Predictions in Hospital Mortality Due to Heart Failure

BAIS Baddies

Contributors: Natalie McCormick, Brittany Fleury, Lauren Jamison, Sophie Selk

1. Executive Summary

Cardiac arrest is the third highest cause of death in America. Hospitals' ability to prioritize patients suffering from cardiac arrest based on the most important features in fatal cases will decrease in-hospital mortality rates. Our dataset is an extract of demographic characteristics, vital signs, and laboratory values data from a 24-hour period of patients admitted to the ICU for heart failure. We targeted the mortality outcomes of patients to predict the most important features and measurements in fatal heart failure cases.

After data preparation and exploration, we completed a Random Forest model that indicated the top 15 features that would be the best for our models. After tuning and running 7 different models, we identified the models that would be the easiest to interpret and had a high AUC. Logistic Regression provided the most accurate and interpretable model with an AUC of 0.9315. The logistic regression described the relationship between each feature and the outcome.

Based on these results, we recommend that doctors implement risk-adjusted mortality. Risk-adjusted mortality can be defined as the ability to prioritize patients based on characteristics that predict the risk of death. This can be implemented by increasing attention to patients with the following factors: occurrence of Atrial Fibrillation, high anion gap levels, high heart rate, increased respiratory rate, high BMI, aged 70-90, high blood pressure (hypertensive), and diabetic patients. By implementing this practice, healthcare providers can decrease in-hospital mortality rates by monitoring the most influential factors in fatal heart failure cases.

2. Problem Description

2.1 Background

According to the National Academies of Sciences, Engineering, and Medicine, cardiac arrest is the third leading cause of death in the U.S. behind cancer and heart disease. Approximately 200,000 U.S. citizens experience cardiac arrest in a hospital each year and on average, only 24% of these patients survive (National Academies, 2015). With a dramatically low survival rate, the need for an increased quality of care and cardiac arrest symptom identification within hospitals is high.

2.2 Business Goal and Data Mining Goal

The business goal of our model is to predict the mortality rate of patients who enter the ICU (Intensive Care Units) due to heart failure. This model will be used to provide doctors with information about what patients they need to prioritize in order to reduce patient-in-house mortality rates. This model will be presented to hospital owners and doctors.

The data mining goal for our data set is to determine what features are most important in determining a patient's chance of in-house mortality. Our prediction target is to determine the

Outcome (0 = Alive, 1 = Death) of a patient. Since our target variable is categoric, this means we are dealing with a classification problem.

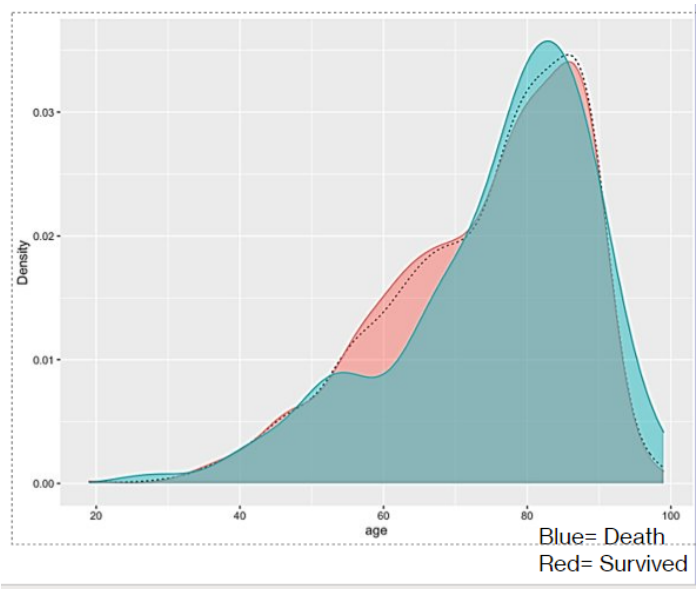
3. Data Description

3.1 Data

This dataset is an extraction of demographic characteristics, vital signs, and laboratory values data from the following tables in the MIMIC III dataset: ADMISSIONS, PATIENTS, ICUSTAYS, D*ICD DIAGNOSIS, DIAGNOSIS*ICD, LABEVENTS, D*LABIEVENTS, CHARTEVENTS, D*ITEMS, NOTEEVENTS, and OUTPUTEVENTS. There are 50 features and 1,176 instances. For our analysis, we utilized feature selection and selected the following 17 features: ID, Outcome, Age, Gender, BMI, Hypertensive, Atrial Fibrillation, Diabetes, Hyperlipemia, Renal Failure, Heart Rate, Respiratory Rate, Urine Output, Platelets, Glucose, Anion Gap, and Bicarbonate. Our target variable is Outcome; whether a patient lives or dies while at the ICU for heart failure.

