2020

ICU Mortality Prediction Scorecard

EBA5005 GRADUATE CERTIFICATE IN SPECIALIZED PREDICTIVE MODELLING CA PROJECT

TEAM FOX

1. Ang Khee Hwa Randi (A0198526N)
2. Ee Meng Hock (A0163456B)
3. Fang Wenlin (A0198417R)
4. Shen Chen (A0058260J)
5. Xiao Jinting (A0198422Y)
6. Zhou Jingyu (A0198413Y)

Contents

1. Executive Summary 3
2. Business Understanding 4
3. Data Overview 6
3.1 Dataset Descriptions 6
3.2 Data Exploration 7
4. Data Preparation 9
4.1 Data Cleaning 9
4.2 Data Balancing 13
5. Modelling 14
5.1 Modelling Overview 14
5.2 Tree-based Model 14
5.3 Multilayer Perceptron Classifier 16
5.4 Logistic Regression 16
5.5 Model Comparison 19
6. Scorecard 20
7. Challenges and Limitations 23
7.1 Data Quality 23
7.2 Modelling limitations 23
7.3 Lack of domain knowledge 24
8. Conclusion 25
References 26
Appendix 27

List of Tables

Table 1: Data Dictionary....6
Table 2: Overview of binary and non-binary datasets....10
Table 3: Imputation Datasets Overview....10
Table 4 Imputation scenarios....10
Table 5: Solution of Data Merging....11
Table 6: Settings Used for Selecting the Worst Value for the Day....11
Table 7: GCS Scoring table....13
Table 8: Summary of Sampling Methods....13
Table 9: Tools and Packages for Models....14
Table 10: Best Parameters of Decision Tree Model....14
Table 11: Key rejected variables....18
Table 12: Summary of Model Performance....19
Table 13: Summary of Model Performance Against APACHE IV mortality predictions....19
Table 14: AUC values of K-Fold Cross Validation with the Logistic Regression Model....19
Table 15: Score Distributions of Scorecard....20

List of Figures

Figure 1: Potential options in handling patients of different acuity (mortality risk) in an ICU setting....5
Figure 2: Average Death Rate by Age and Gender....7
Figure 3: Average Death Rate by Ethnicity and Gender....7
Figure 4: Correlation between Vitals and Labs Blood Tests Variables....8
Figure 5: Data cleaning process....9
Figure 6: WOE Bins of the Score Distribution....12
Figure 7: Decision Tree Model....15
Figure 8: List of variables used....17
Figure 9: WOE binning of variables (body temperature variable used as an example)....18
Figure 10: Notched Box Plot of Score Distributions....20
Figure 11: Density Plot of Score Distribution....21
Figure 12: Binned Scores based on Bad Probability (death rate)....21

1. Executive Summary

With growing aging population, hospitals worldwide are facing the challenge of imbalance in demand and supply for the bed spaces, especially in ICU. ICU mortality prediction scorecards such as APACHE IV emerges as a solution to triage patients on needs basis. Yet due to copyright restrictions, its availability is limited.

Using the dataset is taken from the Global Women in Data Science 2020 Competition, we employed various machine learning techniques namely logistic regression, tree models and neural networks to model an open-source ICU mortality prediction scorecard based on the classification modelling output.

The deliverables of this report include:

1) Classification model output and evaluation
2) Proposed scorecard formulation and limitations
3) Proposed triage plans based on the scoring

Using Weight of Evidence (WOE) binning and logistic regression, we were able to create a preliminary model that is comparable to APACHE IV and other models created by other machine learning methods such as tree-based ensemble methods and neural networks. This model is then used to create a scorecard for triaging patients based of the mortality risk. Going forward, we hope to be able to incorporate more data from other countries and have onboard more healthcare professionals to finetune this scorecard further to allow us to generalise the model to different demographics and geographies globally.

2. Business Understanding

In most hospitals worldwide, Intensive Care Unit (ICU) wards are precious resources to be used judiciously for the treatment of patients due to the high costs (manpower, equipment, infrastructure and consumables) incurred to run the facilities and the need to provide 24-7 monitoring of patients. Thus, only tertiary level care hospitals can afford to support ICU facilities, where only a relatively small number of ICU bed spaces[1] are allocated. In a bid to healthcare costs, high dependency wards have also been created as an intermediary between an ICU and general ward facility.

With the aging population, rising obesity and diabetes rates in many developed nations, there has been an increased in demand for healthcare resources, as more patients suffer from chronic conditions and have more acute conditions requiring hospitalisation, which further exacerbate the **imbalance in demand and supply for the bed spaces**, including ICU. Thus, medical professionals **use an ICU mortality prediction scorecard to triage patients on needs basis (acuity of condition)**. Currently the APACHE II standard is one of the main stay ICU mortality prediction methods used globally. Other versions of APACHE have been proposed such as APACHE IV (whose prediction scores are provided in the dataset), but they do not receive widespread use due to the copyright restrictions and the licensing cost involved for making local customisation (Young, 2019). This is especially problematic since APACHE IV was formulated only with data from ICUs in the USA, making generalisation to other demographics and facilities[2] across different regions difficult.

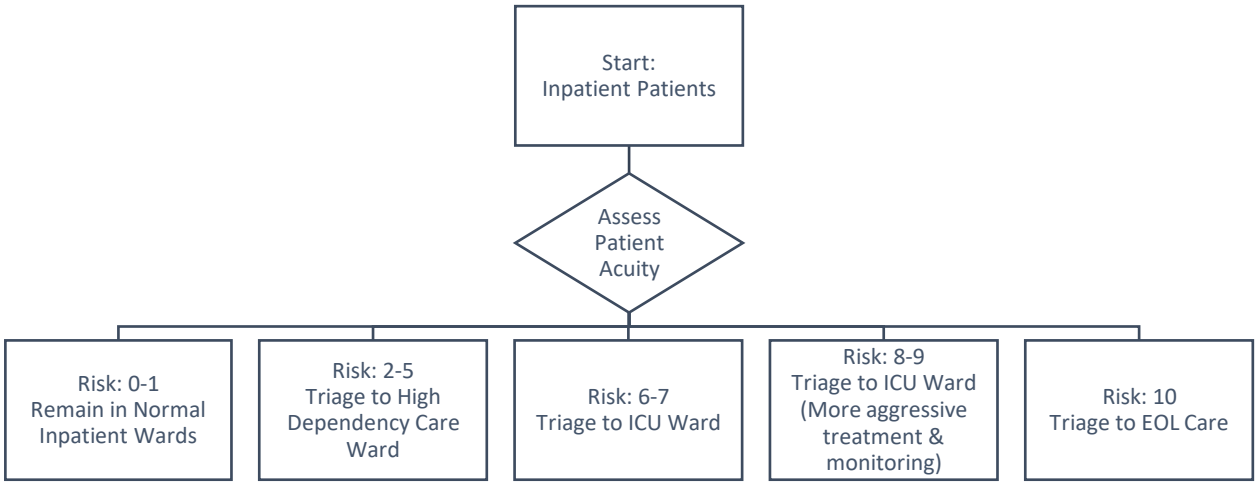
In view of the current situation, we hope to **create an open-source ICU mortality prediction scorecard that is comparable if not superior to APACHE IV** such that the following benefits may be attained for medical professionals globally:

1. **Allow for model customization to local factors for better predictive accuracy** since the model uses open-source creative common license allowing other users to freely make changes where needed**.** Factors such as patient demographics, availability and quality of medical services can affect patient mortality outcomes, hence fine tuning may be required.
2. **Propose better resource allocation of ICU Beds[3]** by classifying inpatient patients via a Patient Acuity Scorecard and assigning them to either an ICU ward, High Dependency Care Ward or remain in normal inpatient ward.
3. **Potential for more timely and aggressive therapeutic interventions for high risk patients** to reduce mortality rates; but in cases where further medical intervention would be futile, End of Life (EOL) care plans could be proposed instead.

