EBA5005

Graduate Certificate –

Specialized Predictive Modelling and Forecasting:

Report

Predicting ICU Mortality and ICU Patients' Length of Stay
using Survival Analysis and White-Box Classifiers

Team Chocolates

Chia Har Teck Alvin – A0027846X

Khoo May Sze – A0198537L

Lim Zhengyi Andy – A0198433W

Ong Kian Eng – A0052270U

Sun Yixuan – A0198428M

11 April 2020

Contents

1.0 <u>Business Understanding</u>

Intensive Care Unit (ICU) patients are critically ill and are continuously monitored for medical abnormalities, disease prognosis and potential complications. ICU patients who are considered highly vulnerable, tend to require more specialized care and are time-sensitive to medical responses by doctors and healthcare professionals.

The demand for ICU hospital beds worldwide has been increasing, with ICU costs having risen to nearly 22% of hospital costs and 5% of the total healthcare cost. Despite the increasing demand for ICU hospital beds worldwide, the availability of hospital beds (less than 10% of hospital beds), medical staff and equipment remain limited resources (Bhattacharya et al., 2017; Ghanvatkar et al., 2019). Decisions made by the management of ICU wards in hospitals can directly impact a patient's survival rate. With limited manpower, equipment, supplies, and bed/ward availability in hospitals, resource allocation is often an issue that hospitals face.

Further, there exists challenges in providing ICU patients with timely intervention as they require more individualized care and attention. Healthcare professionals often use case-based reasoning, relying on knowledge accumulated from similar past medical cases to diagnose and treat their current patient (Morid et al., 2017). As such, the sharing of knowledge and understanding of key risk factors are more likely to help increase the speed and effectiveness of diagnosis for a new patient.

Hospitals currently utilize severity score systems such as SOFA, APACHE to predict ICU mortality. However, these criteria / rubric-based score systems which bin and compare different strata of patient populations have limited predictive accuracy at the individual level (Morid et al., 2017).

Advances in machine learning methodologies have reported classification results with accuracy of above 80% (Johnson et al., 2017). They employ classifiers such as Neural Network and Random Forest which are regarded by healthcare professionals as black boxes. These models do not provide clear links between the input model features and output clinical event (i.e. significant features may not be identified), and hence are less likely to effectively aid in treatment intervention because doctors and healthcare professionals do not know which health indicators to look out for in patients (Tan et al., 2019; Sadeghi et al., 2018; Luo et al., 2016).

This project attempts to build a model which can clearly pinpoint key features (health indicators) that can better predict individual patient's in-hospital mortality. By understanding the type of patients who survived and those who do not, the hospital can better allocate available resources and attention to where care is needed most. This can also be aimed at helping doctors and healthcare professionals plan accurate and effective timely intervention, tailoring specific level of care for an individual ICU patient.

1.1 Project Objectives

Therefore, the objectives of this project are:

- To contribute and fill the gap in existing literature by associating risk factors (of patients) that impact the length of stay of patients who passed away in hospital, as well as those health indicators who survived using Survival Analysis (i.e. Cox Proportional Hazards (CPH) Model)
- To identify significant features (health indicators) that best predict survival outcome (feature selection using white box models such as Logistic Regression and Decision Tree) which will then be used as inputs for classification models (Logistic Regression, Support Vector Classification, Decision Tree, and Random Forest) to predict the survival outcome

The practical implications this project aims to attain for hospitals, doctors and other healthcare professionals are:

- This project aims to help a hospital better manage their ICU wards. The hospital will be able to utilise the length of stay to forecast the amount of resources required (i.e. demands of ventilators or beds/wards utilization rate). As the dataset provided is an aggregation from various hospitals, and the inventory and resource list are not available, the specifics of being able to provide numeric estimates for hospital cannot be further discussed at this stage.
- Doctors and nurses can monitor the relevant health indicators more closely and respond accordingly to render effective and timely medical support to patients with deteriorating medical conditions (e.g. increasing frequency of monitoring health conditions from once hourly to thrice hourly), hence increasing the patients’ survival

2.0 Data Understanding

The dataset is from PhysioNet Challenge 2012, and is already split into train, test and unseen sets – sets A, B and C respectively. There are 4,000 records out of 12,000 in train set A, another 4,000 in test set B. These are made available to participants who entered the competition in 2012. The unseen set C, consisting of another 4,000 records that are previously unpublished during the competition period, is used by the competition organisers to run final scoring and outcomes. These records are of patients with ICU stays of at least 48 hours. While there is a new ICU mortality challenge set released by PhysioNet in 2020[1], the 2012 dataset is used for this project because it allows us to perform feature engineering and our models have benchmark models for us to evaluate against, further to that, there is also an opportunity to utilize Survival Analysis.

This dataset is used to develop methods for patient-specific prediction of in-hospital mortality. There is a maximum of 41 variables (5 on admission, 36 time-series) collected during the first two days of an ICU stay to predict which patients survive their hospitalizations, and which patients do not. Not all patients have complete records for each of the variables, especially the time-based observations[2]. Table 2.1 lists the 41 variables provided in the dataset.

There are 5 variables collected on admission (time-stamp at 0 hour) which identify a patient's demographic such as age, gender and height. This includes the target variable, *In-hospital-death*. There are 36 time series variables which indicate a patient's health status and can have multiple observations per variable. Each observation is associated with a time-stamp (in hours and minutes) indicating the elapsed time of the observation since ICU admission. For example, a time stamp of 35:19 indicates that the observation was made 35 hours and 19 minutes after the patient was admitted to the ICU. The variable *Weight* belongs to both categories of variables because it is recorded on admission as well as measured hourly, for estimating fluid balance.

[1] https://physionet.org/content/widsdatathon2020/1.0.0/

[2] This is further discussed in Chapter 3.2.2 – Missing Values.

Table 2.1 Definitions of Variables – Predictive Models

Variables	Definitions
	Target Variable
In-hospital-death	Indicator variable where 1 indicates that the patient has died in-hospital, and 0 otherwise
	Independent Variables
Patient Demographics	
Age	Number of years
Gender	Indicator variable where 1 indicates male, and 0 indicates female
Height	Centimetres (cm)
Weight	Kilograms (kg)
ICUType	Categorical variable where 1 indicates a patient is warded to the Coronary Care unit, 2 indicates Cardiac Surgery Recovery unit, 3 indicates Medical ICU, and 4 indicates Surgical ICU
Health Indicators (Time-Series)	
Albumin	g/dL
ALP	Alkaline phosphatase (IU/L)
ALT	Alanine transaminase (IU/L)
AST	Aspartate transaminase (IU/L)
Bilirubin	mg/dL
BUN	Blood urea nitrogen (mg/dL)
Cholesterol	mg/dL
Creatinine	Serum creatinine (mg/dL)
DiasABP	Invasive diastolic arterial blood pressure (mmHg)
FiO2	Fractional inspired O_2 (0 to 1)
GCS	Glasgow Coma Score (3 to 15)
Glucose	Serum glucose (mg/dL)
HCO3	Serum bicarbonate (mmol/L)
HCT	Hematocrit (%)
HR	Heart rate (bpm)
K	Serum potassium (mEq/L)
Lactate	mmol/L
Mg	Serum magnesium (mmol/L)
MAP	Invasive mean arterial blood pressure (mmHg)

MechVent	Indicator variable where 1 indicates presence of mechanical ventilation respiration, and 0 otherwise
Na	Serum sodium (mEq/L)
NIDiasABP	Non-invasive diastolic arterial blood pressure (mmHg)
NIMAP	Non-invasive mean arterial blood pressure (mmHg)
NISysABP	Non-invasive systolic arterial blood pressure (mmHg)
PaCO2	Partial pressure of arterial CO_2 (mmHg)
PaO2	Partial pressure of arterial O_2 (mmHg)
pH	Arterial pH (0 to 14)
Platelets	(cells/nL)
RespRate	Respiration rate (bpm)
SaO2	O_2 saturation in hemoglobin
SysABP	Invasive systolic arterial blood pressure (mmHg)
Temp	Temperature in degrees Celsius
TropI	Troponin-I (µg/L)
TropT	Troponin-T (µg/L)
Urine	Urine output (mL)
WBC	White blood cell count (cells/nL)

2.1 Exploratory Data Analysis (EDA)

Data exploration of the dataset helps to reveal any trends and insights of the features. The training dataset, Set A, contains 4000 patients.

2.1.1 Patient Demographics – Age

The age of patients ranges from 15 to 90 years old. As there is only 1 patient who has yet reached adult age (i.e. below 20 years old), this patient will be included in the study. It can be observed that most of the patients are in the age group >60 years old.

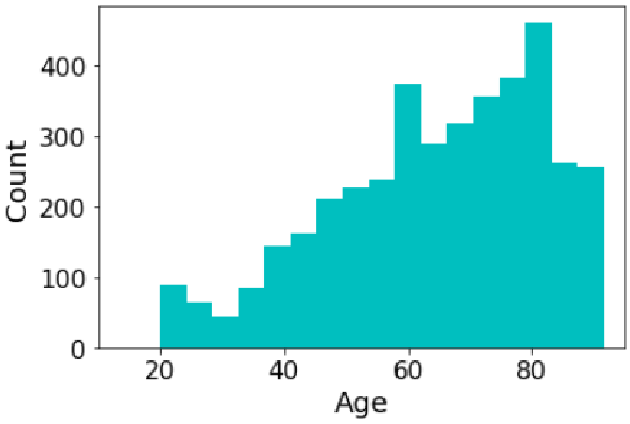


Figure 2.1 Age of Patients

2.1.2 Patient Demographics – Gender

Table 2.2 documents 3 patients with gender -1 (i.e. unknown) in Set A, of which 2 of them have missing data for either height and/or weight. This prevents the possibility of predicting the missing gender. Predicting genders using features (e.g. creatinine or haematocrit) that have different ranges for different genders is also inconclusive. These patients are thus removed from the dataset.

Table 2.2 Participants with Missing Gender Values
The different range of values for males and females are italicized

Patient ID	**Creatinine** *0.7 to 1.3 Male* *0.5 to 1.1 Female*	**HCT** *41 to 51 Male* *36 to 47 Female*	**Height**	**Weight**
135757	1.2 to 3.3 *Male?*	36 to 50.1 *Not conclusive*	180.3	123
137392	1.0 to 1.1 *Not conclusive*	40.3 to 42.9 *Not conclusive*	-1 (Unknown)	-1 (Unknown)
141486	0.3 to 0.4 *Female?*	30.4 to 31.1 *Female?*	-1 (Unknown)	56.7

Figure 2.2 depicts the final gender count of patients in train Set A. There are 2,246 males and 1,751 females.

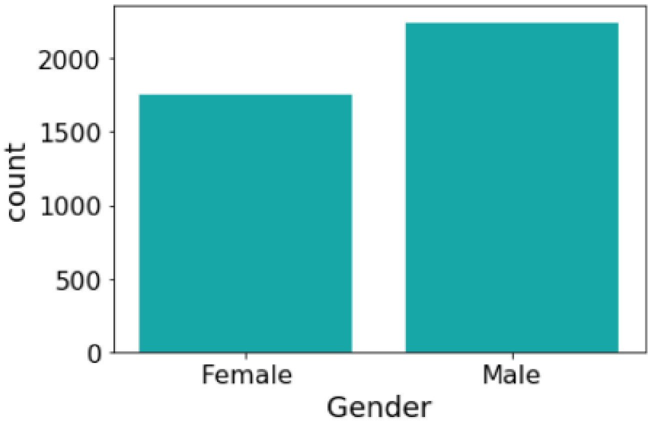


Figure 2.2 Gender Count

2.1.3 Target Variable – *In-hospital death*

Figure 2.3 illustrates the class imbalance of the train set A with 13.9% (553 out of 3997) in-hospital deaths.

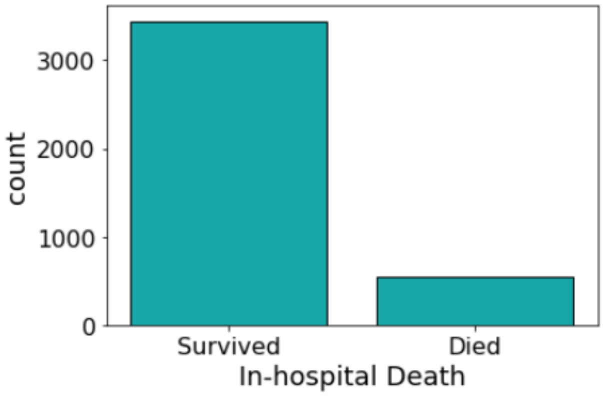


Figure 2.3 Target Variable – In-hospital Death

Table 2.3 shows the proportion of patients who survived and died in the three different datasets at the end of data pre-processing. It also notes that the class imbalance is similar in the test and unseen sets.

Table 2.3 Proportion of Patients who Survived and Died

Outcome (Value)	Train	Test	Unseen
Survived (0)	3444 (86%)	3427 (86%)	3411 (85%)
Died (1)	553 (14%)	568 (14%)	585 (15%)

2.1.4 ICU Type

The variable *ICUType* records 4 different categories. It is noted that a better understanding of these categories can help in the analysis of results and its potential relevance to business applications. Critically ill patients warded in ICU require close supervision and monitoring at various levels of attention from the healthcare professionals. By separating the Medical and Surgical ICU patients, hospitals can make better decisions regarding personnel and bed resources by channelling delicate resources according to the type of care and supervision these patients require.

Figure 2.4 shows the number of patients in each ICU type for train set A, stratified by the number of in-hospital deaths. A large number of patients is observed for Medical ICU wards which treat a large varying range of illnesses and critical conditions - from lung problems, gastrointestinal problems to blood infections. Table 2.4 also notes that its mortality rate is also the highest among all ICU Types at 19%. The second larger group of patients belong to Surgical ICU which wards treat patients who recently had surgery or could potentially need surgery.

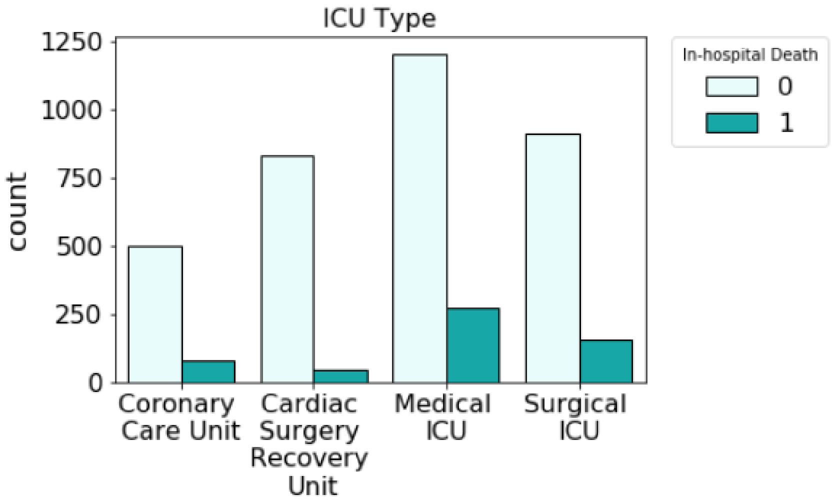


Figure 2.4 In-hospital Death for Different ICU Types

Table 2.4 Mortality Rate for ICU Types

ICU Type	Survived	Died
Coronary Care Unit	319 (84.2%)	60 (15.8%)
Cardiac Surgery Recovery Unit	541 (94.6%)	31 (5.4%)
Medical ICU	790 (80.9%)	186 (19.1%)
Surgical Unit	577 (86.0%)	94 (14.0%)

2.2 Correlation Analysis

Multi-collinearity issues among continuous variables can lead to unstable regression models that are not as effective with the inclusion of highly correlated variables. It can also impact the interpretation of results which may then lead to wrong analysis and implementation of the results. A correlation analysis is undertaken to identify which continuous variables are highly correlated ($r > 0.7$)[3]:

- *ALT* and *AST* – two different enzymes in the blood are highly correlated ($r = 0.83$)
- Mean Arterial Blood Pressure (*MeanMAPfirst*) and Mean Diastolic Arterial Blood Pressure (*MeanDiasABPfirst*) are highly correlated ($r = 0.75$)
- Mean Arterial Blood Pressure (*MeanMAPfirst*) and Mean Systolic Arterial Blood Pressure (*MeanSysABPfirst*) are highly correlated ($r = 0.70$)

Even though *ALT* and *AST* are highly correlated, these variables cannot be removed as they perform different functions in the cells. These enzymes are released into the bloodstream when there is liver damage (Mayo Clinic, 2019).

As for blood pressure measurements, *MeanSysABPfirst* (blood pressure when heart beats) and *MeanDiasABPfirst* (blood pressure during resting phase between heartbeats) are retained as they measure different phases of a heartbeat instead of *MeanMAPfirst* which has been derived from the combination of diastolic and systolic blood pressure readings (Heart.org, 2017).

[3] See Appendix B.

3.0 **Project Approach**

The flowcharts in Figures 3.1 and 3.2 depict the overall data pre-processing and analytical approaches undertaken for this project. The advanced analytical techniques taught within the module have been taken into consideration and applied accordingly.

The project objectives of identifying significant features are achieved using 2 methodologies:

- binary classification approach (i.e. Logistic Regression, Decision Tree)
- time-to-event analysis (i.e. Survival Analysis).

The binary classification approach uses *in-hospital death* as the only outcome variable, whereas Survival Analysis allows us to examine *in-hospital death* whilst considering the *length of stay*. By taking into account time-to-event (in this case, time to in-hospital death), Survival Analysis (using CPH model) can help enhance understanding of the features that influence the length of stay in ICU ward. There is a statistically significant difference between the length of stay of patients who died and those who did not die in the hospital (*t-statistic* = 2.177, *p-value* = 0.03 < 0.05). Hence, by understanding the features that influence extended ICU stays, hospitals can also better manage resources when they ward these patients.

Therefore, there are 2 target variables:

1. *In-hospital-death* (Binary Classification)
2. *Length of stay* (Survival Analysis)

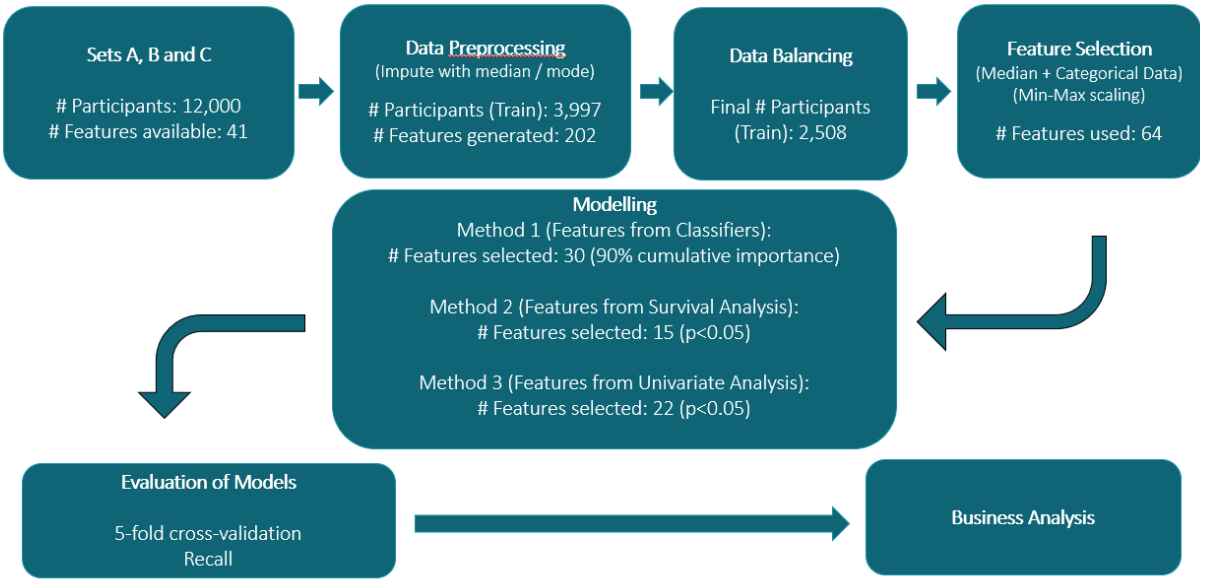

Sets A, B and C
# Participants: 12,000
# Features available: 41
Data Preprocessing
(Impute with median / mode)
# Participants (Train): 3,997
# Features generated: 202
Data Balancing
Final # Participants
(Train): 2,508
Feature Selection
(Median + Categorical Data)
(Min-Max scaling)
# Features used: 64
Modelling
Method 1 (Features from Classifiers):
# Features selected: 30 (90% cumulative importance)
Method 2 (Features from Survival Analysis):
# Features selected: 15 (p<0.05)
Method 3 (Features from Univariate Analysis):
# Features selected: 22 (p<0.05)
Evaluation of Models
5-fold cross-validation
Recall
Business Analysis

Figure 3.1 Project Flow – Binary Classification

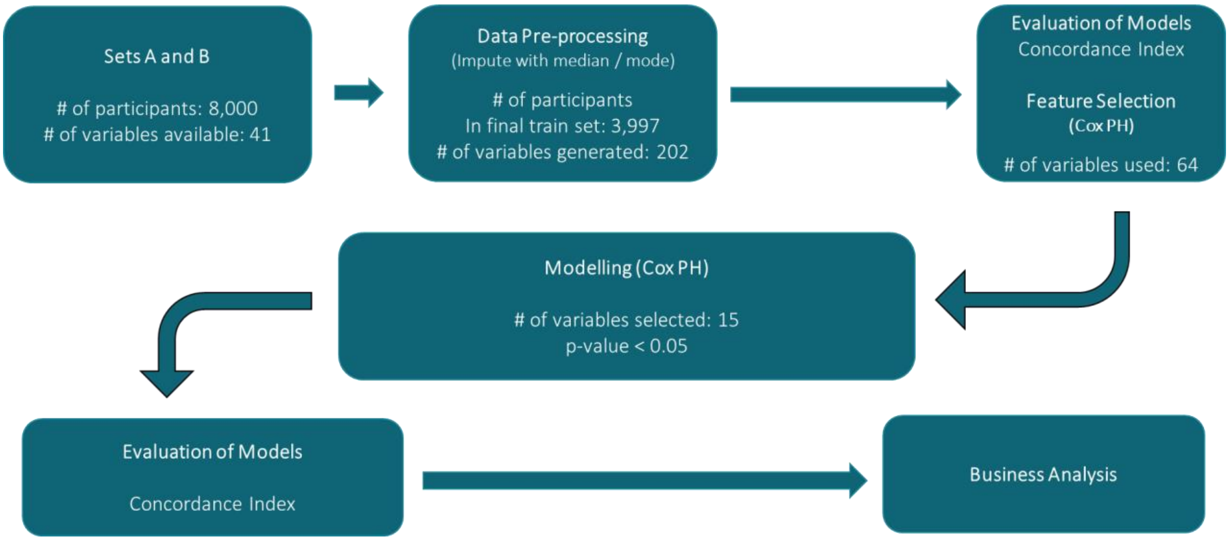


Figure 3.2 Project Flow – Survival Analysis

3.1 Project Software

The project is mainly conducted in **Python**, using ***statsmodels*** for Logistic Regression, ***scikit-learn (sklearn)*** for binary classification and modelling, as well as ***lifelines*** *and* ***scikit-survival*** libraries for Survival Analysis. Common data science libraries such as ***pandas*** and ***numpy*** are used, alongside graph plotting libraries such as ***matplotlib*** and ***seaborn***. Other programs **SPSS** and **JMP** are used for initial modelling using CPH. The omnibus test in **SPSS** is also used to test the significance of the CPH model. The source codes are provided in separate folder as part of the report submission.

3.2 Data Preparation

Data pre-processing is important in understanding, cleaning and getting the data ready for use. The following points state both dataset observations and steps that are undertaken accordingly to prepare the data for modelling purposes:

1. Summary statistics as observations
2. Missing values
3. Outliers
4. Feature scaling/Normalization

3.2.1 Summary Statistics as Features

Summary statistics such as minimum, maximum, median, first and last value of each variable in the first 48 hours of patient's ICU stay have been found useful in other studies (Johnson et al., 2014, Johnson et al., 2017, Lee & Horvitz, 2017, Morid et al., 2017, Nguyen et al., 2017).

The rationale for including such features is due to the significance of high and low values (deviation from the norm / median) that represent medical conditions. For example, high values of blood pressure are severe for patients with the medical condition – hypertension. Whereas, low values of blood pressure are also severe for patients with the medical condition – hypotension. The extraction of minimum and maximum values meant that the model includes both the described effects.

As there are time-series features in this dataset and the distribution of these features are not normal, median values are better estimators. Median values are also able to better capture the key characteristics of a time series that is less affected by outliers as compared to mean.

The assumption in using these variations of the observations is that patients who display greater abnormalities in multiple variables are associated with a high risk of mortality (Luo et al., 2016).

3.2.2 Missing Values

The dataset consists of time-series features where hourly recordings are taken. There are health indicators that are measured more frequently, such as blood pressure, due to their significance in monitoring and diagnosing the patient's condition. Whereas, some tests require less frequent measurements (i.e. non-hourly) such as blood tests, because of their invasive means of measurement and the need to consider the patient's wellbeing, including the necessity and cost of tests. Further, measurements are also likely to be taken less frequently when the patient condition appears stable and vice versa (Aczon et al., 2017). Thus, this affects the availability of measurements for each time stamp in the clinical data sets. A statistical summary of the variables is presented in Appendix A.

Only 15 variables, such as *GCS, Temperature,* and *HR* (heart rate) can be found in at least 98% of the patients. Most of the time-series variables (19 of 37) document missing observations for at least 20% of the patients. Thus, there is no patient with a complete set of variables. The missing values in this dataset cannot be treated the same way as the variables carry different meanings (i.e. height versus blood test variable).

The missing values of *Height, Weight* and *Temperature* are imputed using the median values of patients with similar demographics such as age and gender whose values are not outliers. Any missing value for a medical intervention (e.g. Mechanical Ventilation) is imputed as 0. Missing values of other continuous and binary variables are imputed with the patient's median and mode respectively. If the patient's median value (e.g. respiration rate) is not available throughout the entire stay, the population median value is used instead for missing values imputation.

3.2.3 <u>Outliers</u>

Outlier values are expected of medical data, especially when it is recorded from ICU patients. This is because these patients, who are in critical conditions, are more likely to have health indicator values that deviate from the normal range. Values outside the upper and lower bounds of 1.5 times the interquartile range (IQR) are considered as outliers.

The outlier values of *Height*, *Weight* and *Temperature* are replaced with median value of patients with similar demographics such as age and gender whose values are not outliers.

Some of the variables have a known range of physiologically possible values that makes sense for the human body (American Board of Internal Medicine; Edwards Life Sciences 2009; Merck Manuals 2018; U.S. National Library of Medicine). For example, heart rate ranges from 25 to 300, and pH levels for the human body ranges from 6 to 8. Some of the outlier values are taken to have missing decimal points due to typological error (e.g. pH 700 which possibly mean pH 7.0). The other outlier values of pH are imputed with median.

Continuous variables are subsequently normalised using min-max scaling method which scales them to a range of 0 and 1 (Nielson et al, 2019). This is so that they are scale independent (less affected by outliers as compared to standardization) for modelling.

3.2.4 Feature Engineering

On top of using summary statistics as features in the dataset, other new features are extracted from the original variables. Table 3.1 below summarises these new features.

Table 3.1 Features Engineered

New Features	Description and Rationale
Original Variables: Height and Weight	
BMI	Body Mass Index (BMI) for each patient is calculated by: weight in kg/height in m^2
BMI_Cat	Obese: $BMI \geq 30$ Overweight: $25 \leq BMI < 30$ Normal: $18.5 \leq BMI < 25$ Underweight: $BMI < 18.5$
Original Variables: DiasABP, NIDiasABP, MAP, NIMAP, SysABP, NISysABP	
DiasABP *MAP* *SysABP*	Following existing literature, combine the invasive and non-invasive measurements for each blood pressure measurement – mean arterial blood pressure (*MAP*), systolic arterial blood pressure (*SysABP*), diastolic arterial blood pressure (*DiasABP*) – into single features.
High Blood Pressure	Systolic arterial blood pressure above 140mmHg and Diastolic arterial blood pressure above 90mm Hg based on point of admission to hospital – pre-existing condition
Hypertension	*MeanSysABP* > 140mm Hg or, *MeanDiasABP* > 90 mm Hg
Original Variable: GCS	
GCS_Coma	Patients with GCS scores of 3 - 8 are usually said to be in a coma. Patients with scores of 9 to 15 are not in a coma
Original Variable: HR	
Tachycardia	Heart rate > 100 beats per minute
Original Variables: PaO2, FiO2	
PF ratio	*PaO2/FiO2 - clinical indicator of hypoxaemia*
Original Variable: Urine	
Oliguria (Renal Injury)	An early and sensitive biomarker of renal injury/ kidney damage. $80 < Urine < 400$

Original Variable: Glucose	
Diabetes	As the data dictionary did not specify that patients fasted to have their glucose levels tested, the variable records non-fasting blood glucose level at random for patients. The American Diabetes Association recently published guidelines that recommend a blood glucose target between 140 and 180 mg/dL for critically ill patients in 2010. A random blood sugar level of 200 mg/dL or higher suggests diabetes. Diabetes: *Glucose* ≥ 200 mg/dL Prediabetes: 180 < *Glucose* < 200 mg/dL No diabetes: 140 < *Glucose* ≤ 180 mg/dL
Hyperglycemia	High blood sugar level *Glucose* > 250 mg/dl
Hypoglycemia	Low blood sugar level Hypoglycemia: 54 < *Glucose* < 70 mg/dL Critically low: *Glucose* < 54 mg/dL

3.3 Data Balancing

As there is class imbalance[4], undersampling on the majority class (survived) using Repeated Edited Nearest Neighbours algorithm is conducted to achieve a ratio of 78 (survived) : 22 (died). Due to the trade-off between undersampling which leads to loss of information and oversampling that leads to overfitting, the undersampling approach is adopted in this project to prevent the issue of overfitting.

Table 3.2 Number of Patients in Train Set A after Undersampling

Train Set A	**Before**	**After**
Survived	3444 (86%)	1955 (78%)
Died	553 (14%)	553 (22%)

[4] See Chapter 2.1.3.

3.4 Feature Selection

The process of feature selection ensures that the model does not take on less informative features. Having more features does not necessarily indicate a better model; on the contrary, the elimination of features that are less informative can enhance generalization performance of the models, and as a result, a model with better evaluation metric scores. Further, feature selection also leads to greater computational efficiency, by reducing dimensions and thus computation time. The feature selection process consists of two main ideas, either to use them separately, or combine the features identified to select a group of features. The techniques employed are discussed below.

3.4.1 Binary Classifier-based Selection

Binary classifiers are selected because of their white-box methodology in identifying key features to predict a binary outcome. A regression-based model and a tree-based model are applied in the feature selection process, harnessing the strengths of each model (Kakade et al., 2018; Samanta et al., 2009; Student et al., 2018).

3.4.1.1 Univariate Analysis and Logistic Regression

Logistic Regression is selected because of its white-box methodology using the sigmoid function. Features that are statistically significant at the significance level of 0.05 (i.e. p-value < 0.05) are selected. The lower the p-value, the more significantly a variable influences the model (Student et al., 2018).

Considering the potential issue of multicollinearity that may occur for Logistic Regression, only the median values or value at admission for each feature (where applicable) are selected. This forms the initial set of features that is used as input for the various feature selection approaches (including Decision Tree and CPH), and to ensure fair comparison between various approaches of feature selection.

A univariate analysis using Correlation Analysis and Logistic Regression is undertaken for this project following feedback from the presentation. This method of Logistic Regression individually processes each independent variable independently, against the dependent variable (survival status of patient) (Bursac et al., 2008; Kakade et al., 2018). This method allows significant features to be identified first (p-value < 0.05), without having the interaction of the features impact their coefficients which will make the model less stable (Kakade et al., 2018). The list of 58[5] features and their p-values is appended in Appendix C. The results led to *Hypoglycemia* and *GCS* being removed as their p-values are not statistically significant ($p > 0.05$).

Following which, these features are then processed altogether in Logistic Regression to identify the final set of statistically significant features based on p-values. This results in a final list of 22 features[6].

3.4.1.2 <u>Decision Tree</u>

Decision Tree, widely utilised in healthcare literature is also selected as the tree model is rule-based and can be visualized[7] (Samanta et al., 2009). The output of tree model is easy to understand for doctors who may not be familiar with algorithms and clearly displays key indicators that can be practically implemented in a medical setting.

The importance of a feature in Decision Tree is computed as the total reduction of the Gini index. Gini index, or mean decrease in impurity (MDI), calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits.

The list of features and their variable (feature) importance is appended in Appendix E.

[5] The number of features reported will include dummy variable encoded features.
[6] See Appendix D.
[7] See Appendix E.

Figure 3.3 below shows the cumulative importance of all features. The top 20 features contributed to 70% cumulative importance, and the top 30 features contributed to 90% cumulative importance. As a result, 30 features are selected from the Decision Tree model.

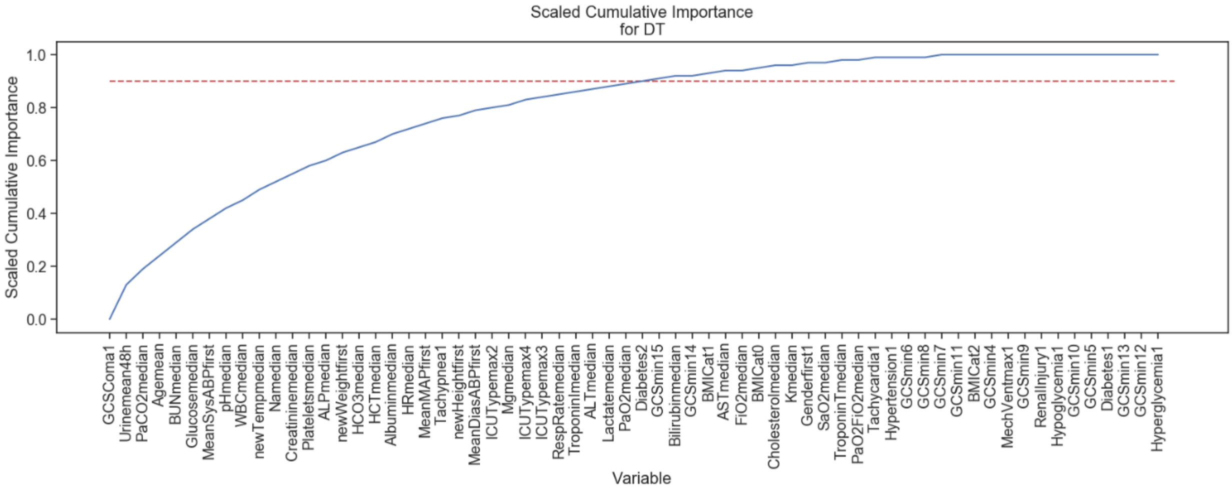


Figure 3.3 Cumulative Importance from Decision Tree

3.4.1.3 Ensemble of Features Identified by Binary Classifiers

The features from Logistic Regression and Decision Tree can either be used individually or combined like an ensemble (Zakharov & Dupont, 2011). Drawing on the different strengths of these models, combining features from different models can ensure corroboration in the final set of features. This is likened to the concept of combining a group of weak learners to become a strong learner in Random Forest. The hypothesis is that if the features are identified in both models, they are likely to be significant. This leads to the use of an intersection set of features (i.e. common to both), as well as a union set (i.e. all features that are significant in either model).

Following feedback from the presentation, another approach using features identified by Logistic Regression processed through another round of selection using Decision Tree is also considered. This resulted in the same set identified in Logistic Regression.

3.4.2 Survival Analysis-based Selection

Besides the Binary classifier-based selection, feature selection using Cox Proportional Hazards (CPH) model is conducted. The CPH model (a time-to-event analysis, in this case time from ICU admission to death) is employed to identify features that have greater impact on the patient's length of stay in hospital while also considering if the patient has died in the hospital. Significant features are selected based on their p-value (p-values < 0.05). Figure 3.4 shows the variable importance of the selected variables. The CPH model is further elaborated in detail in Chapter 4.

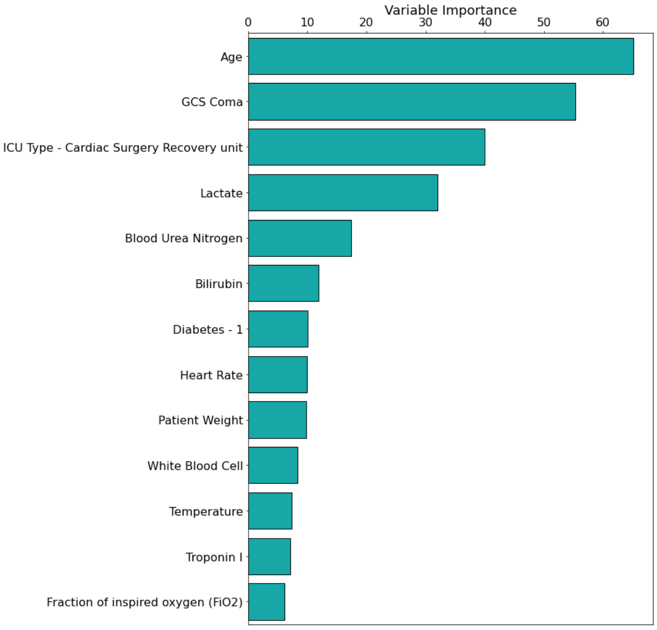


Figure 3.4 Variable Importance from CPH Model

3.4.3 Summary

Features identified by each model can either be used on its own, or in combination with that from another model. The list below shows the various approaches that are adopted:

1. all features (no feature selection)
2. features derived solely from Logistic Regression
3. features derived solely from Decision Tree
4. a union set of features derived from both Logistic Regression and Decision Tree models
5. an intersection set of features derived from both Logistic Regression and Decision Tree models
6. features derived from CPH model

Features identified from CPH model are not used in combination with that of the binary classifiers as the modelling approaches are different.

4.0 Survival Analysis

This project aims to leverage the availability of ICU *length of stay* data in this dataset. Building a Survival Analysis on top of the binary classification problem can help enrich the overall implementation strategy for resource allocation in ICU wards. The limited beds in ICU wards are often occupied and due to the uncertainty of patients' prognosis and new admissions, hospital bed occupancy rates can be difficult to forecast. As Survival Analysis allows the analysis of the impact of different features on the survival outcome, it can help doctors better understand the prognostic health indicators that affect the patient's survival and length of stay in ICU wards (Kim et al., 2019). This can help hospitals identify the type (conditions) of patients admitted to the ICU, who may have a higher probability in a longer ICU stay or vice versa. This will allow hospitals to gauge if any beds may be occupied for an extended period, and to cater for more.

Survival Analysis is thereby, employed to predict the *length of stay* while considering censored data. The censored data, in this case, is represented by the patient's survival status. Figure 4.1 depicts the *length of stay* for 50 patients: the blue line represents patients who survived and the red line patients who died.

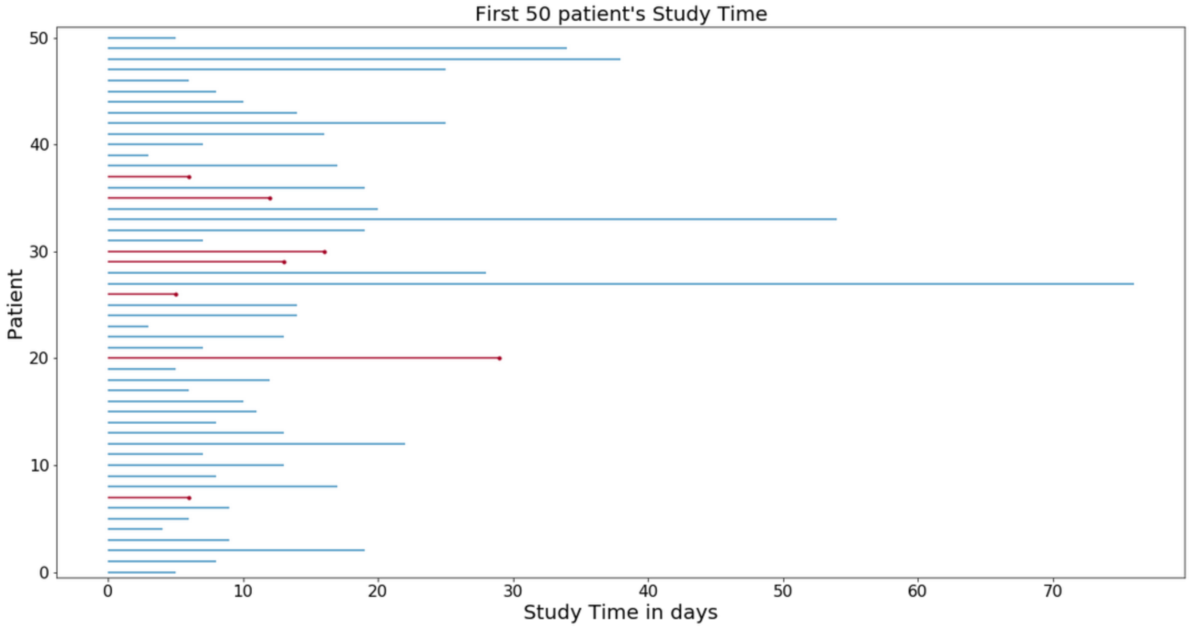


Figure 4.1 Survival Status of Patients

Figure 4.2 below observes the distribution of *length of stay* data. The data is right skewed, which signals challenges in being modelled by traditional statistical methods such as Linear Regression, hence the use of CPH model, a semi-parametric model, can overcome this challenge of skewness.

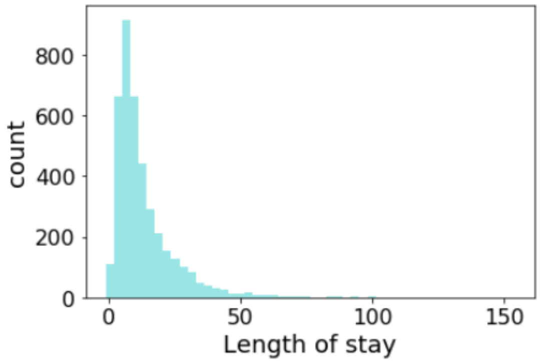


Figure 4.2 Distribution of *Length of Stay* Data

4.1 Analytical Approach

The Kaplan-Meier estimator is the Exploratory Data Analysis (EDA) tool for Survival Analysis. It provides a relatively easy-to-understand approach in visualising time-to-event data. The Kaplan-Meier estimator allows for the examination of different survival probabilities for different groups of patients. For example, Figure 4.3 shows the different survival probabilities for Glasgow Coma Score (GCS) from patients in 'Severe' coma to patients in 'Mild to Moderate' coma. Patients in 'Severe' coma are more likely to have a shorter length of stay for any given survival probability.

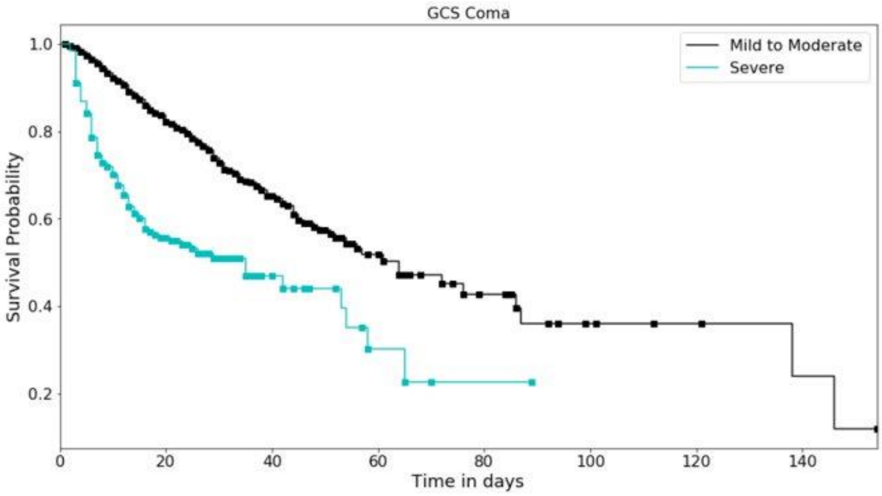


Figure 4.3 Kaplan-Meier Plot for GCS

Binning of continuous variables is required for Kaplan-Meier estimator. The continuous variables are binned into quantiles. Figure 4.4 below observes that patients with heavier weight (dark blue and green lines) are more likely to survive in comparison to patients with lower weight (turquoise and black lines). There are also overlaps observed towards the tail-end of the plot. This may be attributed to the survival estimates becoming less certain as the population at risk declines. By plotting 95% confidence intervals around the Kaplan-Meier lines as shown in Appendix F, the width of the confidence interval will increase.

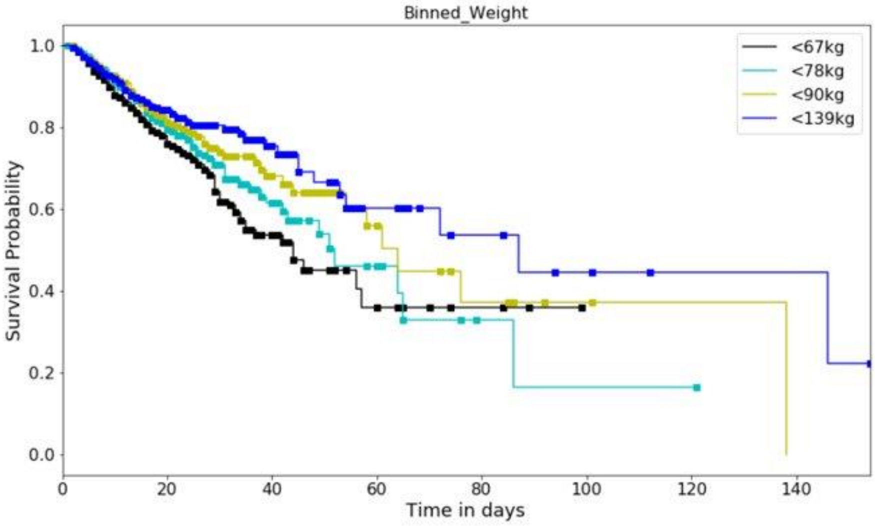
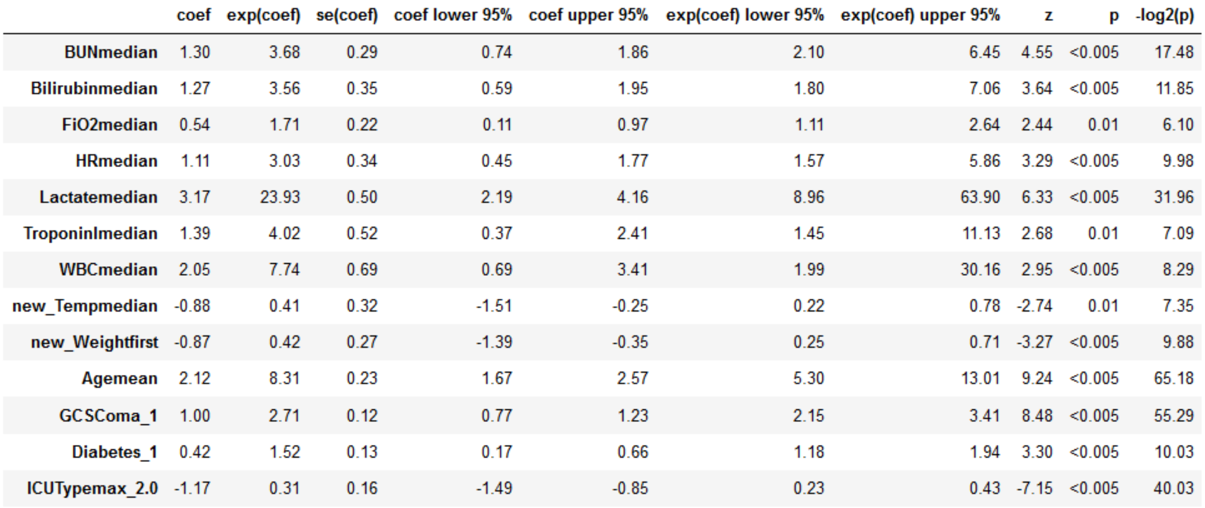


Figure 4.4 Kaplan-Meier Plot for Weight

4.2 Results and Feature Selection

It is discussed above that non-normality and censorship of the data are challenges faced when trying to fit the data into traditional models. Therefore, to overcome these challenges, this project employs a semi-parametric model – Cox Proportional Hazards (CPH) model – based on the assumption of proportional hazards.

The ***lifelines*** library in **Python** allows the generation of statistical summary of the CPH model and hence significant features as shown below in Figure 4.5. These significant features identified will be used for modelling. One thing to note for CPH is that the CPH model will not be able to execute if multicollinearity persists within the variables (Appendix G). Hence the initial set of features for feature selection (i.e. only the median values or value at admission for each feature (where applicable)) are used as inputs for CPH model.



	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
BUNmedian	1.30	3.68	0.29	0.74	1.86	2.10	6.45	4.55	<0.005	17.48
Bilirubinmedian	1.27	3.56	0.35	0.59	1.95	1.80	7.06	3.64	<0.005	11.85
FiO2median	0.54	1.71	0.22	0.11	0.97	1.11	2.64	2.44	0.01	6.10
HRmedian	1.11	3.03	0.34	0.45	1.77	1.57	5.86	3.29	<0.005	9.98
Lactatemedian	3.17	23.93	0.50	2.19	4.16	8.96	63.90	6.33	<0.005	31.96
TroponinImedian	1.39	4.02	0.52	0.37	2.41	1.45	11.13	2.68	0.01	7.09
WBCmedian	2.05	7.74	0.69	0.69	3.41	1.99	30.16	2.95	<0.005	8.29
new_Tempmedian	-0.88	0.41	0.32	-1.51	-0.25	0.22	0.78	-2.74	0.01	7.35
new_Weightfirst	-0.87	0.42	0.27	-1.39	-0.35	0.25	0.71	-3.27	<0.005	9.88
Agemean	2.12	8.31	0.23	1.67	2.57	5.30	13.01	9.24	<0.005	65.18
GCSComa_1	1.00	2.71	0.12	0.77	1.23	2.15	3.41	8.48	<0.005	55.29
Diabetes_1	0.42	1.52	0.13	0.17	0.66	1.18	1.94	3.30	<0.005	10.03
ICUTypemax_2.0	-1.17	0.31	0.16	-1.49	-0.85	0.23	0.43	-7.15	<0.005	40.03

Figure 4.5 Significant Features – CPH Model

Subsequently, the assumptions of CPH model are checked using the Schoenfeld residuals test on the features fitted in the model. Figure 4.6 below depicts a particular feature – GCS – violating the Proportional Hazard model assumption because of its non-zero slope as seen on the right graph. The residuals in the graph have a nonlinear relationship between the residuals and the function of time. Despite its violation of the Proportional Hazards assumption, this feature is retained for further modelling because domain knowledge in existing literature established the feature's importance in ICU patient condition and its predictive ability on mortality (Ting et al, 2010).

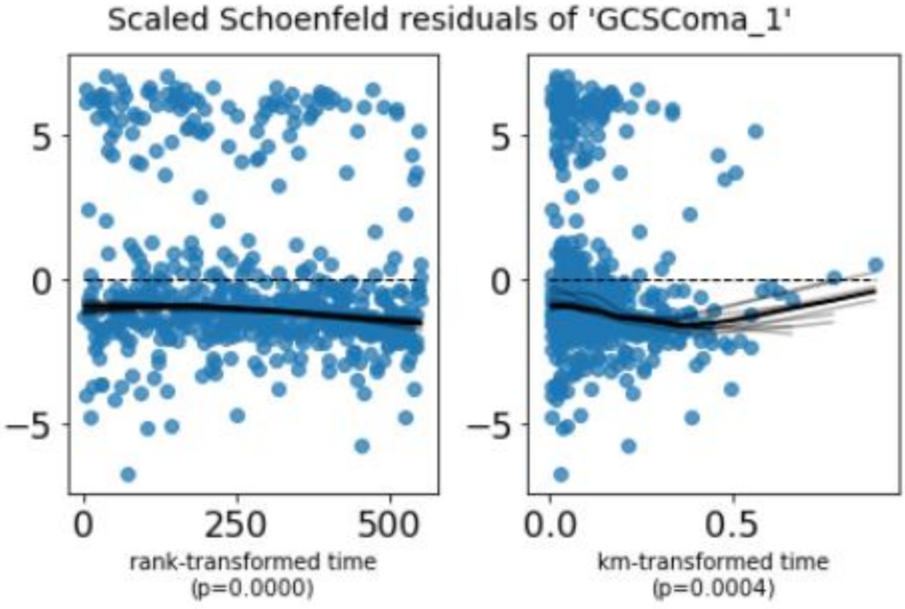
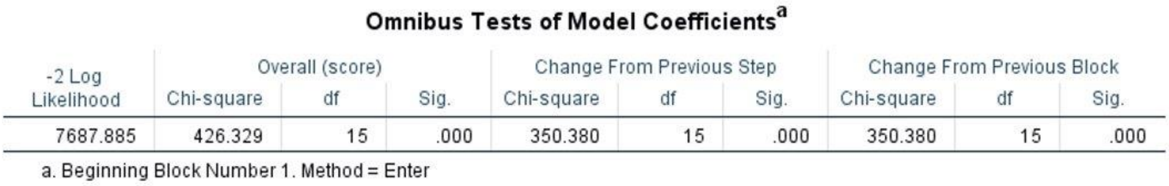


Figure 4.6 Schoenfeld Residuals Test – GCS

4.3 Evaluation Metrics

The adequacy of CPH model is tested using Concordance Index (C-index) and the Omnibus Model Coefficients test. The C-index is a common metric used in Survival Analysis that measures how well the model predicts the order of patient's death times by assigning a score between 0 to 1[8] (Weathers, 2017). A high value for the C-index is indicative of a model than can predict higher probabilities of survival for higher observed survival times. This CPH model achieves a C-index of 0.77.

Figure 4.7 also shows that this CPH model is significant through the Omnibus test run in **SPSS**. The Omnibus test is a likelihood-ratio chi-squared test of the current model versus the null (in this case, intercept) model. The significance value of less than 0.05 indicates that the current model outperforms the null model.



Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
7687.885	426.329	15	.000	350.380	15	.000	350.380	15	.000

a. Beginning Block Number 1. Method = Enter

Figure 4.7 Omnibus Test Run in SPSS

[8] A score of 1.0 is perfect concordance and 0.0 is perfect anti-concordance, whereas 0.5 is expected result from random predictions ("Survival regression", Lifelines Manual, https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html)

Further, this project also considered a black box non-parametric model – Random Survival Forest (RSF) – for this time-to-event prediction to determine if there is a better predictive model that can help hospitals make more effective decisions (Weathers, 2017). RSF is an ensemble of tree-based learners for analysis of right-censored survival data. It also ensures that individual trees are de-correlated by:

i. building each tree on a different bootstrap sample of the original training data, and

ii. at each node, only evaluate the split criterion for a randomly selected subset of features and thresholds.

Predictions of RSF are developed through the aggregation of predictions made by individual trees in the ensemble. The advantages of RSF include:

i. reduces variance and bias by using all variables collected,

ii. automatically assessing non-linear effects and interactions effects, and

iii. no high variance and poor performance issues associated with Decision Trees because a Random Forest builds hundreds of trees and outputs the results by voting.

The plot below on the left in Figure 4.8 shows that the RSF is more conservative in predicting the length of stay (i.e. predicting a shorter length of stay at 50 days as compared to 64 days for patient who died) at the median survival probability. Although the performance of both models is comparable whereby the C-index score for RSF model is the same as the CPH model, this project leans towards the implementation of the CPH model in the medical setting because RSF is a black box model which may not be as well established to be used in the medical context.

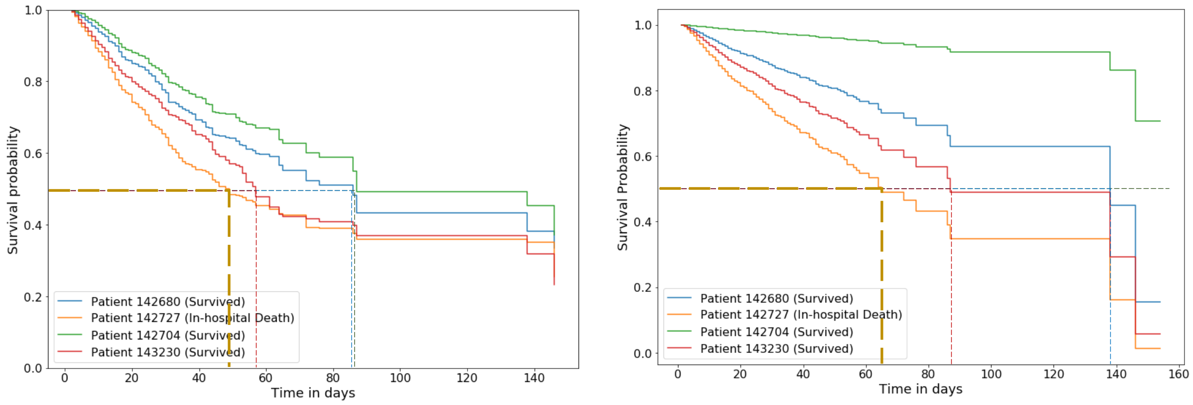


Figure 4.8 RSF (left) and CPH (right) Prediction in *Length of Stay*

With the results from Survival Analysis, hospitals can draw out relevant insights for planning purposes by using the CPH model to predict the length of stay. For example, employing the median survival probability to forecast the number of days that a specific patient is likely to stay in ICU. Another relevant insight includes the proactive diagnosis of patients at different survival risk segments along the timeline. To illustrate, the plot in Figure 4.8 depicts the prediction of the survival function for 4 randomly selected patients. The orange plotline represents a patient who died, while the other coloured plotlines show patients who survived. The patient who died has the lowest survival probability across all *length of stay* based on the observation of the survival functions. This suggests that closer observation is warranted for this patient.

5.0 **Modelling Approaches**

The purpose of this project is to propose classifier models using features (existing and engineered) that can better predict ICU mortality through various approaches, and to validate the models' prediction power through empirical analysis.

5.1 Analytical Approach

The analytical approach undertaken in this project is further discussed below. The features are processed in the following subsets as independent variables for the predictive models:

1. all 58 features (no feature selection)
2. 22 features derived solely from Logistic Regression model
3. 30 features derived solely from Decision Tree model
4. a union set of 48 features derived from both Logistic Regression and Decision Tree models
5. an intersection set of 22 features derived from both Logistic Regression and Decision Tree models
6. 16 features derived from CPH model

The following classifiers are employed to predict in-hospital mortality:

1. Logistic Regression (logR)
2. Decision Tree (DT)
3. Gradient Boosting (GB)
4. Extreme Gradient Boosting (XGB)
5. Adaptive Boosting (Adab)
6. Extra Trees (XT)
7. Random Forest (RF)
8. Support Vector Classification (SVC)

This results in more than 40 model-feature combinations.

5.1.1 Logistic Regression (logR)

Backward elimination is adopted whereby all independent variables are first entered into the regression formula and one variable is progressively removed if its p-value exceeds 0.05 (i.e. not statistically significant). The variables are also checked for multi-collinearity by ensuring that their variance inflation factor (VIF) is less than 5. An example of the result is shown in Appendix D.

5.1.2 Decision Tree (DT)

Decision Trees recursively split features based on their target variable's impurity. The best split is chosen by minimising the Gini index. The Gini index measures how often a randomly chosen attribute from a set is incorrectly labelled. An example of the result is shown in Appendix E.

5.1.3 Extreme Gradient Boosting (XGB), Gradient Boosting (GB) and Adaptive Boosting (AdaB)

Gradient Boosting and Adaptive Boosting are variations of tree-based models where boosting technique is applied to correct the errors made by earlier models.

Gradient Boosting involves creating new models that predict the residuals of prior models. The predictions from new models and that of prior models are then added together to establish the final prediction. It utilises a gradient descent algorithm to minimize the loss when adding new models.

Adaptive Boosting applies greater weights on data points that are difficult to predict. As the tree grows, the weights of difficult-to-classify observations are increased, and the weights of easy-to-classify observations are decreased. The final model is the weighted sum of the predictions made by the previous tree models.

5.1.4 Random Forest (RF) and Extra Trees Classifier (XT)

Random Forest involves building multiple trees by taking bootstrap samples and random sampling of features. Thus, Random Forest is acknowledged as a model that is unlikely to overfit (Kim et al., 2019).

Extra Trees differ from Random Forest in that it introduces more variations into the ensemble by using all data available in the training set and randomly selecting the best split among features also selected at random.

5.1.5 **Support Vector Classification (SVC)**

The methodology of SVC is to construct hyperplanes in a multi-dimensional space that maximizes the margins from different classes.

5.2 Hyperparameters Tuning

Hyperparameter tuning using *GridSearchCV* function in Python is employed to find the parameters that best optimise performance of each classifier. L2 regularization is then used for Logistic Regression while Gini index is used for tree-based algorithms.

5.3 Ensemble

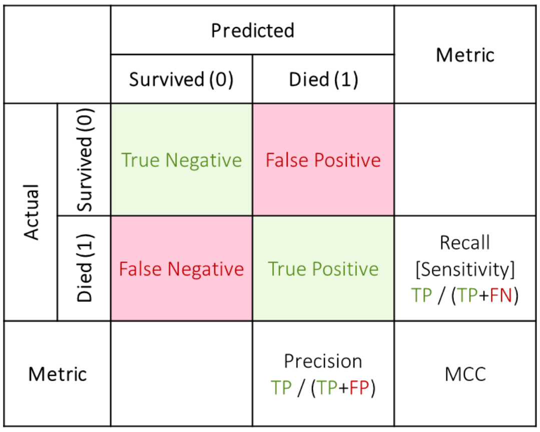
This project also applies the ensemble concept[9] from Random Forest. The best performing base models are aggregated via stacking or voting to create an ensemble model. The model that results in the best Recall score is selected as the final model. The evaluation metrics are further discussed in Chapter 6.

[9] where a group of weak learners can be combined to become a strong learner

6.0 Evaluation of Model Performance and Results

6.1 Model Evaluation

The results from the classifiers are evaluated using a confusion matrix to select the model that best predicts ICU mortality. Figure 6.1 depicts the confusion matrix and evaluation metrics derived from it. Table 6.1 summarises the evaluation metrics that this project will employ in the evaluation process. Recall is the main metric used for model evaluation.



		Predicted		Metric
		Survived (0)	Died (1)	
Actual	Survived (0)	True Negative	False Positive	
	Died (1)	False Negative	True Positive	Recall [Sensitivity] TP / (TP+FN)
Metric			Precision TP / (TP+FP)	MCC

Figure 6.1 Confusion Matrix and Evaluation Metrics

Table 6.1 Definition and Formula for Evaluation Metrics

Metric and Definition	Formula
Recall (Sensitivity) Measures proportion of correctly identified True Positives among actual positives	$\frac{TP}{TP + FN}$
Area Under Curve – Receiver Operating Characteristics (AUC-ROC) Plot of Recall against (1 – Specificity) for different threshold settings	$\frac{Recall}{1 - Specificity}$ Specificity = $\frac{FP}{TN+FP}$
Precision Measures proportion of correctly identified True Positives among predicted positives	$\frac{TP}{TP + FP}$
Area Under Curve Precision Recall (AUC PR) Plot of Precision against Recall for different thresholds settings.	$\frac{Precision}{Recall}$ Precision = $\frac{TP}{TP+FP}$
Matthews' Correlation Coefficient (MCC) Measures how well a model is performance by taking into account True and False Positives and Negatives	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

As the dataset experiences class imbalance (i.e. majority of patients survived), it is highly likely that the models will correctly predict this majority class (in confusion matrix context, True Negative). Nevertheless, since the prediction of ICU mortality in this project takes into account human lives, greater emphasis must be placed on minimising the number of False Negatives (predict the patient will survive but actually died). From the healthcare standpoint, a model resulting with high number of False Negative instances can lead to negligent decisions. One of the business objectives in this project is to provide timely intervention for patients to boost their survival rate. Therefore, it is also crucial that the model selected for implementation can maximise the correctly predictions that patients died. The Recall metric takes these into account and is thus selected. The prediction probability threshold value is set at 0.5.

Another commonly used metric for imbalanced datasets such as mortality prediction is MCC (Chicco and Jurman, 2020; Nielsen et al., 2019). MCC is a more robust indicator for binary classification because it takes into consideration all classes in the confusion matrix (TP, FP, TN and FN). MCC value of 1 indicates a prefect prediction, whereas that of -1 indicates a disagreement between the predicted class and the observation. MCC value of 0 indicates the model predicts randomly.

AUC-ROC is a threshold-independent metric that illustrates the tradeoff between Recall and Specificity. A larger AUC indicates a better overall performance across all thresholds (Ghanvatkar and Rajan, 2019; Harutyunyan et al, 2019). The AUC-ROC metric is also considered to be more informative when faced with highly skewed datasets (Harutyunyan et al, 2019). This is the metric used by past winners of the PhysioNet competition using the same dataset. The use of this measurement is for benchmarking purposes.

Similarly, another threshold-independent metric AUC-PR that illustrates the tradeoff between Precision and Recall is used. This metric is suitable for imbalance dataset as it focuses on the correct prediction of the minority class (True and False Positives) (Harutyunyan et al, 2019). It is used to evaluate the results of k-fold cross-validation.

6.1.1 Evaluation of Base Models

Table 6.2 below reports on the top 5 overall models with highest Recall, out of 48 feature-model combinations. Most of the top-ranked models are variations of boosted trees[10]. The Adaptive Boosting model using features selected from CPH resulted in the highest Recall of 78%, while having a relatively high AUC-ROC of 80%.

Table 6.2 Top 5 Best Performing Models based on Recall

Rank	Subset of Features Processed	Model	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
1	Cox Proportional Hazard	Adaptive Boosting	78.3	80.3	43.5	0.32	2290	1137	123	445
2	Decision Tree	Gradient Boosting	72.4	83.0	46.1	0.38	2662	765	157	411
3	All	Gradient Boosting	72.4	83.0	46.0	0.38	2662	765	157	411
4	Union of Logistic Regression and Decision Tree	Gradient Boosting	72.2	82.7	45.5	0.38	2644	783	158	410
5	Logistic Regression	Gradient Boosting	67.4	83.0	46.1	0.40	2802	625	185	383

[10] See Appendix H for the modelling results.

Following the feedback from our presentation, a comparison of model performance is also made using features derived from univariate analysis (Logistic Regression). Table 6.3 below evaluates the performance of Logistic Regression and Decision Tree models using these features[11]. The model performance evaluated based on Recall for the Logistic Regression and Decision Tree models are not as good as the boosting algorithm mentioned above in Chapter 6.1.1.

Table 6.3 Model Performance Evaluation – LogR and DT

Rank	Subset of Features Processed	Model	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
21	Logistic Regression	Logistic Regression	57.2	81.8	44.4	0.37	2923	504	243	325
39	Logistic Regression	Decision Tree	50.5	61.6	18.9	0.18	2493	934	281	287
35	All	Decision Tree	51.8	64.1	20.6	0.22	2616	811	274	294

[11] 22 features are derived solely from Logistic Regression (using univariate analysis).

6.1.2 Evaluation of Ensemble Models

The respective best model with features selected using Binary classifier-based (i.e. Intersection of Logistic Regression and Decision Tree[12]) and Survival Analysis (i.e. CPH) approaches are used for ensemble.

Table 6.4 below summarises the results of 4 ensemble models. The Stacking classifier results in higher Recall, in comparison to the Voting classifier. As this project drew on different feature selection approaches (i.e. Binary Classification and Survival Analysis), the results show that the use of feature groups from either approach produce relatively similar performance.

[12] Intersection of Logistic Regression and Decision Tree is the first model that appeared twice in the top 10 overall models. See Appendix H.

Table 6.4 Evaluation of Ensemble Models

Subset of Features Processed	Ensemble	Classifier	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
Intersection of Logistic Regression and Decision Tree	Logistic Regression + Adaptive Boosting + Extreme Gradient Boosting	Stacking	67.6	82.59	44.4	0.38	2757	670	184	384
Cox Proportional Hazard	Adaptive Boosting + Extreme Gradient Boosting	Stacking	67.3	80.76	43.5	0.34	2641	786	186	382
Intersection of Logistic Regression and Decision Tree	Adaptive Boosting + Extreme Gradient Boosting	Voting	66.4	82.01	43.5	0.38	2775	652	191	377
Cox Proportional Hazard	Adaptive Boosting + Extreme Gradient Boosting	Voting	66.0	80.85	44.4	0.33	2645	782	193	375

6.1.3 Cross Validation Results

5-folds cross-validation is conducted to increase the stability and generalisation performance of the model. This is performed on the base and best ensemble models discussed above in Chapter 6.1.2. The AUC-ROC and AUC-PR curves for 5-folds cross validation are illustrated in Tables 6.5 and 6.6 respectively. The results show that AUC-ROC of each model is relatively consistent across folds. This indicates that the models are relatively stable.

Table 6.5 AUC-ROC Curves for 5-folds Cross-validation

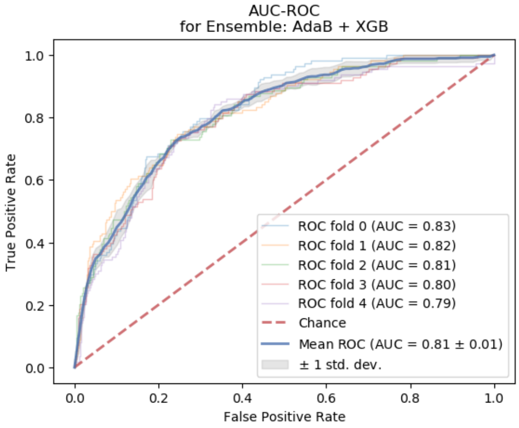
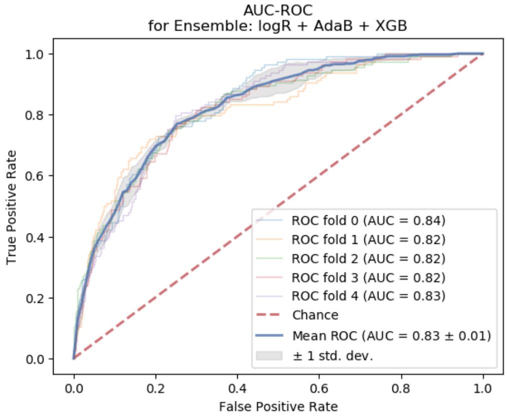
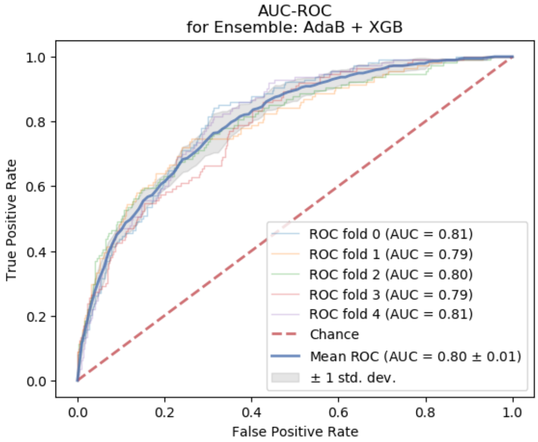
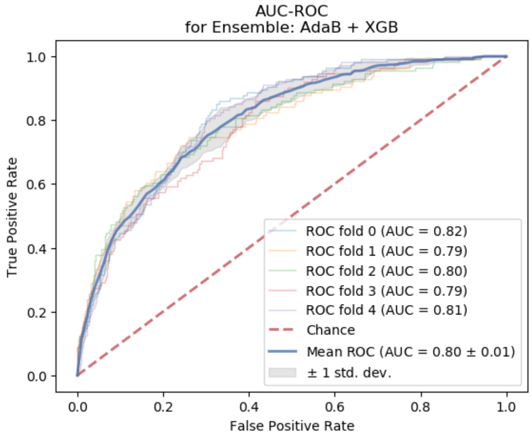
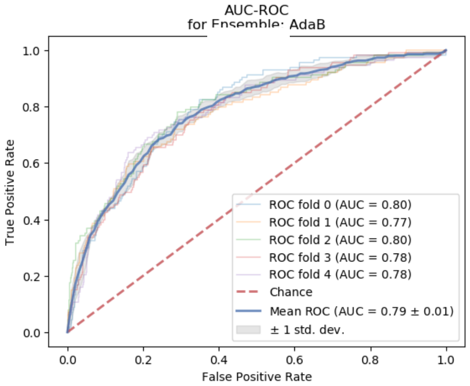
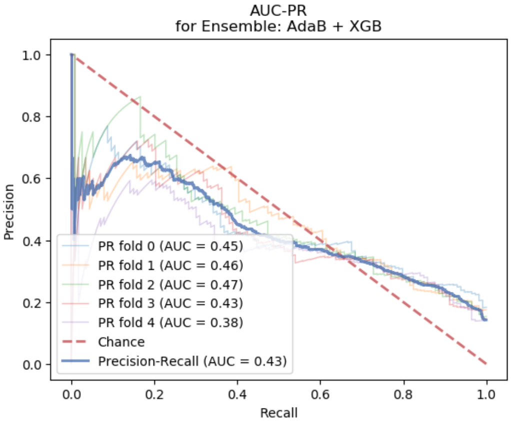
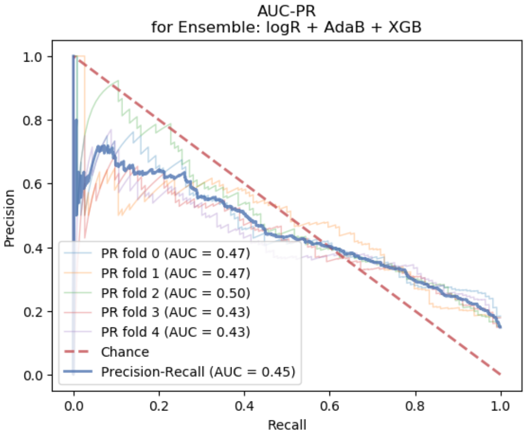
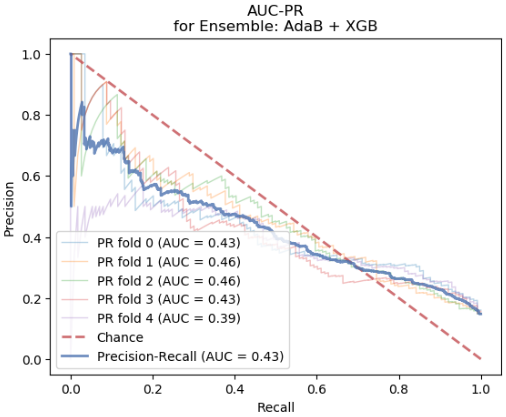
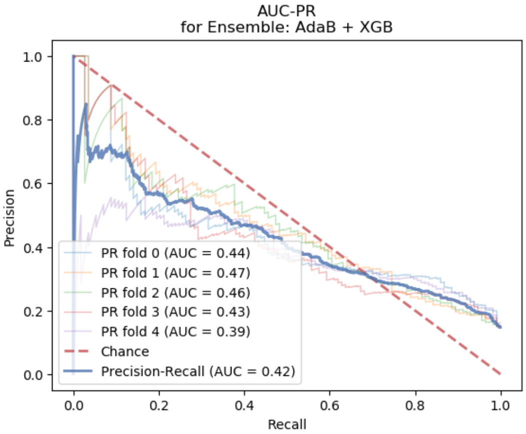
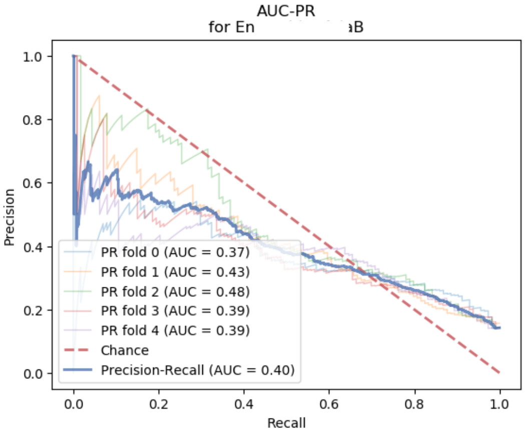
Subset of Features Processed	**Voting**	**Stacking**
Intersection of Logistic Regression and Decision Tree		
CPH		
CPH	Non-ensemble Adaptive Boosting only base model: 	

Table 6.6 shows that AUC-PR curve of each model experiences greater variation across folds, suggesting that all models incorrectly identify Positives and Negatives at different folds. This can possibly be due to the effects of class imbalance.

Table 6.6 AUC-PR curves for 5-folds Cross-validation

Subset of Features Processed	**Voting**	**Stacking**
Intersection of Logistic Regression and Decision Tree		
CPH		
CPH	Non-ensemble Adaptive Boosting only base model: 	

6.1.4 Evaluation of Results on Set C

Table 6.7 summarises the results of ensemble models and base models on set C.

The Adaptive Boosting base model fed with CPH features performed well on both the test set B and unseen set C. It has the best performance given the highest Recall (75% for unseen Set C and 78% for test set B).

The ensemble model (Adaptive Boosting and Extreme Gradient Boosting with Stacking Classifier) fed with CPH features shows the second-best performance given the second highest Recall. This result is relatively consistent with its performance on set B with its Recall still above 65% (previously rank 2, now rank 1 out of 4 for ensemble models).

The Recall of the same ensemble model that uses features selected from the intersection of Logistic Regression and Decision Tree (previously rank 1, now rank 3 out of 4 for ensemble models[13]) decreased from 65% (performance using set B) to around 55%.

Although the AUC-ROC of all models experience slight decrease of 1 to 3%, the other evaluation metrics still suggest that models fed with CPH features are relatively more stable. This shows that the CPH features are useful in predicting ICU mortality.

[13] Not considering the rank of Logistic Regression (Rank 3) and Decision Tree (Rank 5)

Table 6.7 Evaluation on Performance of Ensemble and Base Models on Unseen Set C

Rank	Subset of Features Processed	Model	Classifier	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
1	Cox Proportional Hazard	Adaptive Boosting	None	74.5	77.5	37.3	0.31	2302	1109	149	436
2	Cox Proportional Hazard	Ensemble: Adaptive Boosting + Extreme Gradient Boosting	Stacking	69.1	79.7	43.0	0.34	2572	839	181	404
3	Cox Proportional Hazard	Ensemble: Adaptive Boosting + Extreme Gradient Boosting	Voting	68.5	79.7	42.1	0.34	2581	830	184	401
4	Logistic Regression	Logistic Regression	None	57.2	81.8	44.4	0.37	2923	504	243	325
5	Intersection of Logistic Regression and Decision Tree	Ensemble: Logistic Regression + Adaptive Boosting + Extreme Gradient Boosting	Stacking	56.9	82.0	46.0	0.38	2943	468	252	333
6	Decision Tree	Decision Tree	None	55.1	63.3	19.8	0.20	2453	974	255	313
7	Intersection of Logistic Regression and Decision Tree	Ensemble: Adaptive Boosting + Extreme Gradient Boosting	Voting	54.7	81.5	45.1	0.39	3002	409	265	320

6.1.5 Benchmarking with Competition Results

Different approaches in evaluating model performance are adopted by participants of the competition (Johnson and Mark, 2017). PhysioNet competition organisers also employed another metric – minimum of Sensitivity and Positive Predictive Value (Precision) – to judge. Thus, it is challenging to benchmark our models with existing models because not all existing papers using the same dataset present the same evaluation metrics this project employs (i.e. confusion matrix and AUC-ROC).

The winning model with an AUC-ROC of 0.86, employs a tree-based ensemble approach in a Bayesian framework, with components of the trees being updated using Markov chain Monte-Carlo (Johnson et al., 2012). The runner-up model employs an ensemble of 6 SVM models using balanced subsets of data and same set of all positive outcomes (no record of AUC-ROC) (Citi & Barbieri, 2012). Another study with an AUC-ROC of 0.82, closest to the methodology used in this project, employs a cumulative hazard function with a single governing parameter for 48 hours data (Lee & Horvitz, 2017).

The best model in this project achieved an AUC-ROC of 0.78 for set C, which is not very far off from the above 3 models. The unique method of selecting features using CPH model can thus, be worthy of consideration. The white-box approach is a practical implementation for the medical centres and professionals to employ and still statistically strong.

6.2 Results and Discussion

There are 20 features selected from the intersection of Logistic Regression and Decision Tree. Although there are 13 features selected from CPH model, the CPH approach has an advantage in predicting ICU mortality because there is no need to consider any intersection between another model. This will thus enhance interpretability of the features.

Table 6.8 summarizes the features identified by both approaches. There are 9 common features that are statistically significant. Their commonality enhances their influence in practical implications, giving healthcare professionals greater confidence to consider these variables with higher scrutiny.

Table 6.8 Significant Features

Features from Intersection of Logistic Regression and Decision Tree	Features from Both Models	Features from CPH model
Albumin	Age	Bilirubin
Alkaline Phosphatase (ALP)	Blood Lactate	Diabetes
Serum Carbonate (HCO_3)	Blood Urea Nitrogen (BUN)	Troponin I
Gender	Fraction of Inspired Oxygen (FiO_2)	White Blood Cells
Blood Glucose	Glasgow Coma Scale (GCS)	
Hypertension	Heart Rate	
Mechanical Ventilation	ICU Type	
Oxygen Saturation (SaO_2)	Temperature	
Blood pH	Weight (at admission)	
Renal Injury		
Tachycardia		

Results show that *BUN* and *ICUType* are statistically significant features in the prediction. The presence of increased *BUN* suggests that waste products are not effectively removed from the kidneys or liver. This can prompt healthcare professionals to examine the root cause of kidney or liver damage. This result can be applied to that of the *Renal Injury* indicator in the diagnosis of potential kidney failure.

Figure 6.2 shows that when the (median) BUN level of a patient exceeds the normal range, doctors can respond immediately to further establish if the patient is suffering from kidney or liver damage instead of being misled by other health indicators showing adverse results also.

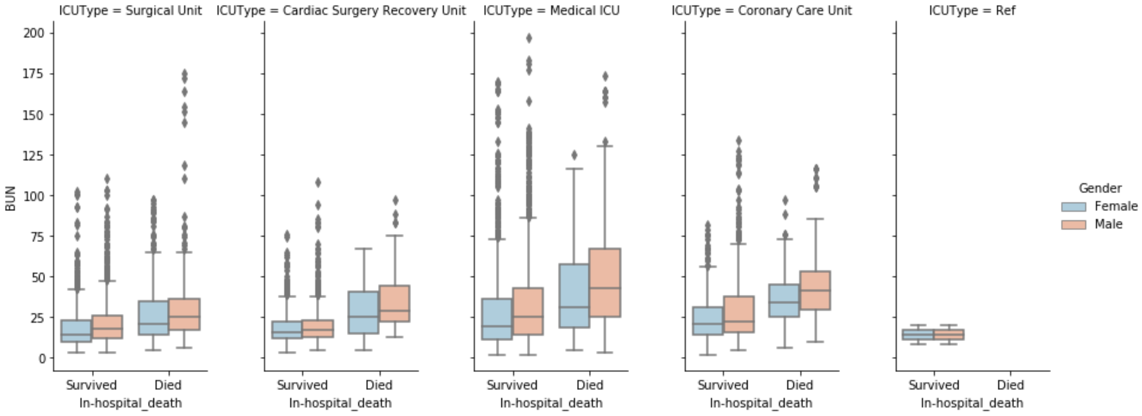


Figure 6.2 Different BUN Levels for Different Types of ICU Patients

Further, Figure 6.2 shows that the median level of BUN is higher for patients warded in Medical ICU in comparison to Surgical ICU. This is in line with the earlier discussion of Medical ICU wards admitting patients with more types of critical health conditions. In a more practical sense, doctors and nurses who have more experience in kidney or liver damage can be deployed to the Medical ICU ward.

Considering the median length of stay of patients (who survived) in Medical ICU is 10 days, the hospital can take this into account when making resource allocation decisions. For example, calculation of the hospital bed turnover rate, the number of hospital beds and medical equipment required. Nevertheless, hospitals also need to consider the expertise of doctors in their roster and the type of patients admitted. Some hospitals may be more recognised in the surgical field and thus may have a larger surgical ICU ward.

Results show that *GCS* score is a statistically significant feature in the prediction. Figure 6.3 shows that patients in Surgical Unit ICU are more likely to be in more severe state of coma. These patients also have a higher risk of passing away in coma. Doctors can use knowledge of past cases to increase the patient's chance of survival by checking on other vital health indicators and administering more timely intervention.

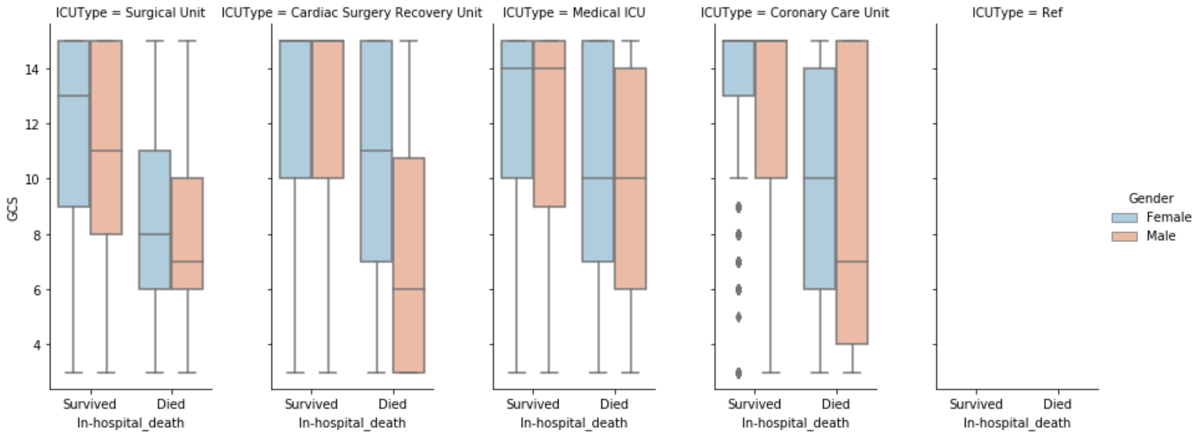


Figure 6.3 Different GCS Coma Score for Different Types of ICU Patients

Other statistically significant features in the prediction include *SaO_2*, *Blood pH* and *Blood Lactate*. The combination of these 3 features is an indication of a critical health condition that requires timely intervention. Specifically, the presence of high *Blood Lactate* levels and low *SaO_2* (oxygen saturation), coupled with acidosis (low *Blood pH* level) signal to the doctors and healthcare professionals that patient is experiencing a probable imbalance between oxygen demand and oxygen delivery to the tissues. Timely supply of oxygen to the patient may drastically impact survival rate.

7.0 <u>Conclusion</u>

The results documented above for the analyses undertaken in this project offer insight into the features that affect the mortality of ICU patients. The following discussion reviews:

i. the practical applications of the results derived from white-box models
ii. how the results contribute to existing literature using Survival Analysis
iii. how the results can practically aid in the decision-making process of ICU wards
iv. limitations of the analysis and discusses the insights for future research.

7.1 Implications of Results

The findings of this project have practical implications for doctors in practice, ICU ward staff and the management teams of hospitals. These key personnel can thus use this information to their advantage, making more confident, data-driven decisions.

7.1.1 <u>Resource Allocation</u>

As the significant features that affect patient's length of stay and mortality can be clearly delineated due to white-box models, the findings help highlight key health indicators that are more vital to observe for mortality.

The findings from the Survival Analysis, unique to this project, helps to discern the health indicators that affect the length of stay in ICU wards. For example, a patient in a critical coma is estimated to have a shorter length of stay. Although Survival Analysis allows for the prediction using multiple health indicators, patients with similar significant health indicators may still be administered different prognosis due to differing survival functions.

The findings suggest that decisions made by hospitals regarding allocating their limited resources to best use can consider the number of beds allocated to the type of ICU wards. For example, Medical ICU wards take in more types of patients and require more beds to cater to increased number of patients.

Resource allocation also include doctors and healthcare professionals. The deployment of personnel with the right expertise to attend to patients when the ICU wards admit patients who show high risk (based on the significant features identified by the above analyses). For example, doctors and nurses who have the right expertise or experience in attending to past similar cases as the current patient's condition.

Further, the results show that seniors and patients in coma are at higher risk of passing away in hospital. Management team of hospitals who makes decisions on departmental assignments can also consider assigning nurses who have more knowledge and skills in caring for such patients to the ICU wards.

7.1.2 Timely and Effective Care

The findings also help highlight health indicators that are more vital to observe than the others for predicting mortality of ICU patients. Decisions doctors and healthcare professionals have to make, such as tailoring the level of critical care required by each individual patient, and prioritization of medical attention, can be reinforced using the above findings. The estimates obtained from the Survival Analysis can be used by doctors to devise suitable treatments or counsel their patient and family members about the prognosis.

These findings can aid doctors and healthcare professionals in narrowing down the type of tests rendered to patients, as well as ensuring the type of tests that are non-negotiable. For example, the results show that blood urea nitrogen is a key health indicator for patient mortality. It is likely that doctors must take care to measure this for patients in critical conditions. This is not only efficient; it can also be more effective because it can cut down the time required to obtain the results and administer timely intervention for the patients' condition. ICU patients care strategies are time sensitive as their conditions are critical and can change drastically in short periods of time. It can also suggest level of critical care required by each individual patient as this project aims to help predict mortality at the individual level.

Providing effective care advice by drawing out key health indicators can also be achieved by drawing on prior experience. Specifically, doctors can better understand the condition an individual is currently experiencing by comparing to past patients who may have similar test results. Doing so and learning from past experience can lead to timely intervention and more fortunately, reduce the risk of mortality.

7.2 Limitations and Future Research

This project is subject to 4 limitations.

First, the modelling and results are subject to limitations of the dataset. This includes:

i. The dataset only provides for 48 hours of data and patients in ICU are likely to have stayed longer, particularly patients in a coma.
ii. The intervention effects of patients' care and medication administered are not captured. For example, doctors and healthcare professionals are likely to prescribe medication if high cholesterol is detected. Future building of datasets can consider taking these effects into account and may be able to establish more stable findings.
iii. The dataset is an aggregation of various hospitals. To help in better forecasting, future research can consider using the dataset of a particular hospital, and the availability of inventory and resources lists.

Second, only one summary statistic (median or first reading) of each feature is included in the model building. This is due to 2 reasons – multicollinearity issues and allowing for a basis of comparison between Survival Analysis and classification models. Future research can consider employing various (if not all) summary statistics for each feature in feature selection to build a more robust model.

Third, there are trade-offs depending on which metric the model performance is evaluated on. Type 2 errors (i.e. less false negative predictions where patients are predicted to survive but died) are to be minimised in the medical context. This may lead to unintended negative psychological consequences on both parties (patient and healthcare professional). For example, the Pygmalion effect of a poor prognosis may result in the patients succumbing to despair or depression. Even though Type 1 errors (i.e. predicting more patients to die but actually survive) may lead to over-provision in terms of resource allocation, the medical context prefers a more conservative approach.

Fourth, the survival outcome of a patient is a complex interaction between multiple factors such as genetic make-up, medical interventions and environmental conditions. While the model may provide statistically good performance given the extant health indicators of a patient, it is challenging to include and capture all the other agnostic factors. This also emphasizes the need for human judgement in medical assessments, in spite of a robust statistical model.

Future research can also consider a nearest k-neighbours clustering approach (clusters of patients in the past cohort with demographics and vitals similar to current patient). This approach can be adopted for case-based prediction to improve on a patient’s prognosis and aid in providing timely intervention and effective care strategy.

References

"*ABIM Laboratory Test Reference Ranges − January 2020*". American Board of Internal Medicine, N.D., https://www.abim.org/~/media/ABIM%20Public/Files/pdf/exam/laboratory-reference-ranges.pdf.

Aczon, M., Ledbetter, D., Ho, L., Gunny, A., Flynn, A., Williams, J., & Wetzel, R. (2017). *Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks*. arXiv preprint arXiv:1701.06675.

Bhattacharya, S., Rajan, V., & Shrivastava, H. (2017). *ICU mortality prediction: A classification algorithm for imbalanced datasets*. In Thirty-first AAAI Conference on Artificial Intelligence.

Bosnjak, A., & Montilla, G. (2012). *Predicting mortality of ICU patients using statistics of physiological variables and support vector machines*. In 2012 Computing in Cardiology (pp. 481-484). IEEE.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). *Purposeful selection of variables in Logistic Regression*. In Source Code for Biology and Medicine, Issue 3, (pp. 17).

Chicco, D., & Jurman, G. (2020). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation*. In BMC genomics, Issue 21, (pp. 6).

Citi, L., & Barbieri, R. (2012). *PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm*. In 2012 Computing in Cardiology (pp. 257-260). IEEE.

Ghanvatkar, S., & Rajan, V. (2019). *Deep recurrent neural networks for mortality prediction in intensive care using clinical time series at multiple resolutions*.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). *Multitask learning and benchmarking with clinical time series data*. In Scientific Data, Issue 6, (pp. 1-18).

Johnson, A. E., Dunkley, N., Mayaud, L., Tsanas, A., Kramer, A. A., & Clifford, G. D. (2012). *Patient specific predictions in the intensive care unit using a Bayesian ensemble*. In 2012 Computing in Cardiology (pp. 249-252). IEEE.

Johnson, A. E., Kramer, A. A., and Clifford, G. D. (2014). *Data preprocessing and mortality prediction: the Physionet/CinC 2012 challenge revisited*. In Computing in Cardiology Conference (CinC) (pp. 157–160). IEEE.

Johnson, A.E., & Mark, R.G. (2017). *Real-time mortality prediction in the intensive care unit*. In AMIA Symposium (pp. 994-1003).

Johnson, A. E., Pollard, T. J., & Mark, R. G. (2017). *Reproducibility in critical care: a mortality prediction case study*. In Machine Learning for Healthcare Conference (pp. 361-376).

Kakade, A., Kumari, B., & Dholaniya, P. S. (2018). *Feature selection using Logistic Regression in case–control DNA methylation data of Parkinson's disease: A comparative study*. In Journal of Theoretical Biology, Issue 457, (pp. 14-18).

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I. H., & Kim, H. J. (2019). *Deep learning-based survival prediction of oral cancer patients*. In Scientific Reports, Issue 9, (pp. 1-10).

Lee, D.H., & Horvitz, E. (2017). *Predicting mortality of intensive care patients via learning about hazard*. In AAAI Conference on Artificial Intelligence.

"*Liver Function Tests*." Mayo Clinic, Mayo Foundation for Medical Education and Research, 13 June 2019, www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595.

Luo, Y., Xin, Y., Joshi, R., Celi, L., & Szolovits, P. (2016). *Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements*. In Thirtieth AAAI Conference on Artificial Intelligence.

Lynn, R., Harvey, J., & Nyborg, H. (2009). *Average intelligence predicts atheism rates across 137 nations*. In Intelligence, Issue 37, (pp. 11-15).

Morid, M.A., Sheng, O.R., & Abdelrahman, S.E. (2017). *PPMF: A patient-based predictive modeling framework for early ICU mortality prediction*.

Nguyen, P., Tran, T., & Venkatesh, S. (2017). *Deep learning to attend to risk in ICU*.

Nielsen, A. B., Thorsen-Meyer, H. C., Belling, K., Nielsen, A. P., Thomas, C. E., Chmura, P. J., & Spangsege, L. (2019). *Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records*. In The Lancet Digital Health, Issue 1(2), (pp. 78-89).

"*Normal Hemodynamic Parameters and Laboratory Values*", Edwards Life Sciences 2009, N.D., http://ht.edwards.com/scin/edwards/sitecollectionimages/edwards/products/presep/ar04313hemodynpocketcard.pdf

"*Normal Laboratory Values: Blood, Plasma, and Serum*". Merck Manuals 2018, N.D., https://www.merckmanuals.com/professional/resources/normal-laboratory-values/blood-tests-normal-values#v8508814.

Sadeghi, R., Banerjee, T., & Romine, W. (2018). *Early hospital mortality prediction using vital signals*. In Smart Health, Issue 9, (pp. 265-274).

Samanta, B., Bird, G. L., Kuijpers, M., Zimmerman, R. A., Jarvik, G. P., Wernovsky, G., & Nataraj, C. (2009). *Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using Logistic Regression and Decision Tree algorithms*. In Artificial Intelligence in Medicine, Issue 46, (pp. 201-215).

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates

Student, S., Płuciennik, A., Jakubczak, M., & Fujarewicz, K. (2018). *Feature Selection Based on Logistic Regression for 2-Class Classification of Multidimensional Molecular Data*. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications (pp. 286-290).

Tan, R., Ding, S., Pan, J., & Qiu, Y. (2019). *ICU mortality prediction based on key risk factors identification*. In International Conference on Health Information Science (pp. 89-97).

Ting, H. W., Chen, M. S., Hsieh, Y. C., & Chan, C. L. (2010). *Good mortality prediction by Glasgow Coma Scale for neurosurgical patients*. In Journal of the Chinese Medical Association, Issue 73, (pp. 139-143).

"*Understanding Blood Pressure Readings*." www.heart.org, 2017, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

"*Vital Signs*", U.S. National Library of Medicine, 23 Mar 2020, https://medlineplus.gov/ency/article/002341.htm

Weathers, B. (2017). *Comparison of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis*.

Worthington, R. L., & Whittaker, T. A. (2006). *Scale development research: A content analysis and recommendations for best practices*. In The Counseling Psychologist, Issue 34(6), (pp. 806-838).

Zakharov, R., & Dupont, P. (2011). *Ensemble Logistic Regression for feature selection*. In IAPR International Conference on Pattern Recognition in Bioinformatics (pp. 133-144). Springer, Berlin, Heidelberg.

Appendix A: Statistical Summary of Independent Variables

Variable	ALP	ALT	AST	Age	Albumin	BUN	Bilirubin	Cholesterol	Creatinine	DiasABP	FiO2	GCS
count	3090	3175	3180	4000	2355	13907	3190	315	13974	145567	32390	61563
mean	116.74	394.61	506.53	64.25	2.92	27.42	2.91	156.52	1.51	59.29	0.55	11.40
std	133.94	1200.53	1516.87	17.56	0.65	23.40	5.91	46.07	1.64	13.32	0.19	3.97
min	12	1	4	15	1	0	0.1	28	0.1	0	0.21	3
25%	59	20	31	52.75	2.5	13	0.5	123	0.7	51	0.4	8
50%	82	43	64	67	2.9	20	0.9	152	1	58	0.5	13
75%	122	162	209	78	3.4	33	2.3	188	1.5	67	0.6	15
max	2205	11470	18430	90	5.3	197	47.7	330	22.1	268	1	15

Variable	Gender	Glucose	HCO3	HCT	HR	Height	ICUType	K	Lactate	MAP	MechVent	Mg
count	4000	13011	13601	18257	228538	4000	4000	14430	8024	143896	31144	13585
mean	0.56	141.47	23.12	30.68	87.52	88.92	2.76	4.14	2.92	79.77	1.00	2.03
std	0.50	67.48	4.71	5.01	18.41	86.53	1.00	0.71	2.58	16.96	0.00	0.42
min	-1	10	5	9	0	-1	1	1.8	0.3	0	1	0.6
25%	0	105	20	27.3	75	-1	2	3.7	1.4	69	1	1.8
50%	1	127	23	30.3	86	152.4	3	4.1	2.1	77	1	2
75%	1	157	26	33.5	99	170.2	4	4.5	3.4	88	1	2.2
max	1	1143	50	61.8	300	431.8	4	22.9	29.3	300	1	9.9

Variable	NISysABP	Na	PaCO2	PaO2	Platelets	RecordID	RespRate	SaO2	SysABP	Temp	TroponinI	TroponinT
count	98331	13560	23293	23268	14095	4000	55043	8185	145650	86202	435	2123
mean	118.59	139.07	40.47	150.42	190.80	137605.12	19.72	96.64	118.70	37.03	7.15	1.20
std	23.26	5.19	9.13	89.30	106.39	2923.61	5.55	3.40	25.02	1.48	9.77	2.72
min	0	98	0.3	0	6	132539	0	26	0	-17.8	0.3	0.01
25%	102	136	35	90	119	135075.8	16	96	102	36.6	0.9	0.06
50%	116	139	39	121	172	137592.5	19	97	116	37.1	2.6	0.2
75%	133	142	45	176	238	140100.3	23	98	133	37.6	10	1.02
max	296	177	100	500	1047	142673	98	100	295	42.1	49.2	24.91

Variable	Urine	WBC	Weight	pH
count	135358	12900	129165	24355
mean	119.40	12.67	83.39	7.49
std	175.18	7.64	25.05	8.24
min	0	0.1	-1	1
25%	36	8.3	67	7.33
50%	70	11.4	80.6	7.38
75%	140	15.4	96	7.43
max	11000	187.5	300	735

Appendix B: Correlation Analysis

	ALPmedian	ALTmedian	ASTmedian	Albuminmedian	BUNmedian	Bilirubinmedian	Cholesterolmedian	Creatininemedian	FiO2median	Glucosemedian	HCO3median	HCTmedian	HRmedian	Kmedian	Lactatemedian	Mgmedian	Namedian	PaCO2median	PaO2median	Plateletsmedian	RespRatemedian	SaO2median	TroponinImedian	TroponinTmedian	WBCmedian	pHmedian	newTempmedian	PaO2FiO2median	Agemean	newHeightfirst	newWeightfirst	Urinemean48h	MeanMAPfirst	MeanDiasABPfirst	MeanSysABPfirst
ALPmedian		0.11	0.16	-0.14	0.15	0.24	0.00	0.13	-0.01	0.00	-0.10	-0.01	0.02	0.02	0.07	0.05	-0.03	-0.04	-0.06	0.07	-0.01	-0.02	-0.01	-0.02	0.08	-0.06	-0.08	0.01	0.00	-0.02	-0.01	-0.01	-0.04	-0.04	-0.06
ALTmedian	0.11		**0.83**	-0.02	0.04	0.10	-0.01	0.07	0.02	0.03	-0.08	0.05	0.09	-0.02	0.28	0.02	0.02	-0.08	-0.03	-0.08	0.01	-0.06	-0.01	0.04	0.01	-0.01	-0.03	0.01	-0.11	0.01	-0.01	0.01	-0.01	0.00	-0.04
ASTmedian	0.16	**0.83**		-0.04	0.05	0.11	-0.01	0.09	0.05	0.06	-0.12	0.03	0.10	0.03	0.40	0.04	0.02	-0.09	-0.03	-0.08	0.01	-0.08	-0.01	0.09	0.03	-0.04	-0.05	0.01	-0.08	0.00	0.00	0.00	-0.02	0.00	-0.06
Albuminmedian	-0.14	-0.02	-0.04		-0.10	-0.09	0.07	-0.03	-0.02	0.04	0.19	0.22	-0.13	-0.06	-0.10	0.02	0.03	0.09	0.08	0.02	-0.03	0.03	0.01	0.04	-0.10	0.10	0.01	-0.02	-0.04	0.02	0.01	0.02	0.13	0.16	0.17
BUNmedian	0.15	0.04	0.05	-0.10		0.18	-0.02	0.68	0.03	0.11	-0.24	-0.09	-0.07	0.28	0.06	0.30	0.03	-0.06	-0.13	-0.03	0.00	-0.03	0.07	0.04	0.10	-0.20	-0.25	0.00	0.23	0.02	0.08	-0.01	-0.10	-0.13	-0.05
Bilirubinmedian	0.24	0.10	0.11	-0.09	0.18		-0.02	0.14	0.02	-0.03	-0.13	-0.04	0.02	-0.03	0.14	0.12	-0.07	-0.10	-0.06	-0.15	-0.02	0.00	-0.01	-0.02	0.02	-0.01	-0.11	-0.02	-0.06	0.04	0.04	-0.02	-0.03	-0.05	-0.06
Cholesterolmedian	0.00	-0.01	-0.01	0.07	-0.02	-0.02		-0.03	0.00	0.03	0.03	0.08	-0.04	0.01	-0.01	0.03	-0.02	0.02	0.02	0.00	0.01	0.02	-0.02	0.03	-0.01	0.01	0.02	0.01	-0.02	-0.03	-0.01	-0.01	0.04	0.03	0.03
Creatininemedian	0.13	0.07	0.09	-0.03	0.68	0.14	-0.03		0.00	0.01	-0.23	-0.06	-0.04	0.30	0.07	0.16	-0.05	-0.09	-0.07	-0.04	-0.02	-0.02	0.03	0.05	0.03	-0.19	-0.16	0.02	0.03	0.05	0.10	-0.03	-0.01	-0.04	0.03
FiO2median	-0.01	0.02	0.05	-0.02	0.03	0.02	0.00	0.00		0.05	0.00	0.06	0.09	0.05	0.07	0.04	-0.05	0.03	-0.17	0.05	0.10	-0.09	0.02	0.01	0.04	-0.07	-0.04	-0.08	0.02	0.02	0.04	0.05	-0.04	-0.04	-0.06
Glucosemedian	0.00	0.03	0.06	0.04	0.11	-0.03	0.03	0.01	0.05		-0.06	0.03	0.04	0.07	0.13	0.04	-0.02	0.00	-0.02	0.00	0.02	0.02	0.04	0.03	0.06	-0.02	0.00	-0.04	0.06	0.04	0.09	0.01	0.03	0.01	0.04
HCO3median	-0.10	-0.08	-0.12	0.19	-0.24	-0.13	0.03	-0.23	0.00	-0.06		0.10	-0.10	-0.10	-0.18	-0.03	0.10	0.62	-0.02	0.07	-0.02	-0.05	-0.02	-0.05	-0.12	0.28	0.06	-0.02	0.00	0.03	0.06	0.09	0.10	0.08	0.13
HCTmedian	-0.01	0.05	0.03	0.22	-0.09	-0.04	0.08	-0.06	0.06	0.03	0.10		-0.07	-0.07	-0.01	0.01	0.11	0.06	-0.10	0.10	0.00	-0.04	0.00	0.08	0.05	-0.03	-0.07	-0.04	-0.05	0.05	0.06	0.03	0.18	0.23	0.12
HRmedian	0.02	0.09	0.10	-0.13	-0.07	0.02	-0.04	-0.04	0.09	0.04	-0.10	-0.07		0.02	0.14	-0.08	-0.02	-0.01	-0.05	0.02	0.17	-0.03	-0.01	-0.02	0.10	-0.11	0.24	-0.03	-0.25	0.02	0.01	0.04	-0.02	0.09	-0.13
Kmedian	0.02	-0.02	0.03	-0.06	0.28	-0.03	0.01	0.30	0.05	0.07	-0.10	-0.07	0.02		0.08	0.25	-0.23	0.13	-0.01	0.00	-0.05	-0.04	0.01	0.01	0.07	-0.30	-0.06	0.03	0.09	0.03	0.12	0.00	-0.15	-0.14	-0.13
Lactatemedian	0.07	0.28	0.40	-0.10	0.06	0.14	-0.01	0.07	0.07	0.13	-0.18	-0.01	0.14	0.08		0.03	0.03	-0.17	0.00	-0.14	0.01	-0.06	-0.01	0.03	0.06	-0.10	-0.01	-0.03	-0.05	0.02	0.02	-0.03	-0.05	-0.04	-0.11
Mgmedian	0.05	0.02	0.04	0.02	0.30	0.12	0.03	0.16	0.04	0.04	-0.03	0.01	-0.08	0.25	0.03		0.00	0.01	-0.03	-0.01	0.02	-0.01	0.00	0.04	0.06	-0.05	-0.12	0.01	0.13	-0.02	0.06	0.02	-0.05	-0.03	-0.05
Namedian	-0.03	0.02	0.02	0.03	0.03	-0.07	-0.02	-0.05	-0.05	-0.02	0.10	0.11	-0.02	-0.23	0.03	0.00		0.06	-0.01	-0.05	0.00	-0.01	-0.01	-0.03	0.00	0.02	0.04	-0.01	0.00	0.01	-0.03	0.03	0.08	0.07	0.08
PaCO2median	-0.04	-0.08	-0.09	0.09	-0.06	-0.10	0.02	-0.09	0.03	0.00	0.62	0.06	-0.01	0.13	-0.17	0.01	0.06		-0.11	0.08	-0.04	-0.14	-0.02	-0.05	-0.03	-0.30	0.02	-0.02	-0.02	0.04	0.11	0.10	0.00	-0.01	0.03
PaO2median	-0.06	-0.03	-0.03	0.08	-0.13	-0.06	0.02	-0.07	-0.17	-0.02	-0.02	-0.10	-0.05	-0.01	0.00	-0.03	-0.01	-0.11		-0.10	-0.07	0.21	-0.03	0.00	-0.05	0.09	0.04	0.09	-0.06	0.00	-0.10	-0.03	0.02	0.01	0.01
Plateletsmedian	0.07	-0.08	-0.08	0.02	-0.03	-0.15	0.00	-0.04	0.05	0.00	0.07	0.10	0.02	0.00	-0.14	-0.01	-0.05	0.08	-0.10		0.07	0.01	0.01	0.00	0.25	-0.01	-0.05	-0.01	-0.02	-0.01	-0.03	0.02	0.04	0.06	0.06
RespRatemedian	-0.01	0.01	0.01	-0.03	0.00	-0.02	0.01	-0.02	0.10	0.02	-0.02	0.00	0.17	-0.05	0.01	0.02	0.00	-0.04	-0.07	0.07		-0.04	0.01	0.03	0.07	0.05	0.02	0.00	0.07	-0.02	-0.04	0.01	-0.04	-0.03	-0.05
SaO2median	-0.02	-0.06	-0.08	0.03	-0.03	0.00	0.02	-0.02	-0.09	0.02	-0.05	-0.04	-0.03	-0.04	-0.06	-0.01	-0.01	-0.14	0.21	0.01	-0.04		0.01	-0.01	-0.03	0.09	0.03	0.05	0.01	-0.01	-0.05	-0.01	0.03	0.03	0.05
TroponinImedian	-0.01	-0.01	-0.01	0.01	0.07	-0.01	-0.02	0.03	0.02	0.04	-0.02	0.00	-0.01	0.01	-0.01	0.00	-0.01	-0.02	-0.03	0.01	0.01	0.01		-0.02	0.00	0.01	0.03	-0.01	0.04	0.00	0.00	-0.01	0.01	0.01	0.00
TroponinTmedian	-0.02	0.04	0.09	0.04	0.04	-0.02	0.03	0.05	0.01	0.03	-0.05	0.08	-0.02	0.01	0.03	0.04	-0.03	-0.05	0.00	0.00	0.03	-0.01	-0.02		0.03	-0.01	0.00	0.00	0.05	-0.01	-0.02	-0.02	-0.01	0.02	-0.07
WBCmedian	0.08	0.01	0.03	-0.10	0.10	0.02	-0.01	0.03	0.04	0.06	-0.12	0.05	0.10	0.07	0.06	0.06	0.00	-0.03	-0.05	0.25	0.07	-0.03	0.00	0.03		-0.09	0.05	-0.02	0.04	-0.02	0.03	0.00	-0.08	-0.08	-0.08
pHmedian	-0.06	-0.01	-0.04	0.10	-0.20	-0.01	0.01	-0.19	-0.07	-0.02	0.28	-0.03	-0.11	-0.30	-0.10	-0.05	0.02	-0.30	0.09	-0.01	0.05	0.09	0.01	-0.01	-0.09		0.10	-0.01	0.03	-0.01	-0.07	0.00	0.12	0.12	0.13
newTempmedian	-0.08	-0.03	-0.05	0.01	-0.25	-0.11	0.02	-0.16	-0.04	0.00	0.06	-0.07	0.24	-0.06	-0.01	-0.12	0.04	0.02	0.04	-0.05	0.02	0.03	0.03	0.00	0.05	0.10		0.01	-0.20	0.03	0.11	0.06	0.05	0.05	0.04
PaO2FiO2median	0.01	0.01	0.01	-0.02	0.00	-0.02	0.01	0.02	-0.08	-0.04	-0.02	-0.04	-0.03	0.03	-0.03	0.01	-0.01	-0.02	0.09	-0.01	0.00	0.05	-0.01	0.00	-0.02	-0.01	0.01		0.01	-0.03	-0.04	-0.01	-0.03	-0.02	-0.02
Agemean	0.00	-0.11	-0.08	-0.04	0.23	-0.06	-0.02	0.03	0.02	0.06	0.00	-0.05	-0.25	0.09	-0.05	0.13	0.00	-0.02	-0.06	-0.02	0.07	0.01	0.04	0.05	0.04	0.03	-0.20	0.01		-0.17	-0.17	-0.11	-0.10	-0.20	0.01
newHeightfirst	-0.02	0.01	0.00	0.02	0.02	0.04	-0.03	0.05	0.02	0.04	0.03	0.05	0.02	0.03	0.02	-0.02	0.01	0.04	0.00	-0.01	-0.02	-0.01	0.00	-0.01	-0.02	-0.01	0.03	-0.03	-0.17		0.35	0.04	0.02	0.08	-0.01
newWeightfirst	-0.01	-0.01	0.00	0.01	0.08	0.04	-0.01	0.10	0.04	0.09	0.06	0.06	0.01	0.12	0.02	0.06	-0.03	0.11	-0.10	-0.03	-0.04	-0.05	0.00	-0.02	0.03	-0.07	0.11	-0.04	-0.17	0.35		0.03	0.00	0.02	-0.03
Urinemean48h	-0.01	0.01	0.00	0.02	-0.01	-0.02	-0.01	-0.03	0.05	0.01	0.09	0.03	0.04	0.00	-0.03	0.02	0.03	0.10	-0.03	0.02	0.01	-0.01	-0.01	-0.02	0.00	0.00	0.06	-0.01	-0.11	0.04	0.03		0.04	0.04	0.03
MeanMAPfirst	-0.04	-0.01	-0.02	0.13	-0.10	-0.03	0.04	-0.01	-0.04	0.03	0.10	0.18	-0.02	-0.15	-0.05	-0.05	0.08	0.00	0.02	0.04	-0.04	0.03	0.01	-0.01	-0.08	0.12	0.05	-0.03	-0.10	0.02	0.00	0.04		**0.75**	**0.70**
MeanDiasABPfirst	-0.04	0.00	0.00	0.16	-0.13	-0.05	0.03	-0.04	-0.04	0.01	0.08	0.23	0.09	-0.14	-0.04	-0.03	0.07	-0.01	0.01	0.06	-0.03	0.03	0.01	0.02	-0.08	0.12	0.05	-0.02	-0.20	0.08	0.02	0.04	**0.75**		0.60
MeanSysABPfirst	-0.06	-0.04	-0.06	0.17	-0.05	-0.06	0.03	0.03	-0.06	0.04	0.13	0.12	-0.13	-0.13	-0.11	-0.05	0.08	0.03	0.01	0.06	-0.05	0.05	0.00	-0.07	-0.08	0.13	0.04	-0.02	0.01	-0.01	-0.03	0.03	**0.70**	0.60	

Appendix C: Results of Univariate Analysis Logistic Regression

Variable	LogR coefficient	LogR t-value	LogR p-value	Wald p-value
Hypoglycemia1	22.57	0.00	**1.00**	0.00
GCSComa1	0.23	1.74	**0.08**	0.00
GCSmin4	0.82	2.27	0.02	0.00
ALPmedian	28.20	32.14	0.00	0.00
GCSmin13	1.62	7.24	0.00	0.00
MeanSysABPfirst	4.21	38.30	0.00	0.00
Hypertension1	1.69	18.41	0.00	0.00
Tachycardia1	1.72	31.77	0.00	0.00
GCSmin5	1.25	3.49	0.00	0.00
GCSmin6	1.47	9.82	0.00	0.00
GCSmin7	1.33	9.08	0.00	0.00
GCSmin8	1.87	8.89	0.00	0.00
GCSmin9	1.91	7.54	0.00	0.00
GCSmin10	2.10	9.10	0.00	0.00
GCSmin11	2.10	6.87	0.00	0.00
GCSmin12	1.97	4.52	0.00	0.00
GCSmin15	2.47	17.76	0.00	0.00
GCSmin14	2.17	12.31	0.00	0.00
ALTmedian	9.21	5.01	0.00	0.00
Diabetes1	1.58	12.37	0.00	0.00
Diabetes2	1.32	15.77	0.00	0.00
Hyperglycemia1	1.13	9.94	0.00	0.00
RenalInjury1	0.93	7.37	0.00	0.00
BMICat0	1.75	31.23	0.00	0.00
BMICat1	2.05	17.96	0.00	0.00
BMICat2	1.99	16.42	0.00	0.00
MechVentmax1	1.69	30.81	0.00	0.00
ICUTypemax2	2.96	18.94	0.00	0.00
ICUTypemax3	1.48	22.13	0.00	0.00
ICUTypemax4	1.77	20.40	0.00	0.00
MeanDiasABPfirst	5.02	38.16	0.00	0.00
Urinemean48h	124.63	28.00	0.00	0.00
newWeightfirst	4.04	36.90	0.00	0.00
newHeightfirst	3.59	38.22	0.00	0.00
ASTmedian	6.10	4.13	0.00	0.00
Albuminmedian	4.17	39.32	0.00	0.00
BUNmedian	7.35	24.87	0.00	0.00
Bilirubinmedian	8.77	7.86	0.00	0.00
Cholesterolmedian	4.41	39.73	0.00	0.00
Creatininemedian	14.63	23.51	0.00	0.00
FiO2median	5.20	34.77	0.00	0.00
Glucosemedian	8.39	35.08	0.00	0.00
HCO3median	4.70	38.90	0.00	0.00
HCTmedian	4.99	38.15	0.00	0.00
HRmedian	3.73	37.01	0.00	0.00

Kmedian	5.58	37.72	0.00	0.00
Lactatemedian	19.33	30.32	0.00	0.00
Mgmedian	10.06	37.06	0.00	0.00
Namedian	3.66	39.27	0.00	0.00
PaCO2median	6.13	38.84	0.00	0.00
PaO2median	8.36	36.67	0.00	0.00
Plateletsmedian	8.08	34.80	0.00	0.00
RespRatemedian	4.29	39.31	0.00	0.00
SaO2median	1.92	39.93	0.00	0.00
TroponinImedian	39.73	35.51	0.00	0.00
TroponinTmedian	36.47	9.90	0.00	0.00
WBCmedian	17.91	34.53	0.00	0.00
pHmedian	3.39	39.20	0.00	0.00
newTempmedian	3.67	37.78	0.00	0.00
PaO2FiO2median	7.46	39.55	0.00	0.00
Agemean	2.39	35.98	0.00	0.00
Genderfirst1	1.87	30.14	0.00	0.00

Appendix D: Results of Logistic Regression using Features Selected by Univariate Analysis

Model Results:

Feature	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
All	59.2	81.18	43.32	0.36	2859	568	232	336

Dependent Variable:	Inhospitaldeathmax	**Df Residuals:**	2480
Model:	Logit	**Df Model:**	22
Method:	MLE	**Pseudo R-square:**	0.3869
Converged:	True	**Log-Likelihood:**	-810.39
Covariance Type:	nonrobust	**LL-Null:**	-1321.8
No. Observations:	2503	**LLR p-value:**	2.713e-202

Variable	Coefficient	Std Error	z	P>|z|	[0.025	0.975]
Const	-2.5090	1.522	-1.648	0.099	-5.492	0.474
BUNmedian	4.4414	0.586	7.573	0.000	3.292	5.591
HCO3median	-1.2290	0.621	-1.980	0.048	-2.445	-0.013
FiO2median	1.0734	0.360	2.980	0.003	0.367	1.779
pHmedian	1.8742	0.669	2.802	0.005	0.563	3.185
Lactatemedian	6.0745	1.333	4.557	0.000	3.462	8.687
ICUTypemax4	0.6833	0.208	3.291	0.001	0.276	1.090
HRmedian	1.1984	0.551	2.173	0.030	0.118	2.279
GCSComa1	3.5626	0.336	10.598	0.000	2.904	4.221
Agemean	2.9331	0.334	8.787	0.000	2.279	3.587

Variable	Coefficient	Std Error	z	P>|z|	[0.025	0.975]
ALPmedian	3.0330	1.147	2.644	0.008	0.784	5.282
SaO2median	-3.2439	1.435	-2.260	0.024	-6.057	-0.431
Albuminmedian	-2.9395	0.694	-4.237	0.000	-4.299	-1.580
Glucosemedian	3.1457	0.729	4.312	0.000	1.716	4.575
newTempmedian	-1.1474	0.432	-2.658	0.008	-1.993	-0.301
ICUTypemax3	0.8269	0.188	4.390	0.000	0.458	1.196
ICUTypemax2	-1.7411	0.252	-6.896	0.000	-2.236	-1.246
Genderfirst1	-0.3053	0.137	-2.231	0.026	-0.574	-0.037
newWeightfirst	-1.5596	0.423	-3.684	0.000	-2.389	-0.730
RenalInjury1	0.6546	0.224	2.923	0.003	0.216	1.094
MechVentmax1	1.0925	0.152	7.207	0.000	0.795	1.390
Tachycardia1	0.6715	0.176	3.823	0.000	0.327	1.016
Hypertension1	0.6981	0.156	4.479	0.000	0.393	1.004

Variance Inflation Factor:

Variable	VIF
Const	796.662133
BUNmedian	1.281025
HCO3median	1.246165
FiO2median	1.047988
pHmedian	1.161157
Lactatemedian	1.110848
HRmedian	1.216764
Agemean	1.188052
ALPmedian	1.075086
SaO2median	1.041976

Variable	VIF
Albuminmedian	1.089538
Tempmedian	1.197796
Glucosemedian	1.067352
Weightfirst	1.094219

Appendix E: Results of Decision Tree

Model Results:

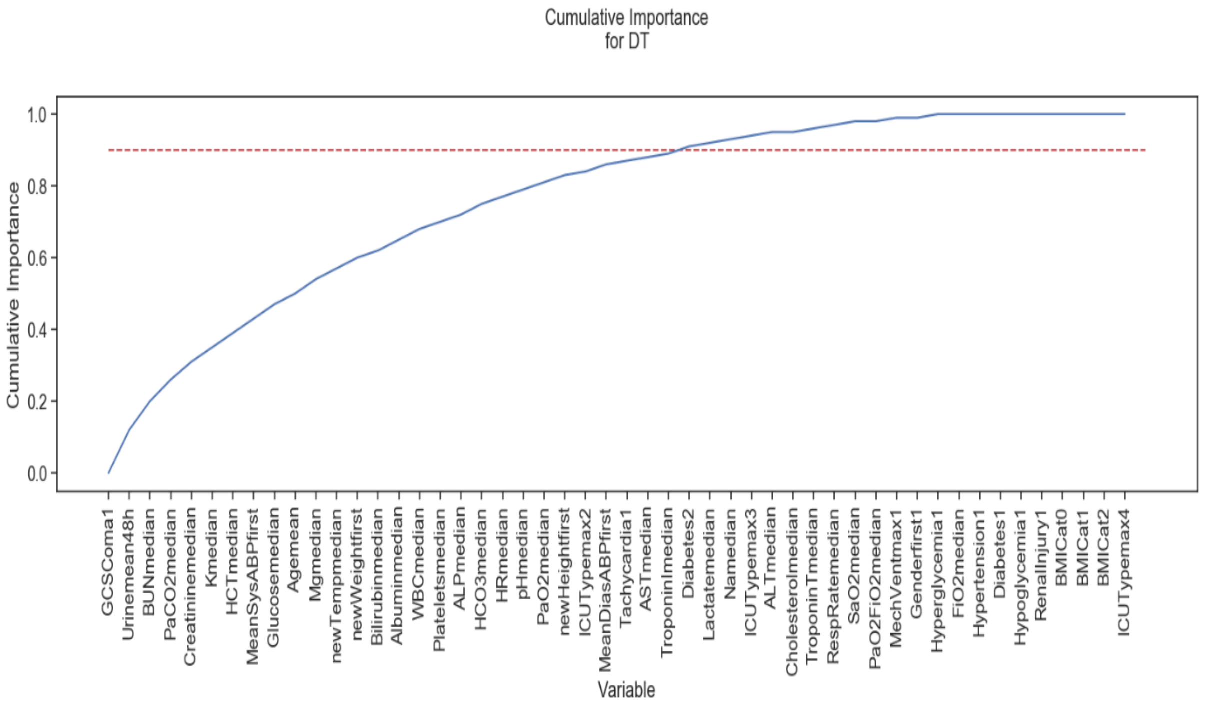
Feature	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
All	51.8	64.05	20.63	0.22	2616	811	274	294

Variable Importance:

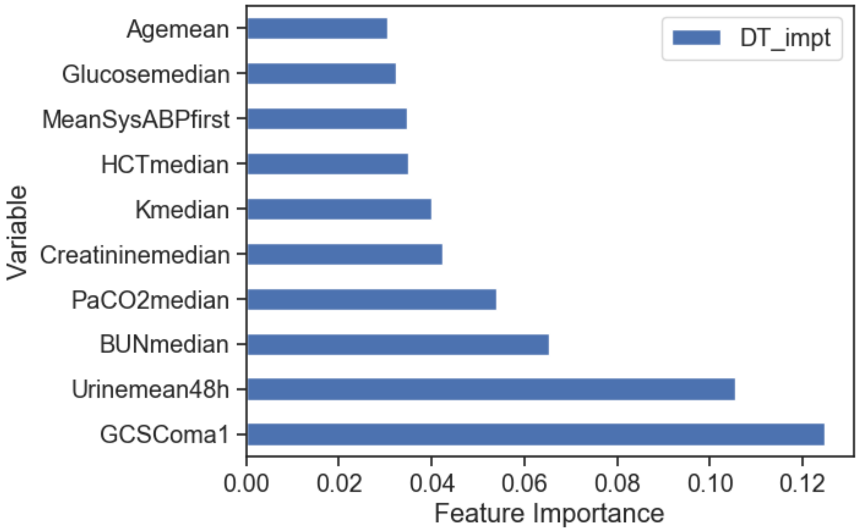
Variable	Variable Importance	Cumulative Variable Importance
GCSComa1	0.12	0.00
Urinemean48h	0.11	0.12
BUNmedian	0.07	0.20
PaCO2median	0.05	0.26
Creatininemedian	0.04	0.31
Kmedian	0.04	0.35
HCTmedian	0.03	0.39
MeanSysABPfirst	0.03	0.43
Glucosemedian	0.03	0.47
Agemean	0.03	0.50
Mgmedian	0.03	0.54
newTempmedian	0.03	0.57
newWeightfirst	0.02	0.60
Bilirubinmedian	0.02	0.62
Albuminmedian	0.02	0.65
WBCmedian	0.02	0.68
Plateletsmedian	0.02	0.70
ALPmedian	0.02	0.72
HCO3median	0.02	0.75
HRmedian	0.02	0.77
pHmedian	0.02	0.79
PaO2median	0.02	0.81
newHeightfirst	0.02	0.83
ICUTypemax2	0.01	0.84
MeanDiasABPfirst	0.01	0.86
Tachycardia1	0.01	0.87
ASTmedian	0.01	0.88

Variable	Variable Importance	Cumulative Variable Importance
TroponinImedian	0.01	0.89
Diabetes2	0.01	0.91
Lactatemedian	0.01	0.92
Namedian	0.01	0.93
ICUTypemax3	0.01	0.94
ALTmedian	0.01	0.95
Cholesterolmedian	0.01	0.95
TroponinTmedian	0.01	0.96
RespRatemedian	0.01	0.97
SaO2median	0.01	0.98
PaO2FiO2median	0.00	0.98
MechVentmax1	0.00	0.99
Genderfirst1	0.00	0.99

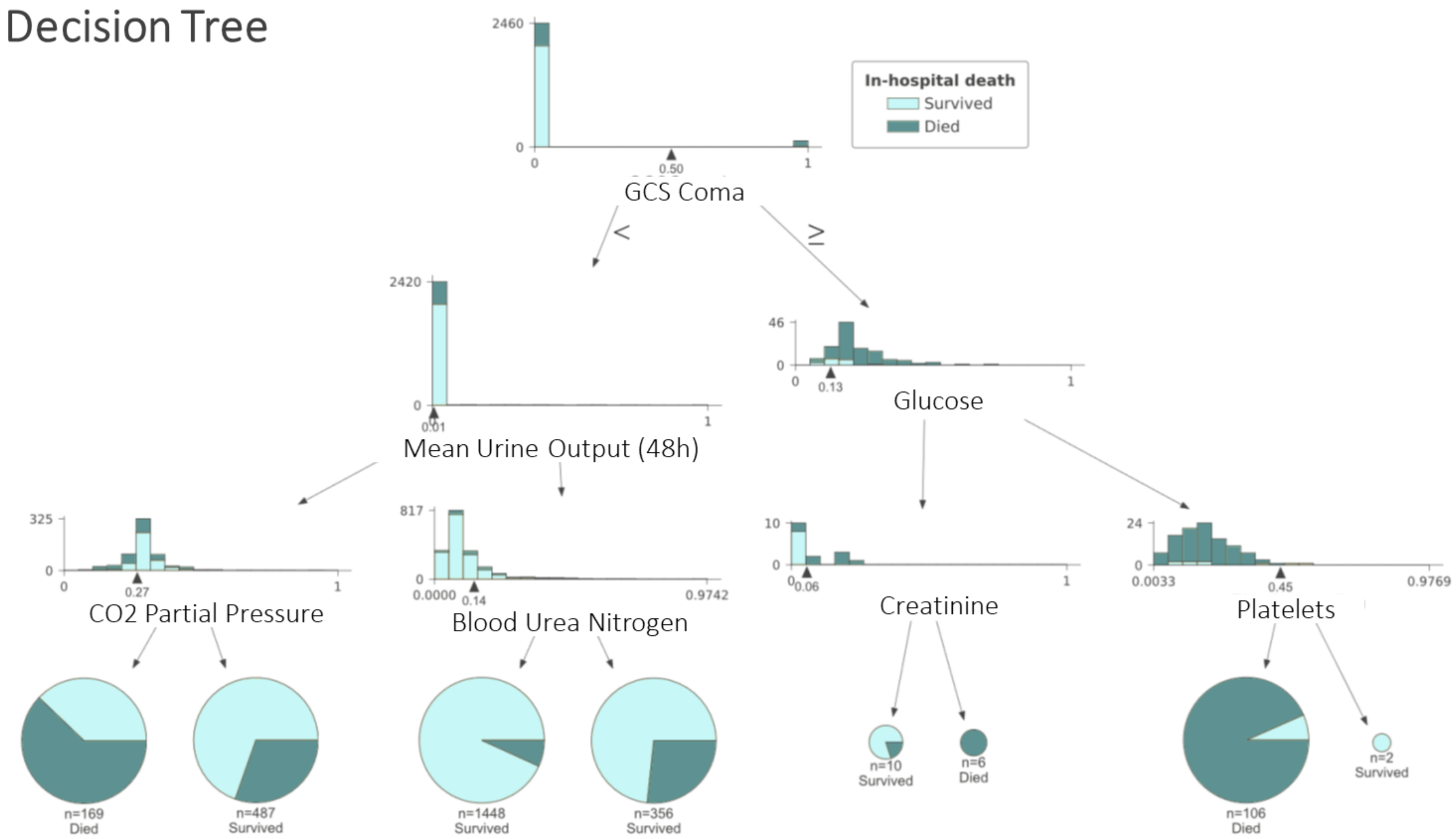
Cumulative Variable Importance:



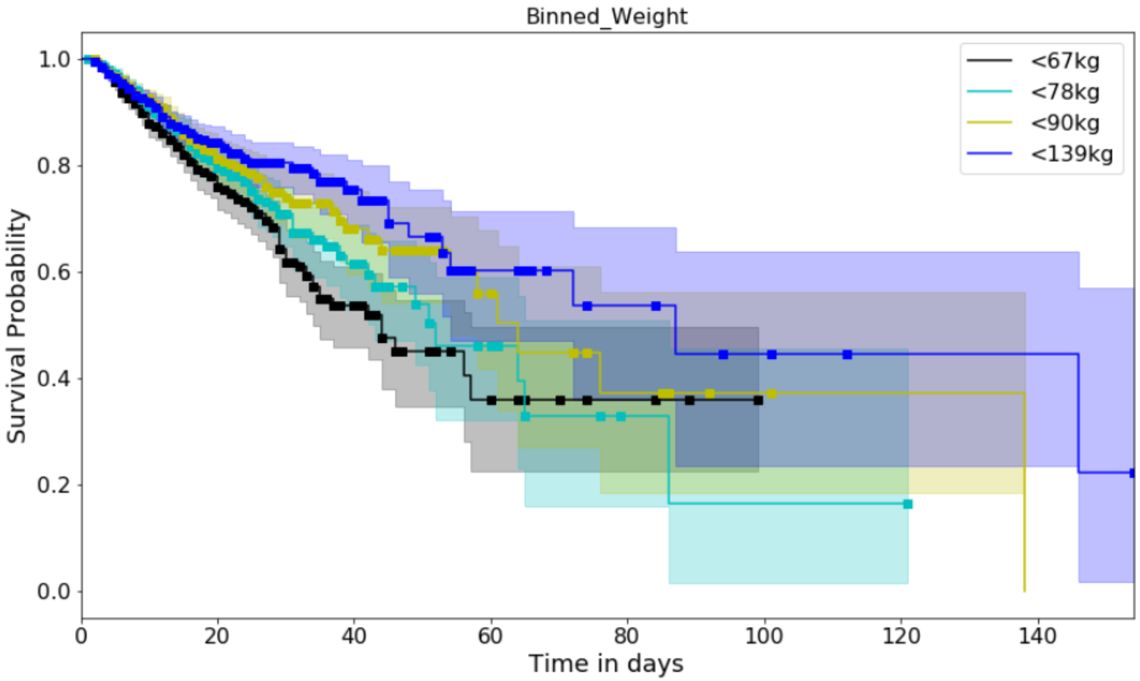
Top 10 Variable Importance:



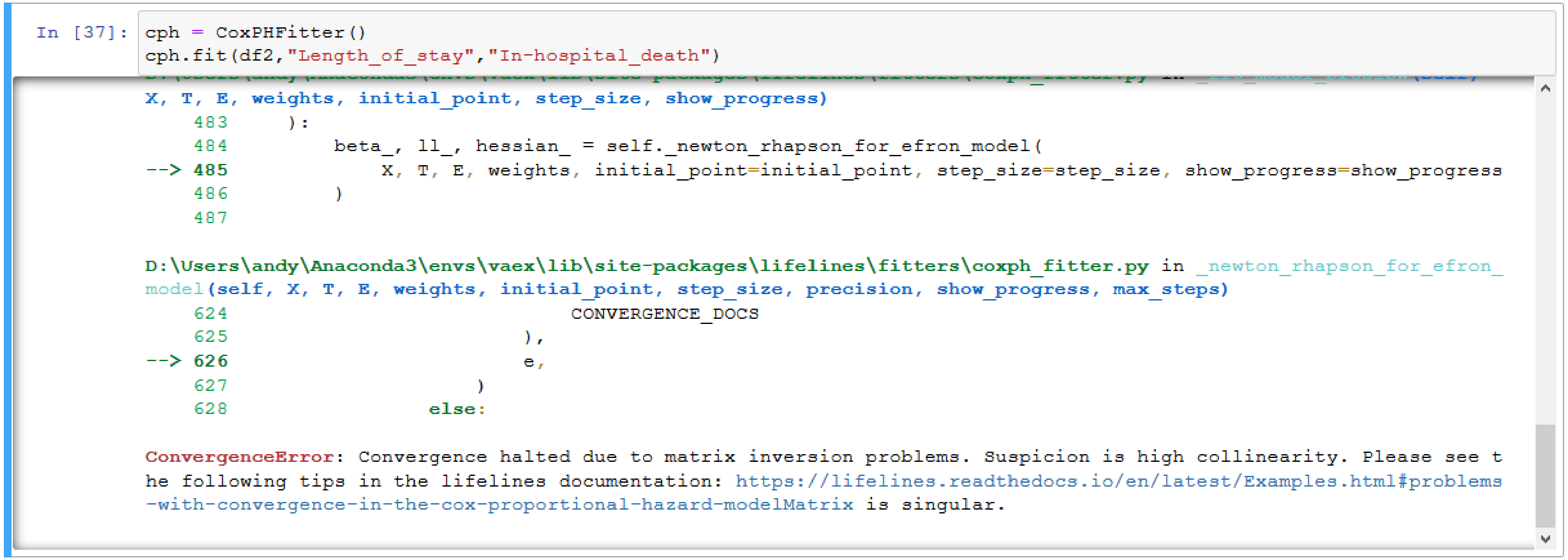
Top 3 Levels of Decision Tree:



Appendix F: Results of Kaplan Meier Estimator with 95% Confidence Interval



Appendix G: Multicollinearity Issues Highlighted By Lifelines CPH Model



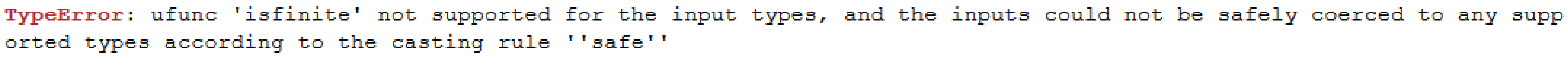
```
In [37]: cph = CoxPHFitter()
         cph.fit(df2,"Length_of_stay","In-hospital_death")
```

```
X, T, E, weights, initial_point, step_size, show_progress)
    483     ):
    484         beta_, ll_, hessian_ = self._newton_rhapson_for_efron_model(
--> 485             X, T, E, weights, initial_point=initial_point, step_size=step_size, show_progress=show_progress
    486         )
    487

D:\Users\andy\Anaconda3\envs\vaex\lib\site-packages\lifelines\fitters\coxph_fitter.py in _newton_rhapson_for_efron_
model(self, X, T, E, weights, initial_point, step_size, precision, show_progress, max_steps)
    624                             CONVERGENCE_DOCS
    625                         ),
--> 626                         e,
    627                     )
    628                 else:

ConvergenceError: Convergence halted due to matrix inversion problems. Suspicion is high collinearity. Please see t
he following tips in the lifelines documentation: https://lifelines.readthedocs.io/en/latest/Examples.html#problems
-with-convergence-in-the-cox-proportional-hazard-modelMatrix is singular.
```

Performing a VIF check on all the variables also gives an error showing that perfect multi-collinearity exists.



```
TypeError: ufunc 'isfinite' not supported for the input types, and the inputs could not be safely coerced to any supp
orted types according to the casting rule ''safe''
```

Appendix H: Results of Models

Table H-1 Top 10 Overall Models with Highest Recall

Rank	Features	Model	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
1	Cox Proportional Hazard	Adaptive Boosting	78.3	80.3	43.5	0.32	2290	1137	123	445
2	Decision Tree	Gradient Boosting	72.4	83.0	46.1	0.38	2662	765	157	411
3	All	Gradient Boosting	72.4	83.0	46.0	0.38	2662	765	157	411
4	Union of Logistic Regression and Decision Tree	Gradient Boosting	72.2	82.7	45.5	0.38	2644	783	158	410
5	Logistic Regression	Gradient Boosting	67.4	83.0	46.1	0.40	2802	625	185	383
6	Cox Proportional Hazard	Gradient Boosting	67.3	80.7	42.1	0.35	2683	744	186	382
7	Decision Tree using features from Logistic Regression	Adaptive Boosting	66.9	81.0	42.6	0.35	2674	753	188	380
8	Intersection of Logistic Regression and Decision Tree	Adaptive Boosting	66.9	81.0	42.6	0.35	2674	753	188	380

Rank	Features	Model	Recall	AUC-ROC	Average Precision Score	MCC	True Negatives	False Positives	False Negatives	True Positives
9	Intersection of Logistic Regression and Decision Tree	Extreme Gradient Boosting	66.4	82.0	42.8	0.38	2771	656	191	377
10	Decision Tree using features from Logistic Regression	Extreme Gradient Boosting	66.2	82.0	43.1	0.37	2768	659	192	376