Optimal Risk Scoring for ICU Admission using Random Forest Models Outperforms the National Early Warning Score Benchmark

Zi Peng Liu

Contents

Abstract

A risk score can help doctors predict whether or not a patient is likely to go to the ICU. To help the hospital operate more efficiently, it is important to find a risk scoring system that predicts well. In this study, a random forest model was built for each time periods to predict patient's risk of going to the ICU. The model uses the patient's status (went to ICU or not) as the outcome variable, and respiration rate, oxygen saturation, air, systolic blood pressure, pulse rate, alert, temperature, sex, and age as the predictor variables. It preforms better than the NEWS as it has better accuracy in the early time periods. The model prediction accuracy increases over time, but the variable importance score goes down over time for all variables due to the nature of health data. However, variable pulse, oxygen saturation, systolic blood pressure, and age are more important than other variables in the early time becausue they contribute more to the prediction. Lastly, the random forest model can act as a risk scoring system. Doctors can input patients data into the model, and the model will predict the risk score. If the score is above 0.5, then the patient is at risk of going to the ICU, and vice versa.

Introduction

Background Information

No matter where people go, hospitals are usually full of patients every day. There is clearly increasing demand for health care. However, hospitals have limited capacity, there are only a few doctors and nurses in each hospital, and they cannot keep track of every patient's health condition every day and every moment. How can hospitals allocate the resources as efficiently as possible and also make sure that the doctors and nurses will be there when the patient needs them, for example, when they need to go to ICU? Hospitals need an scoring system that can help doctors and nurses identify the patients who are at risk of going to the ICU. Some researchers have already been trying to create a scoring system like this. For example, the National Early Warning Score (NEWS) was developed by the Royal College of Physicians which is designed to detect patients whose symptoms are getting worse, especially those who have increased risk of cardiac arrest or death (Bilben, Grandal, Søvik, 2016). However, how well this system works is still unclear, researcher are interested in if there is another scoring system that does a better job.

Research Questions

The main research question in this study is to search for an "optimal" score that helps to identify subjects with an increased risk of going to the ICU. Other relevant questions are:

- Do all components of the score contribute equally to the score construction?
- Identify if a subset of components that do a better job at predicting the risk?
- Search an alternative way of defining the score?

Outline

An overview of the following section in order, first a data summary and cleaning will be presented to show what the each variable represents and what the cleaned data looks like. Then, in the method part, the iterative random forest model will be described, and the result section will relates the model and the research questions. Last section will be conclusion, limitation and following discussion.

Data

Variable Information

Two data sets were provided before the study. Both data sets are collected from 900 patients. The first one is a data log that records each patient's physiological measurements every four hours from the time they entered the hospital until they either went to the ICU or left the hospital. The physiological measurements include:

- Respiration rate (per minute).
- Oxygen saturation (percentage).
- Patient breathes normal air (recorded as 1) or uses supplemental Oxygen (recorded as 0).
- Systolic blood pressure (in mm Hg).
- Pulse rate (per minute).
- Any Alert, when a patient's level of consciousness drop, the device will send an alert to the doctor (1 for alert, 0 for no alert).
- Temperature level (in Celsius).

The National Early Warning Scores (NEWS) was also provided for each patient at each recorded time period. This score can be calculated according to the National Early Warning Score chart below:

Table 1: National Early Warning Score Chart

Physiological Parameters	3	2	1	0	1	2	3
Respiration Rate (BPM)	≤ 8		9-11	12-20		21-24	≥ 25
Oxygen Saturation (%)	≤ 91	92-93	94-95	≥ 96			
Any Supplemental Oxygen		Yes		No			
Temperature ($^{\circ}C$)	≤ 35		35.1-36.0	36.1-38.0	38.1-39.0	≥ 39.1	
Systolic Blood Pressure (mmHg)	≤ 90	19-100	101-110	111-219			≥ 220
Heart Rate (BPM)	≤ 40		41-50	51-90	91-110	111-130	≥ 131
Level of Consciousness				A			V,P or U

The way to use this chart is, given a measurement of one physiological parameter, find the according NEWS increment score base on the range of the measurement. Summing up all the increment scores across all physiological parameters listed in the chart, will give the final NEWS risk score.

The second data set is event data that records each patient's basic information such as gender, age, and event time which is the number of days since admission to the hospital. The last variable is status which records whether or not any serious outcome happened. The definition for the serious outcome is: it is serious when the patient went to the ICU after some period of time, and not serious if they left the hospital.

Data Cleaning

For the data log, each patient at each time period has several different physiological measurements recorded. This data is stored in sequence of rows instead of different columns. To make it easier to visualize and fit into the model, each unique physiological measurement was filtered out, and saved in separate columns. By doing this, the number of rows in the data set had been significantly reduced.

The next task is to normalize the time variable. The times in the data log are recorded in the year-date-time format. The format of recording time series data is correct, however, because each patient entered the hospital at different dates so the time variable should not be used like this directly. Therefore, each patient's first recorded time has been set to 0, and the rest of the times just simply become how many days (in decimals) since admission to the hospital. This will make the unit consistent with the unit of the event time variable in the event data set.