[1] In the measurement of ward utilisation, hospitals would use bed occupancy as the main unit of measurement; though the terms ward space and bed space may be used interchangeably.
[2] ICU standards may differ by region and specialty
[3] Since the data is limited to the first 24hr of admission, we are unable to use the data to assist in assessing the patient's ability to be discharged to a step-down care ward. Also, the APACHE system and its variants are only meant to triage patients at the point of admission, thus is only measured in the first 24 and 72hrs of admission. Patient assessment for discharge is primarily based on their vitals and the care team's assessment of the patient's recovery (stabilization of condition, ability for independent life support without external intervention etc.).

4. **Assess the baseline risk groups being used in clinical trials** for comparison between different treatment groups as shown in Figure 14

Figure 1 : Potential options in handling patients of different acuity (mortality risk) in an ICU setting



In terms of the deployment of the model, this scorecard is intended to be applied in the hospital at the point of ICU admission only. The scorecard can be computed either manually by the clinician assessing the patient or automatically by a calculator embedded in the Electronic Medical Records (EMR) system – the latter is the preferred approach for ease of implementation. It should be noted that this scorecard is not intended for inter-hospital resource allocation of ward facilities. Such events are performed using a different set of factors such as the following (see below):

1. Patient Priority-Status
    a. P0: Dead
    b. P1: Critical life threatening (high priority)
    c. P2: Serious but Non-Life Threatening
    d. P3: Unwell but non-serious (low priority)
2. Hospital proximity to the patient's current location
3. Availability of A&E capacity and capability at the hospital[5]
4. General bed availability[6]

[4] An arbitrary risk scale of 0 (low risk) to 10 (high risk) is used for illustration purposes.
[5] Not all hospitals are able to take in every case of patients as they may lack the specialized staff, equipment and facilities to treat such patients. The ability to stabilize the patient's condition whilst the patient is in transit is also another determining factor.
[6] For serious cases, bed availability is not the main determining factor for assigning hospitals, as the goal is to stabilize the Patient's condition ASAP. Furthermore, there are usually a few ICU and HDC beds reserved as surge capacity to absorb such cases. Though in cases where there is a patient surge overwhelming a hospital's buffer capacity, incoming patients may be allocated to the next nearest available hospital instead. Other lower priority patients may be assigned to temporary "corridor beds" by the A&E or ward corridors till permanent bed space is made available at the various wards.

3. Data Overview

3.1 Dataset Descriptions

The dataset used for the project is taken from the recent **Global Women in Data Science (WiDS) 2020 Competition**[7]. The objective of the challenge is to create a model to **predict the patient mortality in an Intensive Care Ward (ICU) ward,** based on patient profile data taken in the first 24 hours of stay in the ICU ward. The competition made available observations collected from more than 200 hospitals in United States, Argentina, Australia, New Zealand, Sri Lanka & Brazil, and the data was split into:

1. Train data (91.7K) – labelled
2. Test data (39.3K) – unlabelled

Since the test data is still unlabelled at the conclusion of this project, it was not used in our modelling. The train and test datasets used for developing the model is therefore taken only from the train data (91.7k) provided by the competition.

The patient mortality prediction is a **binary classification problem** and the dependent variable is hospital death. The 185 independent variables can be broadly classified into 5 categories, as shown below and Table 1.

Table 1: Data Dictionary

Types	Subcategory	Description	Example	Record Period
Administrative		Unique identifiers of patient, hospital & ICU, admit source, length of stay, ICU type, etc	Encounter id	Admit time
Demographic		Basic info such as age, weight, height, gender, etc.	Age	Admit time
APACHE	Covariate	Some of Labs tests (which results in the highest APACHE III score)	Blood composition	First 24 hours
	Comorbidity	Chronic diseases and other ailments	Diabetes	Admit time
	Prediction	APACHE 4a death probability	ICU death	First 24 hours
Vitals		Max and Min indicators of patient conditions	Temperature	First 24 hours and 1st hour
Labs Blood Tests		Diagnostic tests	Blood composition	First 24 hours and 1st hour

It should be noted that some APACHE covariates are taken from continuous variables that are monitored or tested throughout the day, therefore the values may simply represent the worse value (in terms of patient physiological condition) between the min or max recorded values within the first 24hrs of admission. Feature engineering is applied to those readings recorded from the original source systems at data preparation stage. For instance, for a body temperature with a normal range around 37.5 °C, if a patient have a min temperature reading of 38°C and a max temperature reading of 39 °C, 39 °C would be chosen to be fitted to the model as it is a worse value in terms of physiological outcomes.

[7] https://www.kaggle.com/c/widsdatathon2020/overview

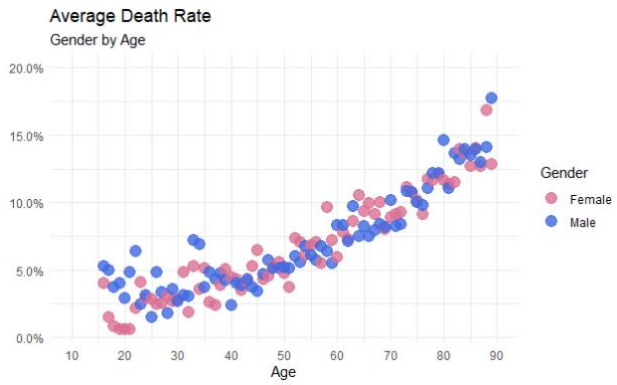
3.2 Data Exploration

The dataset contained large number of missing values. Out of the 90k+ observations for the train data, less than 100 are complete cases without any missing value. To retain as many variables and observations as possible, we had to understand and review each and every variables carefully and group them according to their similarity, and analyse the distribution (against the target variable, as shown in Section 4 to determine the actions to be taken at individual variable or grouped variables level.

3.2.1 Average Death Rate by Age and Gender

Figure 2 shows that the **average death rate increases with age** though there is **no significant difference between genders**.

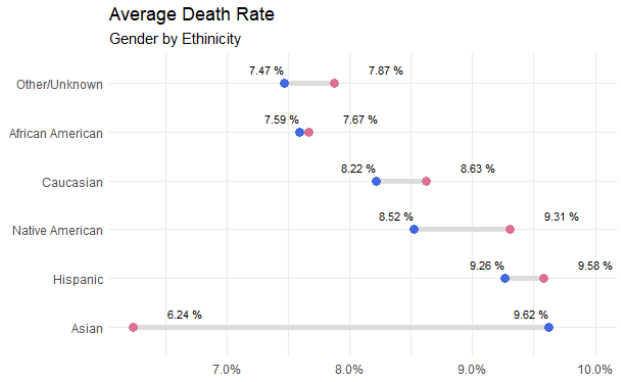
Figure 2: Average Death Rate by Age and Gender



3.2.2 Average Death Rate by Ethnicity and Gender

Figure 3 shows that the **average death rates appear to differ by Ethnicity**, and except for Asian, **average death rate of female tends to be higher** than male across all ethnic group. But since more than 77% of the observations are Caucasian, ethnicity did not show up as significant in the model.

Figure 3: Average Death Rate by Ethnicity and Gender

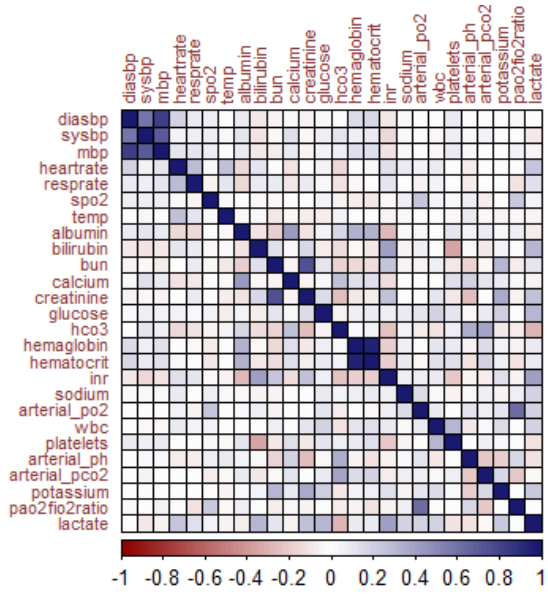


3.2.3 Correlation between Vitals and Labs Blood Tests Variables

Figure 4 shows that in general, there is **low correlation between variables** except for variables belonging to the same medical category. The high correlation between the

variables on the top left (diasbp, sysbp and mbp) are all related to blood pressure and are combined into one variable in later stage. While the variables in the middle (hemoglobin and hematocrit) are both associated with red blood cells[8].

