Survival Rate Prediction: Building a Better Version of the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) score

Aaron Guo [1] **Jiawei Zhu** [1]

Abstract

The need for hospital resource management, especially in the Intensive Care Units (ICU) grows rapidly after the initial outbreak of COVID-19. The Acute Physiology and Chronic Health Evaluation (APACHE) II score is a method utilized by various U.S. hospitals to predict a patient's survival rate, which can be used to improve hospital resource management. However, studies show that the APACHE score has limited ability in terms of predicting the correct outcome. This led to various attempts by the previous researchers to improve this model. In this study, we utilized a binary classification model based on logistic regression to improve the APACHE II model. We selected 30 of the most weighted features for the testing data by obtaining their relative weight through the filter method. The results yield a much higher 0.85-0.98 accuracy compared to the previous 0.74 accuracy. This shows that with our improved model, the APACHE II model would yield an accuracy that is substantial enough for the hospital to understand the actual severity of its patients.

1. Introduction

Hospital resources in the United States are limited. After the initial outbreak of COVID-19, demand for hospital treatment in the U.S skyrocketed. As of Friday Feb. 4, 2022, 14 states have devoted more than one-third of of their ICU beds to Covid patients, according to an NBC News analysis of data from the Department of Health and Human Services. For such reason, the need for a better resource distribution method is on high demand.

There are previous attempts to better predict a patient's survival rate by analysing some of their key features before transferring them into the Intensive Care Unit. APACHE, or the Acute Physiology and Chronic Health Evaluation scores is one of these attempts to make such prediction. It utilizes a group of 83 features, ranging from the patient's age to previous illnesses. It is used as a severity adjustment to diagnosis-related groups (DRG's) or other diagnostic classifications.

Despite its attempt to address the issue of analyzing the patient's severity, the model itself proved to be lacking the accuracy needed for the hospital to make appropriate decisions. Rogers et. al stated in his research paper Use of daily Acute Physiology and Chronic Health Evaluation (APACHE) II scores to predict individual patient survival rate have shown that the "daily APACHE II scores do not predict individual patient mortality. The adjustments needed in the algorithm that was used to avoid a false prediction of death render sensitivity so low that it would be impractical to limit therapy on this basis alone." (Rogers et. al, 1994) Rojek-Jarmula et. al also mentioned in his paper claiming that "APACHE II cannot predict weaning outcome in patients requiring PMV*." (Rojek-Jarmula et. al, 2017)

Machine learning algorithms could be a viable method to improve the result of the prediction. Researchers, such as Luo et. al utilized an extreme gradient boosting algorithm to improve the accuracy of APACHE. The result of the XGBoost method raised APACHE II's accuracy from 0.74 to 0.852, which is a substantial improvement. However, this method is somewhat complicated, and we would like to explore whether we can achieve a similar or even better result using a simpler algorithm.

In this project, we intend to build a binary classification model that would predict the survival of admitted patients with given physiological parameters. We propose to establish a weighted classification model to predict patient survival through a logistic regression model.

*Equal contribution [1]Department of Computer Science, Vanderbilt University, Nashville, United States of America. Correspondence to: Aaron Guo <zaiyang.guo@vanderbilt.edu>, Jiawei Zhu <jiawei.zhu@vanderbilt.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

2. Methods

2.1. Data Utilized

We used the data of the patient survival rate through data from MIT's GOSSIS (Global Open Source Severity of Illness Score) initiative with an emphasis on the chronic condition of diabetes. The raw data is from the Kaggle website (Agarwal, M, 2021). There are 83 different features in this data set, ranging from BMI to age and specific illnesses. We intend to pick around 20-30 most important features through analyzing their relations with the output. During this process, around 90,000 patients' data were collected. The survival rate is defined by whether the patient was discharged during hospitalization. We have made several modifications to allow us to process the data set more efficiently.

2.2. Pre-Processing

First, we have modified the labels for some of the features that are not integer/decimal values. These are values such as race and gender. The modified labels would allow for easier analysis on the data. We have also eliminated the rows that contains at least one null element in any of its features for the same purpose. This would leave us with around 50,000 subjects. The output composition of the 50,000 subjects, which is obtained after the data pre-processing, is very much identical with the initial data set, with around 0.91 (0.914 : 0.908) survival rate on both data sets. We believe that this number, despite being only half of the original number, is sufficient enough for us to generate a model that yields a reasonably high accuracy rate.

2.3. Classification Methods

We have several classification methods at hand, these include Naïve Bayes, GDA, SVM and Logistics Regression. We decided not to proceed with Naïve Bayes and GDA. This is because we have no way to know that the continuous variables are normally distributed, which is an essential precondition to use either of these methods. Additionally, since some variables are categorical (such as ethnicity, gender, and how they were admitted into the ICU), these variables definitely do not follow a Gaussian distribution. Furthermore, Naive Bayes assumes that the features are independent; this is not true for this dataset since the Apache scores are composite scores taking into account other physiological parameters (https://www.merckmanuals.com/professional/critical-care-medicine/approach-to-the-critically-ill-patient/critical-care-scoring-systemsv924713). Thus, the assumption of GDA and Naive Bayes are both violated.

In order to compare the performance of both methods, we used the same train and test set, which is split by an 80-20 ratio, to train and evaluate both methods. We initialized a logistic regression model and an SVM model (with C = 10 and gamma = 0.01) using the existing functions from the sklearn library in Python. Then, we trained and evaluated both models separately, and also timed the time taken to train both models. Using all features, both models can achieve an accuracy of around 0.92. However, the amount of time SVM needed to train (approximately 50s) is much longer than that of the logistic regression model (approximately 1s). It is yet to be seen whether SVM can be more efficient after we complete feature selection and useless features, but logistic regression is clearly much more efficient with no detectable compromise inaccuracy if all 82 features were used.

2.4. Feature Selection

We utilized the filter method from scikit-learn to perform a basic feature selection. The results of the filter method yields the feature with the highest weight as the first in rank and the feature with the lowest weight as the last in rank. We then proceed with the highest value of the feature combination of from the initial filter method and use that as our basis for the training set.

2.5. Initial Validation

After training, we test our model's accuracy by computing the accurate prediction divided by the total number of subjects in the testing set. Because of the data set's existing bias, the total accuracy should be a range of accuracy of all the subsets that we generated in order to reduce the bias in the data set. The generated range of accuracy should be the final outcome of our model's accuracy.

3. Results

3.1. Model Selection

Using the implementation of Gaussian Discriminant Analysis (GDA) from the scikit-learn Python library, the average run time for the entire data set was approximately 50 seconds. Using the implementation of logistic regression from the scikit-learn library (Logistic Regression in sklearn.linear-model), the average run time was approximately 0.4 seconds. Both run times were calculated by running the training code on a laptop made by Apple with an Intel Core i7-9750H processor and 16GB of random access memory (RAM).

3.2. Feature Selection

By choosing the set of features that achieved the highest accuracy, which was achieved with 30 features, we selected the following features as inputs for our logistic regression model:

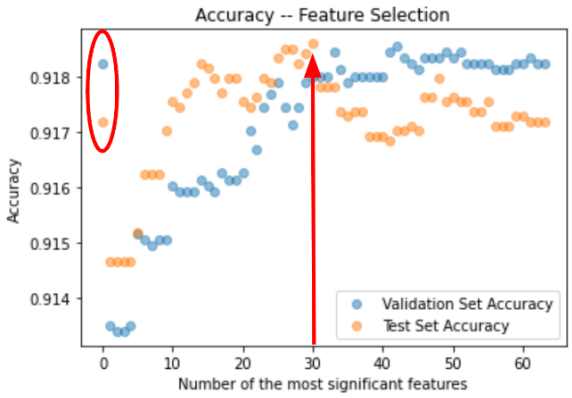

Figure 1. Plot of test and validation set accuracy as a function of number of the most significant features. The accuracy values for both the test and validation reaches the maximum when 30 features are used, and any number of features that is more than 30 causes both validation set accuracy and test set accuracy to decrease, which is most likely due to over-fitting caused by the large size of the training set. One unexpected result was that the model achieved a surprisingly high accuracy with only one feature. This is due to the bias built into the uneven distribution of survival and death in the data set itself.

