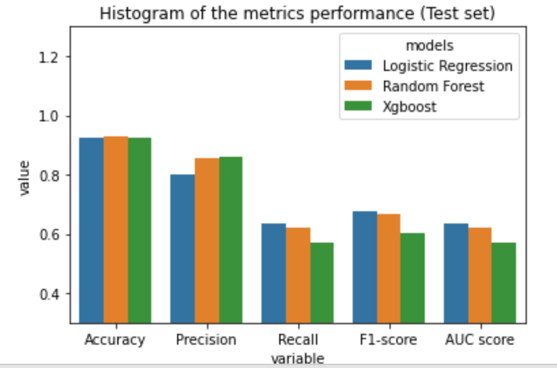
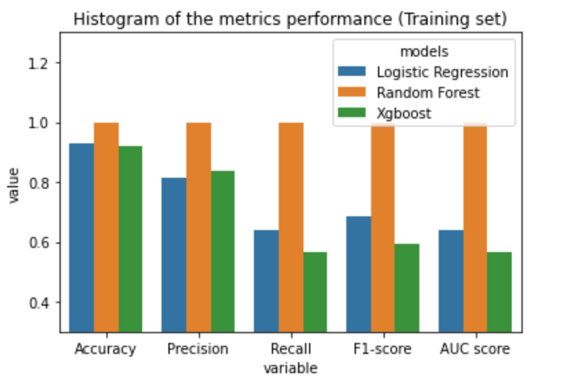
Part 1. Mortality Prediction in the ICU

(f).

By looking at the reports from (c)-(d) and histograms from (e):

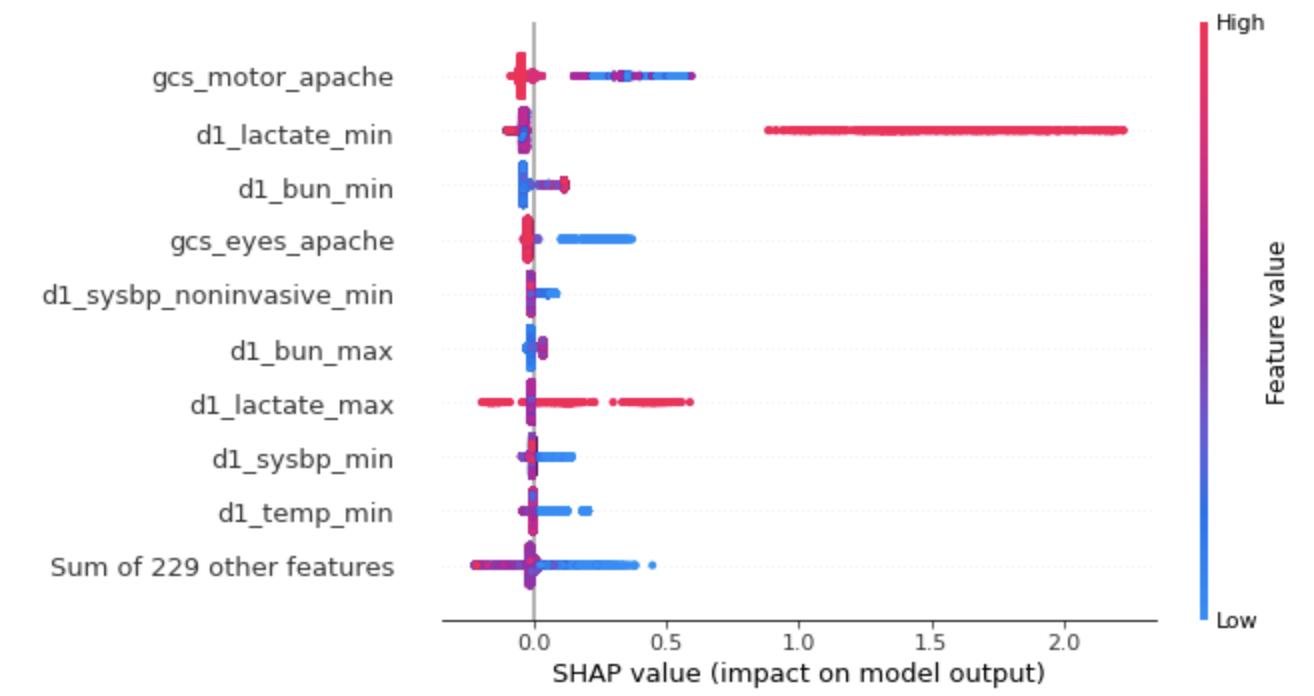


We can conclude that logistic regression model has the best performance among the test set.