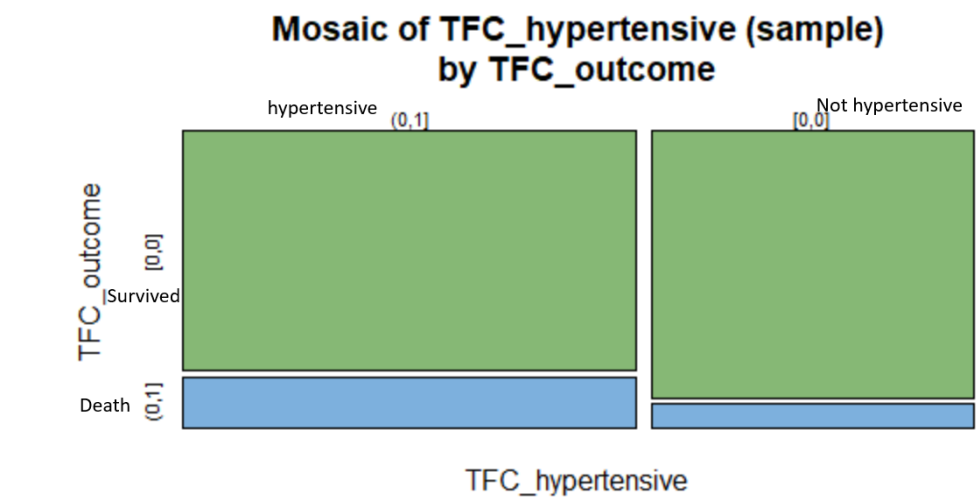
3.2 **Exploratory Analysis**

Age and Morality:



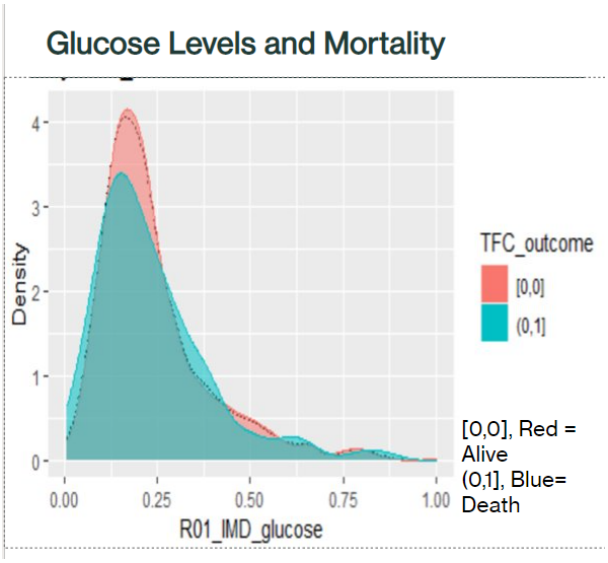
Age with respect to the outcome, this mosaic shows that 70-to-90-year-olds would be more likely to be in the ICU for heart failure and have a higher risk of death.

Hypertension and Mortality:



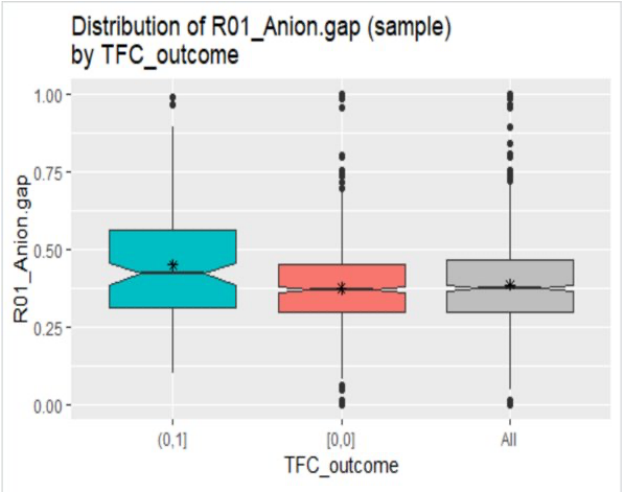
This mosaic shows that hypertensive patients are more likely to die from heart failure. Patients that are not hypertensive then have a lower risk of mortality, therefore, those that are hypertensive with heart failure that enter the ICU should be prioritized.

