Identification of Suitable Predictive Models for Classifying Terminal Patients

Roberto Cancel, Kyle Esteban Dalope, and Nicholas Lee

University of San Diego

Master of Science, Applied Data Science

ADS-503 Applied Predictive Modeling

Section 01

June 27, 2022

**Introduction**

One of the most daunting tasks faced in the medical field is the decision of the order in which patients are admitted. Severity of the same illness can vary from person to person; thus, mortality varies between individuals as well. Early models that attempted to predict for severity and mortality found limited success. For example, Green et al. (1990) the Health Care Financing Administration (HCFA) constructed a model that could not accurately predict for patient mortality across multiple hospitals. Since the release of the study by Green et al., with a deeper understanding of machine learning and medicine, it has been shown that machine learning models can accurately define severity and a patient’s probability of mortality. One study, by Steele et al. (2018), utilized “Cox models, random forests and elastic net regression” models to predict for patient mortality with regards to coronary artery disease. Another study utilized “Support Vector Machine (SVM), Artificial Neural Networks, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor (KNN)” algorithms to accurately identify 89.98% of COVID-19-related mortality rates (Pourhomayoun & Shakibi, 2021). While the models above were able to accurately predict for mortality and severity, those models were limited to a single illness. In cases where illnesses are diverse, such as in the emergency room or intensive care unit, there is a need for a model to rank severity and predict mortality across multiple illnesses at once. To do so, the Acute Physiology and Chronic Health Assessment Evaluation scoring system (APACHE) is commonly used. The goal of this study is to utilize APACHE scores, along with other descriptive features of patients, to create a model that can accurately identify the risk for death in the presence of multiple illnesses.

**Exploratory Data Analysis (EDA)**

Given the objective, the target feature is hospital\_death, a binary feature where zero represents a patient who has not died, and one represents a patient who will pass or has passed. The upper bar chart of Figure 1 illustrates the proportion of instances by the target feature. In addition, the plot shows how imbalanced the data set is, where instances associated with death make up less than 10% of the data set. The subsequent plots help promote insight into the relationship between mortality rate and patient characteristics by plotting the target features against some predictors. For example, the bottom bar chart in Figure 1 shows the total proportion of death by each ICU Stay Type category. While the proportion of death may not be different by category, what this plot is indicating is that there is not much of a difference in mortality risk whether the patient is newly admitted to the ICU, released and readmitted, or transferred to the ICU from another department.

**Figure 1**

*Bar Charts of the Target Feature Isolated and Associated with ICU Stay Type*

Chart, bar chart

Description automatically generated

As mentioned previously, the Apache scoring system is a measure of risk for mortality in the ICU. The Apache system has several categories that patients receive a score from. The scores from each category are summed into a cumulative score, and the overall risk is determined from the sum. In the data set, different versions of the Apache system were taken into consideration, as the system by hospitals globally vary. Apache II versus Apache III vary in that Apache III has been updated; and, according to Hsu et al. (2001), Apache III has slightly more discriminative power, performing better in terms of its ability to predict high risk patients, than Apache II does. In Figure 2, the bar charts show the proportion of death and non-death patients for each category of the two Apache systems. From these bar charts, it can be deduced that patients that receive a score for sepsis with the Apache III system have a higher risk for mortality. Similarly, with the Apache II system, the three categories with the highest proportion of mortality are respiratory, hematologic, and cardiovascular. A plot, such as Fig. 2, can assist in determining whether a patient is at high risk for death if they receive a score for a certain category.

**Figure 2**

*Stack Bar Charts of Categories for Each Apache System*

Chart, bar chart

Description automatically generated

Since physical characteristics play a role in health-related risks, it is important to examine the relationship between the features related to physical traits and the target variable. For example, age can play a significant role in recovery time, illness severity, and general health quality. Figure 3 shows the histogram for age in the data set, colored by the response variable for each bin, and a boxplot. Both the histogram and boxplot show that most of the data lies around the age range of 50 to 75. The data is skewed left; thus, the results from this study may be slightly more applicable to those in that age range, as those outside the age range are not as highly represented. From a logical perspective, this distribution of data makes sense, it is not commonly expected that those in a younger age range would be hospitalized. Where the histogram shows that the data is skewed left, the boxplot reveals four outliers on the lower end of the distribution. While uncommon, there may be instances where a patient falls in the same age range as the outlier points. Therefore, the outliers were not removed from the data set.

**Figure 3**

*Distribution of Age*

Chart, bar chart

Description automatically generated

Like age, height and weight are also physical traits that are used to measure the general health of a patient. While there are some exceptions based on muscle mass and other factors, in general, the weight of an individual can indicate their health status based on their height. Figure 4 shows a scatterplot, colored by the target variable, of the relationship between height and weight in the data set. Typically, an individual that is determined to be unhealthy is either heavier set and shorter or taller and lighter in weight. If this relationship had a confounding effect on the data, it is expected that the points representing death follow a diagonal line, with a negative slope. However, the plot shows that instances where the patient had died are randomly scattered throughout the plot. Thus, the relationship between height and weight, described above, does not influence the data set. Figures in the appendix show the individual distributions of height and weight, with their relationship to the target variable.

**Figure 4**

*Scatterplot of Height and Weight By the Target Feature*

Chart, scatter chart

Description automatically generated

One aspect to health is the patient’s ethnic background. It is known that ethnicity, and therefore genetics, can lead to an inherent predisposition for illnesses and immunity. Thus, when examining the demographics of patients in the data set, it is key to also study the relationship between ethnicity and the target feature. Figure 5 shows two sets of bar charts. On the left, there is a bar chart of frequency for each category in the variable, ethnicity. On the right, the normalized shows the proportion of the target variable by each category in the predictor. The plots reveal that the majority of data is represented by Caucasians. While the differences in total count of each ethnicity may lead to some selective biases, the proportion of death and non-death instances are roughly the same across ethnic classes. Thus, there is limited influence by ethnicity on mortality rate, as the probability of death is approximately equal by each ethic class.

**Figure 5**

*Bar Charts of Ethnicity*

Chart, bar chart

Description automatically generated

Data wrangling and pre-processing (handling of missing values, outliers, correlated features, etc.)

* Data splitting (training, validation, and test sets)
* Model strategies (describing main research questions and appropriate analytics methods)
* Validation and testing (model tuning and evaluation)
* Results and final model selection (performance measures, etc.)
* Discussion and conclusions (address the problem statement and suggestions/solutions could go beyond the scope of the course)

**References**

Green, J., Wintfeld, N., Sharkey, P., & Passman, L. J. (1990). The importance of severity of illness in assessing hospital mortality. *JAMA, 263*(2), 241-246.

Hsu, C. W., Wann, S. R., Chiang, H. T., Lin, C. H., Kung, M. H., & Lin, S. L. (2001). Comparison of the APACHE II and APACHE III scoring systems in patients with respiratory failure in a medical intensive care unit. *Journal of the Formosan Medical Association, 100*(7), 437-442.

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health, 20*, 100178.

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one, 13*(8), e0202344.

* Appendix (reproducible and clean R code)

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated