining power to the models and further pinpointing specific areas to act on. On the other hand, the dashboards provided to doctors are meant to give them a visual and non-technical understanding of administrative values. This could help them understand how the current hospital context could have impacted the prediction of LOS. As such, the dashboard along with the Prediction model would serve as a helpful guideline for doctor's LOS assessment.

With these 2 solutions, we hope to achieve (1) fast and accurate predictions of LOS, (2) better resource allocation, (3) fast operational decision making, (4) greater patient wellbeing.

(1) The model helps provide an accurate guideline to doctors with fast prediction speeds. This could further help reduce visiting times of patients. The dashboards further assist the model by providing a secondary guideline for doctors by supplementing their decision on LOS with the hospital's wider operational aspects. This information is organised and easy to understand for doctors with no technical background.
(2) The variable importance from the models provides the hospital management team with a specific area to target. The dashboards further help provide information on resources needed through the overview of the entire hospital operation.

(3) The dashboards help summarise all information in a concise and structure manner. At one glance, the management team would be able to understand the operations in the hospital and identify potential bottlenecks (EG: with regards to administration delay).
(4) With a lower amount of patient delay, better allocation of manpower, the care provided to patients would improve thus further enhancing their wellbeing.

The concept flowchart of this solution can be found in **Appendix 2.5** for your reference.

12: Limitations/Challenges Faced

12.1: Limited data on higher lengths of stays

Currently, our models have extremely high accuracy for lower Length of Stay since the amount of data on lower lengths of stay are much higher. This would mean that for lower Lengths of Stay, the deviation of the predicted values from the actual values would be smaller as compared to the deviation of predicted values from actual values that are higher.

To predict higher length of stay more accurately, hospitals would have to collect more data with patients of higher stay. This could be done by coordinating with other hospitals nearby to obtain data on patients with higher stay. However, this could take a long time.

An alternative would be to create separate models for different severity types. Since severity is usually linked with higher length of stay, we can analyse and test out how the different models would fair when trained on subsets of data. We can then create models that could more accurately predict based on the specific type of data. It would be an area worth exploring with some preliminary analysis being done in **Appendix 12.1**. From Figure 12.1.2a we can see MARS is the most accurate model for patients after splitting the datasets into different severities.

12.2: Limited Qualitative Predictors

For our machine learning approach, a limited range of predictors were handpicked. Possible areas which were left out include data on a patient's background, such as occupation type, as well as time-related data like readmission rates, and admission and discharge timings. Acknowledging this, a time series forecast for patient inflow would value add to a hospital management's decision making.

12.3: Limited information

We had limited information with regards to the some of the factors such as ward type. With little context on these factors, we were unable to analyse the factors to a greater detail. For example, by knowing the context of ward type, we could possibly have come up with a more detailed solution on patient segmentation within the different ward types.

12.4: Manpower

Our solution facilitates the planning and allocation of resources by hospital management but is limited by the amount of resources hospitals can work with. This hinders the potency of our solution given that South African Healthcare receives one of the lowest proportions of funding among all countries. Thus, much of our solution hinges of each hospital having adequate government funding.

13: References

ABOUT Healthcare. (2021, October 13). Why Is Length of Stay Important? Five Key Reasons | Insights. Why Is Length of Stay Important? Five Key Reasons. *https://www.abouthealthcare.com/insights/blog/length-of-stay/*

Areas of delay related to prolonged length of stay in an emergency department of an academic hospital in South Africa—ScienceDirect. (n.d.). Retrieved April 1, 2022, from https://www.sciencedirect.com/science/article/pii/S2211419X21000124

AV: Healthcare Analytics II. (n.d.). Retrieved March 5, 2022, from https://kaggle.com/nehaprabhavalkar/av-healthcare-analytics-ii

Clemens, M. A., & Pettersson, G. (2008). New data on African health professionals abroad. *Human Resources for Health*, *6*(1), 1. https://doi.org/10.1186/1478-4491-6-1

Cprime Studios. (2022, January 25). *Benefits of the interactive healthcare dashboards for hospitals*. https://cprimestudios.com/blog/benefits-interactive-healthcare-dashboards-hospitals

Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room | Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine | Full Text. (n.d.). Retrieved March 5, 2022, from https://sjtrem.biomedcentral.com/articles/10.1186/s13049-019-0629-z

Digital Health: A health system strengthening tool for developing countries. (2021, March 17). Mobile for Development. https://www.gsma.com/mobilefordevelopment/resources/digital-health-a-health-system-strengthening-tool-for-developing-countries/

Explainability for machine learning models in MATLAB. (2020, May 18). *Opti-Num Solutions*. https://optinum.co.za/explainability-for-machine-learning-models-in-matlab/

Hospital Inpatient Discharges (SPARCS De-Identified): 2015 | State of New York. (n.d.). Retrieved March 5, 2022, from https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8

Kim, S.-H., & Song, H. (2022, January 18). How Digital Transformation Can Improve Hospitals' Operational Decisions. *Harvard Business Review*. https://hbr.org/2022/01/how-digital-transformation-can-improve-hospitals-operational-decisions

Odekunle, F. F., Odekunle, R. O., & Shankar, S. (2017). Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. *International Journal of Health Sciences*, *11*(4), 59–64.

Oleribe, O. O., Momoh, J., Uzochukwu, B. S., Mbofana, F., Adebiyi, A., Barbera, T., Williams, R., & Taylor-Robinson, S. D. (2019). Identifying Key Challenges Facing Healthcare Systems In Africa And Potential Solutions . *International Journal of General Medicine*, *12*, 395–403. https://doi.org/10.2147/IJGM.S223882

Organizational factors affecting length of stay in the emergency department: Initial observational study | Israel Journal of Health Policy Research | Full Text. (n.d.). Retrieved April 3, 2022, from https://ijhpr.biomedcentral.com/articles/10.1186/s13584-015-0035-6

Physicians' Ability to Predict Hospital Length of Stay for Patients Admitted to the Hospital from the Emergency Department. (n.d.). Retrieved March 5, 2022, from https://www.hindawi.com/journals/emi/2012/824674/

Raj, S. (2021, December 11). *Effects of Multi-collinearity in Logistic Regression, SVM, Random Forest(RF)*. Medium. https://medium.com/@raj5287/effects-of-multi-collinearity-in-logistic-regression-svm-rf-af6766d91f1b#:%7E:text=Random%20Forest%20uses%20bootstrap%20sampling,different%20set%20of%20data%20points

Riascos, A., & Serna, N. (2017). Predicting Annual Length-Of-Stay and its Impact on Health. *Proceedings of The First Workshop Medical Informatics and Healthcare Held with the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining*, 27–34. https://proceedings.mlr.press/v69/riascos17a.html

Sisense. (2022, March 18). *Dashboard Design Best Practices - 4 Key Principles*. https://www.sisense.com/blog/4-design-principles-creating-better-dashboards/ https://www.sisense.com/blog/4-design-principles-creating-better-dashboards/

Taking on the Challenges of Health Care in Africa | Stanford Graduate School of Business. (n.d.). Retrieved March 5, 2022, from https://www.gsb.stanford.edu/insights/taking-challenges-health-care-africa

Why Is Length of Stay Important? Five Key Reasons | Insights. (n.d.). *ABOUT*. Retrieved March 5, 2022, from https://www.abouthealthcare.com/insights/blog/length-of-stay/

14: Appendix

The Appendix number corresponds to the section number in the report.

Appendix 2.2: Overview of Business Problem and Approach

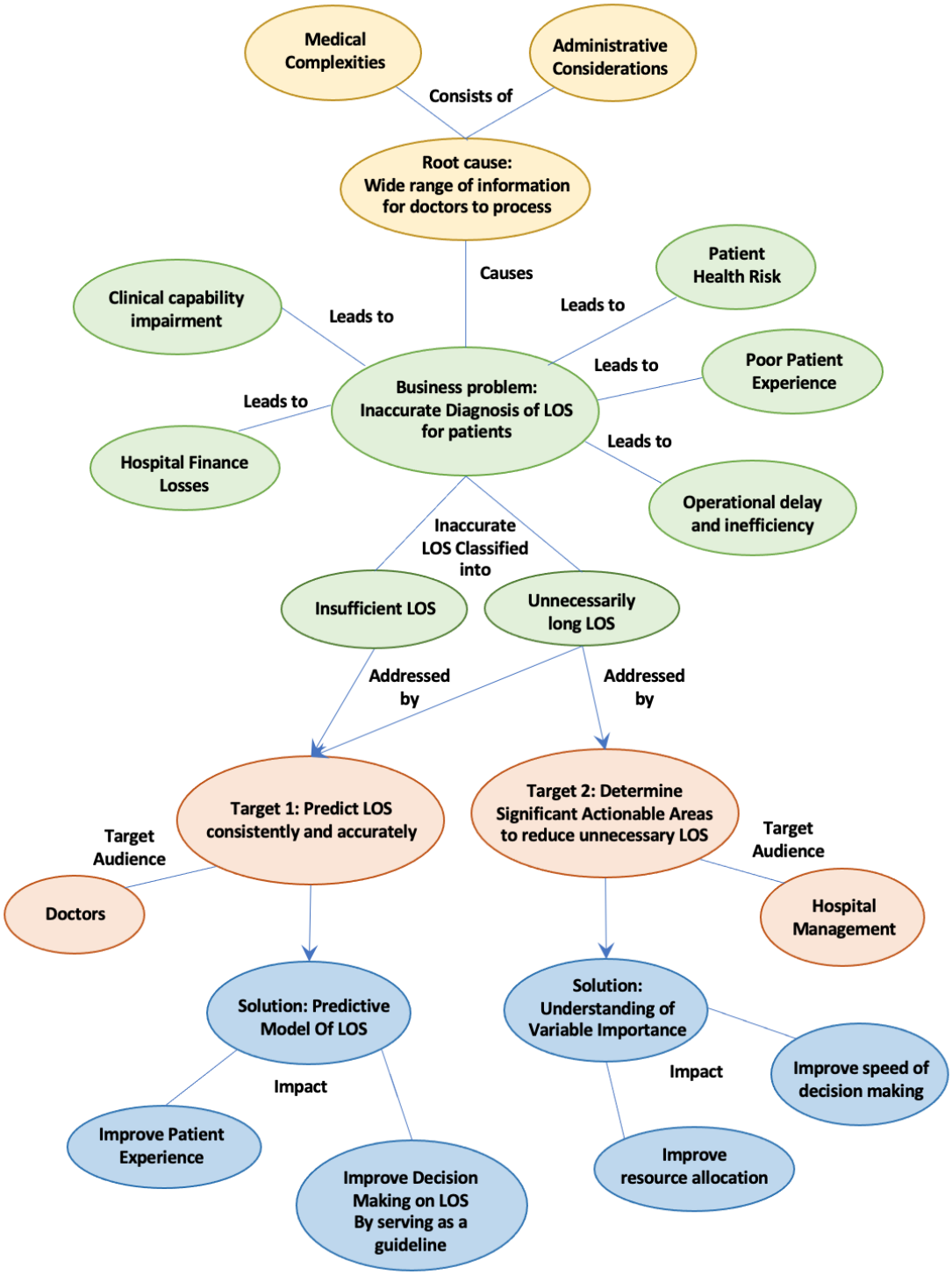


Figure 2.21: Flowchart of Business Problem

[Return to Section 2.2: Business Problem]

Appendix 2.5: Overview of Solution

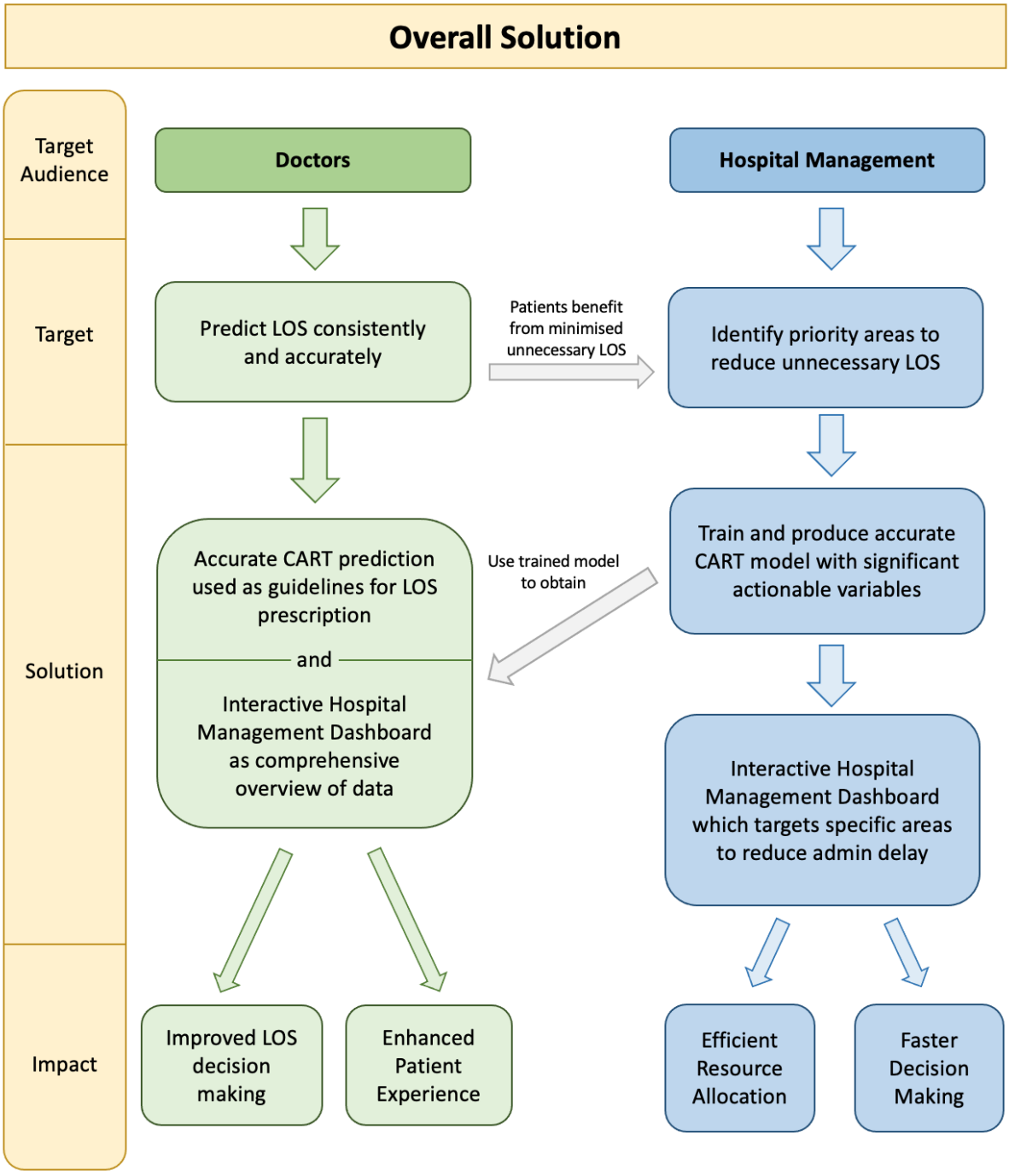


Figure 2: Flowchart of Aims and Solution

[Return to Section 2.5: Project Tasks and Outcomes]
[Return to Section 9.1: Tableau Dashboard]
[Return to Section 11: Summary of Business Opportunity]

Appendix 4.2: Data Dictionary

Data Type	Variable Name	Aspect	Explanation
Cont	*Length_of_Stay*	Clinical, Administrative	Outcome variable. Patient's duration of hospitalisation (in days).
Cont	*Available_Extra_Rooms_in_Ward*	Administrative	Number of rooms that are not fully occupied for the ward allocated to the patient.
Cont	*Admission_Deposit*	Administrative	Deposit amount paid (in South African rand) upon admission into the hospital.
Cont	*Administrative_Delay*	Administrative	Average Delay in administrative processing and certification (in days) experienced by patients in the given ward. Used as a proxy for the delay that this patient would likely face in their stay.
Cont	*Visitors_with_Patient*	Administrative	Number of visitors registered under the patient upon admission. In emergency situations, this number would often be registered by the patient's next-of-kin.
Cont	*Age*	Administrative	Patient's age in years.
Cat	*Type_of_Admission*	Clinical	Category of the patient's preliminary cause of admission (Emergency, Trauma etc.)
Cat	*Ward_Type*	Administrative	Class of ward that the patient is staying in. Each Ward Type differs accordingly in
Cat	*Patient Disposition*	Clinical	Refers to where a patient is being discharged. Categories: Home or Self Care, Home w/ Home Health Services, Skilled Nursing Home, Others
Cat	*Severity_of_Illness*	Clinical	Severity of illness diagnosed by the doctor. Categories: Minor, Moderate, Extreme
Cat	*APR_Risk_of_Mortality*	Clinical	All Patient Refined-Risk of Mortality as an indicator of the patient's likelihood of dying. Categories: Minor, Moderate, Major, Extreme
Cat	*APR_Medical_Surgical_Description*	Clinical	All Patient Refined-Medical-Surgical Description indicates whether the patient primarily undergoes surgical or medical treatment. Categories: Medical, Surgical

Cat	*Payment_Typology*	Administrative	Mode of Payment for Hospital Charges. Certain payment schemes include subsidies. Categories: Medicare, Medicaid, Private Health Insurance etc.

Return to Section 4.2: Data Preparation

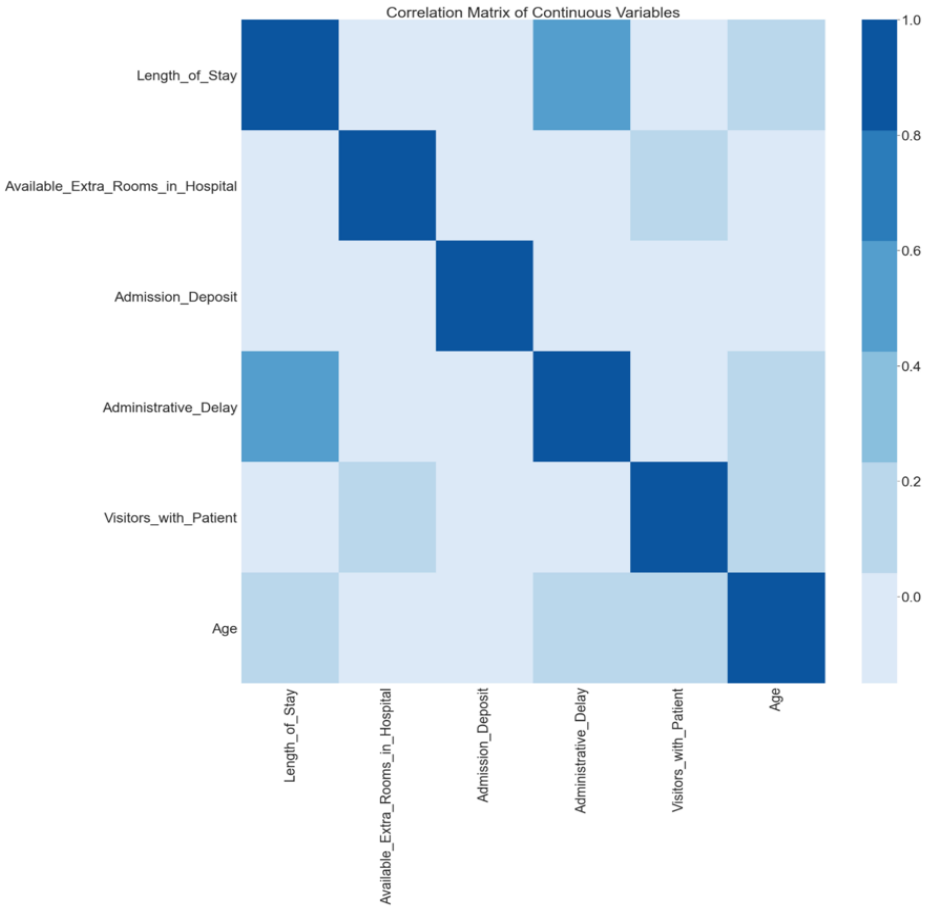
Appendix 4.4: Explanation and importance of predictors

Available extra rooms	• No. of beds available in the room • Bed availability contributes to the prolonged LOS. • https://www.sciencedirect.com/science/article/pii/S2211419X21000124
Admission deposit	• A deposit is normally collected at the time of the patient's admission to cover his/her estimated hospital bill. The amount varies, depending on the type of cases and the classes of ward chosen https://www.sgh.com.sg/patient-care/inpatient-day-surgery/day-of-admission-singapore-general-hospital • A deposit is required at the time of admission. The amount will vary depending on the choice of accommodation, type of operation/procedure, estimated length of stay • https://www.thomsonmedical.com/birth-at-thomson/finance-payment/deposit/
Administrative delays	• The main causes of delayed discharges were faulty organisational management, inadequate discharge planning • https://www.ijhpm.com/article_3844_df28ede79c2e8edf0f1b4cefb8bd05ef.pdf
Visitors with patient	• The number of visitors could help reduce length of stay since it reduces anxiety felt by patients • However, too many visitors may increase length of stay since it might increase the workloads of nurses impeding their ability to provide the best care. • https://www.clinicalcorrelations.org/2018/08/03/do-hospital-visitors-impact-patient-outcomes/

Return to Section 4.4: Train-Test Set Preparation for Machine Learning

Appendix 5.1: Exploratory Analysis

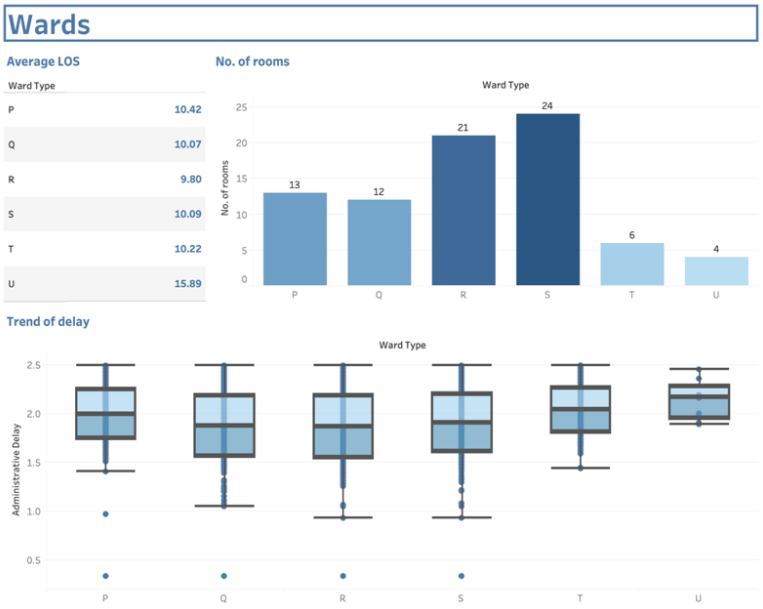
Correlation between variables



Return to Section 5.1: Data exploration of correlation between continuous variables

Appendix 5.3: Additional Exploratory Analysis

Ward type visualisation



The lower the number of available rooms per ward, the length of stay and administrative delay tends to increase.

Further Data Exploration:

Surgical Description	75.5% of the patients were admitted to the hospital due to medical issue while 25.5% of the patients were admitted for surgical operation The average LOS of patients under medical are 9 days while patients undergoing surgical operation stayed for an average of 12.5 days. The increase in LOS might be because patients required a longer time to recover from surgical operation.
Risk of mortality	Half the patient's population has minor risk of mortality There is a trend showing the increase in risk of mortality result in a longer LOS
Age (bin:20)	The average LOS for adults above the age of 40 is generally higher implying that age does play a role in length of stay.
Department	The average LOS for patients across all departments are around 10 days except for patients in the surgery department with a average LOS of 13 days Similar to the surgical description, this could indicate that patients take a longer time to recover from surgery as compared to other departments.
Patient Disposition	Majority of the patients choose home/self-care as their disposition method. Patients dispose to skilled nursing home generally have a higher LOS of 18 days and patients dispose back home/self-care have a lower LOS 7 days. This might be because patients dispose to a nursing home would require more administration process between the hospital and the nursing home which increase the administrative delay which hence results in a higher LOS.
Severity of Illness	There is a trend showing the increase in severity of illness resulting in a longer LOS However, a sharp increase of 15 days can be observed from patients with moderate illness to extreme illness.
Type of admission	Almost half of the patients are admitted due to trauma. Average LOS across all types of admission is roughly about 10 days.
Available extra rooms	There is a decreasing trend of LOS as number of available extra rooms increases
Administrative delay	There is an increasing trend of LOS as administrative delay increases

Return to Section 5.3: Data exploration of Independent Variables

Appendix 6.2: MARS Evaluation

MARS Degree 1

Earth Model

Basis Function	Pruned	Coefficient
(Intercept)	No	-59.8922
Administrative_Delay	No	29.8776
Severity_of_Illness	No	3.08154
Payment_Typology_1_Medicaid	No	1.53382
APR_Risk_of_Mortality	No	1.50742
Patient_Disposition_Skilled Nursing Home	No	4.14387
Payment_Typology_1_Medicare	No	-2.56967
Patient_Disposition_Others	No	1.71616
APR_Medical_Surgical_Description_Medical	No	1.27148

MSE: 175.5937, GCV: 175.6245, RSQ: 0.2654, GRSQ: 0.2653

MARS Degree 2

Earth Model

Basis Function	Pruned	Coefficient
(Intercept)	No	288.922
Administrative_Delay	No	-271.061
Administrative_Delay*Administrative_Delay	No	63.7315
Severity_of_Illness*Administrative_Delay	No	9.45217
Severity_of_Illness	No	-20.4108
Patient_Disposition_Skilled Nursing Home*Severity_of_Illness	No	2.04776
Payment_Typology_1_Medicaid*Severity_of_Illness	No	1.2557
APR_Medical_Surgical_Description_Medical*Administrative_Delay	No	36.9541
APR_Medical_Surgical_Description_Medical*Severity_of_Illness	No	-2.33418
Severity_of_Illness*Severity_of_Illness	No	2.48579
APR_Medical_Surgical_Description_Medical	No	-71.9991
APR_Risk_of_Mortality*APR_Medical_Surgical_Description_Medical	No	-7.53646
APR_Risk_of_Mortality	No	19.0507
APR_Risk_of_Mortality*Administrative_Delay	No	-5.79567
Patient_Disposition_Home or Self Care*APR_Risk_of_Mortality	No	-1.08234
Payment_Typology_1_Medicare*Severity_of_Illness	No	-1.08598

MSE: 148.8784, GCV: 148.9264, RSQ: 0.3772, GRSQ: 0.3770

Log MARS Degree 1

Earth Model

Basis Function	Pruned	Coefficient
(Intercept)	No	-4.56486
Administrative_Delay	No	2.79514
Severity_of_Illness	No	0.147343
APR_Medical_Surgical_Description_Medical	No	0.305183
Patient_Disposition_Skilled Nursing Home	No	0.278822
APR_Risk_of_Mortality	No	0.0844392
Payment_Typology_1_Medicaid	No	0.087837
Payment_Typology_1_Medicare	No	-0.066778

MSE: 0.3417, GCV: 0.3417, RSQ: 0.5317, GRSQ: 0.5316

Log MARS Degree 2

Earth Model

Basis Function	Pruned	Coefficient
(Intercept)	No	13.264
Administrative_Delay	No	-12.384
Severity_of_Illness*Administrative_Delay	No	0.424842
Severity_of_Illness	No	-0.672632
APR_Medical_Surgical_Description_Medical	No	-3.55379
Administrative_Delay*Administrative_Delay	No	3.15502
APR_Medical_Surgical_Description_Medical*Severity_of_Illness	No	-0.166313
APR_Medical_Surgical_Description_Medical*Administrative_Delay	No	1.91807
Patient_Disposition_Skilled Nursing Home	No	0.26886
APR_Risk_of_Mortality*APR_Medical_Surgical_Description_Medical	No	-0.307755
APR_Risk_of_Mortality	No	0.252954
Patient_Disposition_Home w/ Home Health Services*APR_Risk_of_Mortality	No	0.0628701
Payment_Typology_1_Medicaid*Administrative_Delay	No	0.0290302

MSE: 0.2799, GCV: 0.2800, RSQ: 0.6164, GRSQ: 0.6163

MARS Feature Importance by GCV	In [75]:
	`varImpt = mars2.summary_feature_importances(sort_by='gcv')`
	`varImpt = varImpt.splitlines()`
	
	`# varImpt = pd.DataFrame(varImpt)`
	`varImpt`
	
	Out[75]:
	`['                                                gcv',`
	`'Administrative_Delay                                0.66`
	`'APR_Medical_Surgical_Description_Medical             0.15`
	`'APR_Risk_of_Mortality                                 0.12`
	`'Severity_of_Illness                                   0.05`
	`'Patient_Disposition_Skilled Nursing Home              0.01`
	`'Payment_Typology_1_Medicare                           0.01`
	`'Patient_Disposition_Home or Self Care                 0.00`
	`'Payment_Typology_1_Medicaid                           0.00`
	`'Age                                                   0.00`
	`'Ward_Type_S                                           0.00`
	`'Visitors_with_Patient                                 0.00`
	`'Type_of_Admission_Emergency                           0.00`
	`'Type_of_Admission_Trauma                              0.00`
	`'Type_of_Admission_Urgent                              0.00`
	`'Ward_Type_P                                           0.00`
	`'Admission_Deposit                                     0.00`
	`'Ward_Type_Q                                           0.00`
	`'Ward_Type_R                                           0.00`
	`'Ward_Type_U                                           0.00`
	`'Ward_Type_T                                           0.00`
	`'Patient_Disposition_Others                            0.00`
	`'APR_Medical_Surgical_Description_Surgical            0.00`
	`'Payment_Typology_1_Blue Cross/Blue Shield             0.00`
	`'Payment_Typology_1_Department of Corrections          0.00`
	`'Payment_Typology_1_Federal/State/Local/VA             0.00`
	`'Payment_Typology_1_Managed Care, Unspecified          0.00`
	`'Payment_Typology_1_Miscellaneous/Other                0.00`
	`'Payment_Typology_1_Private Health Insurance           0.00`
	`'Payment_Typology_1_Self-Pay                           0.00`
	`'Payment_Typology_1_Unknown                            0.00`
	`'Patient_Disposition_Home w/ Home Health Services      0.00`
	`'Available_Extra_Rooms_in_Hospital                    0.00`

Summary of Model Accuracy

	Model	RMSE	Normalised RMSE
0	MARS Degree 1	13.396318	7.483976
1	MARS Degree 2	12.309775	6.876969
2	Log MARS Degree 1	0.582010	12.918202
3	Log MARS Degree 2	0.525214	11.657563
4	MARS Degree 1 (Extreme)	5.533991	3.091615
5	MARS Degree 2 (Extreme)	4.959099	2.770447
6	Log MARS Degree 1 (Extreme)	0.495330	10.994253
7	Log MARS Degree 2 (Extreme)	0.463305	10.283442
8	MARS Degree 1 (Moderate)	16.323869	9.119480
9	MARS Degree 2 (Moderate)	15.370330	8.586776
10	Log MARS Degree 1 (Moderate)	0.605260	13.434250
11	Log MARS Degree 2 (Moderate)	0.561226	12.456870
12	MARS Degree 1 (Minor)	9.176864	5.126740
13	MARS Degree 2 (Minor)	8.616500	4.813687
14	Log MARS Degree 1 (Minor)	0.531139	11.789072
15	Log MARS Degree 2 (Minor)	0.495780	11.004250