S.No.	Feature name
1	age
2	d1_temp_min
3	d1_sysbp_max
4	d1_mbp_min
5	d1_sysbp_noninvasive_max
6	d1_heartrate_max
7	h1_mbp_noninvasive_min
8	d1_mbp_noninvasive_min
9	d1_spo2_min
10	d1_sysbp_min
11	h1_resprate_min
12	d1_potassium_max
13	h1_heartrate_max
14	d1_glucose_max
15	diabetes_mellitus
16	h1_diasbp_noninvasive_min
17	solid_tumor_with_metastasis
18	h1_sysbp_min
19	d1_diasbp_noninvasive_max
20	h1_heartrate_min
21	d1_diasbp_noninvasive_min
22	pre_icu_los_days
23	d1_resprate_max
24	h1_sysbp_noninvasive_min
25	gender
26	cirrhosis
27	d1_temp_max
28	d1_resprate_min
29	d1_mbp_max
30	h1_sysbp_max

3.3. Model Performance on Test Data

By using 100 randomly divided datasets with 114 entries in each individual set, we achieved a mean accuracy of 0.921 ± 0.024, with a maximum accuracy of 0.982 and the minimum accuracy of 0.842.

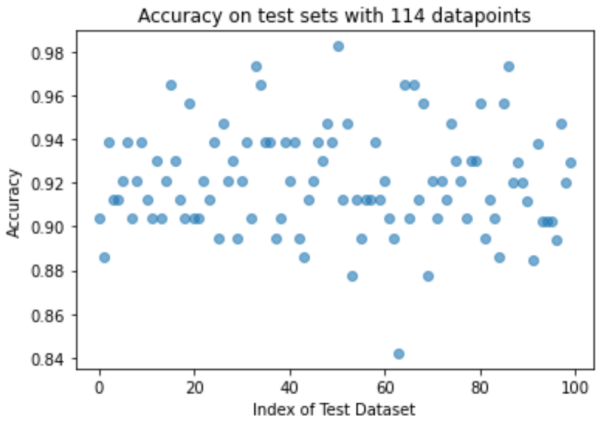

Figure 2. Accuracy of the model on 100 datasets. The 100 datasets were randomly divided from the test set, and each small data set contains 114 entries. The accuracy range from approximately 0.84 to 0.98, which is overall higher than the accuracy of APACHE II reported in literature (Luo, Y., Wang, Z., amp; Wang, C. 2021).

4. Discussion

We have shown that with a logistic regression model using 30 features, we are able to achieve a prediction accuracy that is in the range between 0.84 and 0.98. Additionally, the model can be trained with a relatively short run time of less than a second with over 50,000 data entries on a laptop. However, despite the apparent success of our model, there are important limitations that have to be addressed.

4.1. Data Bias

The most significant source of error originates from the inherent bias of the data set. We calculated the ratio of patients that passed in the entire data set, and found that this group consists of only 0.0863 of the entire data set. Removing all data entries that contain at least one NaN in its feature array does not have a significant effect, resulting in a data set with 0.0859 of patients who eventually passed. As a result, the model is prone to always predicting survival rather than death, as predicting the survival without considering any features could give an overall accuracy of 0.92 based purely on the distribution of the data set itself. Plotting the confusion matrix confirmed our suspicion: The model has a very high false negative rate of 0.88, indicating that the model is biased in favor of predicting survival.

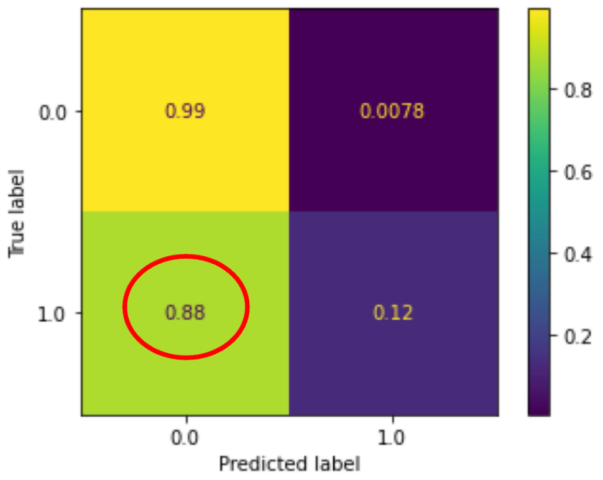

Figure 3. Confusion matrix of the model. The model gives a high false negative prediction rate (circled cell). This is primarily because the dataset is unevenly distributed with a very high (0.92) survival rate. As a result, the model is trained to predict survival in order to achieve high accuracy.