The next thing is to spot the missing values. The event data has no missing values, therefore no action is required. The data log has a decent amount of missing values. Because the data log is time-series data, there exists a correlation between the adjacent values, therefore the imputation method is used to replace the missing values. If the current value is a missing value, simply check if the value at the previous time lag is a missing value. If not, replace the current missing value with the previous one. There might be cases that the previous value belongs to a patient with a different ID. If that is the case, replace the current missing value with the one after. For the missing values in the NEWS column, simply calculate the expected NEWS score based on the National Early Warning Score chart, and use it to replace the missing NEWS value.

The last thing is to merge the event data into the data log. New columns were created to record the information into the data log. This will provide more information when fitting the model. The cleaned and combined data set looks like this:

Table 2: Combined ICU Data

id	time	Respirationrate	SpO2	Air	SBP	Pulse	Alert	Temperature	NEWS	Sex	Age	Status
1	0.00	16	92.1	1	134	69.6	1	36.7	2	0	68	0
1	0.17	17	92.4	1	133	70.8	1	36.5	2	0	68	0
1	0.34	17	93.4	1	128	63.3	1	36.7	2	0	68	0
1	0.51	17	93.7	0	131	67.8	1	36.4	3	0	68	0
1	0.68	17	94.3	1	131	64.6	1	36.5	1	0	68	0
1	0.85	17	92.7	1	127	73.8	1	36.5	2	0	68	0

Methods

Model

A decision tree is a flow chart model that use useful variables from the data and the feature of those variables to make a decision path in order to predict the value of a target variable. This tree takes data as input and go through a path of true or false decisions split, when the path reaches the end, it will gives the prediction or classification. Random forest model achieves the same outcome, but instead of one decision tree, it contains many decision trees that have distinct structure of classification. All these decision trees will make a prediction on their own. The final prediction of the random forest model is the average of all the predictions made by those decision trees.

Compare to a single decision tree, the random frost model does not have the problem of over-fitting. Other advantages of random forest models are:

- It is robust to extreme values.
- It works well for non-linear data.
- It can handle large data set.
- It gives good predictions.

The data log is multidimensional, this means the data has different patients, and each patient has measurements taken at different time periods. The duration of stay for each patient varies, as well as their entry time to the hospital. This dependency structure will make most regression models unsuitable because the model assumptions will be violated. However, the random forest model does not have many assumptions. We only assume there are no formal distributions in this case.

Because physiological measurements change over time, it will be unreasonable to just fit one model for all the time periods. The approach here is to first select the unique time periods in the data. By time period, it means the number of days each patient's measurements got recorded since they entered the hospital. Because those measurements were taken every four hours, therefore we can select a finite number of the time periods. The first time period will just be 0, and the last time period will be the last recorded time period for the longest staying patient. There are a total of 80 different time periods in the data. The next step is to loop over each unique time period. For each time period, select data only from that time, split 70% of this data as the training data, and 30% of this data as the testing data, then build a random forest model based on the training data with the patient's status (went to ICU or not) as the outcome variable, and respiration rate, oxygen saturation, air, systolic blood pressure, pulse rate, alert, temperature, sex, and age as the predictor variables. Because there are 80 different time periods, so 80 random forest models have been made for each time. In each loop, variable of importance was also collected, and model accuracy was tested using the testing data. Both information was collected for each time period.

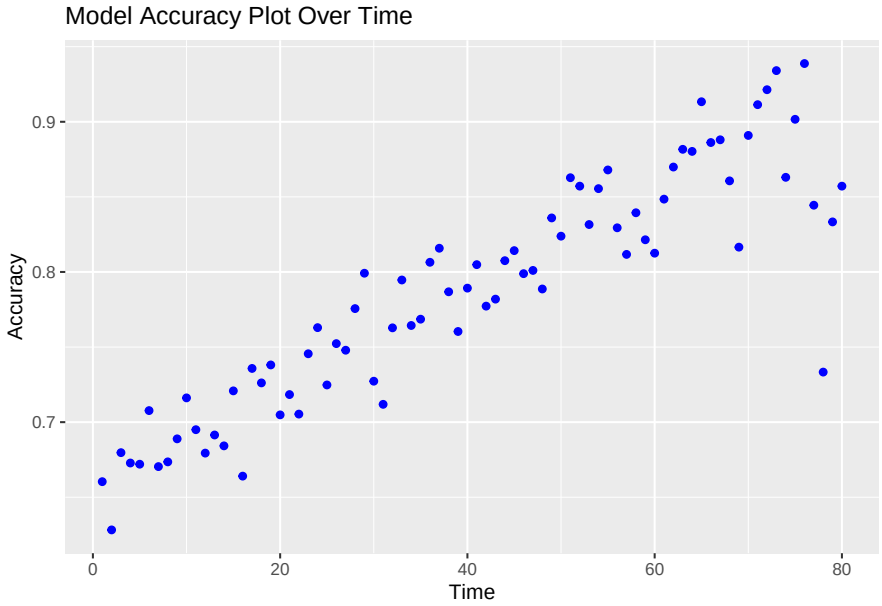
One thing to note is that the **Distribution of Patient Counts at Different Time Periods** histogram from the appendix shows that the amount of data is decreasing as time moves on. This happens due to the

nature of health data, the number of patients who is really sick and stayed until the end is usually very few. That is why there are not much data left for the last few time periods. This information will be taken into consideration in the analysis.

Model Accuracy

In the previous part, for each time period, the random forest model will produce a confusion matrix. This confusion matrix has two columns and two rows. The columns represent the actual outcomes, the patient went to ICU or not. The rows represent the predicted outcomes. Therefore, a confusion matrix can give the counts for the number of true positive, false positive, true negative, and false negative. The model accuracy can be calculated by summing up the number of true positive and true negative, then divide by the total number. This proportion represents the number of patients that are correctly predicted.

The testing data were used to test the model accuracy, and the accuracy results were collected for each time period. The plot below is the model accuracy over time.



As time pass, the model will get better and better at predicting the outcome, because the model accuracy increases over time as shown in the plot. The reason is in the early stage, patient's symptoms have not yet developed, therefore there will not be too much information related to the outcome. Notice the model accuracy estimate dropped for the last few time periods, this happens because the sample size are small in those periods due to the nature of the data. Overall the model is good at predicting the outcome, even in the first time period with lack of information, the model accuracy is already more than 65%.

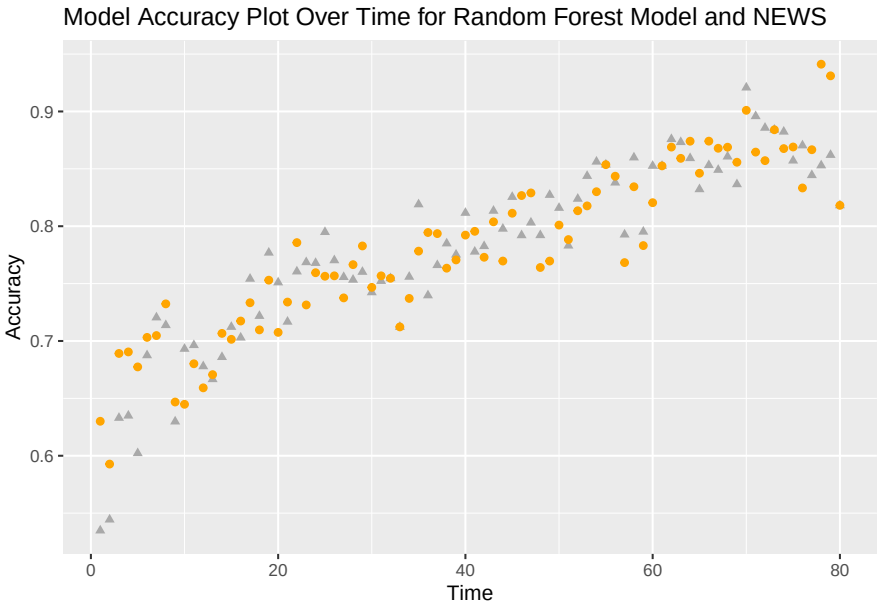
Results

Variable importance represents how much a variable contributes to the model in terms of make accurate predictions. The more a model relies on a variable to make predictions, the more important it is for the model. The model built in the previous section uses mean decrease Gini (based on the Gini impurity index used for the calculation of splits in trees) to measure the variable importance. The higher the mean decrease Gini score, the higher importance a variable is for the model prediction.

After looping over the model, the variable of importance for 80 time periods were all collected, and plotted based on each predictors in the model over time. There are nine predictors, therefore nine variable of importance plots have been plotted. They are shown in the appendix. The interesting insight from those plots is, the importance of all variables decrease over time. This seems odd at first, but it is reasonable. Because the nature of health data, patients who are getting more sick will have a higher chances of going into

