In-Hospital Mortality Prediction for ICU Patients

Final Report: CKME 136 Capstone Project

Prepared by: Linda Kaleis

November 25th, 2016

In-Hospital Mortality Prediction for ICU Patients

Abstract

In recent time, electronic medical record (EMR) data has allowed for predictive analyses of patient outcomes, demonstrating the potential support in clinical decision-making. By having access to information such as physiological measurements, medication history, demographics, and diagnoses, the health factors that best predict mortality outcome can be identified. This study evaluates EMRs of patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center which are housed in an open-access critical care database, known as the Medical Information Mart for Intensive Care (MIMIC-III). After selecting relevant baseline features from the dataset, prediction models are constructed using logistic regression and support vector machine (SVM) algorithms after being trained on the patient population. The aim is to compare performance between the models that only use baseline and demographic data with those that are trained on severity score and demographic data, with no baseline features considered. The evaluation consists of assessing their performance in predicting the binary outcome variable for ICU patients – in-hospital mortality.

Introduction

An EMR contains important medical information pertaining to patient demographics, drug prescriptions, procedure and intervention history, as well as dynamic physiological data. Leveraging these known variables from patient records allows for researchers to identify the health factors that best predict mortality and thus construct appropriate machine learning models. In order to improve patient outcomes, care quality, and reduce healthcare expenditures, in-hospital mortality rates should be analyzed and minimized. It is possible to use different binary classifiers to build predictive models to address this problem. Overall, integrating a quality metric for mortality risk into primary care facilities will enable physicians to deliver timely and quality care to patients in the long-run,

Literature Review

In an aging population that is experiencing more and more comorbidities, health expenditures continue to skyrocket. In particular, it has been shown that about 22% of costs in hospital settings are due to patient stays in the intensive care units (ICU) (Pirracchio et al., 2014). One main reason for the ICU's overwhelming share of costs is in-hospital mortality, which has been hovering around an average of 11-12% in recent years in the United States. As a result, machine learning algorithms are being tested and validated to try to best capture mortality prediction in current ICU patients. Although scoring mechanisms such as the SAPS II, SOFA, and APACHE II scores exist, there is still a gross overestimation of mortality (Pirracchio et al., 2014).

Consequently, having a high prediction performance for mortality enables clinical decision makers to allocate resources appropriately and prepare for new treatment plans. More often than not, existing methods of mortality prediction use a variety of classification algorithms on individual patient cases. Several studies have shown that exploring different binary classifiers for regression analysis is needed to find the best predictive model with high performance (Rouzbahman et al, 2016; Rouzbahman & Chignell, 2014).

1. Dataset

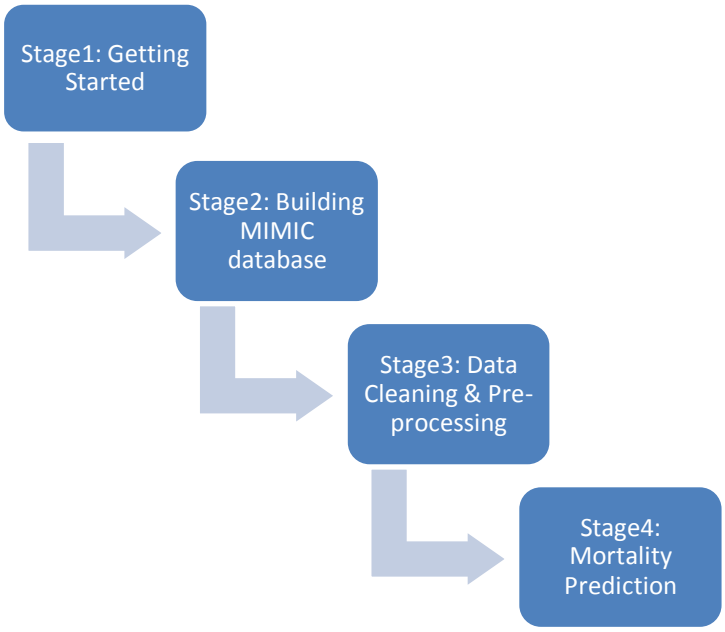
The Medical Information Mart for Intensive Care (MIMIC-III) dataset is publicly accessible and contains over 40,000 individual critical care patient records dating back from 2001 to 2012. Specifically, the patient data is from the Beth Israel Deaconess Medical Centre in Boston, Massachusetts. The de-identified records features demographic information, temporally-resolved physiological data, clinical measurements, medication history, diagnostic codes, unstructured nurse notes, diagnostic imaging, and lab results. The data can further be classified into static (ie. death date) and dynamic (ie. heart rate) data. For simplicity, this analyses will not be focused on high temporal resolution events such as blood pressure changes over time. Instead, the variables of interest pertain to demographic data (birth date, ethnicity, gender, weight, etc), high level ICU stay data (diagnosis, length of stay, admission type, number of procedure events, medication), and scoring metrics. For example, the simplified acute physiology score (SAPS) will be calculated to show patient severity of illness whereas the sequential organ failure assessment (SOFA) will

be used to measure organ failure. The code for calculating severity scores can be found in the MIMIC-III code repository here: https://github.com/MIT-LCP/mimic-code/tree/master/severityscores.

To access this database, a PhysioNetWorks account needed to be created and a Collaborative Institutional Training Initiative (CITI) program had to be completed to ensure compliance with working with sensitive data and confidentiality.

Tables	Description of Attributes
ICU Stay Detail	This includes: hospital identifiers, length of ICU stay in days, admission type, hospital expire flag (live vs die), and demographics such as age, gender, and ethnicity.
Height & Weight	This includes: the first maximum height (cm) and weight (kg) recorded for each patient.
Vitals First Day	The vital measurements include: heart rate (bpm), systolic blood pressure (BP), diastolic BP, mean BP, respiration rate, temperature (Celsius), spo2, and glucose.
Ventilation First Day	This includes one binary variable to show if patient was ventilated on the first day of ICU stay.
Urine Output	This included one measurement of total urine output for first day.
Severity Scores	The constructed severity scores are: 1) OASIS – Oxford Acute Severity of Illness Score 2) SAPS – Simplified Acute Physiology Score 3) SOFA – Sepsis-related Organ Failure Assessment Score All of these scores were calculated using baseline measurements from a patient's ICU stay.
Labs First Day	This includes lab values taken in first 24hr such as: albumin, bicarbonate, bilirubin, creatinine, chloride, hematocrit, hemoglobin, lactate, platelet, potassium, sodium, blood urea nitrogen (BUN), and white blood cell count (WBC).
Glasgow Coma Scale	This includes one score value for the Glasgow Coma Scale (GCS).

2. Approach Overview & Detailed Steps



Stage1: Getting Started

1. Complete required training course

-complete CITI "Data of Specimens Only Research" course ~1.5/2 hours

@ *https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0*

2. Request access to MIMIC-III

-create PhysioNet account

-apply with CITI training certificate & project outline

-confirmation given by research superviser

3. Download MIMIC-III Database

-download CSV files

-unzip them to flash drive

4. Download & Install PostgreSQL on Windows

-create account

-set up filesystem permissions

-learn how to use PgAdmin tool

-create mimic database

Stage2: Building MIMIC Database in Postgres

1. Create tables, load data, create indexes and constraints

-(1) create 26 tables with defined attributes and types

-(2) load the data into tables

-(3) create indexes for easier ability to work with data

-(4) create constraints

Scripts: (1) https://github.com/MIT-LCP/mimic-code/blob/master/buildmimic/postgres/postgres_create_tables.sql

(2) https://github.com/MIT-LCP/mimic-code/blob/master/buildmimic/postgres/postgres_load_data.sql

(3) https://github.com/MIT-LCP/mimic-code/blob/master/buildmimic/postgres/postgres_add_indexes.sql

(4) https://github.com/MIT-LCP/mimic-code/blob/master/buildmimic/postgres/postgres_add_constraints.sql

---

3. Subset ICU patients

-(5) create materialized icustay_detail

-(6) copy table to CSV file and import into R

-(7) extract patients over age of 18 and who are first time in ICU & hospital

-(7) create new dataframe where patient IDs from (3) match original ICUSTAYS csv file patient IDs; this is a new version of the original ICUSTAYS csv file that was provided in MIMIC-iii

Scripts: (5) https://github.com/MIT-LCP/mimic-code/blob/master/demographics/postgres/icustay_detail.sql

(6)https://github.com/lkaleis/mimic_capstone/blob/master/psql_queries/stage2_subset_ICUpatients.psql

