HDAT 9910 Capstone Project

Background

Intensive Care Unit (ICU) accommodates patients who require intense support and treatment, monitoring and specialised care. ICU often requires teams of trained specialists providing 24/7 care support to patients admitted from anywhere in critical health condition (Healthdirect, 2021). For the high requirement of medical resources (both medical professionals and equipment) and scarcity of available beds at any point in time, being able to accurately predict or understand impact on ICU mortality can help improve quality of ICU care and ICU resources allocation. The first 24 hours in ICU is very crucial for patients in terms of mortality. Machine learning has been widely applied in the healthcare industry in recent years. It has shown promising performance in predictive models for assisting medical decision-making such as disease diagnosis. The models developed in this project are not compared with any of the existing ICU Scoring systems such as APACHE II, SAPS II and SAPS III, although they have known limitations.

MIMIC-III database contains de-identified health-related data of the ICU stays at the Beth Israel Deaconess Medical Center between 2001 and 2012. Access has been provided for 16 tables comprising pre-processed data from the MIMIC-III database.

Project aims to build a predictive model for ICU mortality using data collected in the first 24 hours in ICU, using machine learning algorithms. The wrangled dataset for model building consists of 51,065 ICU stays of 47,820 hospital admissions opf 37,530 patients, including 61 Because of data unbalance, class weight has been considered for certain algorithms such as logistic regression. No feature reduction is considered in this project due to time scarcity. Best model will be evaluated in a list of hand-picked metrics including areas under the Precision-Recall Curve (PR_AUC) and under the receiver operating characteristic curve (ROC_AUC). For the high stake of false negatives (patients' lives), models will also be criticised harshly in terms of false negative counts.

Project will also investigate the impact of weekend admission to ICU on risk of ICU mortality by conducting contingency table analysis (odds ratio, risk ratio and chi-squared test of independence) and by fitting a simple logistic regression model and a multilevel linear model.

Methods

The primary table of interest is `pt_icu_outcome` (ICU Stay table) with each data point representing a unique ICU stay. `patients` (Patients table) and `admissions` (Admissions table) were extracted for patient demographics. `vitals_hourly` (Vitals table) consists of hourly vital sign measurements made at the bedside. `labs_hourly` (Labs table) contains pre- and post-ICU admission laboratory test results. `gcs_hourly` (GCS table) contains hourly GCS' including scores (overall, eye, motor and verbal).

Data Preparation[1]

Because the research interest is the ICU mortality, it is decided that the ICU Stay table will be the primary data table of interest. The units of analysis are each ICU stays which are nested within higher level units like hospital admissions and patients.

(Primary Table) ICU Stay Table

Data from the ICU Stay table has three levels: ICU stay, hospital admission and patient. contains three useful ID columns: “icustay_id” is the unique identifier/ the primary key of the table; “hadm_id” is the unique identifier of/ the foreign key to the Admissions table; “subject_id” is the unique identifier of/ the foreign key to the Patients table.

There are 61,532 unique ICU stays of 57,786 hospital admissions for 46,476 ICU patients. 3 out of the original 79 columns of the original table (“dod”, “ttd_days” and “hosp_deathtime”) have more than 50% of the values missing. No action was taken for missing values because: “hosp_deathtime” is dropped because it contributes to “dod” column with “dod_ssn” column according to Psysionet[2]. “ttd_days” is a pre-process calculated column using “dod” and “intime”, representing time to death in days (period between “intime” and “dod” in days).

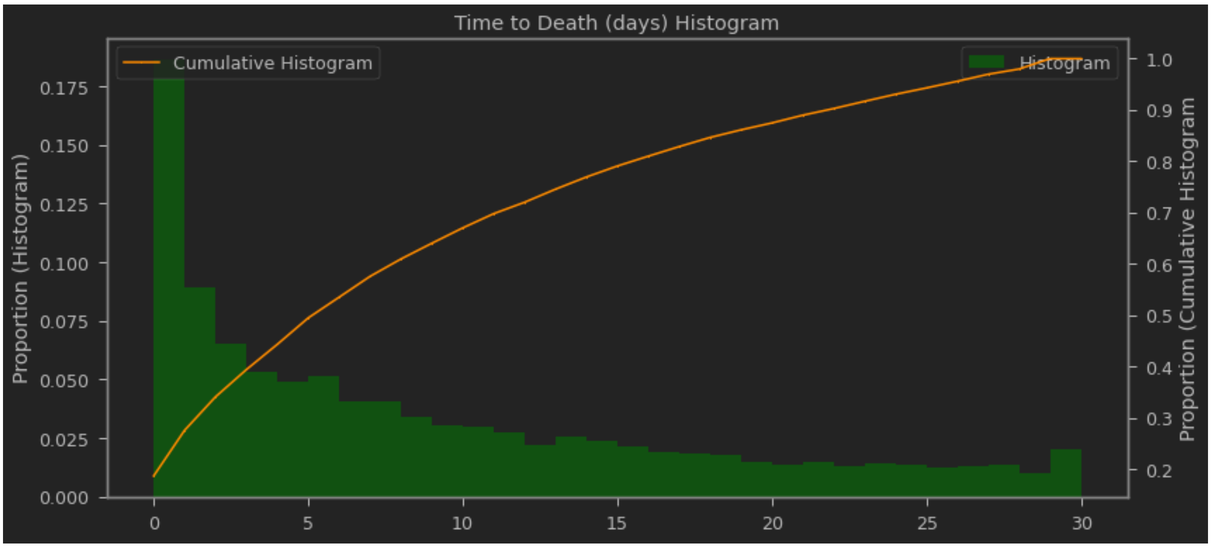
Records with missing ICU in-time or out-time have been excluded from the analysis, together with duplicated records for ICU stay ID 229922 and records with negative “ttd_days” values. There are 61,516 ICU stay records left in the data set after the extraction step.

`intime_weekday` is created to extract the day of week of ICU admission, whilst `icu_adm_weekend` recognises if an ICU admission occurred on the weekend. `age_years` is already transformed all ages over 89 years old to 94. It is a common practice to categorise age, and so `age_years` is further transformed and divided into 5 groups: “under 44”, “45-54”,”55-64”,”65-74” and “over 75”.

Outcome: ICU Mortality Definition

The outcome variable, ICU mortality[3], is not clearly defined. Out of the 61,516 records left in the dataset, 39% ICU stays had recorded deaths. However, time to death in days (TTD) varies from 0 to 4332 (11+ years from ICU discharge) with median 132 and 25-percentile at 14 days. Death recorded over 30 days after the ICU discharge are considered not an associated outcome of the ICU admission. As observed from histogram, the majority (over 60%) of ICU-associated deaths have recorded TTD under 7 days. Therefore, an ICU mortality in this analysis is standardised as “the death of an ICU patient that is recorded under 7 days of the patient’s ICU discharge (i.e. `ttd_days` < 7). Label is created and named `standard_mortality_label` (later renamed to `ICU_mortality` or `ICU_Mortality`). Label balance will be discussed in the model development step.

[1] Link to flowchart map of full ETL steps
[2] https://mimic.physionet.org/mimicdata/datatypes/
[3] ICU mortality is different from ICU mortality rate which is calculated by the number of deaths of ICU patients divided by the number of patient discharges from the ICU.



Patients Table

Only the gender and subject ID column are extracted because the rest could be found in the pre-processed ICU Stay table. The original patients table contains 46,520 patient records (not all of them have been admitted into ICU). An inner join between the wrangled Patients table and the wrangled ICU Stay table on \`subject_id\`.

Admissions Table

The admissions table contains 58,976 unique hospital admissions with over 50% values missing in the \`death_time\` column. Variables such as \`marital_status\` and \`ethnicity\` can also possess high predictive power for the model's purpose. But excluded in this project due to scarcity of time, as those variables require more cleaning/wrangling. Routinely collected data such as \`admission_type\`, \`insurance\`, \`language\` and admission timestamps are considered useful and relevant. As a patient can have multiple hospital admission/readmission records and therefore multiple languages on record. \`english_speaker_True\` is 1 for patients who have at least one of their hospital admissions records as an English-speaker (i.e. language is "ENGL"). \`re_adm_in30d\` is a binary column which equals 1 when the same patient was readmitted to the hospital within 30 days of previous hospital discharge. \`len_of_adm\` represents the length of hospital admission in days, from being admitted to being discharged. The wrangled Admissions table will be inner-joined with the wrangled ICU Stay table on \`hadm_id\`.