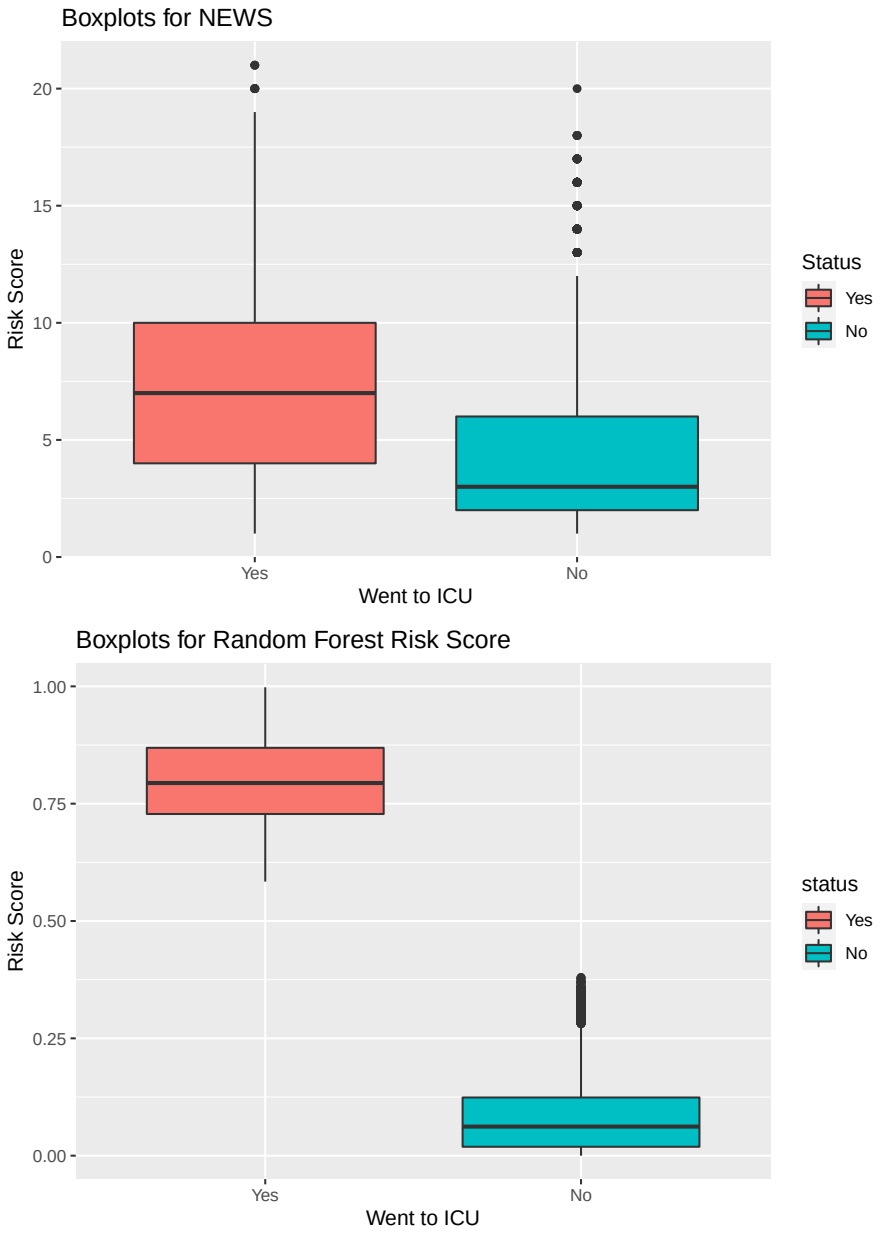
the ICU when the time passes, then those predictors will not matter too much in the prediction, that is why their importance or mean decrease Gini score are going down. From those plot, another insight is, there is a subset of components that does a better job at predicting than others. Especially in the early time period, it is clear that pulse, oxygen saturation, systolic blood pressure, and age have more importance than other variables. They do a better job at predicting the outcome in the early time periods.

Next, the same random forest model is compared with a model that only has NEWS as the predictor. This is to compare their accuracy at predicting whether or not a patient will go to the the ICU. The plot is shown below.



The dark gray color represents the model for NEWS and orange color represents the random forest model. Overall, these two model are both very good at predicting the outcome, and both model accuracy are increasing over time. However, in the early time periods, like the first six time periods, the random forest model has a higher accuracy than just NEWS. This proves that the random forest model is actually better, because the early time periods are very important. If doctors can successfully predict the patient who needs to go to ICU in the early stage, it can save lives and help the hospital operates much more efficiently.

Therefore, the random forest models for different time period can act as a better alternative risk scoring system to help doctors predicting the risk of a patient. However, it needs one more modification. Instead, of classify patients into 1 or 0, change it to the probability, which is a number between 1 and 0. This will act as a risk score.



Both risk scores, random forest risk score and NEWS have been plotted as side by side boxplots. On the x-axis, is the status of a patient went to the ICU or not. The NEWS risk score means the higher the score the more risk the patient, and more likely go to the ICU. If this scoring system is good, there will be significant difference in distribution for both outcomes. However, on average the patients who went to ICU has a higher risk score than the patients who did not, but this gap is small, and the range of the distributions for both outcome are very wide. This means sometimes a paitent has very high risk score but did not go to the ICU. On the other side, the random forest risk score does much better job. The range of the box plot for each outcome does not overlay, which means it can easily predict the outcome based on the score range.

The actionable advice for using this new scoring system is to use 0.5 as a cutoff. If the risk score is below 0.5, most likely the patient will not go to the ICU. If the risk score is above 0.5, most likely the patient will go to the ICU.

Conclusion

Summary

In conclusion, the random forest models built on different time periods can produce the "optimal" risk score compare to the NEWS, as it can give much accurate prediction in the early period of time. The variable importance plots over time periods showed that not all components of the score contribute equally to the score construction, some variables do a better job at predicting the risk. Variable pulse, oxygen saturation, systolic blood pressure, and age have better predictability than other variables in the early time periods. As time passes the importance of all variables goes down due to the nature the data. lastly, the alternative way of defining the risk score is the random forest model itself. The doctor can input the patient's data into the model and predict a risk score. If this score is below 0.5, then most likely the patient will not go to the ICU, and vice versa.