(7):https://github.com/lkaleis/mimic_capstone/blob/master/R_scripts/stage2_subsetICUpatients.R

---

4. Create materialized views:

-*urine output first day
-*ventilation first day
-*vitals first day
-*gcs first day
-*blood gas first day
-*ventilation duration
-*labs first day
-*blood gas first day arterial
-*echo data

Scripts: (8) https://github.com/MIT-LCP/mimic-code/tree/master/etc/firstday

---

5. Extract Severity Scores

-(9) SOFA: Sepsis-related Organ Failure Assessment

-(9) SAPS: Simplified Acute Physiology Score

-(9) OASIS: Oxford Acute Severity of Illness Score

-(10) create materialized view for all severity scores

Scripts: (9) https://github.com/MIT-LCP/mimic-code/tree/master/severityscores

(10) https://github.com/lkaleis/mimic_capstone/blob/master/severityscores_psql

---

Stage3: Data Cleaning & Pre-processing

1. Create materialized icustay view

-(1) extracts all features that will be predictors in model for all patients

-(1) exclude features that have null values in lab results and vital measurements

-(2) copy as csv file to import into R

Scripts: (1)
https://github.com/lkaleis/mimic_capstone/blob/master/psql_queries/final_ICUpatients.psql

(2) https://github.com/lkaleis/mimic_capstone/blob/master/psql_queries/export2csv.psql

2. Import csv file into R for cleaning

-(3) find all missing values

-(3) remove appropriate columns

-(3) omit incomplete records

Script: (3)
https://github.com/lkaleis/mimic_capstone/blob/master/R_scripts/stage3_clean_mimic.R

3. Visualize descriptive statistics

-(3) data summary, structure

-(3) histograms

-(3) correlation matrix

Stage4: Mortality Prediction

1. Feature selection

-*first split data into train and test sets (70:30 split)

-*convert categorical variables to factors

-*use Boruta feature selection only on training set

-*plot variable importance

-*remove rejected variables from train and test sets

Scripts: (1)
https://github.com/lkaleis/mimic_capstone/blob/master/R_scripts/stage4_boruta_featureselection.R

2. Train Logistic Regression & Support Vector Machine (SVM) Models

-there are two feature sets to be tested:

-(2) **first feature set:** tune, train one SVM and logistic regression classifier for patient baseline data

(using features that were left after selection from above)

-(2) **second feature set:** tune & train another SVM and logistic regression classifier using only patient demographic/administrative data, SOFA, and SAPS scores

Scripts: (2)
https://github.com/lkaleis/mimic_capstone/blob/master/R_scripts/models.R

