MLND Capstone Project Proposal

ICU Mortality Prediction

Andrew Lee

Feb 2017

Proposal

Domain Background

The intensive care unit (ICU) cares for the most critical patients who are admitted to the hospital for emergent issues that require immediate action and close management. Commonly requiring intubation and assisted ventilation, patients are vulnerable and their health statuses can change precipitously. Timing of clinical intervention is therefore paramount to attenuating morbidity and avoiding mortality, making mortality prediction an approach for identifying patient risk early enough for the right mortality-preventing intervention. Fortunately, the current paradigm of clinical decision-making in the ICU is supported by temporal data generated by a host of monitoring systems and lab assays, allowing machine learning to potentially solve mortality prediction.

Published literature provides a few examples of successful predictive models that can discriminate the risk of mortality with high confidence 4[1, 2] and 5[3] hours in advance.

Problem Statement

The goal of this project will be to predict the risk of mortality in the ICU using temporal frameworks different from those already elucidated by literature. Timing of intervention is extremely important in mitigating the risk of mortality, and thus novel temporal frameworks must be investigated for incremental utility. The target will be the incidence of mortality during the admit (1=death, 0=no death) and predictors will be collected from the first day of the admit. The admits will be filtered down to a specific set in order to maintain independence between observations (explained in the next section). The prevalence of mortality in this subset is 14%.

Datasets and Inputs

The MIMIC-III dataset is a critical care database containing all ICU admits at the Beth Isreal Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. There are 53,423 unique admits for 38,597 unique adults, as well as associated monitoring sensor and lab data, which are timestamped and will serve to test how early mortality risk can be discriminated.

Inclusion/exclusion criteria will be examined to create the cleanest and most representative dataset for modeling and prediction. Patients can have multiple admits in the MIMIC-III data. To maintain independent observations, only the first admit for each patient will be used. Furthermore, admits can be surgical or non-surgical. Surgical admits will be excluded because they represent a type of patient admitted for pre-planned procedures. Excluding these patients will improve the homogeneity of the resulting cohort given the target. Additionally, only adult patients older than the age of 18 will be considered. 24,418 unique patient-ICU stay observations are left once subsequent admit, surgical, and non-adult exclusions are applied.

Sensor data from the monitoring systems provide the vitals data for sustaining life such as blood pressure, heart rate, and oxygen saturation. Lab data contain information ordered by clinicians to further understand health status. Biochemical markers such as serum albumin concentration can show the degree of malnutrition, inflammation, or metabolism. Some labs must be read together to understand the state of organ dysfunction and impending failure, for example. These data will be collected from the first day of the admit to represent the earliest information possible. The predictors will be aggregated as averages, minimums, and maximums across the first day, as well as the averages, minumums, and maximums of change across sequential measurements the first day. These aggregations should be indicative of health status as the standard of care is founded on absolute readings (eg, albumin less than 3.5 is low-normal, and less than 3.0 is low). The delta readings will provide relative information that may provide additional information to the model.

There are approximately 40 sensor and lab variables, which will be aggregated into avg, min, max; then, another set of avg, min, max aggregations on the deltas of the 40 raw variables, leading to a total of 240 initial predictors. All variables should be continuous.

The dataset will be split into 70% training and 30% testing, so that there are about 1,000 target observations in the test set.

Solution Statement

Classification models will be attempted given the binary target. The target prevalence is 14% and is not rare (<10%), but building models using a balanced dataset may be attempted. The interpretability of the model will be extremely important since the context of the problem is healthcare. The model will need to generally align with clinical thinking to be accepted and will need to be mechanically understood since a cost of missed predictions will be patient deaths. Thus, decision tree models consisting of basic decision trees, random forest, and gradient boosted trees will be attempted. A successful model will stratify risk such that patients identified as high-risk will have the highest probability of dying.

Unpacking the classification models will mostly guide feature selection. Standardization of features may be utilized to improve the normal distribution of data. Missing values may get filled with median values or pulled forward from previous values (for delta features).