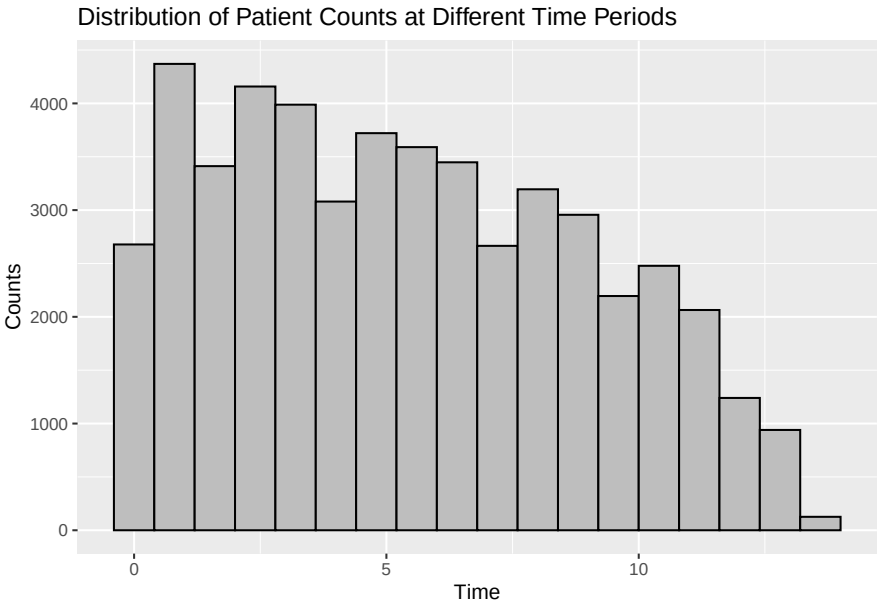
Limitations

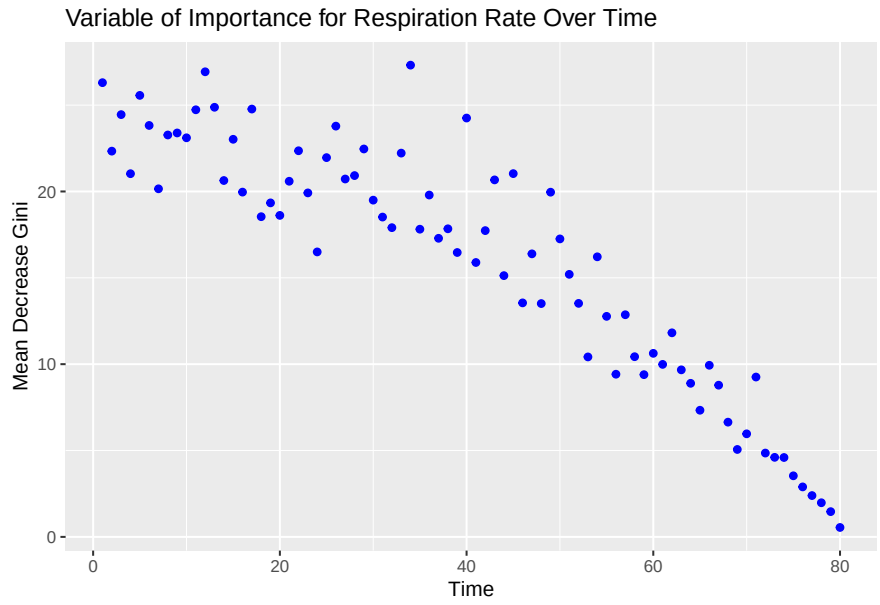
All models have limitations. Even though the random forest model is good at predicting, it has shortcomings. One is that random forests are biased towards the categorical variables, especially the one that has many levels. In our data set, air and alert are the two categorical physiological parameters we are interested in, but because they are categorical, the random forest model might not estimate them without bias. Based on the results, this issue seems not too big. Another issue is that, in these data sets, each patient's data was recorded at different times, they have different entry time and different duration of stay in the hospital. Another method to model these is to use survival analysis model or joint model. To deal with missing values, imputation has been used, this might affect the model a little, but the amount of missing values are small, so this affect can be ignored. Last issue is, due to the nature of the data, for the last few time periods, the amount of data recorded was really small, therefore the model accuracy got affected. To improve, more data can be collected, or using past data. In the future research, models like extended cox proportional model, or joint model can also be used for this study.