For models predicting the mortality rate in the test set, although the logistic regression model has the lowest precision among all 3 models, it have the highest recall, F1-score and AUC score. Thus, we can conclude that logistic regression model has the best performance.

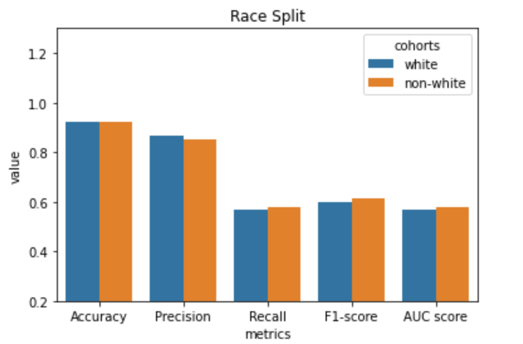
(g).

To get a high-level overview of which features contribute the most to our models’ pre- dictions, we made a beeswarm plot:

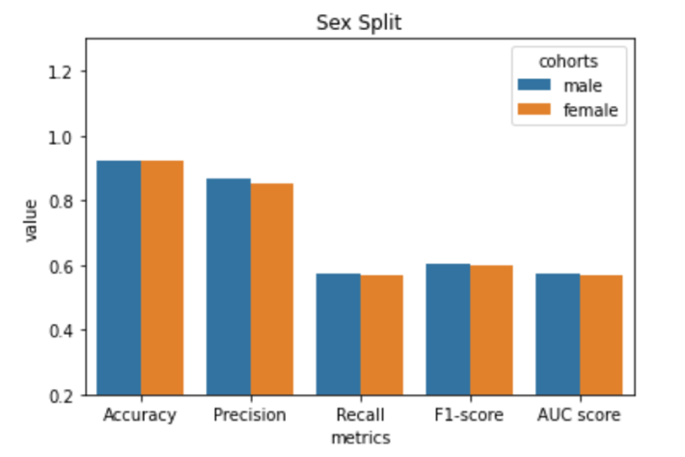


From the beeswarm plot, feature “gcs\_motor\_apache” contributes the most to the model’s predictions. This feature stands for the motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score. It is a reasonable feature that the model can rely on: generally, lower GCS scores are correlated with higher risk of death. In this beewarm SHAP plot, a high “gcs\_motor\_apache” (motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score) lowers the predicted death probability.

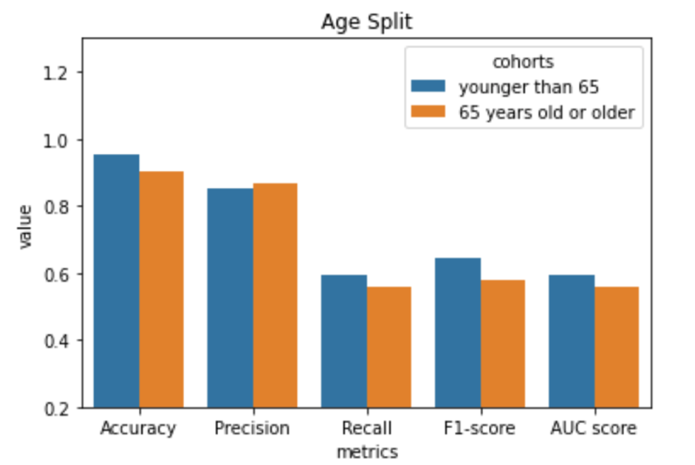
(j).



For the white/non-white split, the xgboost model perform better in the non-white split, for which it has higher Recall, F1-score and AUC score.