3. Predict Mortality & Compare Performance

-(2) predict mortality outcomes on test set data for both feature sets using SVM

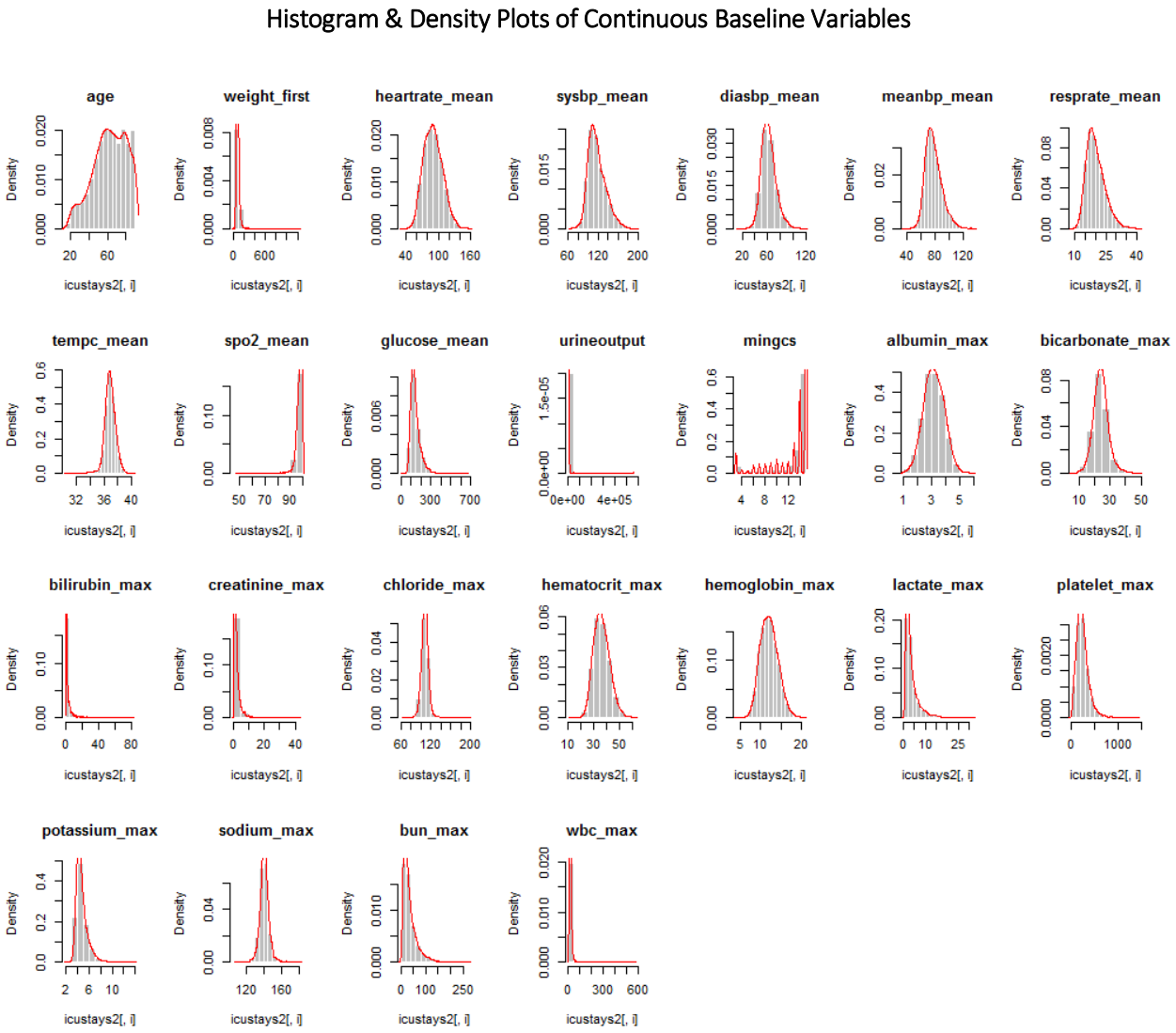
-(2) predict mortality outcomes on test set data fro both feature sets using logistic regression

-(2) create confusion matrixes and assess performance

3. Results

Data Visualization

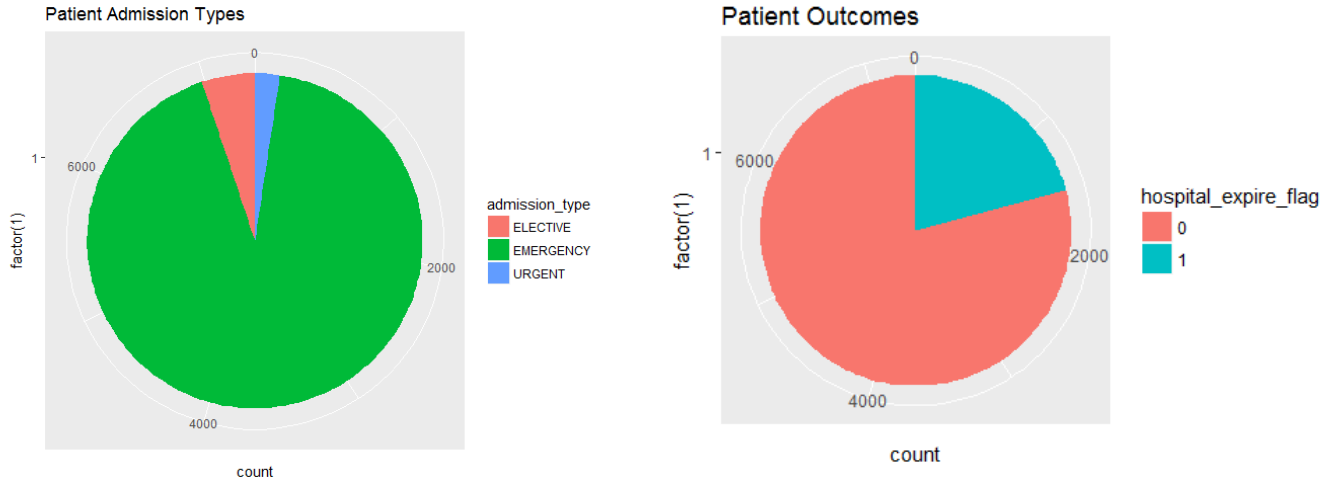
Data was first cleaned and pre-processed (Stage 3) so that appropriate visualizations could be made. The following shows histograms and density lines presented together for baseline variables from lab, vitals, and other physiological and physical measurements.