Return to Section 6.2: MARS model
Return to Section 8.1: Medical + Actionable Variables

Appendix 6.3: CART Evaluation

CART Trained using Original Dataset Results	Train Set Errors Root Mean Squared Error: 11.82817 Normalized Root Mean Squared Error: 6.608% Test Set Errors Root Mean Squared Error: 12.12861 Normalized Root Mean Squared Error: 6.776%
CART Trained using Log Dataset Results	Train Set Errors Root Mean Squared Error: 0.55191 Normalized Root Mean Squared Error: 12.25% Test Set Errors Root Mean Squared Error: 0.55157 Normalized Root Mean Squared Error: 12.243%

Return to Section 6.3: CART model

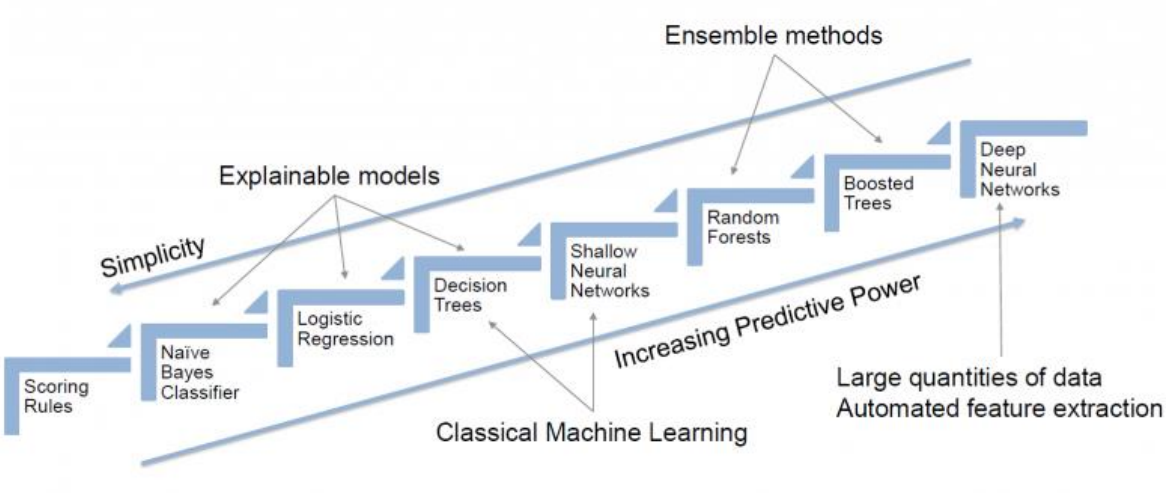
Appendix 6.4 Random Forest Cross Validation

Random Forest optimal Parameters (Top 15 out of 75 in excel sheet)

	mean_fit_	std_fit_tin	mean_sco	std_score_	param_ma	param_ma	param_n_	params	split0_test	split1_test	split2_test	split3_test	split4_test	split5_test	split6_test	split7_test	split8_test	split9_test	mean_test
0	2.02271	0.05467	1.00012	0.00975	10	500	500	{'max_feat	-139.698	-153.505	-146.248	-153.911	-142.545	-142.583	-146.257	-151.163	-162.173	-145.769	-148.385
1	4.39116	0.45292	2.16387	0.16787	10	500	1000	{'max_feat	-139.344	-153.523	-146.566	-153.729	-142.552	-142.546	-146	-151.197	-162.281	-145.706	-148.345
2	6.47872	0.33567	3.28377	0.1464	10	500	1500	{'max_feat	-139.229	-153.564	-146.518	-153.809	-142.434	-142.674	-145.946	-151.145	-162.253	-145.717	-148.329
3	8.68561	0.40619	4.44724	0.13286	10	500	2000	{'max_feat	-139.153	-153.624	-146.354	-153.746	-142.272	-142.68	-145.88	-151.15	-162.304	-145.622	-148.278
4	10.345	0.75196	5.19682	0.23635	10	500	2500	{'max_feat	-139.155	-153.63	-146.37	-153.693	-142.303	-142.618	-145.875	-151.137	-162.486	-145.669	-148.293
5	3.0542	0.19441	1.23521	0.11562	10	1000	500	{'max_feat	-136.782	-150.855	-143.642	-150.973	-139.753	-140.414	-143.328	-148.779	-159.685	-142.703	-145.691
6	6.08986	0.25438	2.53652	0.1516	10	1000	1000	{'max_feat	-136.441	-150.738	-143.571	-150.836	-139.449	-140.236	-143.126	-148.334	-159.4	-142.454	-145.459
7	9.03068	0.31007	3.56075	0.11255	10	1000	1500	{'max_feat	-136.318	-150.752	-143.417	-150.667	-139.364	-140.166	-143.055	-148.289	-159.371	-142.707	-145.411
8	11.7658	0.57776	4.74469	0.18553	10	1000	2000	{'max_feat	-136.263	-150.782	-143.311	-150.728	-139.417	-140.156	-142.975	-148.238	-159.242	-142.717	-145.383
9	14.9427	0.76106	5.89319	0.2393	10	1000	2500	{'max_feat	-136.237	-150.686	-143.388	-150.681	-139.472	-140.091	-143.009	-148.24	-159.237	-142.761	-145.38
10	4.71453	0.12051	1.42175	0.08674	10	2000	500	{'max_feat	-134.46	-148.647	-141.768	-148.538	-137.808	-138.638	-141.859	-146.63	-157.657	-140.74	-143.674
11	9.39452	0.24488	2.9699	0.29652	10	2000	1000	{'max_feat	-134.361	-148.635	-141.593	-148.449	-137.738	-138.545	-141.471	-146.193	-157.381	-140.408	-143.477
12	14.0095	0.34832	4.30698	0.26674	10	2000	1500	{'max_feat	-134.483	-148.587	-141.492	-148.313	-137.554	-138.581	-141.18	-146.216	-157.384	-140.548	-143.434
13	19.1666	0.84367	5.76051	0.43629	10	2000	2000	{'max_feat	-134.358	-148.476	-141.434	-148.31	-137.617	-138.499	-141.126	-146.145	-157.285	-140.595	-143.385
14	24.5097	0.93503	7.55617	0.78707	10	2000	2500	{'max_feat	-134.233	-148.493	-141.378	-148.288	-137.545	-138.408	-141.173	-146.088	-157.287	-140.561	-143.345
15	7.07938	0.2998	1.70055	0.16939	10	3000	500	{'max_feat	-133.97	-148.065	-140.641	-147.624	-136.967	-137.855	-140.77	-145.063	-156.211	-139.95	-142.712