Benchmark Model

Throughout modeling, naive (default hyper-parameter settings) decision trees, random forests, and gradient boosted trees will serve as benchmarks for each other as the data is changed and the models are tuned. Ultimately the trees will be visualized and unpacked to ensure interpretability.

Additionally, the resulting set of promising models (including the final model) will be benchmarked against models from literature to compare the utility of varying data (feature) strategy and temporal frameworks.

Evaluation Metrics

Several performance metrics will be used to compare between models, including AUC, cumulative lift, precision, and recall.

* AUC is the area under the ROC curve is a measure of how well a model can discriminate between a 1 or 0 in the target, or the probability that the classifier will rank a randomly chosen positive observation higher than a negative one. Graphically, it is the integral area between a curve constructed as the true positive rate vs true negative rate and a linear curve representing random chance of discrimination (or

50/50 guessing). AUC can be between 0 and 1. A value greater than 0.5 indicates that a model is better at discriminating than random guessing. A value less than 0.5 indicates that a model is worse than random guessing. AUC will be important for comparing performance between models.

* Cumulative lift shows the concentration of predicted targets when ranked by their propensity, or probability of being the target. This is important because in the ICU setting since a high cumulative lift will translate to stratification of patients with high risk.

* Precision, or the positive predictive value, is the proportion of predicted targets that correctly identified as targets. This is important because a higher precision will ensure that ICU resources are not being wasted on incorrectly predicted patients.

* Recall, or sensitivity, is the proportion of all true targets that are predicted correctly by the model. A higher sensitivity means that the model is generalized.

Project Design

1. Data exploration: The ICU data contains patients afflicted with many types of conditions as evidenced by the plurality of ICD-9 diagnosis codes. The frequencies of the conditions should be understood to ensure that prediction happens on a homogenous group, so there may be further exclusions on the disease level. For example, there are thousands of potential drivers of mortality, but only a handful may represent the majority of cases, so refining the problem set to only those specific disease may improve model results.

2. Data prep: Predictors (already pulled in aggregated forms: average, minimum, maximum, and as deltas) will be analyzed for completeness using descriptive statistics. Predictors with high missing values will be dropped according to threshold to be determined later by researching what allowances should be expected in the ICU setting. Most data should be continuous, so data will be visualized to understand distributions, outliers, and aberrations. Decisions will be made to standardize the data to improve normality while considering how it will affect interpretability. Missing values may get filled with median values or pulled forward from previous values (for delta features). Some patients may get dropped for having too many null values row-wise.

3. Naive model benchmarks: Models will be built using default hyper-parameters as a benchmark for feature derivations and model tuning. Decision tree, random forest, and gradient boosted trees will be built and visualized. Some adjustment of default parameters may need to be done in case the default settings are hurtful to results (eg, default max_depth is too high and leads to overfit results). Additionally, the variable importance will be searched to understand which metrics in their raw forms are most promising.

4. Feature selection and engineering: The splits in the naive trees will be used to understand how the data is being segmented and whether any variables should be binned to reduce overfitting or nonsensical splits. The highest importance variables will be visualized univariately and bivariately with the target to further understand how the trees are selecting features and splitting. Second and third degree polynomial terms may be created among the most important features to understand if there are any interactions.

5. Rerun naive models: Rerun the naive models using updated features to see differences that feature engineering may have added. Ultimately, the most parsimonious or least contrived predictors should be used unless there are considerable improvements.

6. Tune models: Use iterative methods to optimize evaluation metrics given varying parameter values and/or use GridSearchCV. Trees will be visualized again to understand the splits and compared with the naive trees to ensure parsimony and generalizability.

7. Choose a final model based on evaluation metrics, parsimony, generalizability, and interpretability, then compare with results reported by literature.

References

1. Calvert, J. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. 2016. http://www.sciencedirect.com/science/article/pii/S2049080116300413

2. Desautels, T. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065680/

3. Calvert, J. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. 2016. http://www.annalsjournal.com/article/S2049-0801(16)30131-5/pdf