Figure 4: Correlation between Vitals and Labs Blood Tests Variables



[8] Hemoglobin is an oxygen carrier compound found in red blood cells, hence the natural correlation

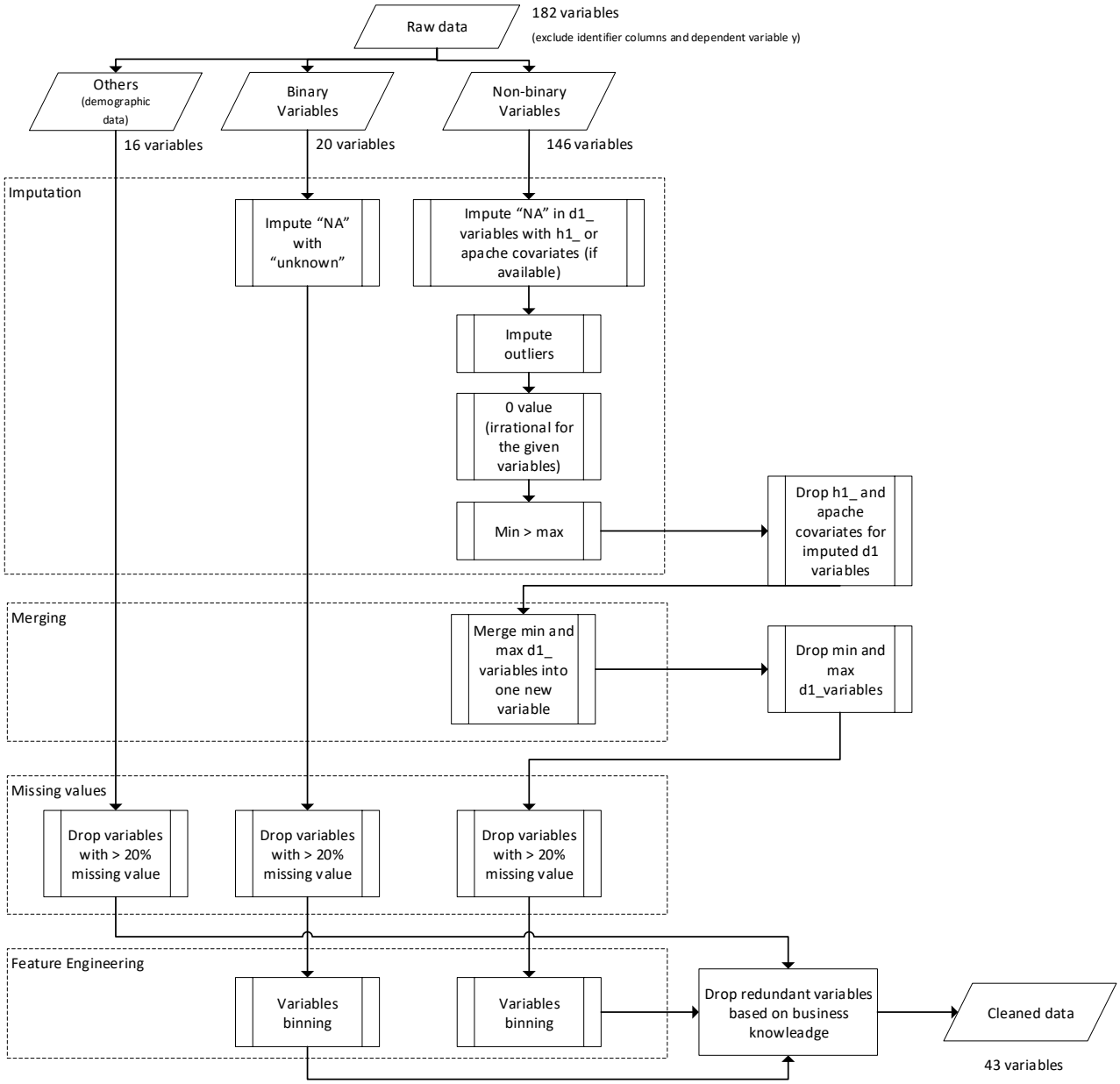
4. Data Preparation

4.1 Data Cleaning

Our data cleaning steps are depicted in Figure 5 following a 4 staged process:

(1) **imputation** – impute values where possible
(2) **merging** – merging of variables from the same category
(3) **missing values handling** – removal of columns with excess missing values
(4) **feature engineering** – categorise (bin) the numeric variables according to business understanding

Figure 5: Data cleaning process



4.1.1 Imputation

Different imputation rules were adopted for **binary** variables and **numeric** variables as shown in Table 2.

Table 2: Overview of binary and non-binary datasets

Category	**Variable Group**	**Imputation Rational**	**Assumption**	**Imputation rules**
Binary	Apache Comorbidity	Handle only missing value	"NA" indicates that there is no record for the patient	Impute "NA" with "unknown"
Numeric	Vital and Lab test	Check missing values, outliers, 0 values and logic test (min/max)	Some variables correspond to APACHE covariate measured at different time points. Example: for PH data, there are *ph_apache* *d1_arterial_ph_max* *h1_arterial_ph_max* *d1_arterial_ph_min* *h1_arterial_ph_min*	d1_ data → h1_ data → apache covariate → median (d1_data) Summarized in Table 3

Table 3 : Imputation Datasets Overview

Data Level	**Subcategory**	**Content**	**Order of preference**
APACHE covariate			3rd
Vital	**d1 (1st day)**	Min & Max	1st
	h1 (1st hour)		2nd
Lab Test	**d1 (1st day)**	Min & Max	1st
	h1 (1st hour)		2nd

For imputing the numeric variables, we used **d1_data** as the base value for the following reasons:

1. Comparing to hourly *h1_ data*, daily *d1_ data* can reflect **risker status** of the patients.
2. The missing value percentage of *d1_ data* is much fewer than that of *h1_ data.*
3. The min and max values of *h1_ data* are almost the same, and there are no min and max values in the *apache_data*.

Table 4 summarised the imputation process and scenarios for the numeric variables.

Table 4 Imputation scenarios

Scenario	**D1_data**	**H1_data**	**Aapche Covariates**	**Imputation outcome**	**Assumption**
(1) Missing values	Available	Available or Missing	Available or Missing	No action required	Values are not recorded
	Missing	Available	Available	H1_data	
	Missing	Missing	Available	Apache	
	Missing	Missing	Missing	"unknown"	
(2) Outliers	Same imputation rational as (1)				Values are unreasonable, for example, 300 for heart rate
(3) Dubious 0 values	Same imputation rational as (1)				0 is not possible for the variable given the patient survived, for example 0 for heart rate
(4) Min > Max	Same imputation rational as (1)				Either the Min or the Max value is wrong. Impute for the wrong value (determined by eyeballing)

4.1.2 Merging

The purpose of merging is **to combine the variables with D1 min and D1 max values into one new column**. For model development, there is a need to **select the worst value** recorded in the first 24 hrs of the patient's stay[9]. We first define the **Goldilocks' Point** and compare the absolute deviation of the D1 min and D1 max values from that point, selecting whichever value with the higher deviation, as summarised in Table 5 and Table 6.

If there is a tie in the absolute deviation, a bias factor is applied as a tie breaker, based on the relative severity if there is a deficiency (low) or excess (high). The reason that all biological organisms functioned best at a certain optimal point maintained by homeostasis. When the balance is disrupted in times of stress, it could lead to permanent cell damage and even death[10]. Hence, without in-depth medical knowledge, we believe that this is a reasonable compromise for variable selection in the model.