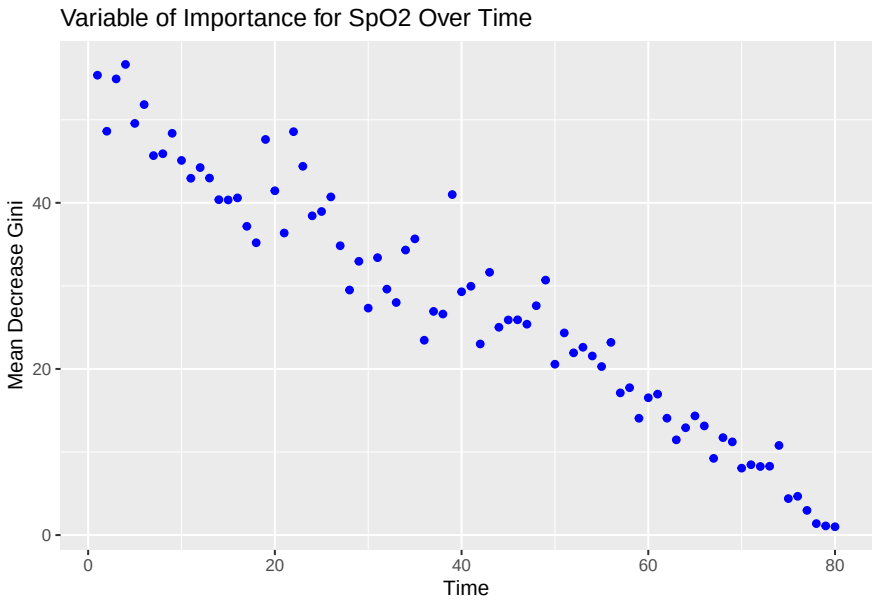
References

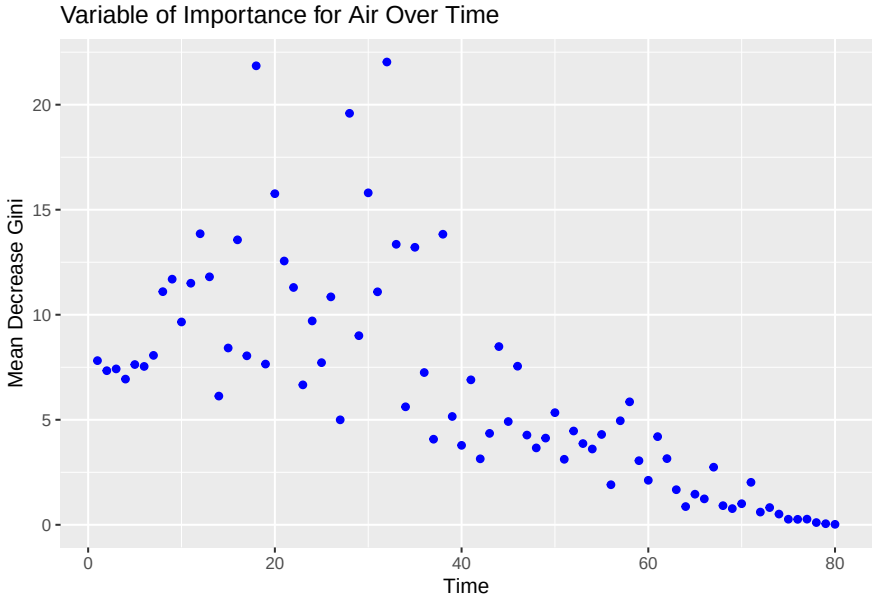
Bilben, B., Grandal, L., & Søvik, S. (2016). National Early Warning Score (NEWS) as an emergency department predictor of disease severity and 90-day survival in the acutely dyspneic patient - a prospective observational study. Scandinavian journal of trauma, resuscitation and emergency medicine, 24, 80. https://doi.org/10.1186/s13049-016-0273-9