Most baseline features have normal distribution. In some cases, such as with weight, spo2, urine output, bilirubin, creatinine, lactate, BUN, and wbc, there are a few outliers leading to a skewed distribution. However, for this experiment, they were not removed due to lack of domain knowledge about the possibilities of these extremely high or low values. In the case of “mingcs”, it is normal to have most scores closer to 15 as this indicates the patient is conscious and not in a coma. This is the one variable that is

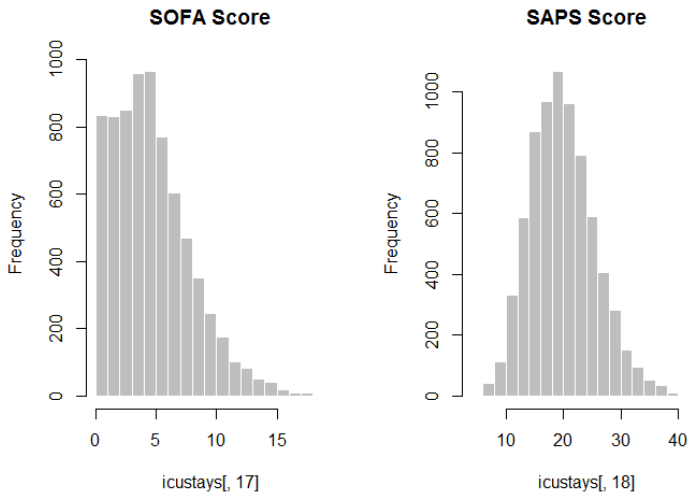
discrete here. In terms of age, the mean age of patients is 72.6 years; patients who were older than 90 had their true age masked and therefore their age was set to 90.

Below in the first pie chart, it is shown that 92.5% of patients (6802 out of 7354) are admitted as "emergency type" through the ER. 5.1% are admitted as elective and 2.4% are admitted as urgent. The second pie chart shows that 26.2% of patients in this cohort died in hospital (1528/5826).



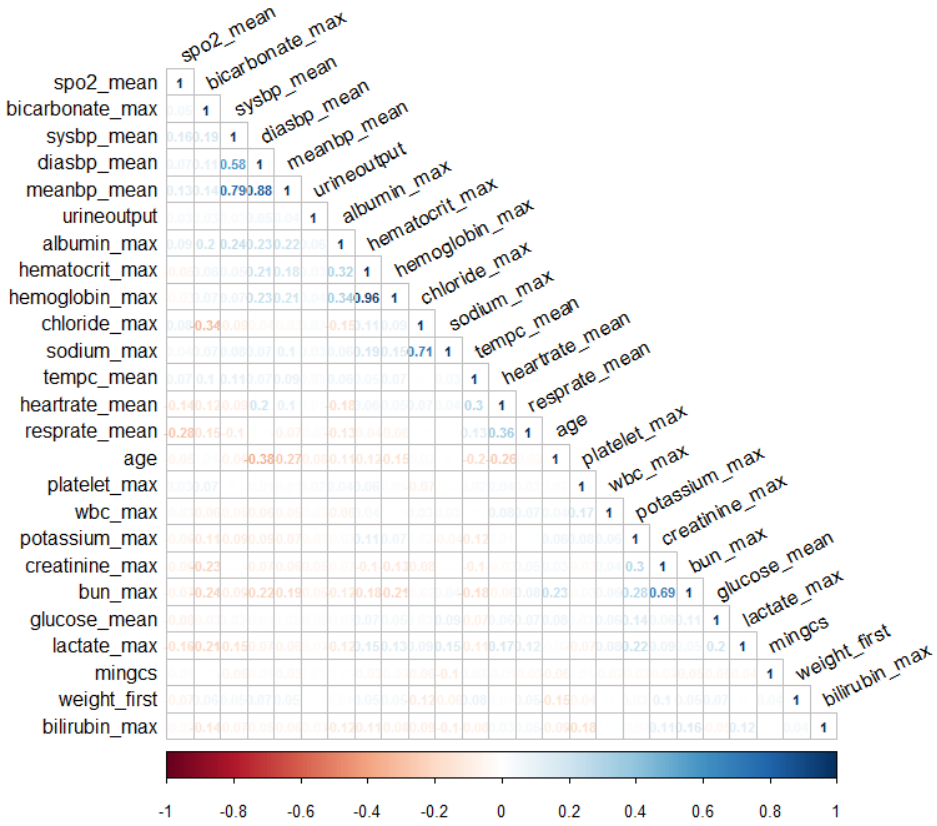
Histograms for Severity Scores

The Sepsis-related Organ Failure Assessment Score (SOFA) has a mean of 5.15 and a median of 5. Its calculation is based on the function of different organ systems (coagulation, nervous, hepatic, cardiovascular, renal, and respiratory) (Vincent & Moreno, 2010). The higher the SOFA score (out of 24), the worse the condition of these organ systems and a higher chance of patient death. The Simplified Acute Physiology Score (SAPS) has a mean of 20.35 and a median of 20. The SAPS score uses 14 variables for calculating risk of death for patients; it uses data collected from the first day in the ICU.