Glucose levels and mortality:



When interpreting glucose levels and mortality, we can see that most patients with lower glucose levels have lower chances of death.

Anion gap and mortality:



If the patient had a higher anion gap, there was a higher chance of death. If a patient has a higher anion gap, it is a sign of a disorder in your lungs, kidneys, or other organ systems.

3.3 Data Pre-Processing

First, we created a random forest with all 50 features and based on the importance of each feature, we selected the 15 most important features. These features are listed above in the data dictionary. We then decided to look at the features with the missing variables. We re-coded the numeric features into mean and the categorical features into mode. We then rescaled the numeric features 0 to 1 and categorical features as indicator variables. Finally, we deleted the observations missing.

The following are the features that were ignored based on very low importance when running random forest: atrial fibrillation (yes/no), CHD with no MI (yes/no), depression (yes/no), BMI, Systolic blood pressure, glucose.

4. Data Mining Solution

4.1 Models

For Random Forest, we first had to find the appropriate number of variables that resulted in a higher AUC with a high number of trees. Once we found the appropriate number of variables that returned a higher AUC, we kept that number of variables (at 7) and then tuned the number of trees. Both AUC results were found with the validation set. Once we found the appropriate number of trees with the number of variables at 7, we evaluated both results with the testing set and found that the higher AUC was found with the 400/7 random forest with an AUC of 0.9352.

For the decision tree model, we set the defaults of the min split and min bucket to 0, and the max depth to 30. The complexity of the model was set to 0.000 to find the lowest xerror. With the lowest xerror, we found the corresponding CP value (0.0265) to then use in the Complexity value to find the evaluation of an AUC of 0.8710 on the testing set.

For the boost models, we evaluated both ADA Boost and Gradient Boost models. For the Gradient Boost model, we tuned the model based on the max. depth amount. We evaluated multiple numbers for the max depth with validation set and found that the highest AUC was associated with 30 - with an AUC of 0.9051 in the testing set. For the ADA Boost model, we set the max depth to 30 and min split to 0. For tuning this model, we keep the number of trees at default of 50 and see the changed of complexity. The highest AUC was associated with a complexity of 0.010. With this complexity we then tuned the number of trees – finding the highest AUC at 30 trees. After evaluating these two results in the testing set, the highest AUC was 0.9264 with the number of trees at 30 and the complexity of 0.010.

For the logistic regression model, we found that Atrial Fibrillation, Anion Gap, Heart Rate, Respiratory Rate, BMI, Age, Hypertensive and Diabetes were all significant features in the linear regression with a positive coefficient. For this model, the AUC was 0.9315 on the testing set.

For the SVM models, we evaluated both the Radial Kernel model and the Polynomial model. For the SVM Radial Kernel model, we chose three different parameters to tune the model and ran this through three different seeds to find the best AUC at 0.9056 at C=0.1 with the random seed of 4324. For the Polynomial SVM model, the parameter took into account degree and complexity for tuning the model. We chose three different parameters for both the degree and complexity with three different random seeds and found the highest AUC (at 0.9202) to be with a degree of 1, complexity of 0.01, and random seed at 4324.

For the ANN model, we tuned the model by changing the number of hidden layer nodes and found that the highest AUC was at 0.7869 with 11 hidden layer nodes.

After using the elbow method for clustering, we found that the K-means number would be most appropriate at 2. For Cluster One we determined that this cluster consisted of mostly males who were more likely to be hypertensive, less likely to be diabetic, and less likely to have hyperlipemia and have renal failure. For Cluster Two we determined that this cluster had mostly females that were less likely to be hypertensive, more likely to be diabetic, and more likely to have hyperlipemia and have renal failure.

After running all of the models, we decided to use logistic regression because it was the easiest to interpret and had a high AUC.

4.2 **Performance Evaluation**

We used AUC to evaluate our models. Our top 3 AUCs were from random forest (0.9352), logistic regression (0.9315), and gradient boost (0.9280). However, we used our 2nd highest AUC

model (Logistic Regression) because it was easier to interpret and was still very accurate. Also, the random forest model is not interpretable.

5 Conclusion

5.1 Recommendations

Insights obtained from 3&4: Based on the logistic regression, we found that the variables with the positive intercepts had a relationship with the mortality outcome. We looked at the strongest positive intercepts to rank the most influential factors that influence mortality. Looking at the exploratory analysis, we indicated the relationship certain variables and outcome. We compared outcome and gender, glucose levels, age and anion gap.