Table 5: Solution of Data Merging

Data Level	Subcategory	Solution
Demographics		Drop redundant variables.
APACHE covariate		
Vital	d1 min	1. Select values with a larger absolute distance from the Goldilock's zone. 2. If the min and max have the same absolute distance from the Goldilock's zone, assign a bias factor
	d1 max	
Lab Test	d1 min	
	d1 max	

Table 6: Settings Used for Selecting the Worst Value for the Day

SN	Variable	Description	Goldilocks' Point	UOM	Bias (Low / High)	Bias Rationale
1	d1_diasbp_fin	Diastolic Blood Pressure	90	mm/Hg	High	Extreme stress
2	d1_sysbp _fin	Systolic Blood Pressure	140	mm/Hg	High	Extreme stress
3	d1_map_fin	Mean arterial blood pressure	85	mm/Hg	Low	Weakened state of health
4	d1_heartrate_fin	Heart rate	80	Beats per min	High	Possible cardiac arrest
5	d1_resprate_fin	Respiration Rate	80	Breaths per min	High	Extreme stress
6	d1_spo2_fin	Oxygen Saturation	95	%	Low	Breathing issues
7	d1_temp_fi	Body Temperature	37	Degrees Celsius	High	Permanent cell damage from fever
8	d1_bun_fin	Blood urea nitrogen	15	mg/dL	High	Possible kidney disorder
9	d1_calcium_fin	Blood calcium level	9.4	mg/dL	High	Possible thyroid disorder or cancer
10	d1_creatinine_fin	Blood creatinine level	0.95	mg/dL	High	Possible kidney disorder
11	d1_glucose_fin	Blood glucose level	100	mg/dL	High	Diabetes
12	d1_hco3_fin	Blood bicarbonate level (oxygen saturation)	26	mmol/L	Low	Breathing issues
13	d1_hemoglobin_fin	Blood hemoglobin level	14	g/dL	Low	Respiration issues; possible hemorrhaging
14	d1_hematocrit_fin	Red blood cell concentration	40	%	Low	Respiration issues; possible hemorrhaging
15	d1_sodium_fin	Blood sodium level	140	mmol/L	High	Irreversible dehydration
16	d1_wbc_fin	White blood cell concentration	7	%	Low	Possible immunosuppression

[9] We considered using methods like average, however, since a patient's vitals may fluctuate throughout the day and most medical scorecards use only the worst value of that time span, we needed to be consistent.
[10] It should also be noted that the degree of damage per unit of change varies from variable to variable and may differ for values below and above the Goldilocks' point.

17	d1_platelets_fin	Blood platelet concentration	275	10^9/L	Low	High risk of excessive hemorrhaging
18	d1_potassium_fin	Blood potassium level	4.4	mmol/L	High	Possible kidney disorder, trauma and cell damage

4.1.3 Missing Values Handling and Dropping Redundant Variables

At this stage, columns with large proportion of missing values, in particular those that fall under the blood gas lab tests category, were dropped. We have chosen to remove these variables as they might be incomplete information and there is no good way to derive a reasonable estimation of the values. We set the missing proportion threshold at 20% and remove all columns that falls outside the threshold.

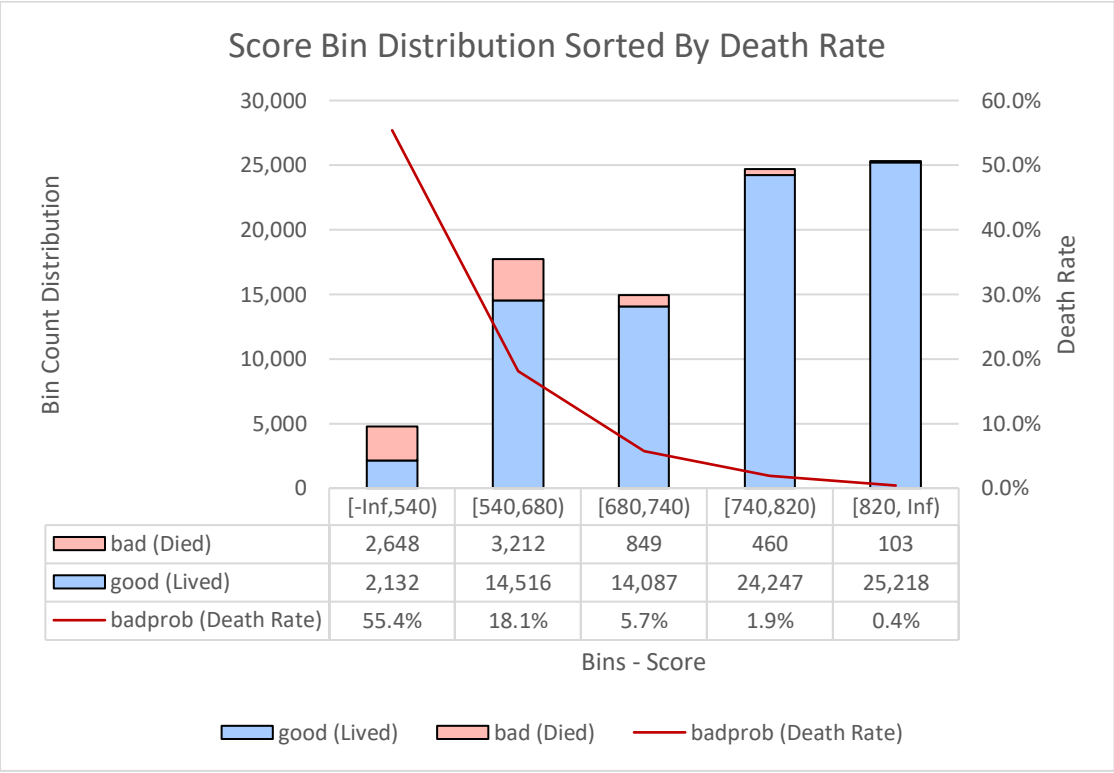
We also dropped the redundant original columns after merging the D1 min or D1 max value columns for each variable handled in the previous section.

4.1.4 Feature Engineering

In this stage, we **categorise the numeric variables**. The WOE package from R is used.

Weight of Evidence (WOE) group a continuous independent variable into a set of bins based on dependent variable distribution (survived or died). It is been used as a benchmark to screen variables in the credit risk modelling projects such as probability of default. Since our scorecard (discussed in Section 6) is derived based on a credit scorecard model, we adopted this methodology to transform the variables. An outcome example is shown in Figure 6.

Figure 6: WOE Bins of the Score Distribution



For features such as the **Glasgow Coma Scale** (GCS) for the eyes, verbal and motor responses, **we combined the scores together** as an overall GCS score with a range of 0 to 24. This is akin to how the older versions of APACHE such as APACHE II uses this variable. The GCS ranges are as follows:

1. Eye: 0 to 4
2. Verbal: 0 to 5
3. Motor 0 to 6

Table 7: GCS Scoring table

GCS Variable	Scale Description	Points
Best Eye Response	Spontaneously	4
	To verbal command	3
	To Pain	2
	No eye opening	1
	Not testable (NA)	0
Best Verbal Response	Oriented	5
	Confused	4
	Inappropriate words	3
	Incomprehensible sounds	2
	No verbal response	1
	Not testable (NA)	0
Best Motor Response	Obeys commands	6
	Localizes pain	5
	Withdrawal from pain	4
	Flexion to pain	3
	Extension to pain	2
	No motor response	1
	Not testable (NA)	0

4.2 Data Balancing

The minority class (hospital death = 1) of **the dataset seemed imbalance**, at only 8% of the total, which may pose problem to classification algorithms like Decision Tree. To have a fair comparison between selected modelling techniques, we used **5 methods to balance** the train data and validate the results with **the test data for all models. A total of 6 sets of train data were created as shown in the table below.**

Table 8: Summary of Sampling Methods

Method	Definition	No. of Non-death	No. of death	Total rows	Death ratio
Over Sampling	Randomly select minority examples to increase incidences of 1 (death)	41842	32350	74192	43.6%
Under Sampling	Randomly reduce the number of incidences of 0 (non-death)	3515	3515	7030	50%
Over& Under	Randomly increase the number of death cases and reduce the non-death cases	25103	15465	40568	38.1%
Rose	Uses smoothed bootstrapping to draw artificial samples from the feature space neighbourhood around the minority class	24919	15649	40568	38.6%
SMOTE	draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbours in the feature space	33222	23730	56952	41.7%
None	Original death and non-death cases in training datasets	56140	5090	61230	8.3%

5. Modelling

5.1 Modelling Overview

For ease of interpretation and deployment, we planned to use **Logistic Regression as the primary model** for the mortality prediction and creation of Scorecard; whilst **the other models** were used for **benchmarking** and discovery of new or enhanced features. These models were separately tuned by iterating different parameters to generate best ROC-AUC scores for each of the datasets. The selection of the threshold value is dependent on the resulted ROC curve and vary between different datasets and different models. It should be noted that existing versions of the APACHE mortality prediction scorecards also use logistic regression at its core for the same reason.