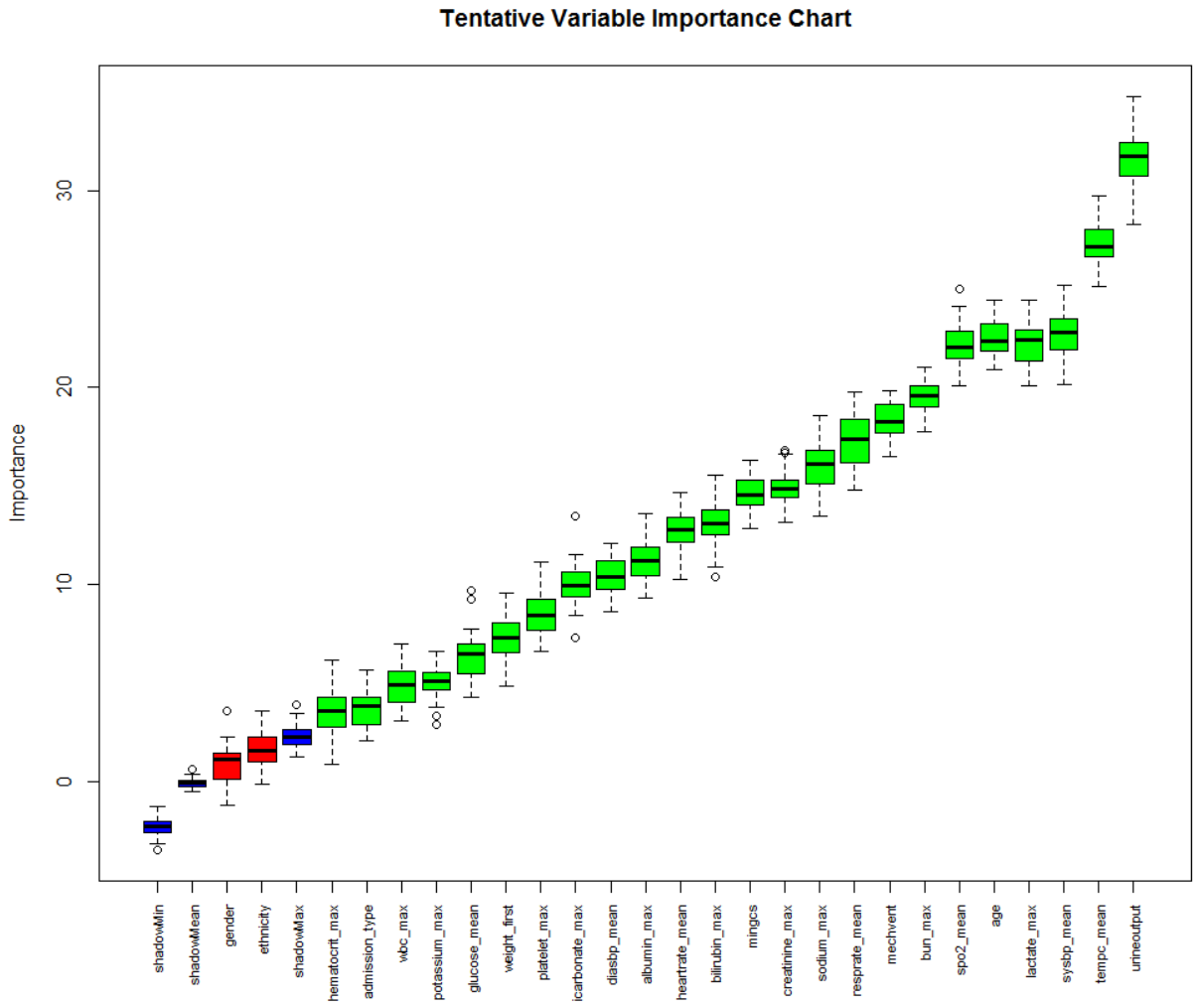


Feature Selections

A correlation matrix was constructed to show the correlation between numerical attributes (this did not include hospital expire flag variable, severity scores, or categorical attributes). A correlation coefficient cutoff of 0.70 was used to select features to remove from the training set. It was decided that meanbp_mean, hemoglobin_max, and chloride_max would be removed due to high correlations with: sysbp_mean and diaspbp_mean; hematocrit_max; and sodium_max, respectively.

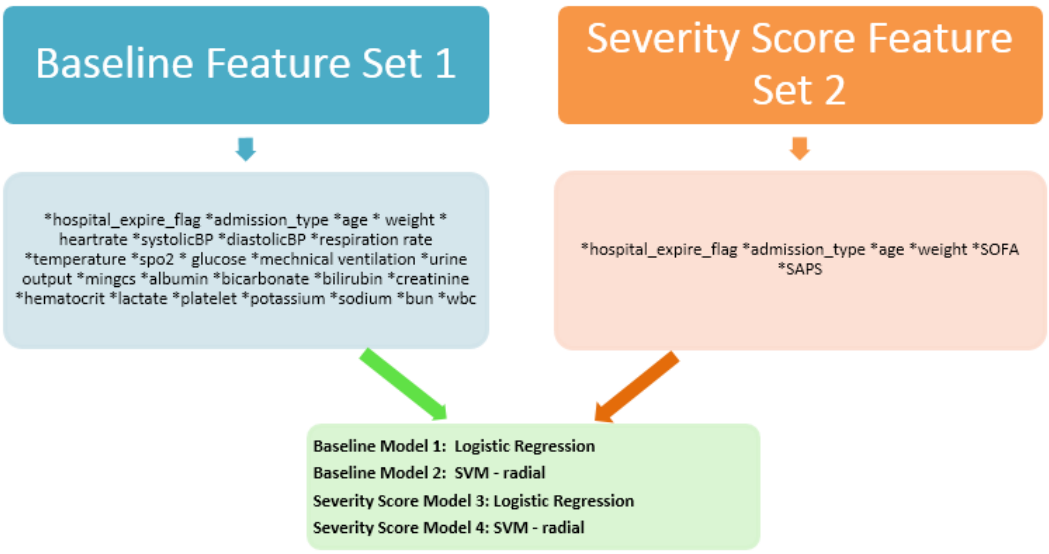


Further, the Boruta feature selection algorithm was used to judge whether each remaining feature is important. It is a wrapper algorithm over random forest method that is used on the data. The following features in the Tentative Variable Importance Chart below which are in green were confirmed to be important and included for future model building. Those in red were deemed unimportant and rejected (gender, ethnicity) and therefore removed from being included in the model.



Predicting Mortality - Models

Two binary classification algorithms were chosen to predict mortality outcome for ICU patients in their first 24 hours of stay – 1) logistic regression (glm - binomial) and 2) support vector machine (SVM - radial). These algorithms have been widely used and proven to have good performance in predicting in-hospital death. The data used for training and testing was split 70:30. Both predictive models were trained on two different feature sets chosen from the same data. The first feature set included demographic and baseline data whereas the second feature set only had demographic data and the two severity scores, SAPS and SOFA, with no baseline data included. This is because the severity scores are constructed using baseline data.