Return to Section 6.4: Random Forest model

Appendix 7.2: Explainability of Machine Learning Models



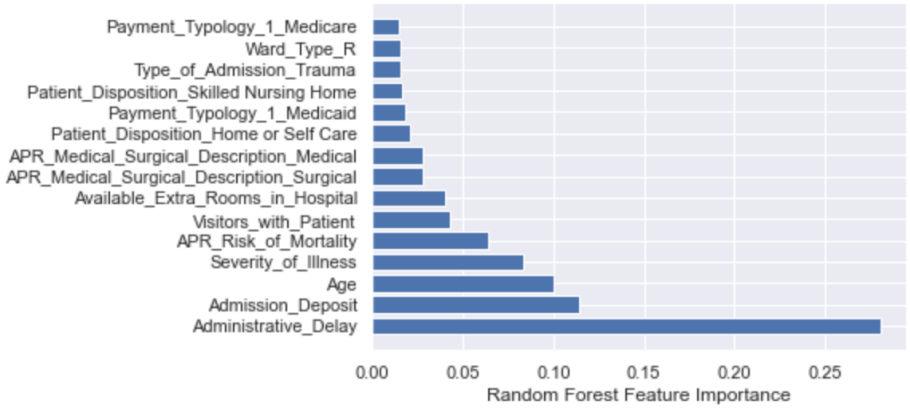
Return to Section 7.2: Explainability

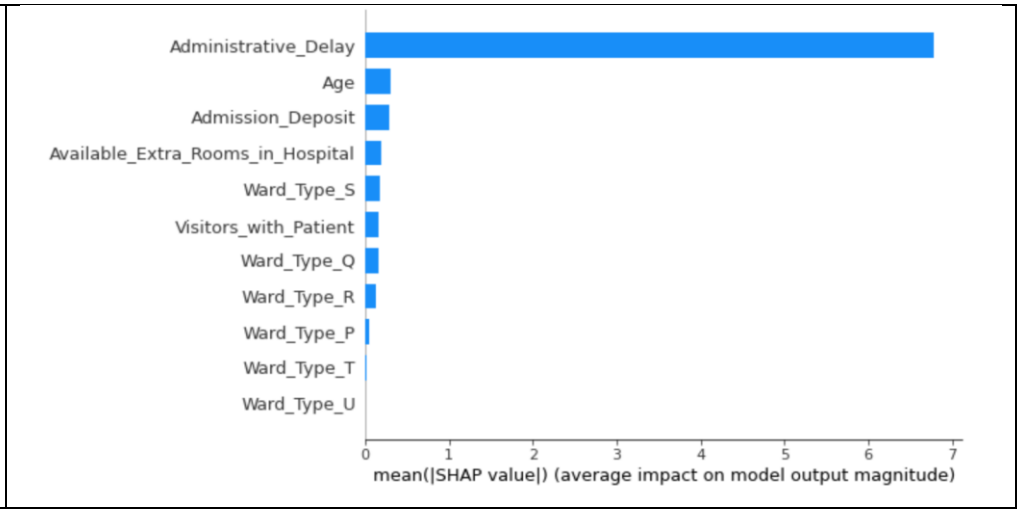
Appendix 8.2.1. Shortlisted Actionable Variables for Further Analysis

Variable	Possible Justification
Administrative Delay	Administrative Delay results from inefficiencies within the hospital, such as insufficient manpower to care for patients or process documents. If deemed significant, the hospital may choose to use it as a measure for internal bottlenecks and operational inadequacies.
Admission Deposit	Admission Deposit is estimated based on the type of cases and classes of wards chosen. If deemed significant, the hospital may choose to relook the categorisation of Admission Deposit.
Age	If deemed significant, the hospital management may want to investigate the significance of patient age groups and consider segmenting their inpatient wards around age.
Available Extra Rooms in Hospital	The number of available rooms could be important to the quality of care experienced by patients. Hospital management may want to explore alternative segmentation for their wards to improve resource efficiency.
Visitors with Patient	Hospital management could revise their visitor policy if the number of visitors holds a significant influence over LOS.
Ward Type	If deemed significant, hospitals could directly control how patients are segmented into their wards and possibly categorise patients into wards by LOS.

Return to Section 8.2: Actionable Variables

Appendix 8.2.2. Other Variable importance of models

Random Forest Original Dataset	

Random Forest Feature importance Permuted	

[Return to Section 8.1: Medical + Actionable Variables]
[Return to Section 8.2: Actionable Variables]

Appendix 9.1: Dashboards

Appendix 9.1.1: Hospital dashboards:

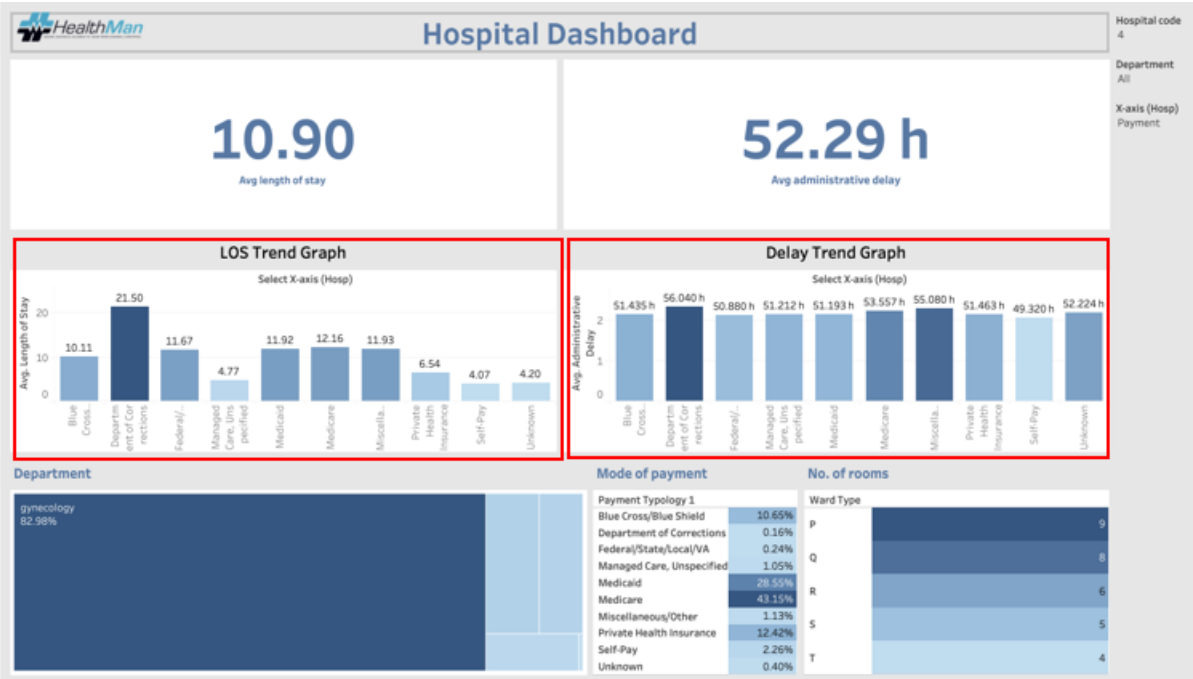
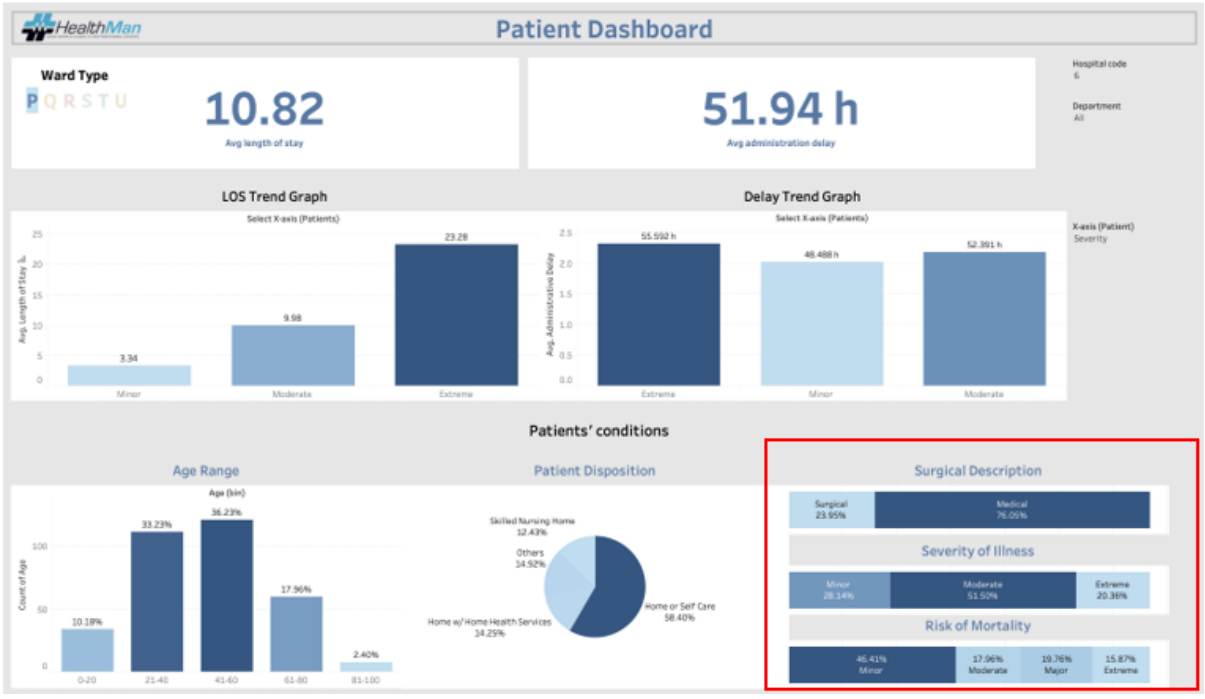
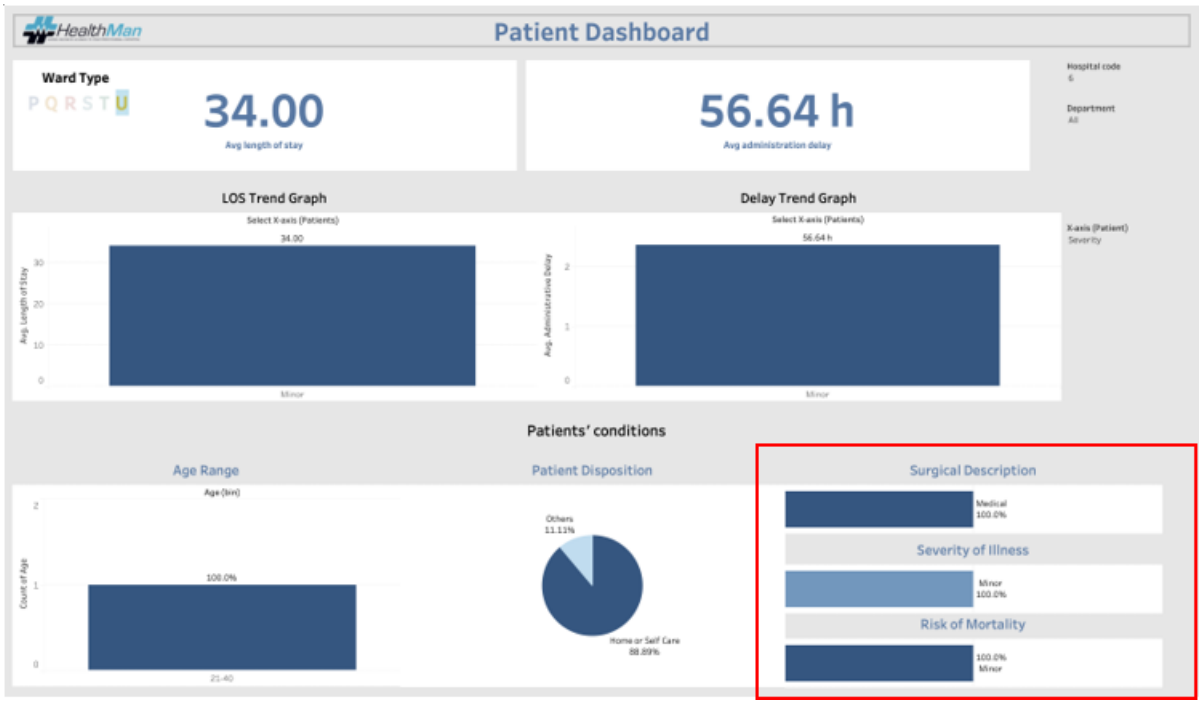