Tables Contain Hourly Records

As one of the research questions is only interested in using data from the first 24 hours in the ICU, all three tables with hourly records are cleaned to include only the records of the first 24 hours of each ICU stay ID (i.e. 0<\`hr\`<=24).

Vitals Table

The original vitals table contains over 7 million hourly vital sign bedside measurements for 61,522 unique ICU stays (less records than the ICU Stay table). 3 out of the 11 columns (`fio2`, `glucose` and `temperature`) have more than 50% values missing. Acknowledging that Vitals table may comprise "artificial" or error entries and that units of measurements recorded may vary by records, in addition to the high volume of missing values in the three columns, it is better to transform and aggregate than to drop or to immutate. The aim of wrangling this table is to identify if any health-threatening medical condition was signalled by the bedside measurements during the first 24 hours of each ICU stay.

It is recognised that some temperature entries ranging from 96 to 108 may have been recorded in Fahrenheit. Their `temperature` values are converted to Celcius accordingly. Rows with "highly unlikely" values in measurement columns are excluded (for example: `temperature` below 0°C or `heartrate` over 480BPM)[4]. New columns are created for each row with their original columns dropped. These new columns are to identify medical conditions at the time measurements were taken, including stages/severity of hypertension, signs of hypoxaemia or respiratory failure, fever, significant risk of heart attack/stroke, tachycardia or bradycardia, or hyperglycemia. The table is then grouped by `icustay_id` with all remaining values (except `hr`) summed by group. These grouped sum columns are then transformed into binary columns (whose values equals 1 if the grouped sum is greater than 0, indicating a medical condition ever signalled by the bedside measurements taken in the first 24 hours of the ICU stay). For the final dataset for predictive model development, the aggregated Vitals table is at ICU stay level, and will be inner-joined to the wrangled ICU Stay table on `icustay_id`.

Labs Table

The original vitals table contains 928,195 hourly laboratory test results for 60,313 unique ICU stays (less records than the ICU Stay table). 19 out of the 21 columns have more than 50% values missing. Due to the extremely high volume of missing values and lack of knowledge of hospital procedures (when and how the urinal/blood tests were taken), to transform and aggregate is again preferred over dropping or immutatation. The aim of wrangling this table is to identify if any health-threatening medical condition was signalled by the laboratory test results recorded during the first 24 hours of each ICU stay.

Instead of dropping rows with "highly unlikely" values, new columns are created for each row with their original columns dropped to identify abnormal readings that signal dangerous health conditions at the time when samples were taken, which includes hepatocyte damage/acute hepatitis, dehydration, hyperglycemia, diabetes, impaired kidney functionality, anaemia, leukemia and etc[5]. The table is then grouped by `icustay_id` with all remaining values (except `hr`) summed by group. These grouped sum columns are then transformed into binary columns (whose values equals 1 if the grouped sum is greater than 0, indicating a medical condition ever signalled by the lab results taken in the first 24 hours of the ICU stay). For the final dataset for predictive model development, the aggregated Labs table is at ICU stay level, and will be inner-joined to the wrangled ICU Stay table on `icustay_id`.

[4] More details on ranges applied and references see Supplementary Information section

[5] More details on ranges applied and their medical condition indications see Supplementary Information section

GCS Table

The original vitals table has over 1.5 million hourly GCS scores and its component scores for 52,355 unique ICU stays (less records than the ICU Stay table). No column contains more than 50% values missing and all scores vary within the GCS scale. The aim of wrangling this table is to identify if any of the GCS scores (overall or individual assessment) show deep unconsciousness of the patient or under mechanical ventilation during the first 24 hours of each ICU stay.

Column 'endotrachflag' indicates if the patient was under mechanical ventilation. The total score on GCS scale varies from 0 to 15: under 8 indicates severely impaired consciousness and 8-12 indicates moderately impaired consciousness. Although the scores for assessing the ICU patients ability to perform eye movements, speak, and move their body use different scales, a 0 score in any of the three assessments means no response/deeply unconscious. New columns are created with the original columns dropped to identify severely and moderately impaired consciousness, and unresponsiveness to eye, verbal and movement stimuli. The table is then grouped by `icustay_id` with all remaining values (except `hr`) summed by group. These grouped sum columns are then transformed into binary columns (whose values equals 1 if the grouped sum is greater than 0, indicating the patient was ever dangerously unconscious according to GCS scores taken, or was ever under mechanical ventilation, in the first 24 hours of the ICU stay). For the final dataset for predictive model development, the aggregated GCS table is at ICU stay level, and will be inner-joined to the wrangled ICU Stay table on `icustay_id`.

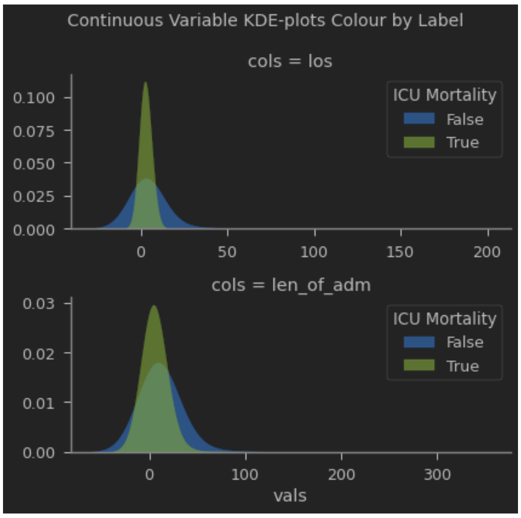
ML Model Development

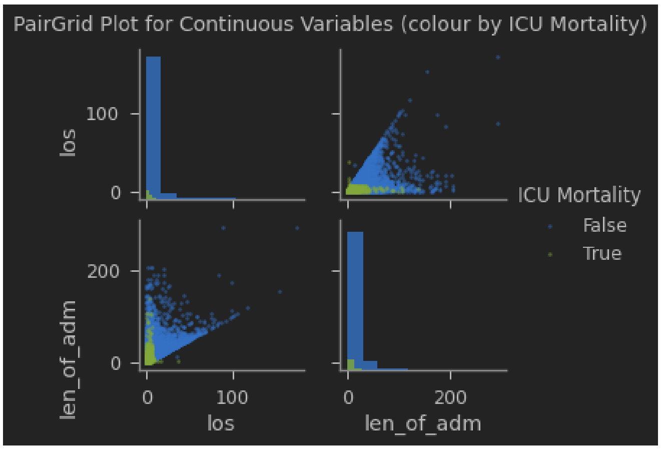
The development happened in a a well set-up virtual environment using Python 3.8 Anaconda[6]. The ML model is to perform a classification task. Its sole objective is to predict the ICU mortality for an ICU stay using 61 inputs collected during the first 24 hours in the ICU. Linear and tree-based models are commonly used models for the classification tasks, known for their promising performance.

Variables for Machine Learning (ML) Model

The final table for developing the ML model is formed by inner-joining the primary table with patients, admissions and hourly-record tables on respective primary-foreign keys. To prepare the data for ML algorithms, one-hot-encoding is applied to all the categorical and few boolean[7] columns, while the rest of the boolean columns are transformed into binary integer columns. After this step, the table is left with 51,065 rows of unique ICU stays (of 47,820 hospital admissions for 37,530 ICU patients). ID columns (`subject_id`, `hadm_id` and `icustay_id`) are dropped before saving the final input dataset in the local directory. Candidate variables in the final table for ML model development were carefully examined, screened and engineered on the basis of reviewing published literature and existing published ICU scoring systems (such as SAPS ⅔ and APACHE II). 61 features were selected.