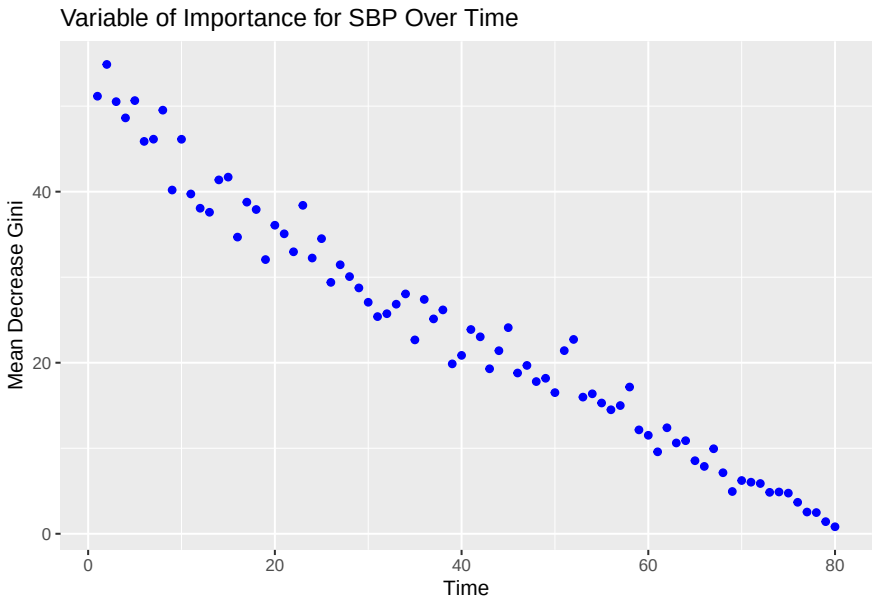
Appendix

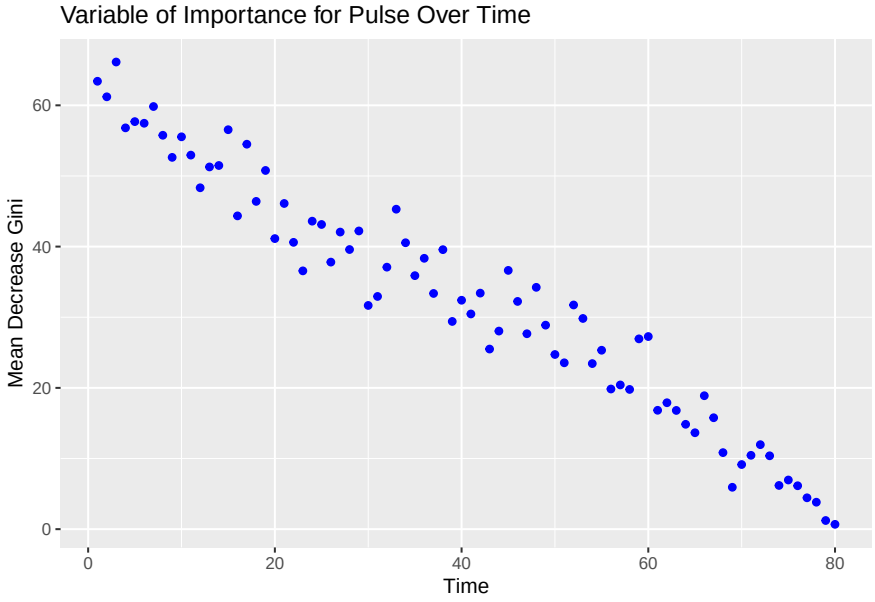

Distribution of Patient Counts at Different Time Periods
Counts
4000
3000
2000
1000
0
0
5
10
Time

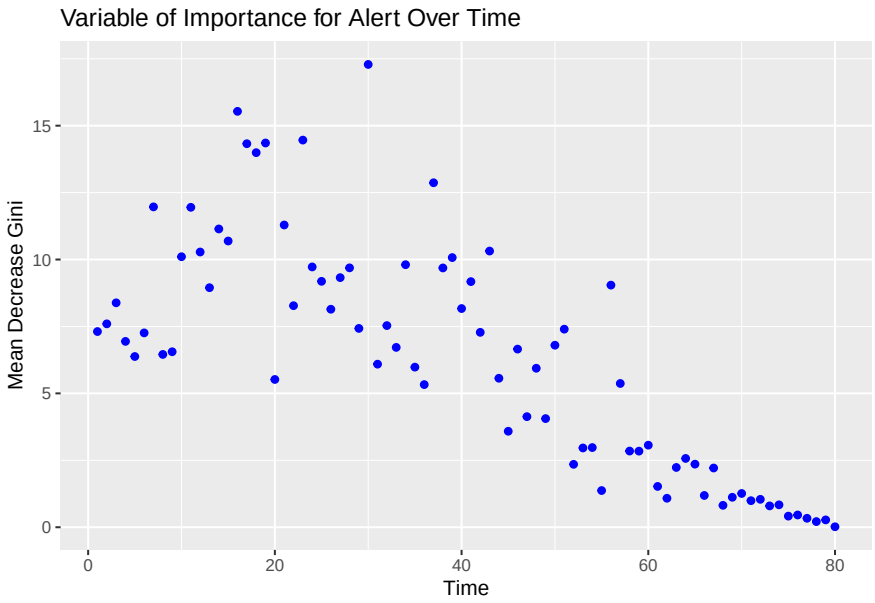

Variable of Importance for Respiration Rate Over Time
Mean Decrease Gini
20
10
0
0
20
40
60
80
Time

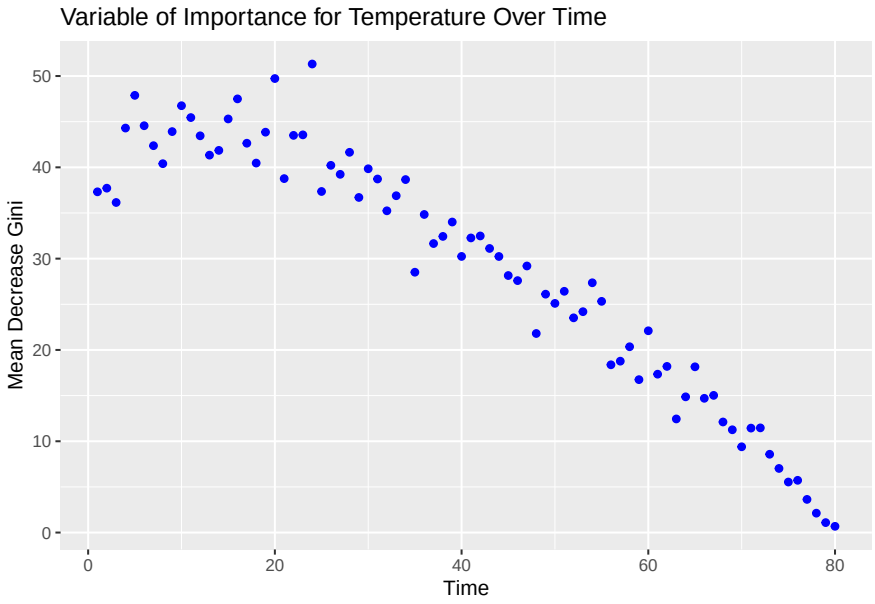

Variable of Importance for SpO2 Over Time
Mean Decrease Gini
0
20
40
0
20
40
60
80
Time