The programming languages and packages used for the selected modelling algorithms are summarized in Table 9

Table 10: Tools and Packages for Models

Modelling Algorithms	Programming Language	Key Package / Framework
Decision Tree	Python	sklearn
CatBoost	R	Catboost, caret
LightBoost	Python	LGBM
Multilayer Perceptron Classifier	Python	Keras
Logistic Regression	R	glm, caret
Scorecard	R	scorecard

5.2 Tree-based Model

5.2.1 Decision Tree

Decision Tree Model is a supervised machine learning model for classification.

Steps:

1. Tuned parameters by assigning **different parameters** to generate best AUC score for each dataset.
2. Used **1920 combinations** of the parameters to find the best parameters which can maximize the AUC score
3. Used ROC curve to select **optimal threshold** which maximized AUC score

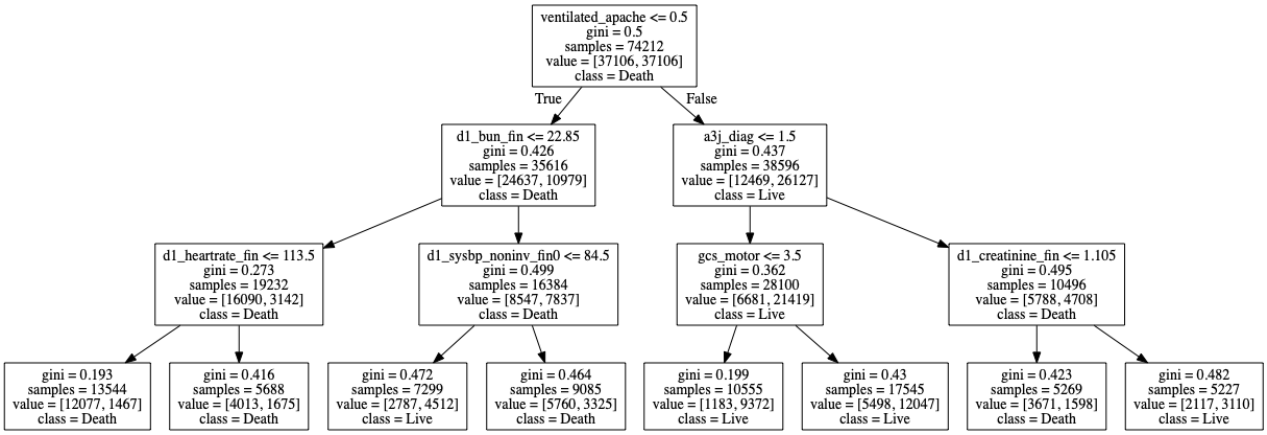
Outcome:

1. Best Model: Decision Tree model on oversampling dataset beats other sampling methods.
2. Best Parameters for oversampling dataset, summarized in the Table 9:

Table 11: Best Parameters of Decision Tree Model

Model	Best Parameters		
Decision Tree	Max Depth	Min Samples Split	Min Samples Leaf
	3	0.1	0.07

Figure 7: Decision Tree Model



5.2.2 CatBoosting / LightBoositng

CatBoost and LightBoost are very similar in their boosting algorithm and tuning parameters. The main difference roots in how the categorical data are handled: CatBoost will transfer the data to one-hot encoding while LightBoost adopts algorithm to split and group the categorical data.

Steps:

1. Tuned parameters by **grid search** to maximize the AUC score and avoid overfitting.
2. Iterated the model 500 times to find the best model with early stopping set

Outcome:

1. Best Model: Both CatBoost and LightBoost model on oversampling datasets beat other sampling methods.
2. Best Parameters for oversampling datasets, summarized in Table 12:

Table 11: Best Parameters of Boost Models

Model	**Best Parameters**				
LightBoost	Num of leaves	Learning rate	Max bin	Feature fraction	N estimators
	40	0.05	60	0.7	60
CatBoost	L2 regularization	Learning rate	Depth	Bagging temperature	Border count
	5	0.3	3	3	32

5.3 Multilayer Perceptron Classifier

Multilayer Perceptron is a supervised learning algorithm that trains using backpropagation.

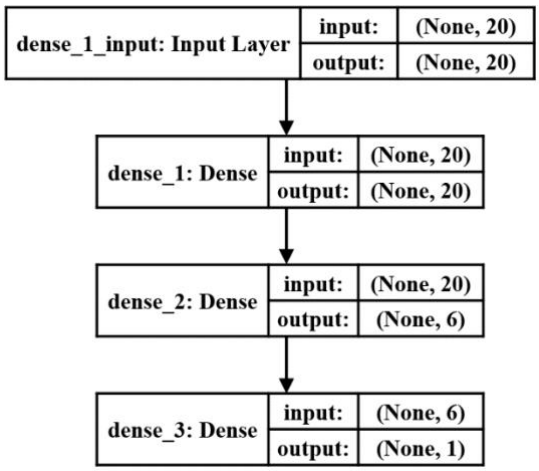
Steps:

1. Searched for the best model by **adjusting hyperparameters** to generate the best AUC score for each dataset.
2. The number of hidden layers, learning rate, activation functions and other hyperparameters were optimized.
3. Use **regularization**, which involves the addition of a penalty term to the error function, to ensure that the model does not overfit.

Outcome:

1. Best Model: Multilayer Perceptron on under-sampling dataset beats other sampling methods.
2. Figure 12 shows the structure of the NN model with a group of parameters achieving the best model performance. The model has 3 layers (dense 1 to 3). In the right column, "None" indicates that there are no intercept and the numeric value denotes the number of active (in the case of output) or passive (in the case of input) nodes.

Figure 12: Model structure of the best MLP classifier



5.4 Logistic Regression

For the logistic regression model, initial set of variables were selected manually by removing aliasing and multi-collinearity between variables. The outcomes show no significant difference between no sampling and different methods of over-/under-samplings. Below is the resulted list of variables ranked by information values (and relative importance) from the model using dataset without sampling.