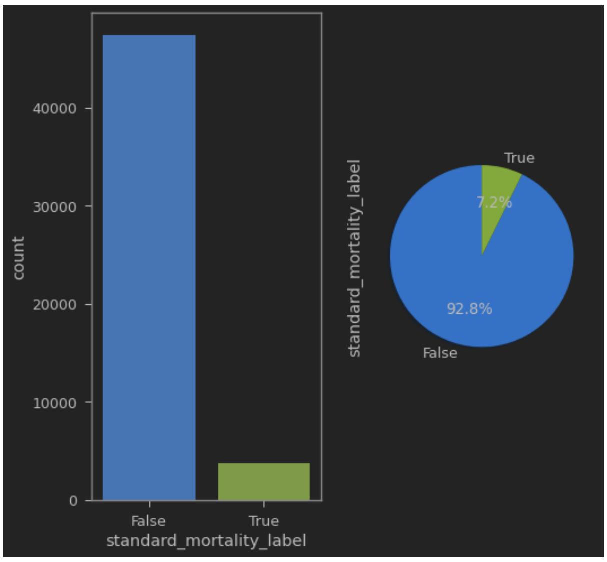
[6] Please find exported environment *.yml file in provided repository
[7] `gender_M` flagging male patients and `english_speaker_True` flagging English-speaking patients

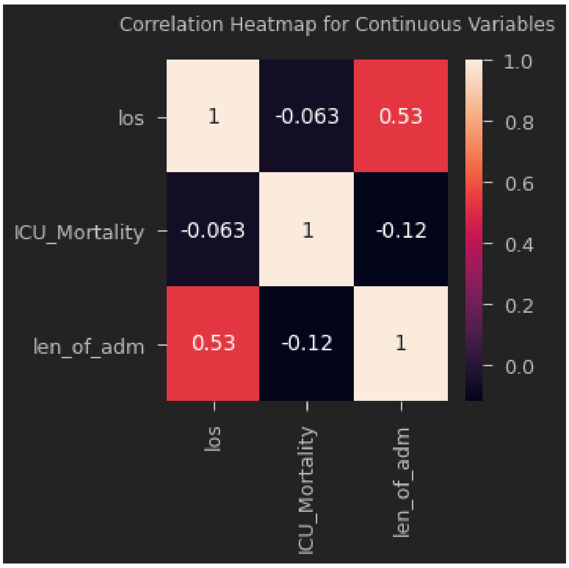




Label Imbalance

The final dataset has 47368 negative cases and 3697 positive cases. The ratio is approximately 13:1 which is very imbalanced (~ 93% vs 7%). There is no strong correlation between any feature and the target label.





For the seriously imbalanced data, pruning will be conducted to improve F1-score, Average Precision (AP), Matthews correlation coefficient (MCC) and Cohen's Kappas (CK), primarily refit on F1-score. MCC is a more reliable metric compared with F1-Score and accuracy dealing with imbalanced dataset, as it considers the ratio between positive and negative cases (Chicco & Jurman, 2020). It is also advised to use CK when facing imbalance (Brownlee, 2020). Misclassifying an ICU stay as high risk of mortality (i.e. a false positive) results in wasteful medical resource. Misclassifying an ICU stay as low risk of mortality (i.e. a false negative) puts the patient’s lives at stake (much higher cost). As false negative is much higher at stake than a false positive, False Positive count will be kept reasonably low however can be compromised under circumstances.

For model performance comparison, accuracy, although important, will not be used as the primary metric to evaluate model performance, for its insensitivity to false positives and false negatives. In general, the higher the F1-score and the accuracy on both training and test sets, the better. Areas under the Precision-Recall Curve (PR_AUC) and under the receiver operating characteristic curve (ROC_AUC).

Training and Pruning

Four off-the-shelf algorithms are selected for this project: Logistic Regression (LR) for linear options, and Random Forest Classifier (RFC), Gradient Boosted Classifier (GBT) and eXtreme Gradient Boosting Classifier (XGB) for tree-based options.

It is decided to leave 25% of the data that are randomly selected to be the test set. Due to imbalance, stratification is turned on when using the function under `sklearn`, so that the function distributes the identical percentage of classes across the training and test set. Although tree-based models are almost immune to unnormalised data, features require scaling/normalisation before putting into LR models. Therefore, features for both training and test sets are scaled using the StandardScaler from sklearn package. Feature normalisation is decided to be used for all models trained and pruned in this project.

Baseline models were then created for each classifier and showed greatly promising results with an accuracy range of 94.4% to 95.3% and a F1-score range of 44.9% to 58.9%. The best performing baseline model is the one trained under XGB, also has the highest recall when predicting the positive cases.

(Table of baseline model performance summary)

Baseline Model Algorithms	**Accuracy**	**F1 score**	**PR_AUC**	**ROC_AUC**	**Recall-** **predicting positive**	**Precision-** **predicting positive**	**F1 score-** **predicting positive**
LR	0.944	0.449	0.554	0.869	0.31	0.78	0.45
RFC	0.950	0.488	0.631	0.887	0.33	0.94	0.49
GBT	0.952	0.547	0.647	0.902	0.40	0.87	0.55
XGB	0.953	0.589	0.648	0.899	0.47	0.79	0.59

When cross-validating using the Grid Search method, to compensate for label imbalance, `StratifiedShuffleSplit` is applied to create randomised stratified folds for cross-validation. This method combines `StratifiedKFold` and `ShuffleSplit`, which appears to be the most suitable in this use case for the label unbalance and robustness (Scikit-learn.org, 2020). The cross-validation pipeline has `StandardScaler` as the default `Transformer` and assigning `Estimator`'. GridSearchCV was looking for the best parameters for each model with the best scores in F1-score, AP, MCC and CK whilst refitting on F1-score.

Statistical Model Development

Statistical analysis and model development is conducted in R-studio IDE using R 3.6.3.

Variables for Investigating Weekend ICU Admission and Risk of ICU Mortality

The final table for modelling weekend ICU admission and risk of ICU mortality is different from the table used for the ML model. There are more records in this table as it is an inner-join between wrangled ICU Stay table and wrangled Admissions table only. This table at ICU stay level, contains 61,516 unique ICU stays of 57,771 hospital admissions for 46,464 ICU patients.

The MIMIC-III dataset can be seen as a longitudinal multilevel table:

- Level 3 ICU stay
- Level 2 Hospital Admission
- Level 1 Patient

In this analysis only level 2 and level 3 are investigated for time scarcity.

Variable Summary:

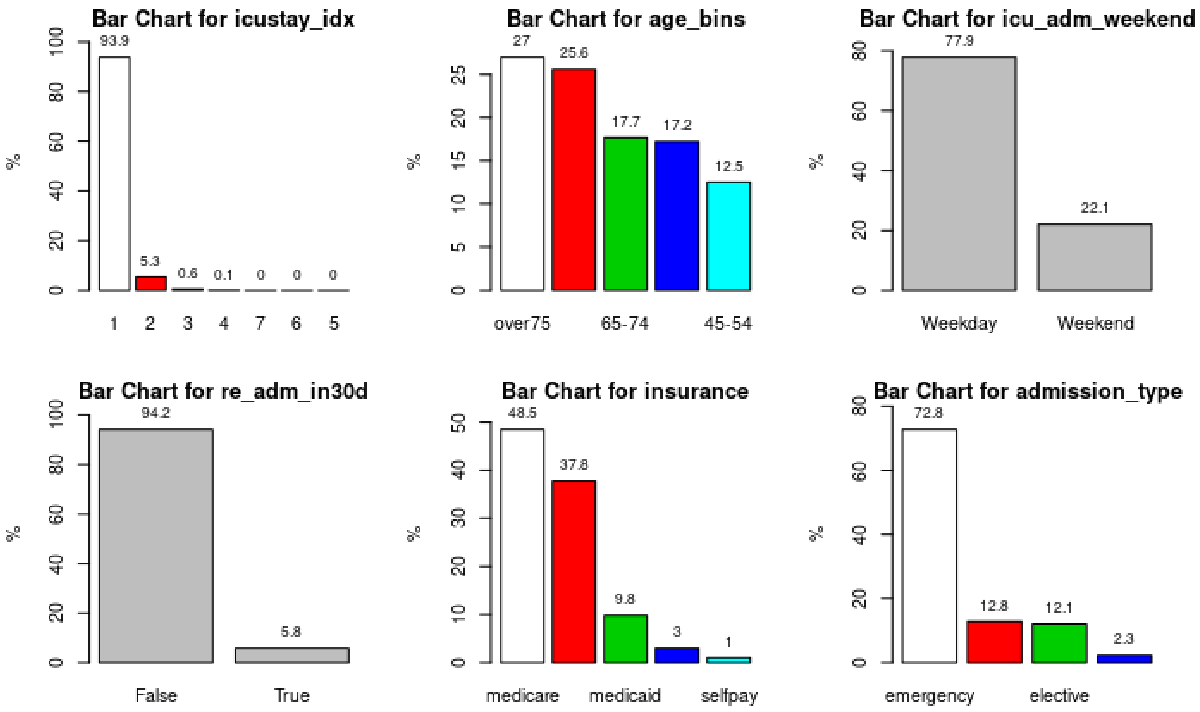
- Indicators:
  - `hadm_id` hospital admission identifier
  - `icustay_idx` ICU stay identifier grouped by hospital admissions[8]
- Dependent Variable:
  - `ICU_mortality` 1=ICU mortality, 0=No ICU mortality
- Independent variable of interest:
  - `icu_adm_weekend` 1=ICU admission on weekend, 0=ICU admission on a weekday
  - `age_group` ordinal variable from 1 to 5 representing age groups “under 44”, “45-54”, “55-64”, “65-74” and “over75”

Some characteristic of the sampled data can be extracted from the bar charts below:

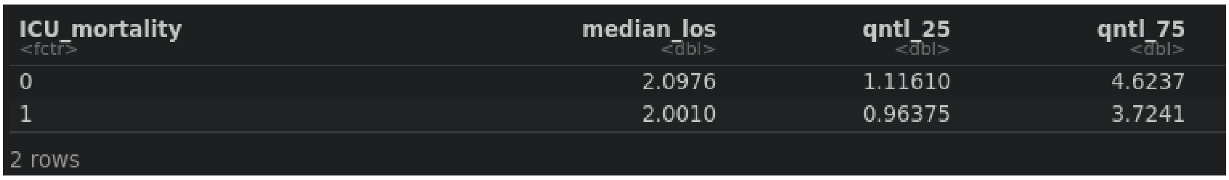
- 93.9% of the ICU stays has only one “parent” hospital admission (i.e. an 1-on-1 relationship between `icustay_id` and `hadm_id`)
- The highest number of ICU readmissions of any hospital admission is 7

[8] Original ICU stay IDs are discarded as it cannot be used for longitudinal analysis. New ICU stay IDs are created at hospital admission level to recognise the n-th ICU stay of a hospital admission.

- 27% of the ICU admissions occurred when patients were over 75
- 22.1% of the ICU admissions happened during weekends
- 5.8% of the ICU stays belong to a hospital 30-day readmission
- 48.5% of the ICU stays were paid by Medicare
- 37.8% of the ICU stays were paid privately
- 72.8% of the ICU stays had been admitted as type “Emergency”

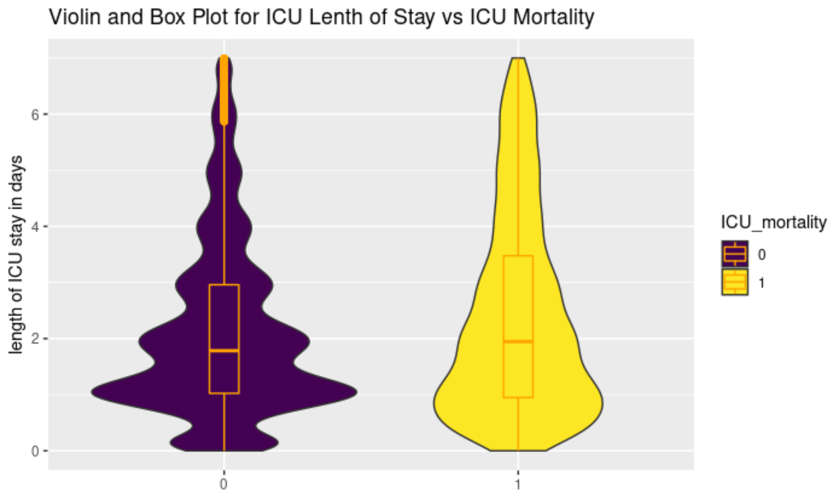


As shown in the violin and box plot, length of ICU stay in days (LoS) is slightly shorter for stays that resulted in ICU mortality than stays that did not result in ICU mortality.



ICU_mortality <fctr>	median_los <dbl>	qntl_25 <dbl>	qntl_75 <dbl>
0	2.0976	1.11610	4.6237
1	2.0010	0.96375	3.7241

2 rows



Statistical Analysis for Weekend ICU Admission and ICU Mortality

To investigate the effect of weekend ICU admission on risk of ICU mortality, we start by drawing a contingency table for the two variables.The odds ratio and risk ratio (and their 95% confidence intervals) are examined. Chi-squared test of independence is then conducted to determine if an association between the two categorical variables exists.

Before training any statistical model. It is noted that there is a serious imbalance in data. Undersampling method is chosen to tackle the imbalance issue (undersampled same number of positive and negative cases for model training).

Because both variables are binary and categorical, firstly, a simple logistic regression is examined via `glm()` function, as the response variable is non-normal. Since you can have multiple ICU stays per hospital admission, independence of responses is questioned. Therefore, mixed linear models (MLM) are used and evaluated (achieved by using `glmm()` function with family='binary' and link function 'logit'). Because ICU stays within the same hospital admissions are likely to result in the same response, all the MLMs will use level 2 and 3 (hospital admissions and ICU stays). The models are then compared and discussed to draw a final conclusion.

Results

ML model for predicting ICU mortality based on data from the first 24 hours in ICU

ML Model Input Data Description

Characteristics of ICU patients, their stays and hospital admissions

The input data contains 51,065 ICU stays with length of stay varying from 0.18 to 173.07 days. The sampled ICU stays belong to 37,530 patients with

- 57% male and 43% female;
- 56% ever reported at hospital admissions as an English speaker.

These ICU stays belong to 47,820 hospital admissions which:

- 48% admission occurred when patient was married
- 83% admissions are of emergency type and 14% of elective type
- 44% admissions patients were ever under mechanical ventilation
- 56% of the admissions paid using medicare with 32% privately paying

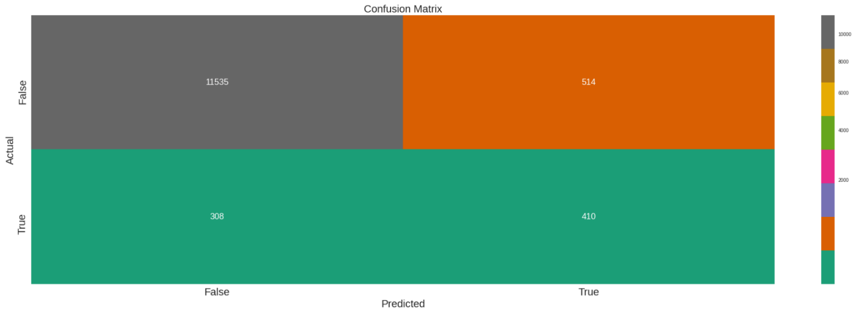
ML Model Selection and Evaluation

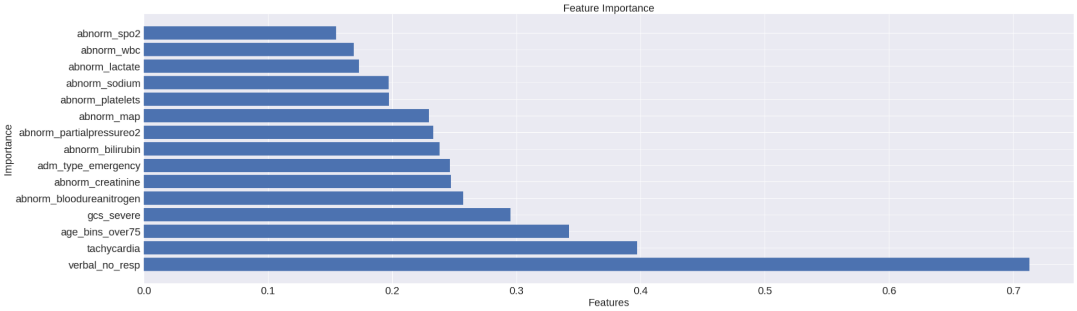
(Table of pruned model performance summary: values in brackets are of baseline models)

Pruned Model Algorithms	**Accuracy**	**F1 score** **test**	**F1 score** **(best cv score)**	**Recall**	**PR_AUC**	**ROC_AUC**	**Recall-** **predicting positive**	**Precision-** **predicting positive**	**F1 score-** **predicting positive**
LR	0.936 (0.944)	0.499 (0.449)	0.528	0.571/ 0.444	0.538 (0.554)	0.872 (0.869)	0.44 (0.31)	0.57 (0.78)	0.45 (0.50)
RFC	0.940 (0.950)	0.597 (0.488)	0.609	0.585/ 0.610	0.665 (0.631)	0.922 (0.887)	0.61 (0.33)	0.58 (0.94)	0.60 (0.49)
GBT	0.957 (0.952)	0.635 (0.547)	0.641	0.726/ 0.513	0.726 (0.647)	0.937 (0.902)	0.51 (0.40)	0.83 (0.87)	0.63 (0.55)
XGB	0.955 (0.953)	0.664 (0.589)	0.666	0.729/ 0.609	0.728 (0.648)	0.937 (0.899)	0.61 (0.47)	0.73 (0.79)	0.66 (0.59)

Logistic Regression (LR)

Baseline LR model was the worst performing baseline model. For hyperparameter tuning, the classifier has its solver set as liblinear for it supports both Lasso (L1) and Ridge (L2) penalty (Scikit-learn.org, 2020d.). Grid Search cross-validation is performed on parameters class_weight, C and penalty. The best performing post-tuning model has parameters: C = 10, class_weight of {1: 0.7, 0: 0.3}, and penalty='l2'.

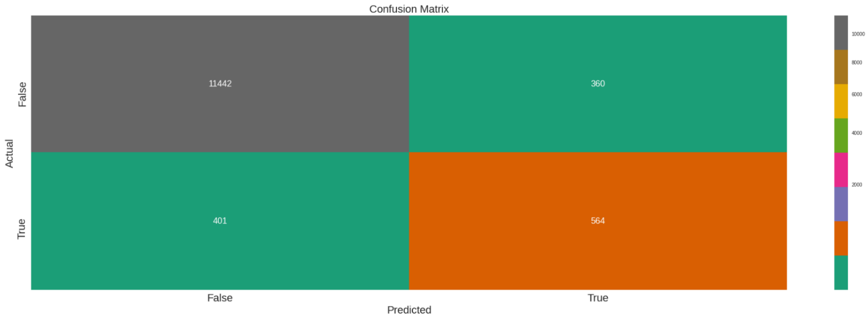


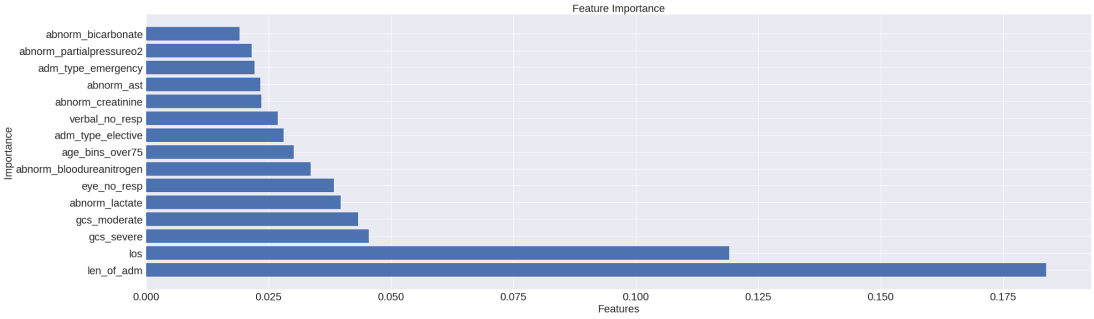


When predicting for test set the tuned model only mildly improved the F1-score, and recall when predicting the positive labels (ICU Mortality). The best cross-validation f1-score 0.530 is better than it of test set prediction (0.499), potentially suggesting that there may be a slight overfitting problem with the tuned model. Hyperparameter tuning did not improve LR model by much. There are 308 false negatives and 514 false positives after pruning.

Random Forest (RFC)

The baseline RFC model is the second worst baseline model. Grid Search cross-validation was performed on class_weight, max_depth and n_estimators, and found the best parameters with class_weight as {0: 0.9, 0: 0.1}, max_depth of 10 and n_estimators of 100. Hyperparameter tuning improved the baseline RFC models performance in F1-score, PR_AUC, ROC_AUD and recall when predicting positive cases.





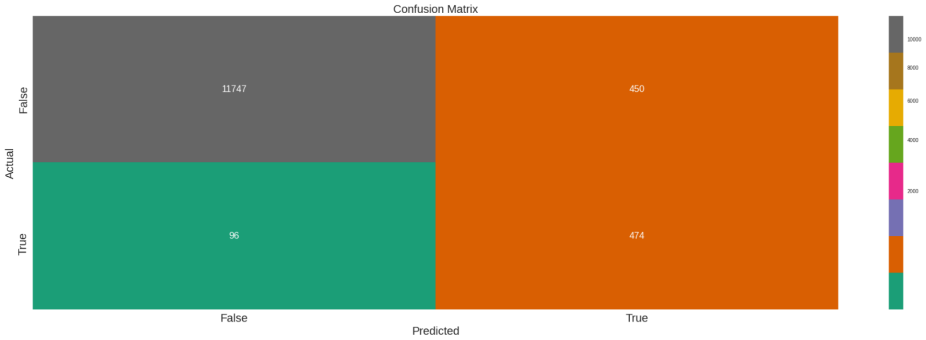
There are 401 false negatives and 360 false positives after pruning.

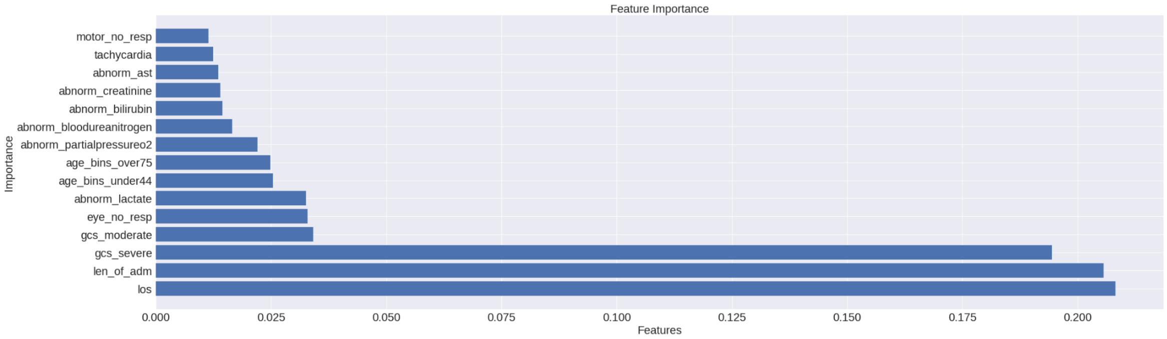
Gradient Boosted Classifier (GBT)

The baseline GBT model already showed satisfying results. Grid Search cross-validation was performed on learning_rate, max_depth, and n_estimators.The best performing model has parameter learning_rate of 0.1, max_depth of 6 and n_estimators of 100. Compared with the baseline GBT model, the pruned model has improved in accuracy, F1-score, PR_AUC, ROC_AUD, and recall when predicting positive cases.