Based on the insights gained from above, we recommend that doctors implement risk-adjusted mortality by increasing attention to patients with the following factors: occurrence of Atrial Fibrillation, high anion gap levels, high heart rate, increased respiratory rate, high BMI, Aged 70-90, high blood pressure (hypertensive), diabetic. By implementing this practice, healthcare providers can decrease in-hospital mortality rates by monitoring the most influential factors in fatal heart failure cases.

5.2 Limitations

The limitations of the data set are the data was collected in only through a 24-hour long collection period. This means that we are missing both important data of those who may have passed after the time frame data was collected. This experiment is also missing important data that affects a person's health like background, family medical history and lifestyle. This type of data is found in medical documents even in cases in the emergency rooms.

Appendix

Data Dictionary

Feature	Description
ID	Patient Identifier
Outcome	Whether patient lives or dies while in ICU(0 = Alive, 1 = Death)
Age	Age of patient
Gender	Gender of patient
BMI	Body Mass index of patient (<18.5 = underweight, 18.5-24.9 = normal, 25-29.9 = overweight, 30-34.9 = obese, >35 =extremely obese)
Hypertensive	Whether the patient has high blood pressure (0= normal blood pressure, 1= high blood pressure)

Atrial fibrillation	Whether the patient has Atrial Fibrillation (irregular and often fast heartbeat) (0 = no, 1= yes)
Diabetes	Whether the patient has Diabetes (0 = no, 1= yes)
Hyperlipemia	Patient has Hyperlipemia (A condition in which there are high levels of fat particles (lipids) in the blood.) (0 = no, 1 = yes)
Renal Failure	Whether the patient experiences kidney failure (0 = no, 1= yes)
Heart Rate	Heart rate of patient (>90 is considered high)
Respiratory Rate	Respiratory rate of patient (A respiration rate under 12 or over 25 breaths per minute while resting is considered abnormal)
Urine Output	Urine Output of patient (The normal range of urine output is 800 to 2,000 milliliters per day)
Platelets	Platelet levels of patient (A normal platelet count range is 140 to 400 K/uL)
Glucose	Blood glucose level of patient (Your blood glucose level should be no higher than 99 mg/dL)
Anion Gap	The anion gap is a measurement of the difference-or gap-between the negatively charged and positively charged electrolytes. (Normal results are 3 to 10 mEq/L)
Bicarbonate	Bicarbonate level in patient (Normal bicarbonate levels are: 23 to 30 mEq/L in adults.)

Random forest:

Number of Trees	**Number of Variables**	**AUC**
400	3	0.8917
400	4	0.8976
400	5	0.8972
400	7	0.8985
40	7	0.8883
30	7	0.8956

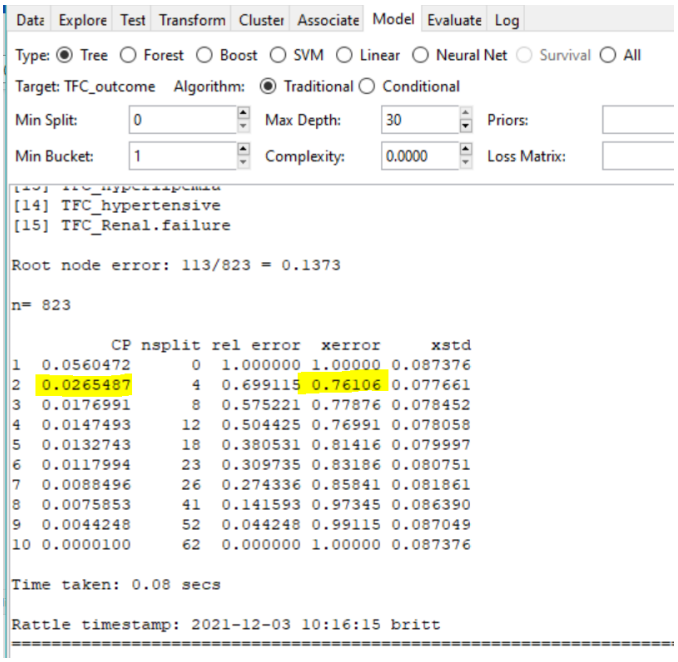
25	7	0.8995

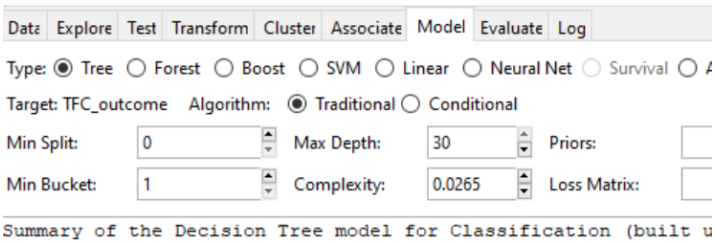
*Validation Set

Number of Trees	Number of Variables	AUC
400	7	0.9352
25	7	0.9317

*Testing Set

Decision trees:





AUC= 0.8710 on testing set

Gradient boost:

Max. Depth	AUC
30	0.9051
10	0.9021
9	0.8977
8	0.8959
7	0.8913

6	0.8885

*Validation

Max. Depth	AUC
30	0.9280

*Test

ADA boost:

Number of Trees	Complexity	AUC
50	0.000	0.9097
50	0.001	0.9067
50	0.005	0.9072
50	0.010	0.9123
40	0.010	0.9115
30	0.010	0.9144

*Validation Set

Number of Trees	Complexity	AUC
50	0.010	0.9245
30	0.010	0.9264

*Testing set

Logistic regression:

```
Coefficients:
                             Estimate Std. Error
(Intercept)                -23.61499  803.33400
TFC_gender(1,2]             -0.32522    0.26674
TFC_hypertensive(0,1]        0.89988    0.29649
TFC_atrialfibrillation(0,1] 19.61208  803.33279
TFC_diabetes(0,1]            0.13893    0.27819
TFC_Hyperlipemia(0,1]       -0.04524    0.27025
TFC_Renal.failure(0,1]      -0.04445    0.27842
R01_age                      1.10779    0.87425
R01_Platelets               -3.52567    1.42079
R01_Anion.gap                3.07078    1.21316
R01_Bicarbonate             -1.95177    1.16022
R01_IMD_heart.rate           2.25119    0.91381
R01_IMD_Respiratory.rate     1.99165    1.01762
R01_IMD_Urine.output        -3.26022    1.19276
R01_IMD_glucose             -0.51234    0.90125
R01_IMD_TNM_BMI              1.77399    0.51467
                            z value Pr(>|z|)
(Intercept)                  -0.029 0.976549
TFC_gender(1,2]              -1.219 0.222752
TFC_hypertensive(0,1]         3.035 0.002404 **
TFC_atrialfibrillation(0,1]   0.024 0.980523
```

AUC= 0.9315 on testing set

SVM: Radial Kernel

Parameters	Random seed=42	Random seed=100	Random seed=4324	Average	AUC
C= 0.1	14.8%	17%	13.6%	15.1%	0.9056
C=0.01	14.8%	17%	13.6%	15.1%	0.9012
C=0.001	14.8%	17%	13.6%	15.1%	0.9038

AUC= 0.9056 using random seed 4324 and C=0.1

SVM: Polynomial

Parameters	Random seed=42	Random seed=100	Random seed=4324	AUC
D=3 C= 0.1	19.9%	14.2%	14.2%	0.7824
D= 1 C=0.01	14.8%	17%	13.6%	0.9202
D=2 C=0.001	13.7%	14.7%	11.9%	0.8974

AUC= 0.9202 using random seed 4324 and D=1 and C=0.01

ANN:

Validation:

# of Hidden Layer Nodes	Error	AUC
10	19.3%	0.6562
9	14.3%	0.7131
11	19.9%	0.7869
12	17.1%	0.7309

Testing set: AUC with 11 hidden layers= 0.7779

Sources:

"Cardiac Survival Rates Around 6 Percent for Those Occurring Outside of a Hospital, Says IOM Report." *Nationalacademies.org*, The National Academies of Sciences Engineering

Medicine, 30 June 2015, https://www.nationalacademies.org/news/2015/06/cardiac-survival-rates-around-6-percent-for-those-occurring-outside-of-a-hospital-says-iom-report.

2021 Heart Disease and Stroke Statistics Update Fact Sheet ... https://www.heart.org/-/media/phd-files-2/science-news/2/2021-heart-and-stroke-stat-update/2021_heart_disease_and_stroke_statistics_update_fact_sheet_at_a_glance.pdf?la=en.

Centers for Disease Control and Prevention. (2021, September 27). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved December 3, 2021, from https://www.cdc.gov/heartdisease/facts.htm.

Goodacre, S., Campbell, M., & Carter, A. (2015, March 1). *What do hospital mortality rates tell us about quality of care?* Emergency Medicine Journal. Retrieved December 3, 2021, from https://emj.bmj.com/content/32/3/244.