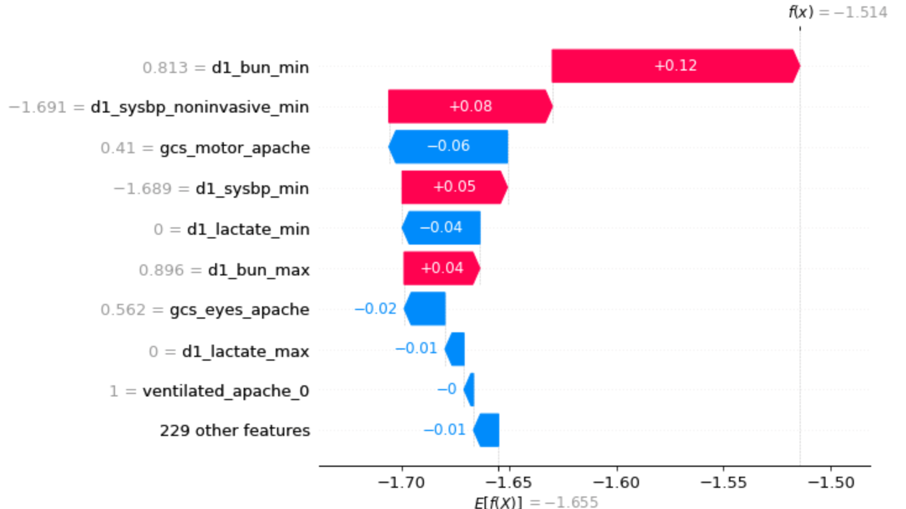
For the age split, the xgboost model perform better in the younger than 65 split, for which it has higher Accuracy, Recall, F1-score and AUC score.



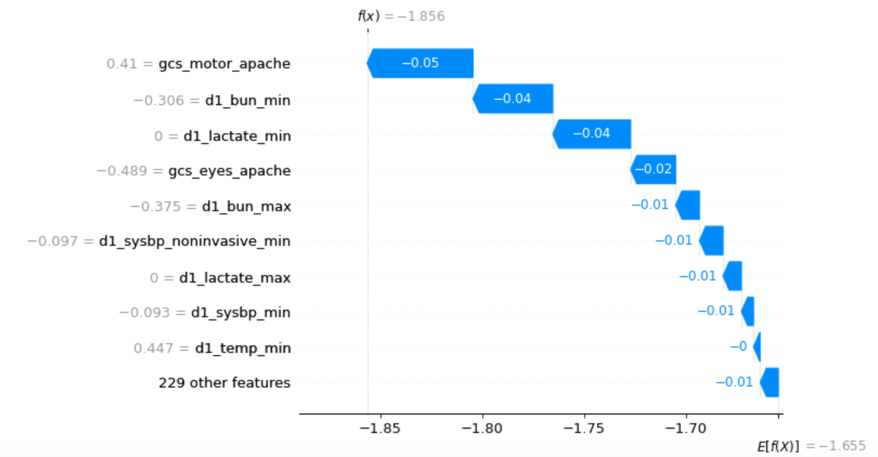
(k).

- Race

white patients:

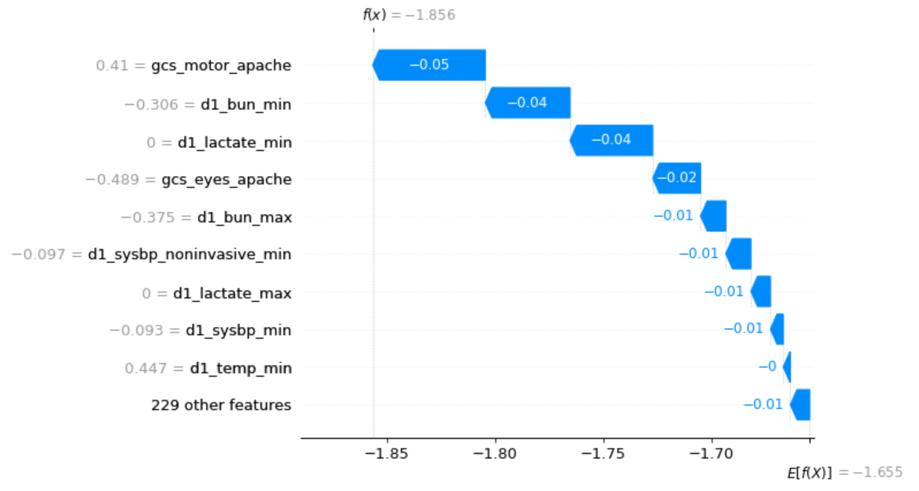


non-white patients:

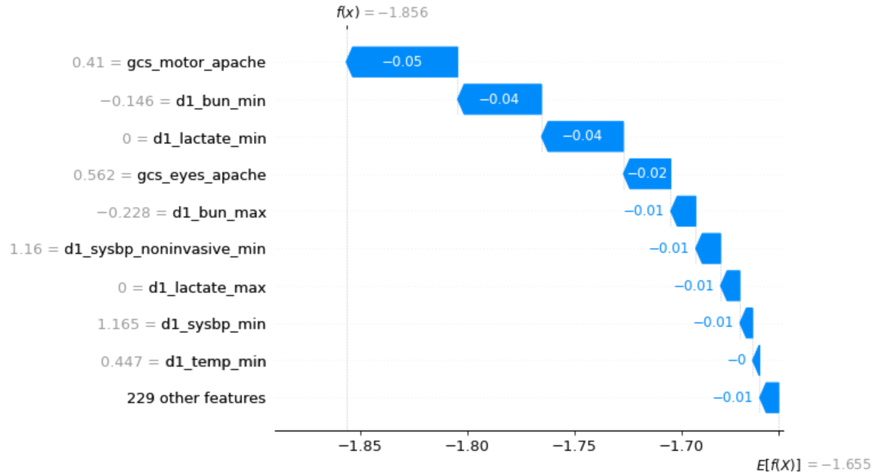


- Sex

male patients:

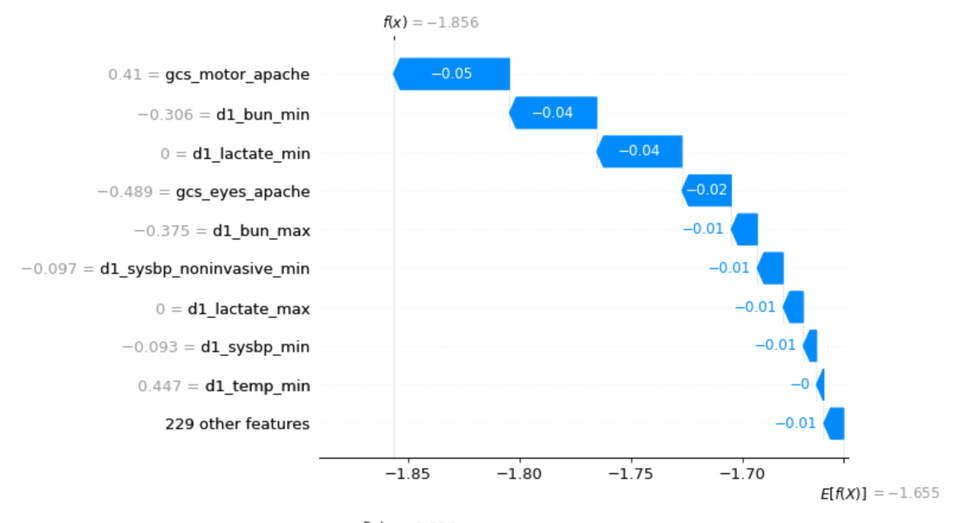


female patients:

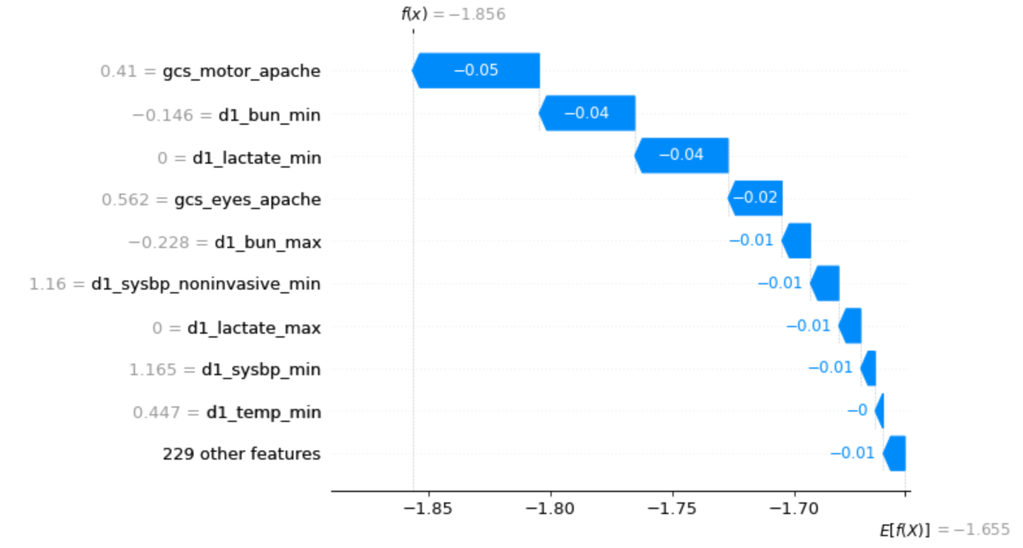


- Age

patients younger than 65:



patients 65 years old or older:



(l).

I’ve noticed discrepancies in the features used by the xgboost model to make predictions for the white/non-white split. Among the white patients,  the importance order of contributing features (from high to low) is: d1\_bun\_min > d1\_sysbp\_noninvasive\_min > gcs\_motor\_apache > d1\_ sysbp\_min > d1\_lactate\_min > d1\_bun\_max > gcs\_eyes\_apache > d1\_lactate\_max > ventilated\_apache\_0 > (other 229 features). Among the non-white patients,  the importance order of contributing features (from high to low) is: gcs\_motor\_apache > d1\_bun\_min > d1\_lactate\_min > gcs\_eyes\_apache > d1\_bun\_max > d1\_sysbp\_noninvasive\_min > d1\_lactate\_max > d1\_ sysbp\_min > di\_temp\_min > (other 229 features). Feature “d1\_bun\_min” (the lowest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay) contributes to push the model output from the base value (the average model output over the training dataset we passed) to the model prediction output higher in white patients, while it pushes the prediction lower in non-white patients. Similarly, features “d1\_sysbp\_noninvasive\_min” (patient's lowest systolic blood pressure during the first 24 hours of their unit stay, non-invasively measured), “d1\_ sysbp\_min” (patient's lowest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured) and “d1\_bun\_max” (the highest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay) all contribute to push the model output from the base value to the model prediction output higher in white patients, while they push the prediction lower in non-white patients.

Such discrepancies were not observed among male/female split and patients younger/older than 65 split, which may due to there’s not too much imbalance among these splits.

Part 2. Delving into Disparities

(a).

Seeing from the 6 pie plots, we do not see too much differences between the train and test cohorts. For the ethnicity, 78% of the patients are Caucasian, which is a huge imbalance. For the gender, 54% among the patients are male while 46% among the patients are female. For the age, younger patients are slighter than the elder, although it make sense for the normal age distribution in the population.

(c).

Reducing the number of female patient datapoints does not affect the performance of the model. This may because our female and male patients are similar in most of the features’ distributions.

(d).

I would expect to see the similar results when varying the degree of missingness of some populations in patients younger/older than 65 split. Because from 1(k) and 1(l), we’ve notice that discrepancies of features’ importance for prediction were not observed among this split.

However, I would expect to see the different results when varying the degree of missingness of some populations in patients white/non-white split. Because from 1(k) and 1(l), we’ve observed obvious discrepancies of features’ importance for prediction among this split.

(e).

Increasing the degree of missingness of test results in some training datapoints would lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain less error.

(f).

A potential way of handling missingness of test results is to prevent the missingness by well-planning the study and collecting the test data carefully. To be more specific, before the beginning of the clinical research, form a detailed documentation of the study should be developed in the form of the manual of operations, which includes the tests to screen the participants, protocol to train the nurses, methods to communicate between the research investigators and doctors/nurses, implementation of the treatment, and procedure to collect, enter, and edit data.