Figure 9.1.1: Hospital Dashboard Model

[Return to Section 9.1.1: Hospital Dashboards]

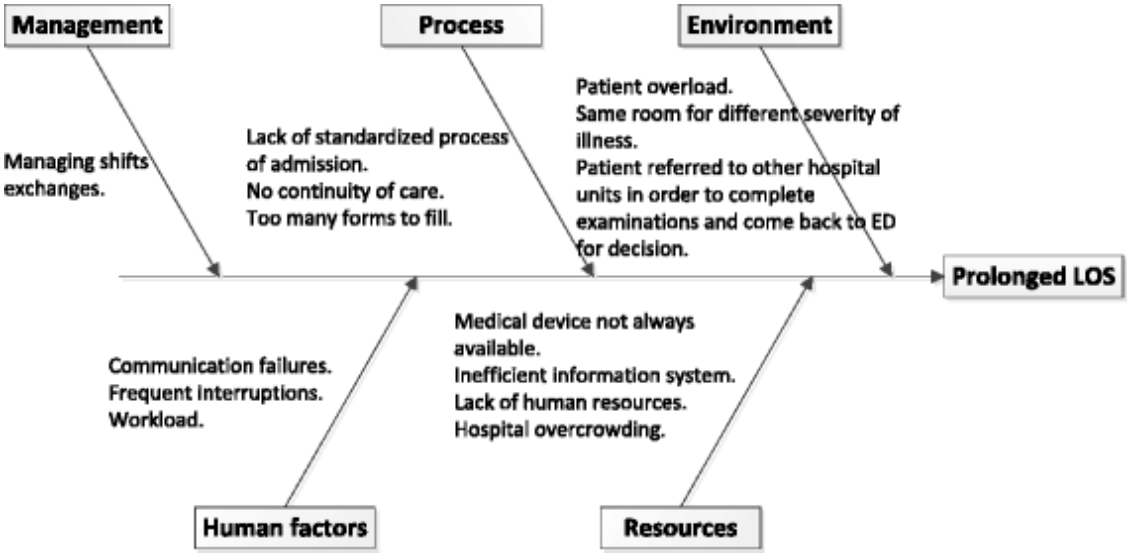
Appendix 9.1.2: Patient dashboards:





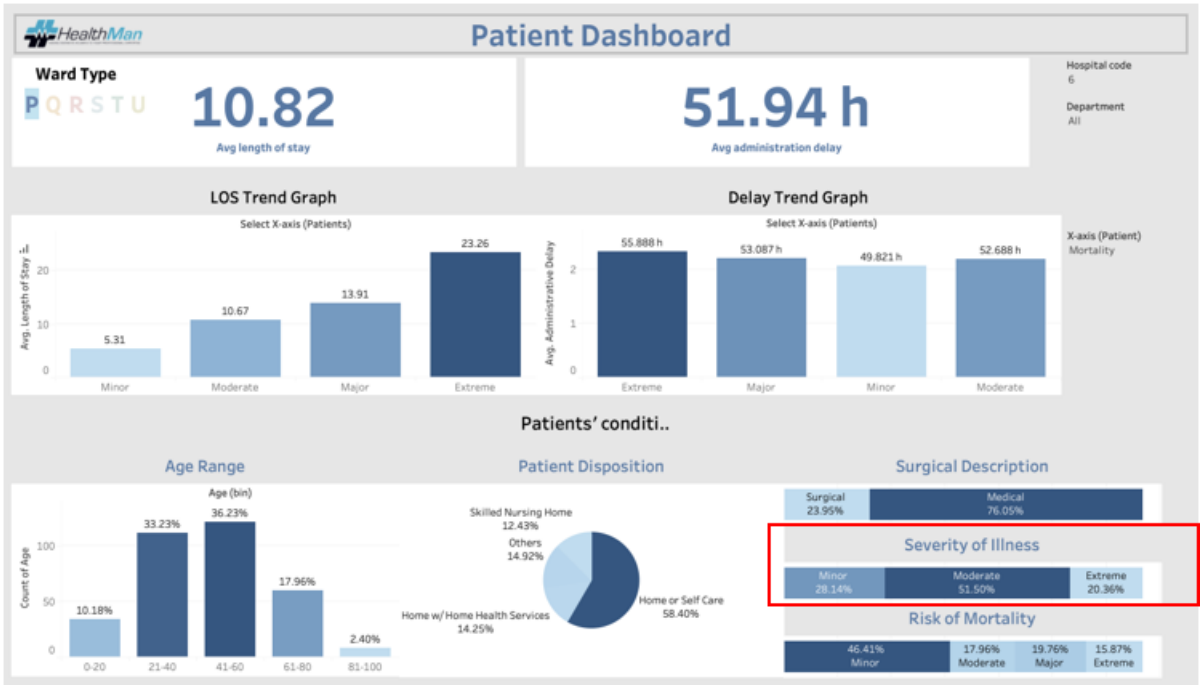
Return to Section 9.1.2: Patient Dashboards