4.2. Health Conditions

In addition to the bias, the model also does not take into account the specific conditions patients may have. One limitation of APACHE is that it fails to consider pre-existing health conditions, such as the need for mechanical ventilation. Because we wanted to reduce the dimension of the feature space, we had to eliminate features that are less relevant. However, this also means that specific health conditions are discarded: such specific conditions are not universal and therefore only appear for certain patients; as a result, its correlation with survival will be low by nature. Although our model achieves a higher accuracy than the accuracy of APACHE reported in literature (Luo, Y., Wang, Z., amp; Wang, C. 2021), it is still a general model with few consideration of specific pre-existing conditions.

4.3. Correlation and Causation

Finally, this model cannot provide a causative explanation of patients' survival outcomes. As of any machine learning algorithm, logistic regression only captures the correlations between individual features and the prediction label. The weight of any features does not provide a physiological rationale of why a particular feature indicates patient survival / death. It would be possible that some features only have a positive / negative correlation with the output but not necessarily a causal relation.

4.4. Solutions to the Limitations

There are possible solutions to the aforementioned limitations. We can build a synthetic dataset using the original data entries such that the synthetic dataset has a 50-to-50 ratio of patients who survived / passed. This would make sure that the model will not be skewed to predicting survival. If we want to build a model that accounts for specific health conditions, regularized regression might be used to assign higher weights to features that contain the conditions of interest. Further, it should be kept in mind that this model has poor explanatory power and thus should not be used for providing a basis of patient outcomes, but rather just as a prediction tool.

4.5. Ethical Issues

Apart from the technical limitations of this model, the potential ethical issues that arise with this model cannot be neglected either. As discussed earlier, our model is more prone to predicting survival. While this, to a small extent, defeats the purpose of helping to allocate ICU resources, the cost is overshadowed by the benefit that more treatment will be allocated to patients who have a lower chance of surviving (as they might come out as positive when being predicted by the model). From an ethical standpoint, the accuracy of a machine learning model should never overshadow the importance of the right to life and access to medical treatment. Additionally, this model essentially inherits any ethical disputes that APACHE II may evoke. The main difference of our model is that we use a more complex mathematical model to predict survival, resulting in higher accuracy overall. The ultimate goal, which is to help allocate medical resources, remains consistent. Thus, as long as the moral principles of medical practices are adhered to, the model will not give rise to additional ethnic dilemmas.

4.6. Future Improvements

To further improve the model, we would need to use a synthetic dataset with a 50-to-50 distribution of survival / death to avoid bias; this allows us to evaluate the true accuracy cleared of interference from the uneven distribution. Additionally, we could evaluate more classification algorithms, although we must also keep in mind that few algorithms achieves a decent accuracy with simplicity that logistic regression possesses.

Acknowledgements

This is a project done by Aaron Guo and Jiawei Zhu. We have worked together for the majority of the project. This includes but is not limited to coding, (data analysis and preprocessing), as well as brainstorming. We believe that it is easier for both of us to focus on the same code and continue

at the same pace instead of us working at a different pace, as this could make debugging more difficult when we integrate our code, considering that this project is not intended to have multiple different modules. We divide our work for the report writing. Usually, we will separate our tasks of the report into two parts that require a similar length of writing. There were many complex issues that are beyond the scope of this class. For example, the unexpected result of high accuracy with only one feature, how to use confusion matrix to explain the cause of this issue, and what solutions we can use to have a better estimate of the true accuracy of our model. For this reason, all the work were done collaboratively during our in-person meetings.

References

ROGERS, J. A. M. E. S., amp; FULLER, H. U. G. H. D. (1994). Use of daily acute physiology and Chronic Health Evaluation (apache) II scores to predict individual patient survival rate. Critical Care Medicine, 22(9), 1402–1405. doi:10.1097/00003246-199409000-00008

Hombach, R., amp; Krzych, Ł. J. (2017). Apache II score cannot predict successful weaning from prolonged mechanical ventilation. Chronic Respiratory Disease, 14(3), 270–275. doi:10.1177/1479972316687100

Luo, Y., Wang, Z., amp; Wang, C. (2021). Improvement of apache II score system for disease severity based on XG-Boost algorithm. BMC Medical Informatics and Decision Making, 21(1). doi:10.1186/s12911-021-01591-x

Agarwal, M.(2021, December 26). Patient survival prediction. Kaggle. https://www.kaggle.com/datasets/mitishaagarwal/patient/metadata. Accessed 21 April 2022