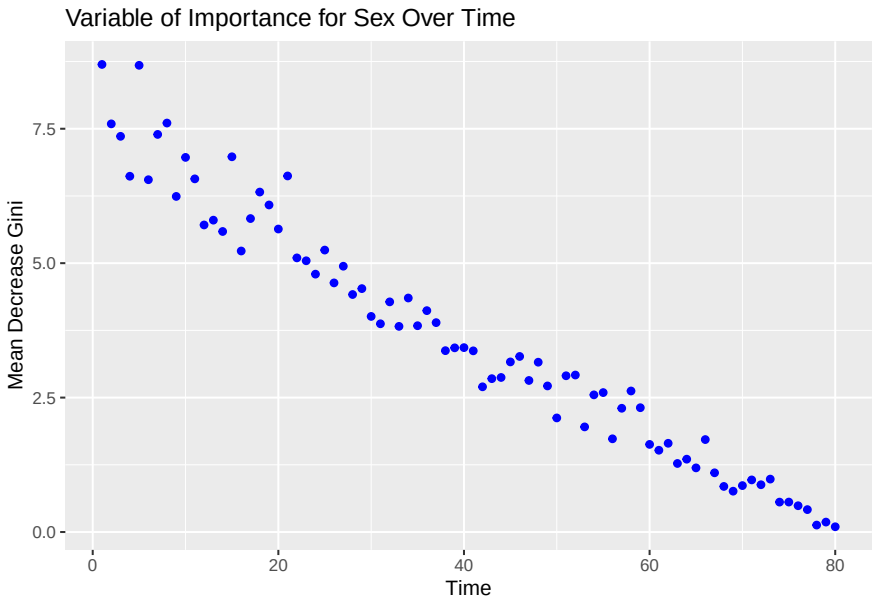

Variable of Importance for Air Over Time
Mean Decrease Gini
0
5
10
15
20
0
20
40
60
80
Time

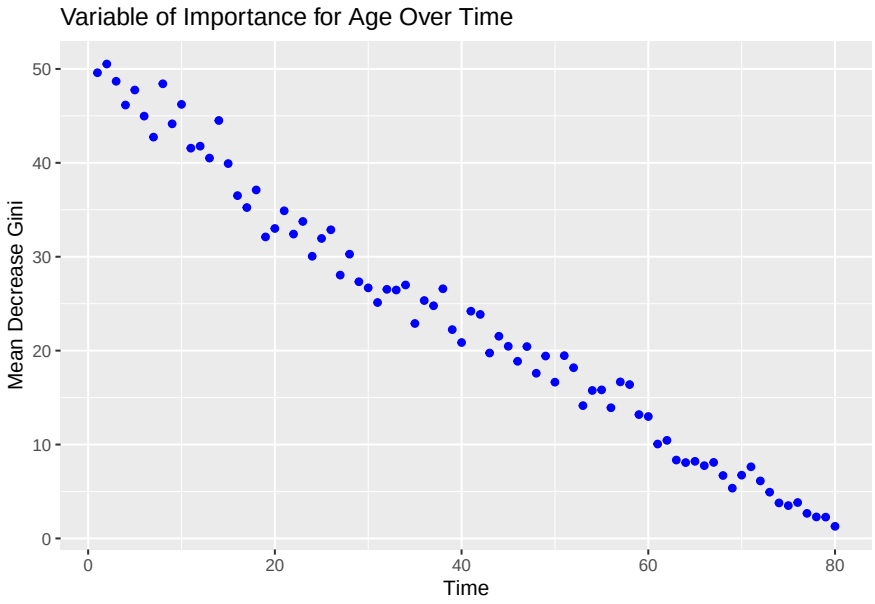

Variable of Importance for SBP Over Time
Mean Decrease Gini
0
20
40
0
20
40
60
80
Time


Variable of Importance for Pulse Over Time
Mean Decrease Gini
0
20
40
60
0
20
40
60
80
Time


Variable of Importance for Alert Over Time
Mean Decrease Gini
0
5
10
15
0
20
40
60
80
Time


Variable of Importance for Temperature Over Time
Mean Decrease Gini
0
10
20
30
40
50
0
20
40
60
80
Time


Variable of Importance for Sex Over Time
Mean Decrease Gini
7.5
5.0
2.5
0.0
0
20
40
60
80
Time


Variable of Importance for Age Over Time
Mean Decrease Gini
50
40
30
20
10
0
0
20
40
60
80
Time