There are 96 false negatives and 450 false positives after pruning. The top 10 important features of the pruned GBT model are:

1. days of ICU stay
2. days of hospital admission
3. severely impaired consciousness
4. moderately impaired consciousness
5. no eye response (GCS score)
6. abnormal lactate level
7. age under 44
8. age over 75
9. abnormal partial pressure O2 level
10. abnormal blood urea nitrogen level



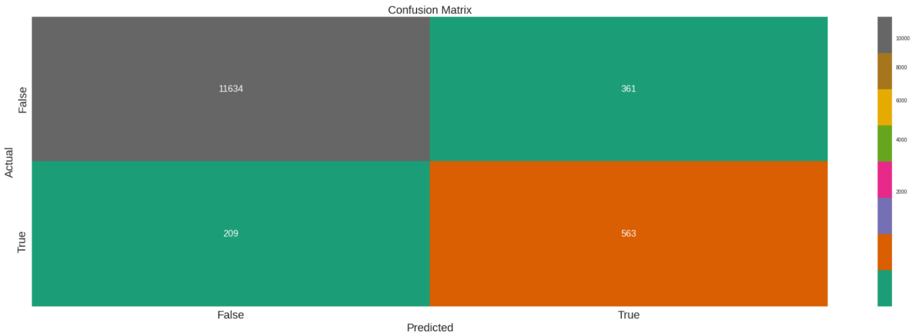


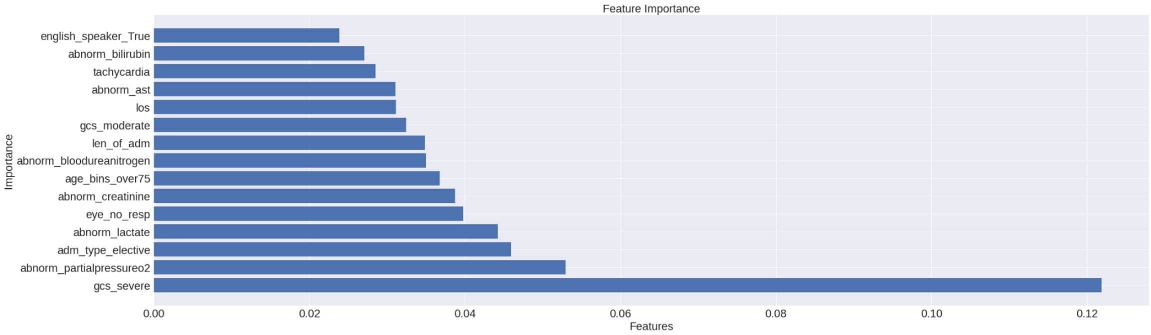
eXtreme Gradient Boosting Classifier (XGB)

The baseline XGB model was the best performing baseline model by any means. Classifier has “booster” set as “gbtree”, “objective” set as “binary:logistic” and “eval_metric” set as error, AUC_ROC and AUC_PR for the type of task it’s performing and label imbalance. “tree_method” is set to “gpu_hist” to utilise GPU for fast computation, which is only available in version 1.5.0. Grid Search cross-validation was performed on almost all parameters tuned separately in a specific order: scale_pos_weight, learning_rate, subsample, colsample_bytree, colsample_bylevel, max_delta_step, n_estimators, sampling_method. The best performing model has parameters:

- subsample=1
- colsample_bylevel=0.9
- colsample_bytree=0.6
- learning_rate=0.3
- scale_pos_weight=2
- gamma=0.2
- max_depth=4
- reg_alpha=0
- reg_lambda=1

Compared with the baseline XGB model, the pruned model has improved in accuracy, F1-score, PR_AUC, ROC_AUD, and recall when predicting positive cases.





There are 209 false negatives and 361 false positives after pruning. The top 10 important features of the pruned XGB model are:

1. severely impaired consciousness
2. abnormal partial pressure O2 level
3. hospital admission type elective
4. abnormal lactate level
5. no eye response (GCS score)
6. abnormal creatinine level
7. age over 75
8. abnormal blood urea nitrogen

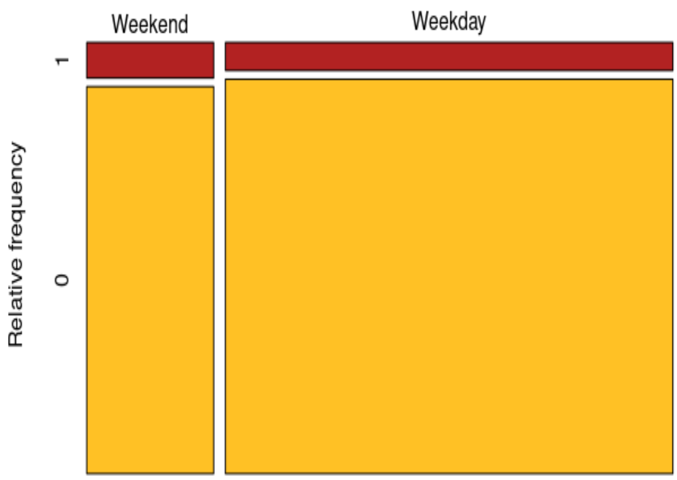
9. days of ICU stay
10. days of hospital admission

Statistical model to estimate the effect of weekend ICU admission on the risk of mortality

Contingency Table Analysis for Association

The 2x2 contingency table of the dependent and independent variables:

	Weekend	Weekday	Total
ICU Mortality	1143	3180	4323
Non ICU Mortality	12453	44740	57193
Total	13596	47920	61516



- Estimated odds ratio = 1.29 with 95% C.I. (1.20, 1.39) with small p-value
- Estimated risk ratio = 1.21 with 95% C.I. (1.15, 1.28) with small p-value
- Chi-squared test of independence shows a very small p-value (1.006 * 10^{-12})
- G-test for Contingency Tables also shows a very small p-value (2.77 * * 10^{-12})

All results confirm that there is sufficient statistical evidence that there is association between ICU admission on the weekend and ICU mortality. The odds ratio and risk ratio are relatively close with very narrow confidence intervals. This suggests that ICU admissions on the weekend are approximately 1.29 times more likely to result in ICU mortality, compared with weekday ICU admissions.

Fitting a Simple Logistic Regression (SLR)

Without considering the multilevel nature, assuming that all responses are independent and that error is normally distributed with mean 0 and some standard deviation σ, a simple logistic regression model between the two variables is constructed. The SLR is fitted using the undersampled dataset, using formula (formula = `ICU_mortality` ~ `icu_adm_weekend`), presenting below results:

	Estimate	95% CI	P-value
Intercept	-0.0678	(-0.116, -0.0194)	0.006
ICU Admitted on Weekend	0.285	(0.186, 0.384)	1.93e-08

The model achieved accuracy around 74.6%[9] when predict ICU mortality for the dataset before undersampling,

Fitting a Mixed Linear Models (MLM)