Confusion Matrices

```
model_pred_mortality    0    1
                   0 1680  294
                   1   72  161

               Accuracy : 0.8342
                 95% CI : (0.818, 0.8495)
    No Information Rate : 0.7938
    P-Value [Acc > NIR] : 9e-07

                  Kappa : 0.3817
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9589
            Specificity : 0.3538
         Pos Pred Value : 0.8511
         Neg Pred Value : 0.6910
             Prevalence : 0.7938
         Detection Rate : 0.7612
   Detection Prevalence : 0.8944
      Balanced Accuracy : 0.6564
```

Baseline Model 1 >

```
model_pred_mortality_ss    0    1
                      0 1674  327
                      1   78  128

               Accuracy : 0.8165
                 95% CI : (0.7997, 0.8324)
    No Information Rate : 0.7938
    P-Value [Acc > NIR] : 0.004193

                  Kappa : 0.2969
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9555
            Specificity : 0.2813
         Pos Pred Value : 0.8366
         Neg Pred Value : 0.6214
             Prevalence : 0.7938
         Detection Rate : 0.7584
   Detection Prevalence : 0.9067
      Balanced Accuracy : 0.6184
```

Severity Score Model 3

```
    true
pred    0    1
   0 1709  307
   1   43  148

               Accuracy : 0.8414     ***
                 95% CI : (0.8255, 0.8564)
    No Information Rate : 0.7938
    P-Value [Acc > NIR] : 7.198e-09

                  Kappa : 0.383
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9755
            Specificity : 0.3253
         Pos Pred Value : 0.8477
         Neg Pred Value : 0.7749
             Prevalence : 0.7938
         Detection Rate : 0.7744
   Detection Prevalence : 0.9135
      Balanced Accuracy : 0.6504
```

Baseline Model 2 >

```
    true
pred    0    1
   0 1713  365
   1   39   90

               Accuracy : 0.8169
                 95% CI : (0.8002, 0.8329)
    No Information Rate : 0.7938
    P-Value [Acc > NIR] : 0.003569

                  Kappa : 0.2389
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9777
            Specificity : 0.1978
         Pos Pred Value : 0.8244
         Neg Pred Value : 0.6977
             Prevalence : 0.7938
         Detection Rate : 0.7762
   Detection Prevalence : 0.9415
      Balanced Accuracy : 0.5878
```

Severity Score Model 4

Data splitting was used to partition the data into a 70:30 train and test set and the same sets were used to train all models. This was done to ensure that the performance can be compared easily and accurately at the end. Above, it is shown that when comparing logistic regression models (top two squares), the classification yielded a higher accuracy when all baseline features were used (Model 1, ACC = 0.8342). When comparing the bottom two models that used SVM, the same result was seen. The baseline model 2 performed better in terms of accuracy (ACC = 0.8414) than the SVM that used the severity score feature set. It is also noted that both models using SVM had lower false positives and higher sensitivity rates than logistic regression. Overall, baseline model 2 performed with the highest accuracy of all 4 models.

4. Conclusions

Researchers and clinicians alike are continuing to examine new ways of how to best make use of available patient data to predict critical information such as mortality outcome. This predictive health project demonstrated how using binary classifiers can help in this prediction. It has also revealed that by using all of the available baseline data about the patient rather than relying on traditional scoring systems is far better for having a higher accuracy in predicting whether the patient lives or dies.

Further research into finding the most accurate prediction model will aid clinicians in delivering high quality patient care, improving resource allocation, and in prioritizing the highest risk patients in the hospital.

References

Barros, R., Carvalho, A., Basgalupp, M., Quiles, M. (2011). A clustering-based decision tree induction algorithm. Intelligent Systems Design and Applications (ISDA).11th International Conference.

Karunarathna, M. (2016). Predicting ICU Death With Summarized Patient Data. University of Cambridge.

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): A population-based study. The Lancet Respiratory Medicine, 3(1), 42–52.

Rouzbahman, M., Chignell, M. (2014). Predicting ICU Death with Summarized Data: The Emerging Health Data Search Engine., KMD.

Rouzbahman, M., Jovicic, A., Chignell, M. (2016). Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU. IEEE Biomedical and and Health Informatics, PP(99):1–8.

Vincent, J. L., & Moreno, R. (2010). Clinical review: scoring systems in the critically ill. Critical care, 14(2),1.