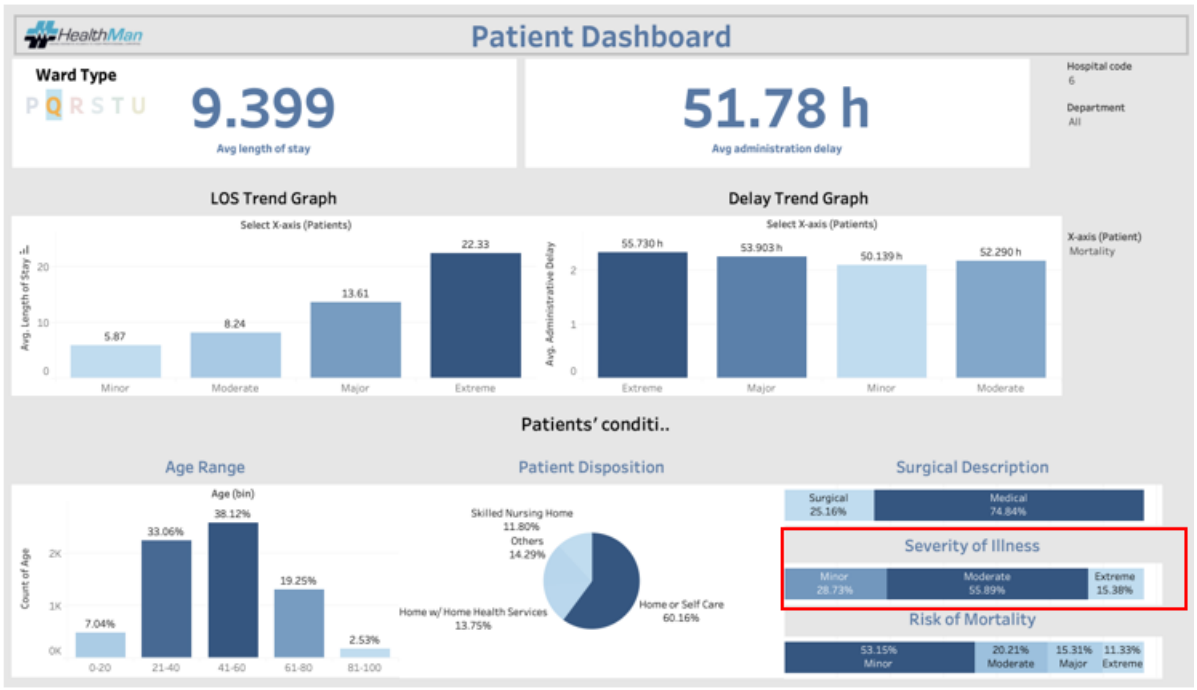
Appendix 10.1.1: Recommendations

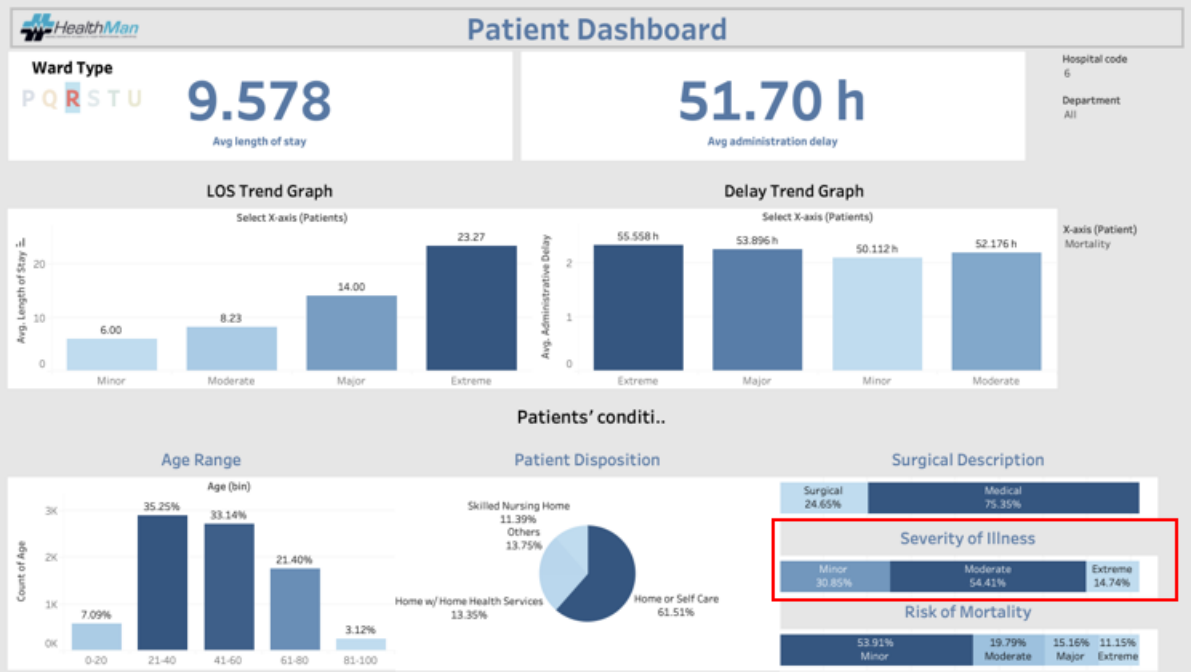


Return to Section 10.1: Recommendations

Appendix 10.1.2: Recommendations – Patient Segmentation in Wards







Return to Section 10.1: Recommendations

Appendix 12.1: Limitations – Severity Split Analysis

Appendix 12.1a: Segmented Analysis on Severity of Illness

Given that a patient's Length of Stay is often tied to the severity of their case, we further investigate the predictive accuracy of each model across individual *Severity of Illness.*

Model	Severity of Illness		
	Minor	Moderate	Extreme
MARS Degree 1 (Original Dataset)	4.250%	7.678%	2.863%
MARS Degree 2 (Original Dataset)	3.979%	6.086%	2.505%
MARS Degree 1 (Log Dataset)	11.161%	10.999%	10.930%
MARS Degree 2 (Log Dataset)	10.435%	10.525%	10.175%

Figure 12.1a: Comparison of Normalised RMSE for differing Severity Cases

Figure 12.1a suggests that running MARS models separately on datasets split accordingly Severity of Illness further helped to reduce Normalised RMSE and improve prediction accuracy.

Appendix 12.1b: CART Segmented Analysis on Severity of Illness

Given that a patient's Length of Stay is often tied to the severity of their case, we further investigate the predictive accuracy of each model across individual *Severity of Illness.*

Model	Severity of Illness		
	Minor	Moderate	Extreme
CART Optimal Tree (Original Dataset)	4.643%	8.556%	2.558%
CART Optimal Tree (Log Dataset)	10.993%	12.426%	10.092%

Figure 12.1b: Comparison of Normalised RMSE for differing Severity Cases

Figure 12.1b: suggests that running CART models separately on datasets split according to Severity of Illness also helped to reduce Normalised RMSE and improve prediction accuracy.

Appendix 12.1c: Random Forest Segmented Analysis on Severity of Illness

Given that a patient's Length of Stay is often tied to the severity of their case, we further investigate the predictive accuracy of each model across individual *Severity of Illness*.

Model	Severity of Illness		
	Minor	Moderate	Extreme
Optimal Random Forest (Original Dataset)	6.163%	8.509%	2.553%

Figure 12.1c:Comparison of Normalised RMSE for differing Severity Cases

Figure 12.1c suggests that running Random Forest models separately on datasets split according to Severity of Illness also helped to reduce Normalised RMSE and improve prediction accuracy.

Return to Section 12.1: Challenges/Limitations