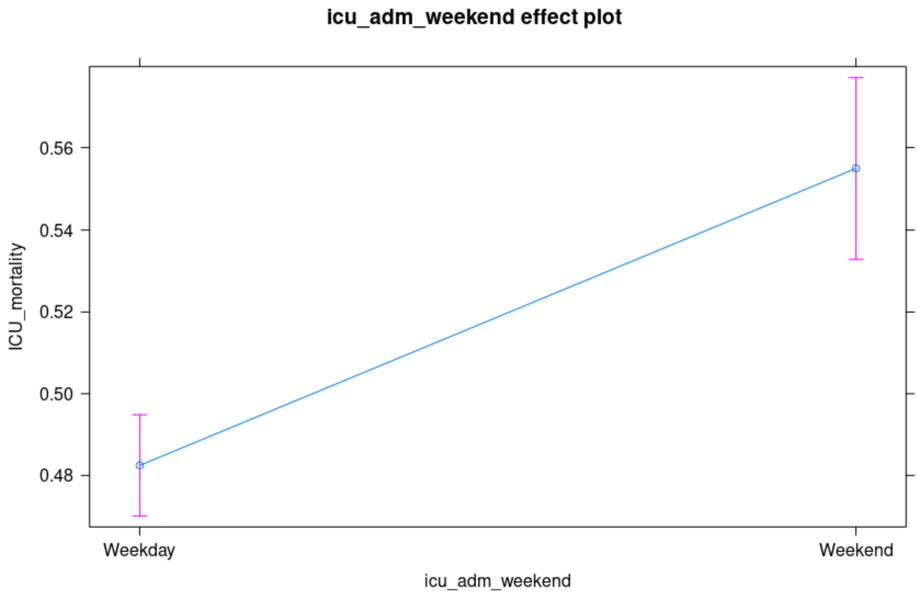
By comparing (log-likelihood of) single-level and empty random intercept models at hospital admission level and patient level, and with a VPC of 0.6, it is suitable to apply Hierarchical Multilevel Model to this data considering `icu_adm_weekend` as a random effect. It suggests that there is an impact on `ICU_mortality` by `hadm_id`, i.e. risk of `ICU_mortality` varies among `hadm_id`s.

First a random intercept only model is created to investigate the effect of weekend ICU admission on ICU mortality (ICU_mortality ~ icu_adm_weekend + (1|hadm_id))

	Estimate	95% CI	P-value
Intercept	-0.0702	(-0.120, -0.0207)	0.00555
ICU Admitted on Weekend	0.291	(0.189, 0.393)	2.26e-08

The variance in random effect (`hadm_id`) is 0.0932. The p-values are very small for each of the coefficient estimates. The results indicates that with `ICU_mortality` = -0.0702 + 0.291 * `icu_adm_weekend`

[9] 95% CI = (0.7424, 0.7493)

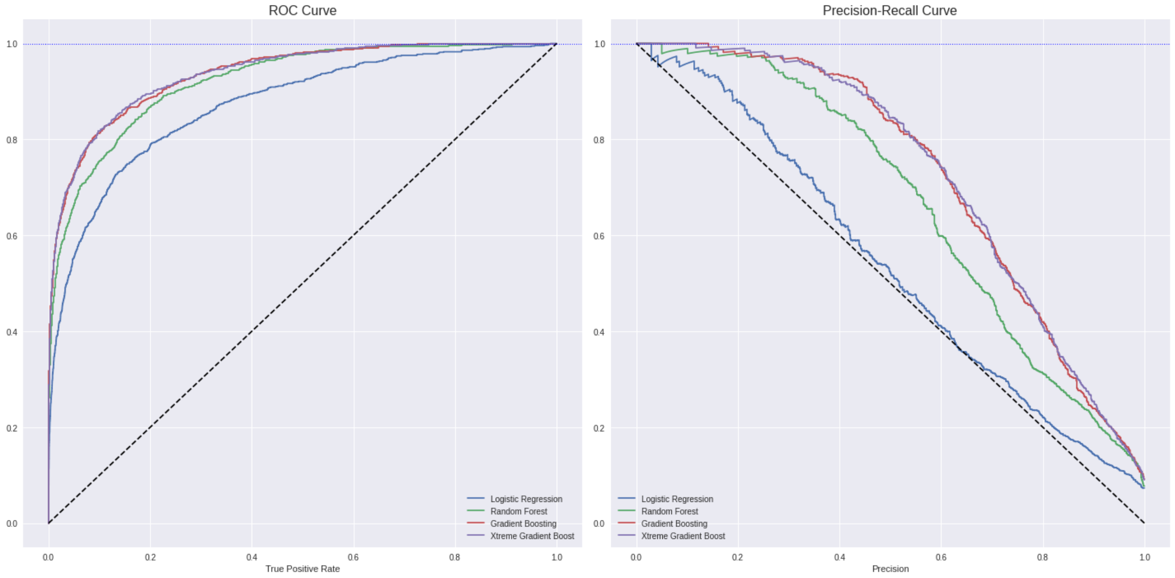


The estimate for `icu_adm_weekend` is positive, which means that it is estimated to increase ICU mortality on average, across hospital admissions. The estimate for `icu_adm_weekend` is significant and gives confidence in the “positive” effect. The effect plot above using the `effect` package visualises the effect of weekend ICU admissions on ICU mortality by this MLM. The MLM has log-loss of -5975.974 (df=3) which is greater than -5977.097 (df=2) of the SLR model, which also agrees that MLM has a better fit than SLR.

Conclusion and Discussions

It is possible to accurately predict mortality based on data from the first 24 hours in ICU

By looking at the ROC and PR curve plotted for all four models, it is clear that GBT and XGB models outperformed the LR and RFC models.

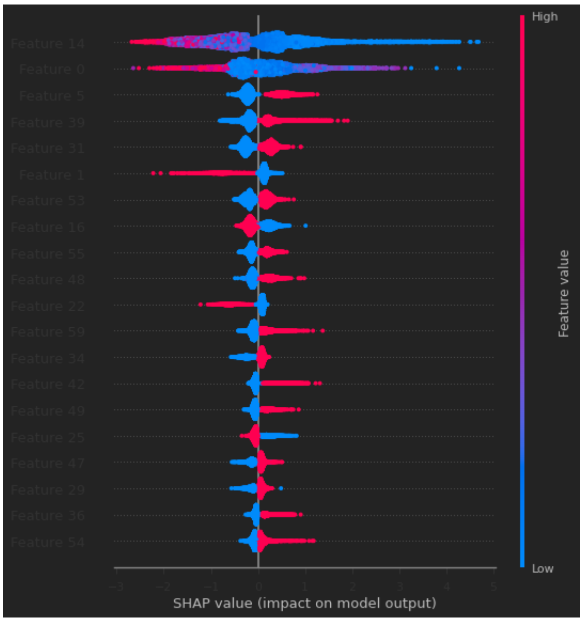


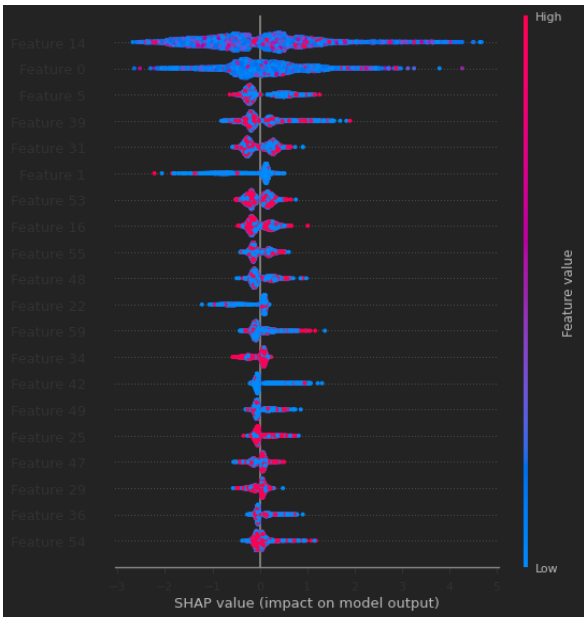
The XGB model slightly outperformed in all measures, except in accuracy and false negative counts. It is difficult to decide which model is the best. Under situations where ICU resources are extremely scarce and madly expensive, the XGB model may be considered better. Because with only 0.2% compromise in accuracy, it contains less bias when predicting ICU mortality and produces approximately 20% less false positives than the GBT model. However, in most cases, the stake of false negatives outweighs it of false positives. Improvement of XGB in MCC and AP are so insignificant. GBT produced 96 false negatives (less than half of what XGB produced). Therefore, the GBT model is arguably the best performing model for predicting ICU mortality.

The best model uses Gradient Boosted tree-based algorithm, achieving accuracy 95.7%, F1-score 63.5% and least number of false positives.