Figure 8: List of variables used

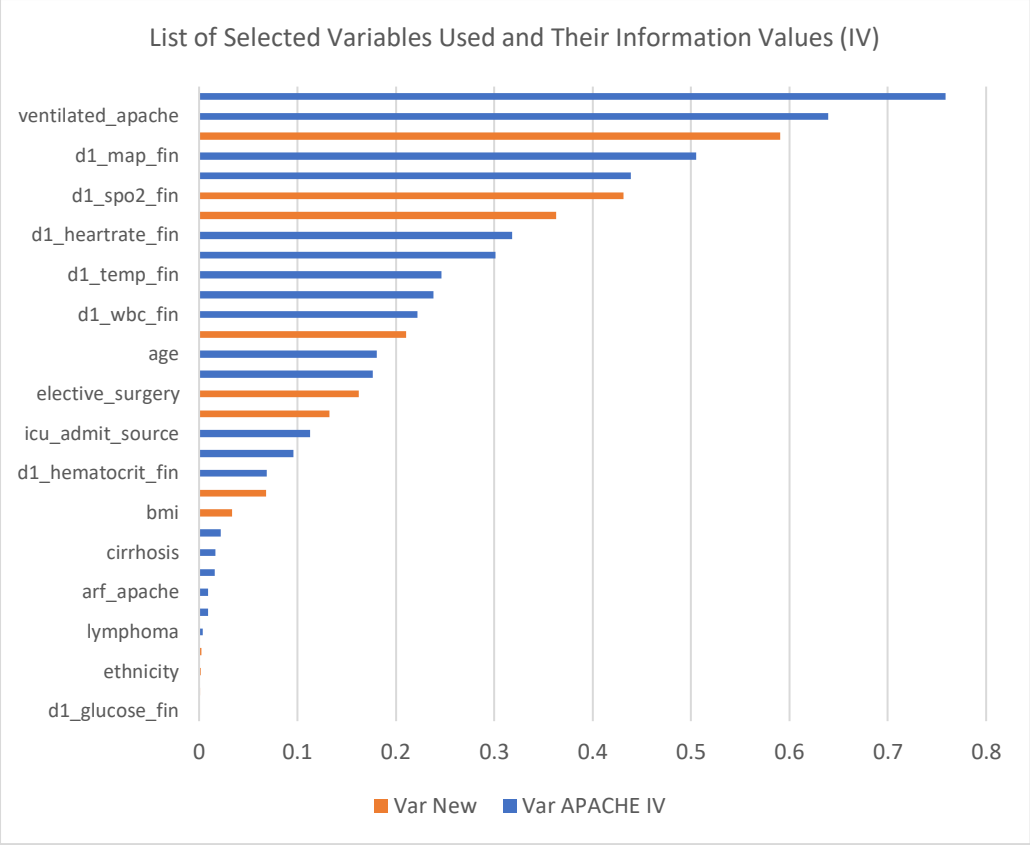


Figure 8 show the list of variables used in the scorecard. "Var New" are variables not used in APACHE IV, whilst "Var APACHE IV" refers to variables currently used in the APACHE IV scorecard. These variables were selected based on the following criteria:

1. Variable significance
2. Variable independence
3. Long term model robustness
4. Ease of generalization
5. Ease of data collection in actual operations

On this note, we believe that the reason the H1 values suffer from a lot of missing data was due to the operational issues on the ground, where clinicians are busy trying to stabilize the patient's condition and settling the administrative work of warding a patient. Hence much of the data is not recorded. It should also be noted that PaCO2 and other blood gas measurements are usually quite critical for assessing patient mortality, particularly that for patients on mechanical ventilation support who are often in critical condition. However, much of the data was missing, forcing us to junk the variable. Interviews of ICU doctors revealed that a possible reason was due to these readings are often collected and calculated by hand, and they are often not entered an EMR platform, hence the missing values.

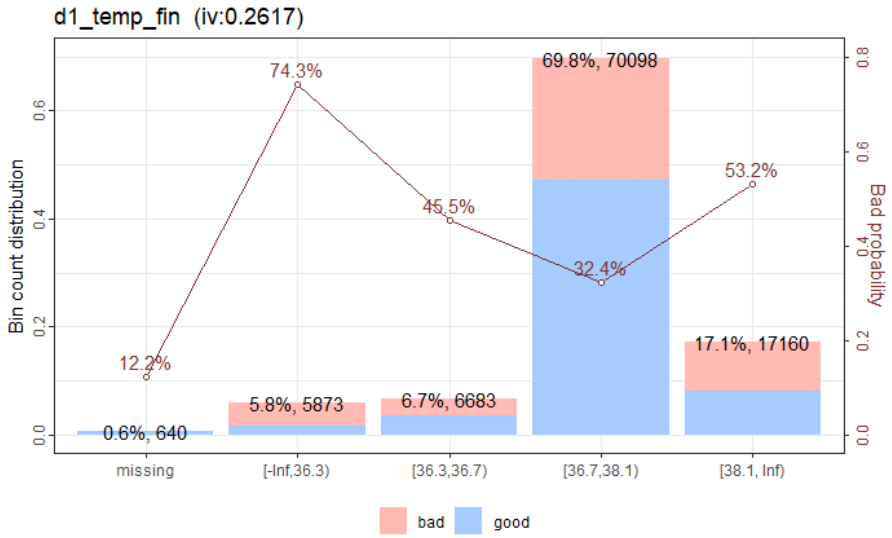
The following variables have been discarded due to issues with aliasing (collinearity of x-variables), low predictive power and lack of interpretability.

Table 12: Key rejected variables

Variable	Rationale for discarding	Details
d1_potassium_fin	Poor predictive power	NA
d1_creatinine_fin	Poor predictive power	NA
gender	Poor predictive power	NA
d1_diasbp_fin	Poor predictive power + aliasing	This variable is a component for calculating the mean arterial blood pressure (MAP)
aids	Poor predictive power	Low number of aids cases observed in sample pool; possibly due to more advanced treatment options of HIV available from developed nations (which the sample population is derived from), preventing the progression of HIV infection to AIDS. Furthermore, AIDS does not kill the host, but rather the cancers and opportunistic secondary infections as a result of the severely compromised immunity system because of AIDS. On this note, the immunosuppression variable is also used as a predictor, so this gap has been covered indirectly.
apache_post_operative1	Poor predictive power + lack of interpretability	No details of medical codes used provided, making interpretation impossible
gcs_eyes	Aliasing	Data has been merged as a combined GCS score
gcs_motor	Aliasing	Data has been merged as a combined GCS score
gcs_verbal	Aliasing	Data has been merged as a combined GCS score
d1_hemoglobin_fin	Aliasing	This is analogous to hematocrit, albeit different units of measure used

Weight of Evidence (WOE) binning is then applied on elected variables to get the WOE values for each bin. Doing so would also address the issue of missing values; and allow a linear set of values to be fitted to the model (namely the WOE values). This is especially since, most of the continuous variables do not behave in a linear fashion in influencing mortality risk. It should be noted that the actual Bad probability past the ± infinity points may be steeper than what is shown, as the sample used in the model simply do not have significant number of observations past those thresholds, resulting in a truncated result. Nevertheless, the WOE values of the binned values will then be used in the creation of the logistic regression model. The variable coefficients from the model would subsequently be used for the scorecard creation.

Figure 9: WOE binning of variables (body temperature variable used as an example)



The WOE-binned variables were then put through logistic regression model again with stepwise feature selection followed by 10-fold cross-validation with performance metric set to maximise the AUC. The outputs show consistent AUCs between train and validation datasets.

5.5 Model Comparison

For test datasets:

Table 13: Summary of Model Performance

Model	Threshold	AUC	Accuracy	Sensitivity	Specificity
Decision Tree	0.595	0.74	0.742	0.75	0.97
LightBoost	0.476	0.79	0.784	0.25	0.85
CatBoost	0.49	0.89	0.787	0.78	0.82
MLP	0.5	0.86	0. 782	0.28	0.97
Logistic Regression	0.08	0.876	0.755	0.751	0.755

Table above shows the summary of model performances. AUC is our main evaluation criteria. While the AUC of Logistic Regression model is a tick lower than CatBoost model on the test data, the 10-Fold cross validation of the model has consistently yielded high AUC values on both the train and validation data, with an average of 0.9 for both, which further validated the model's robustness.

Table 14: Summary of Model Performance Against APACHE IV mortality predictions

Model	Bal. Accuracy	Threshold	Accuracy	Sensitivity	Specificity
L.Reg Model (Balanced)	0.801	0.08	0.8	0.801	0.801
L.Reg Model (Sensitivity Biased)	0.798	0.06	0.754	0.851	0.744
L.Reg Model (Specificity Biased)	0.793	0.1	0.831	0.747	0.839
APACHE IV Hospital Death (Sensitivity Biased)	0.769	0.087	0.719	0.828	0.709
APACHE IV ICU Death (Specificity Biased)	0.768	0.087	0.84	0.681	0.856

Table 15: AUC values of K-Fold Cross Validation with the Logistic Regression Model

Data Set	Train - AUC	Validation - AUC
1	0.903	0.897
2	0.902	0.905
3	0.903	0.900
4	0.903	0.902
5	0.903	0.901
6	0.902	0.906
7	0.903	0.901
8	0.902	0.903
9	0.902	0.906
10	0.903	0.901
Average	**0.903**	**0.902**

The dataset also came with the APACHE IV mortality prediction probabilities for hospital death and ICU death which we did not use in our model built. We applied the threshold values (percentage of minority class) to the probabilities to compute the confusion matrix and use as a rough benchmark against our logistic regression model, as shown in Table 1Table 14. We found that our model performance is comparable which is a good sign in terms of validating the robustness of our model.