Model	**Best F1-score CV**	**AUC_ROC**	**Accuracy**	**PR_ROC**	**F1-score test**	**MCC**	**CK**	**False Negatives**	**False Positives**
GBT	0.641	0.937	0.957	0.726	0.635	0.633	0.613	96	450
XGB	0.665	0.937	0.955	0.728	0.664	0.643	0.640	209	361
Diff. %	3.74%	0%	-0.21%	0.28 %	4.57%	1.58%	4.4%	118%	-20%

By using the `shap` package and 10,000 rows in both training and test data, it is made possible to visualise feature importance when predicting negative and positive cases in both training set (graph on the left) and test set (graph on the right). It is unclear in case of test set data, but for training set los, len_of_adm, age_bins_under44 were the most important when predicting negative cases, whilst los, len_of_adm, gcs_severe were the most important when predicting positive cases (ICU mortality).

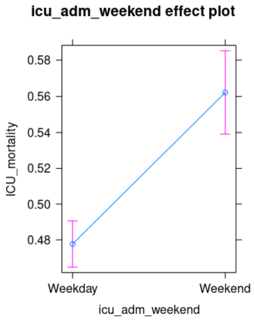


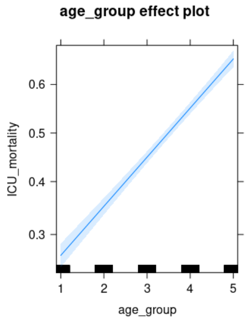


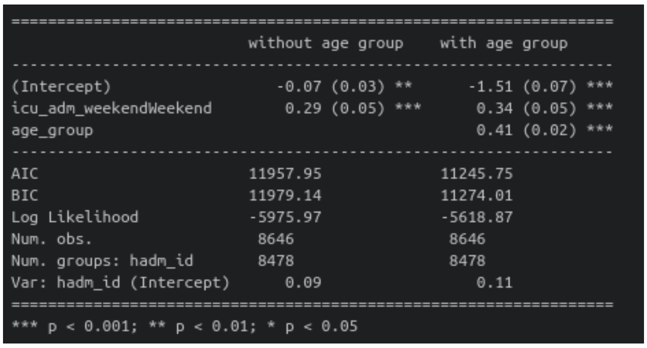
Admission to ICU over the weekend increases the risk of mortality

Contingency table analysis provided statistically significant evidence in the existence of an association between admission to ICU over the weekend and risk of mortality. By fitting undersampled data with a MLM model (ICU_mortality ~ icu_adm_weekend + (1|hadm_id))[10], we are confident that weekend ICU admission increases the risk of ICU mortality. The MLM model has a better fit than the SLR model.

To find a MLM with better fit, we also fitted a model with both icu_adm_weekend and age_group[11] in fixed effect. Compared with the MLM above, the new model has lower AIC, lower BIC, higher log likelihood and with high confidence in coefficient estimates. Both coefficient estimates (0.34 and 0.41 for icu_adm_weekend and age_group respectively) are "positive" effects on ICU mortality. In a word, weekend ICU admission and older in age at admission increases risk of ICU mortality.







```
==================================================================
                              without age group    with age group
------------------------------------------------------------------
(Intercept)                      -0.07 (0.03) **      -1.51 (0.07) ***
icu_adm_weekendWeekend            0.29 (0.05) ***      0.34 (0.05) ***
age_group                                              0.41 (0.02) ***
------------------------------------------------------------------
AIC                           11957.95             11245.75
BIC                           11979.14             11274.01
Log Likelihood                -5975.97             -5618.87
Num. obs.                      8646                 8646
Num. groups: hadm_id           8478                 8478
Var: hadm_id (Intercept)          0.09                 0.11
==================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

[10] `hadm_id` as random effect and `icu_adm_weekend` as fixed effect using "logit" link function under binomial family
[11] `age_group` is ordinal categorical (e.g. "under 44":1, "45-54":2, "55-64":3 ...)

References

Healthdirect.gov.au. 2021. What to expect in intensive care. [online] Available at: https://www.healthdirect.gov.au/what-to-expect-in-intensive-care [Accessed 9 May 2021].

Nickson, C., 2020. *Standardised Mortality Ratio • LITFL • CCC Research*. [online] Life in the Fast Lane • LITFL. Available at: <https://litfl.com/standardised-mortality-ratio/> [Accessed 9 May 2021].

Scikit-learn.org. 2020e. Sklearn.Model_Selection.Stratifiedshufflesplit — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html?highlight=stratified#sklearn.model_selection.StratifiedShuffleSplit [Accessed 8 May 2021].

Scikit-learn.org. 2020d. Sklearn.Linear_Model.Logisticregression — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [Accessed 8 May 2021].

Developers, XGBoost. (2020). XGBoost Tree Methods — xgboost 1.5.0-SNAPSHOT documentation. XGBoost Documentation. https://xgboost.readthedocs.io/en/latest/treemethod.html [Accessed 8 May 2021].

Supplementary Information

Item 0: GitHub Repository for all relevant files (code, notebooks, pickle files)
Item 1: Detailed Schematic Diagram for ETL process

Item 2: Summary table for all hourly measures/lab results, their normal range, and potential medical condition indication by anomaly

Table Name	Measures	Normal Range	Potential Medical Conditions Indication by Anomaly
labs_hourly	bicarbonate	23-29	metabolic alkalosis
labs_hourly	albumin	3.5-5	malnutrition/ acute infections, heart attack
labs_hourly	troponin	<0.4	heart attack or other heart damage

labs_hourly	Blood urea nitrogen	7-20	kidney injury or disease/ heart failure
labs_hourly	Partial pressure o2	75-100	emphysema. chronic obstructive pulmonary disease
labs_hourly	bilirubin	0.1-1	liver damage or disease
labs_hourly	alanine transaminase	7-56	liver damage or diseases
labs_hourly	aspartate transaminase	5-40	liver damage/ other organ damage
labs_hourly	Hemoglobin	116-166 (regardless of gender)	a blood disorder
labs_hourly	hematocrit	35.5-48.6 (regardless of gender)	anemia/ loss of blood
labs_hourly	White blood cell	3.4-9.6	leukaemia/ bacterial infection/ cancer
labs_hourly	platelets	135-371	Thrombocytosis
labs_hourly	sodium	135-145	hypernatremia
labs_hourly	chloride	95-110	hyperchloremia
labs_hourly	creatinine	0.6-1.3	a sign of kidney disease

labs_hourly	glucose	Under 200	hyperglycemia
labs_hourly	neutrophils	45-75	neutrophilic leukocytosis (neutrophilia)
labs_hourly	C reactive protein	Under 10	a marker of inflammation
labs_hourly	lactate	Under 1.0	heart failure, a severe infection (sepsis), or shock
labs_hourly	International normalised ratio	2-3	blood clotting problem
vital_hourly	Diastolic blood pressure	Normal: Under 80 High BP Stage 1: 80-89 High BP Stage 2: 90-119 Hypertension Crisis: Over 120	hypertension
vital_hourly	Systolic blood pressure	Normal: Under 120 Elevated: 120-129 High BP Stage 1: 130-139 High BP Stage 2: 140-180 Hypertension Crisis:	Hypertension

		Over 180	
vital_hourly	SpO2	>=95 <95: high risk of hypoxemia	hypoxaemia
vital_hourly	temperature	Under 38C > 38C: fever	fever
vital_hourly	Heart rate	<60: bradycardia 60-100 >100: tachycardia	bradycardia/tachycardia
vital_hourly	glucose	Under 200	hyperglycemia
vital_hourly	Mean arterial pressure	60-100	heart attack/ blood clots
gcs_hourly	Glasgow Coma Scale Scoring	Mild: 1-8 Moderate: 9-12 Severe: 12-15	severe injury
gcs_hourly	GCS Eyes	No response: 1	severe injury
gcs_hourly	GCS Motor	No response: 1	severe injury
gcs_hourly	GCS Verbal	No	severe injury

		response: 1	