6. Scorecard

In terms of the scorecard development, scores would be assigned to each WOE bin via the formulae listed below:

$$Score = -\left(\sum_{j,\,i=1}^{k,\,n}(woe_j * \beta_i) + a\right) * factor + offset$$

Where,

1. Points to Double Odds (pdo) = 50
2. Factor = pdo / ln(2)
3. Offset = Target score – (factor * ln(odds))
4. WOE = weight of evidence for each variable level
5. β = regression coefficient for each variable
6. a = intercept term from logistic regression
7. n = number of variables in model
8. k = number of bins of each variable
9. Target score = 1000
10. Odds (Bad rate) at Target Score = 1 / Target Score

The resulted scorecard has a min and max values of -80 and 1115 respectively, whereby the lower the score, the higher the odds of dying. It should be noted that depending on the sampling method chosen, the score distributions much like threshold values for the confusion matrix may differ. The score distribution are as follows:

Table 16: Score Distributions of Scorecard

Score Dist.	P-00	L.Limit	P-05	P-25	P-50	Mean	P-75	P-95	U.Limit	P-100
Pop.	-45	203	259	424	539	536	648	785	875	1367
Lived (0)	150	369.5	407	532	617	618	697	834	865	1367
Died (1)	-45	368	192	326	412	406	492	602	866	1147

Figure 10: Notched Box Plot of Score Distributions

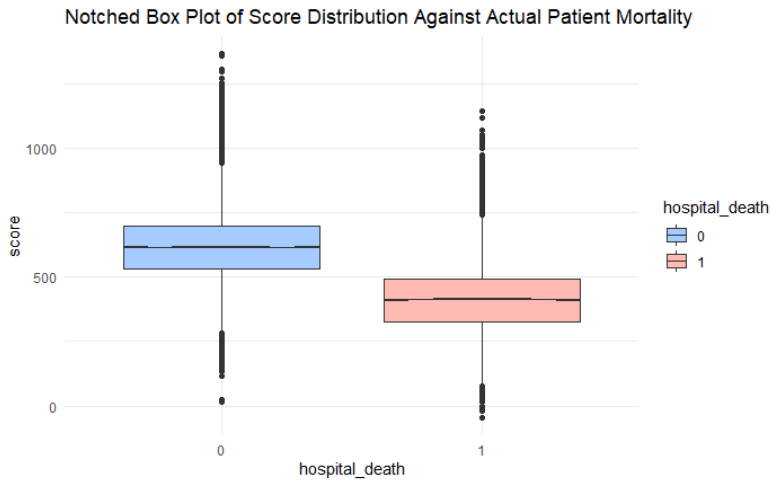


Figure 11 show the density distribution of survivors and deaths. The median values of survivors and deaths have been colour coded in blue and red respectively. The Intersect point has been marked in black.

Figure 11: Density Plot of Score Distribution

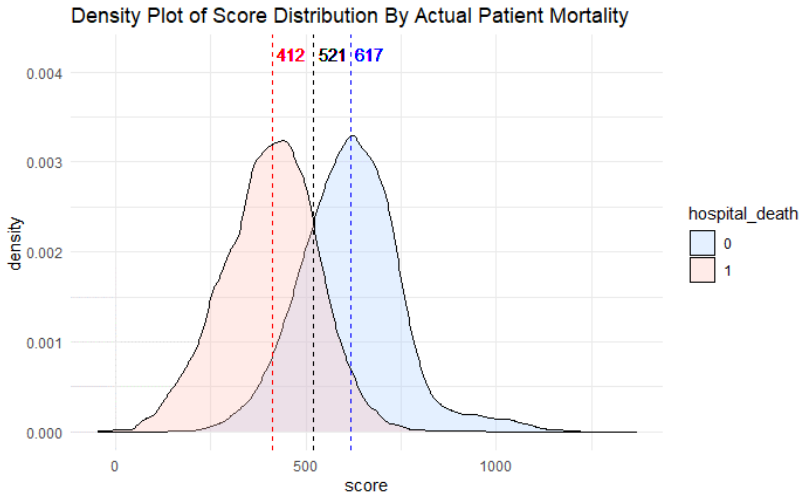
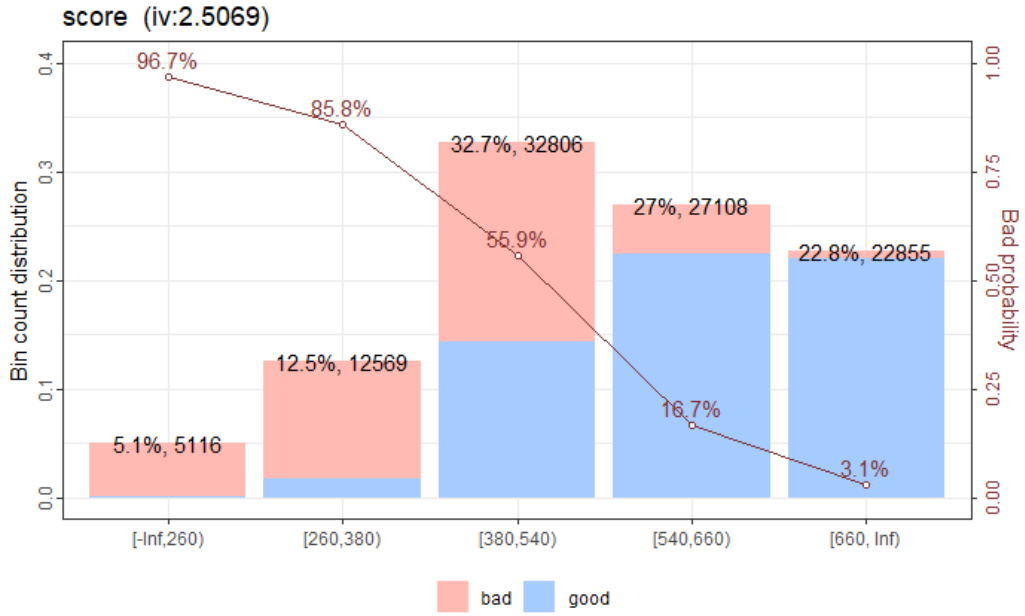
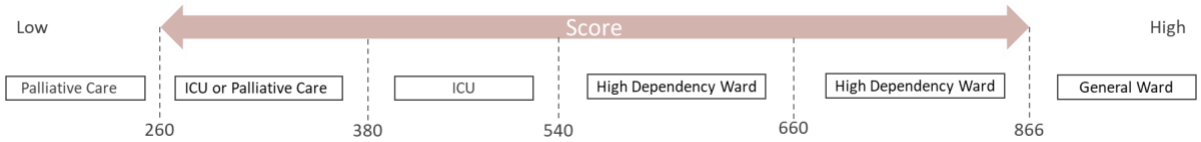


Figure 12: Binned Scores based on Bad Probability (death rate)



Based on the distribution plots and statistics above, we would hypothetically propose the following threshold levels with their corresponding action plans:



The choice of the threshold levels was primarily based on the degree of separation between the good and bad rates, to avoid misclassification of the patients. This is also considered a risk-based approach in terms of delivering appropriate care to the patients.

For instance, using the lower and upper outlier threshold points, patients with a low score (extremely high risk of dying) such as those with a terminal illness with a near 100% chance of dying would be better off seeking palliative care for the EOL, rather than more aggressive treatments which would just add to more suffering and financially costly as well. Furthermore, ICU resources will also end up being squandered on patients who would not benefit from such care, instead of being used by other patients who might benefit from it. Conversely, patients with a high score (low risk of dying) such as those with mild illnesses, would not benefit from ICU care too, and thus should be filtered out. Patients with scores in between may be triaged to high dependency or ICU wards depending on their mortality risk level.

Ultimately, the scorecard and the corresponding threshold levels and classifications are intended to be used as guidelines to support decision making, and not as the final verdict. This is due to the complexities in modelling different combination of patient conditions, demographics, facilities standards etc, which would make handling each (patient) case relatively unique. Hence it will still take the judgement call of a trained physician to be the final arbiter.

Also, for the ease of deployment, this scorecard should be embedded as a calculator within a centralised EMR used by the hospital, so that latest values from the lab and ward may be entered whenever available and have the score automatically computed. That said, having it as a scorecard format also have the advantage of being able to manually compute it by hand offline, especially as a contingency measure in the hospital. Prediction models created by other methods will be much harder to implement without the assistance of computer.

7. Challenges and Limitations

The challenges and limitations of our project can be summarized in the following three aspects:

1. Data quality
2. Modelling limitations
3. Lack of domain knowledge

7.1 Data Quality

1) Large number of **missing data** limit the ability to use and re-engineer some of the variables. Take for example, new features derived from the differences between H1 and D1 vitals or lab-test results.
2) Model **may not generalise well** as it is currently dominated by Caucasian ethnicity.
3) Healthcare systems and healthcare quality outcomes in different nations and regions vary quite widely.
    a. The data provided comes mostly from Argentina, Australia, New Zealand, Sri Lanka, Brazil and the United States, however due to the strong dominance of the Caucasian ethnic group and the data source coming from GOSSIS which is primarily based in the US, we suspect that most of the data comes primarily from the USA.
    b. As such, whilst the USA may boast some of the most advanced medical care facilities in the world, its actual patient care outcomes is somewhat lackluster with mixed results compared to other developed nations (Tikkanen & Abrams, 2020).

7.2 Modelling limitations

1) While complex interactions between variables can be modelled by logistic regression, the outcomes cannot be easily interpreted and used in the creation of scorecard. Take for example, some vital signs and lab-test results are inter-related. Even if we can use the various combinations of these variables in the modelling to improve the prediction, the outcomes may not be easily translated and manually computed (by hand), which is one of the benefits of scorecard. Current versions of APACHE scoring systems do not have provision for multivariate interactions too.
    a. Even with domain knowledge, many of the vital signs data are interlinked, making accurate modelling by hand to account for those combinations exceedingly difficult. Even current versions of APACHE also do not make provision for multivariate interactions.
2) The model may not predict well for new or rare cases of illnesses, firearms and CBRE injuries; and times when hospital resources are severely stressed (e.g. pandemic and mass casualty scenario).
    a. Avoidable deaths which would not normally occur would happen when hospitals are overwhelmed and unable to allocate sufficient resources to each patient, resulting in lower care outcomes and higher patient deaths. (Ellis, Schneider, Caswell, & Posner, 2020)

   b. If there are simply insufficient cases documented for rare injuries and illnesses, accurate modelling of those cases would be difficult.
3) As the min age in this dataset is 16, the model may not predict well for patients younger than 16-year-old.

7.3 Lack of domain knowledge

Domain knowledge is required in the following areas:

1) To understand how the independent variables are used for the diagnosis of diseases and body function; and any relationship between them.
2) Determine the thresholds for the Scorecard category and the medical/business decisions and actions for each category.
   a. Currently, the exact cost of misclassifying a patient wrongly and standard of critical care facilities available is unknown, making model tuning difficult.
   b. Multiple stakeholders from different departments, including both clinicians and hospital administrators, need to be consulted in order to determine the best thresholds which would optimize hospital resource allocation and maximise patient care outcomes.
3) To determine whether data from a variable can be readily collected in actual practice, since there is a need to avoid overburdening clinicians and patients with unnecessary tests which can lead to waste and missing data issues.
   a. Whilst effort has been made to consult ICU medical practitioners in the Singapore setting to ensure that the chosen variables are viable, local conditions in other countries could still prevail.
4) To make decisions on possible missing values imputation.

8. Conclusion

Overall, we believe that our scorecard model is robust and can serve as the first step to creating an effective open source ICU mortality prediction scorecard system that can rival the APACHE IV scorecard. That said, data from more hospitals world-wide will be needed to better generalize the scorecard to be used in more countries. Hospitals should also adopt EMR platforms to digitise the data and to reduce data errors, and also participate in open source repositories such as Global Open Source Severity of Illness Score (GOSSIS) to facilitate broad-base healthcare analytics. This will allow the model to incorporate more data, new or rare patient conditions, to better predict the mortality risk associated with them. Lastly, in any future development of this scorecard, we hope to have more clinicians and hospital administrators onboard in the development of the scorecard, especially for its deployment in a real-world setting.

References

Ellis, S., Schneider, M., Caswell, E., & Posner, J. (Directors). (2020). *Why fighting the coronavirus depends on you* [Motion Picture]. Retrieved from https://www.youtube.com/watch?v=dSQztKXR6k0

Janin, P. (2020). *ICU Medical Calculators*. Retrieved 01 Apr, 2020, from Intensive Care Network: https://intensivecarenetwork.com/124-icu-calculators/

Knaus, W., EA, D., Wagner, D., & Zimmerman, J. (1985). APACHE II: a severity of disease classification system. *Critical Care Medicine*, 818-829. doi:10.1097/00003465-198603000-00013

Knaus, W., Wagner, D., Draper, E., Zimmerman, J., Bergner, M., Bastos, P., . . . Damiano, A. (Nov, 1991). The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 1619-1636. doi:10.1378/chest.100.6.1619

MD Calc. (2020). *APACHE II Score*. Retrieved 01 Apr, 2020, from MD Calc: https://www.mdcalc.com/apache-ii-score

Merck Manual. (2019). *APACHE II Scoring System and Mortality Estimates*. Retrieved 01 Apr, 2020, from Merck Manual: https://www.msdmanuals.com/professional/multimedia/clinical-calculator/apache%20ii%20scoring%20system

Siddiqi, N. (2017). *Intelligent Credit Scoring* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons Inc.

Tikkanen, R., & Abrams, M. K. (30 Jan, 2020). *U.S. Health Care from a Global Perspective, 2019: Higher Spending, Worse Outcomes?* Retrieved 01 Apr, 2020, from The Commonwealth Fund: https://www.commonwealthfund.org/publications/issue-briefs/2020/jan/us-health-care-global-perspective-2019

WiDS. (2020). *WiDS Datathon 2020*. Retrieved 01 Apr, 2020, from WiDS Conference: https://www.widsconference.org/datathon.html

Young, A. (01 Mar, 2019). *A New Way to Calculate Global Intensive Care Unit Mortality Risk*. Retrieved 04 Apr, 2019, from Institude for Medical Engineering & Science: http://imes.mit.edu/mit-gossis-algorithm-calculates-global-icu-mortality-risk/

Zimmerman, J., Kramer, A., McNair, D., & Malila, F. (May, 2016). Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 1297-1310. doi:10.1097/01.CCM.0000215112.84523.F0

Appendix

Confusion Matrix for Train data		
Decision Tree	Predict 1	Predict 0
Actual 1	29041	8065
Actual 0	11585	25521
CatBoost	Predict 1	Predict 0
Actual 1	33722	6549
Actual 0	7282	6657
LightBoost	Predict 1	Predict 0
Actual 1	33496	3610
Actual 0	7294	29812
NN	Predict 1	Predict 0
Actual 1	1688	346
Actual 0	4886	17121
Logistic Regression	Predict 1	Predict 0
Actual 1	29295	7811
Actual 0	7611	29495

Confusion Matrix for Test data		
Decision Tree	Predict 1	Predict 0
Actual 1	1532	650
Actual 0	6129	17931
CatBoost	Predict 1	Predict 0
Actual 1	1670	364
Actual 0	4759	17248
LightBoost	Predict 1	Predict 0
Actual 1	1779	403
Actual 0	5253	18807
NN	Predict 1	Predict 0
Actual 1	1626	594
Actual 0	4161	19861
Logistic Regression	Predict 1	Predict 0
Actual 1	1640	542
Actual 0	5884	18176