SI 670 Final Project Report: Predicting Mortality of Intensive Care Unit (ICU) patients with Respiratory Complications

Elishua K. Shumpert
elishuas@umich.edu

Stephen Toner
srtoner@umich.edu

Abstract

The emergence of electronic intensive care units (eICUs) has not only allowed for physicians to treat critically ill patients from afar with remote access to vital signs, drug intake statistics, and lab results, but also created a wealth of data for analysis and machine learning. In our project, we leverage the eICU Collaborative Research Database to design and evaluate a suite of models to predict mortality of patients with respiratory related problems. Because of the ethical considerations inherent in healthcare and the relative incidence of mortality in the data set, we place great focus on evaluation metrics and model calibration. We conclude with a comparison of our results with those of prevailing research and a discussion of areas for future work.

1. Introduction

The intensive care unit (ICU) consists of patients with severe, and often life-threatening, conditions that are constantly monitored by specialized healthcare professionals. ICUs have both the greatest rates of mortality and the highest costs of treatment in hospitals; as such, significant research has been conducted to develop best practices and treatment strategies for the ICU specifically to reduce patient mortality. One such strategy is the creation of eICUs, in which healthcare professionals can observe patient vitals remotely and interact with patients via telemedicine, which allows for highly specialized practitioners to share their expertise and helps with the knowledge transfer of a patient's condition in between hospital shifts. An added benefit is that researchers have access to a vast amount of data from which they can draw insights about the predictive factors of patient mortality in an ICU environment.

We make use of the eICU Collaborative Research Database, a collaborative project between MIT and Philips Healthcare. Using data from patient demographics and their admission and discharge details, time series of vital signals, lab results, and Acute Physiology Age Chronic Health Evaluation (APACHE) measurements and predictions, we apply various machine learning algorithms including logistic regression, random forest, XG-Boost and neural networks to predict patient outcome (mortality) for patients with respiratory complications including pneumonia, asthma, laryngeal and lung cancer, bronchitis, and many others. These models are evaluated relative to a dummy classifier baseline and assessed on metrics including area under the precision-recall curve (AUPRC), precision, recall, and F2-score.

2. Methods

2.1. Data

The data used in this project is provided courtesy of the eICU Collaborative Research Database

[1]. This database consists of data tables of statistics for patients treated by the Philips eICU program across intensive care units in the United States from 2014 to 2015, covering over 200,000 ICU admissions. The features span data of patient vital signs, laboratory measurements, medications, APACHE severity of illness measures, care plan information, admission diagnosis, patient history, and respiratory charting and treatment. Each instance in the base patient records table corresponds to a unique ICU stay for a patient. The target population for patient outcome prediction consists of patients with respiratory complications, thus, we subset the data to only include patients with admission for respiratory problems which accounts for 21,528 unique patients or 15.4% of patients in the full data set.

2.2. Data pre-processing

For this analysis, we consider only the first ICU visit for each patient and we drop patients with duplicate records in the first ICU visit. To develop features for our predictive models, we utilize several tables of the research database and combine them with the base patient records table. The tables that we used include data on the APACHE score for patients composed of variables used to calculate the Acute Physiology Score (APS) for patients and predictions of mortality generated by the APACHE, the hospital that the patient is admitted to, drug intake, lab tests, vital signs and respiratory measurements. Given that some of tables including data on drug intake, lab tests, vital signs and respiratory measurements contain multiple measurements over a period of time for each patient stay we aggregate this data with statistics including the minimum, mean and maximum for the variables in these respective tables.

One percent of patients in the data had missing values for discharge status, or in other words, whether the patient lived or died at the end of the ICU stay was not available. We simply decided to remove these cases. In terms of missing values for other variables, we dropped any variables where more than 80% of the data was missing. Many of the dropped variables were related to respiratory measurements and drug infusions. For any remaining variables with missing values, categorical variables were imputed with the most frequent category and numerical variables imputed with the median after splitting the data set into a training and testing set to prevent data leakage. Furthermore, we one-hot encode all categorical variables before training our predictive models. After pre-processing, the final data set contains 16,989 respiratory patients and 238 predictor variables.

To predict patient mortality, we train and evaluate four functional forms including logistic regression, random forest, XG-Boost, and a multilayer perceptron. We compare all models against a baseline dummy classifier that simply classifies the target variable with the most frequent label, which in this case is that a patient is alive at the end of their ICU stay. Our choice of modeling forms seeks to compare linear approximation with logistic regression with more flexible, non-linear models in predicting patient mortality. Given that the feature space was large after data pre-processing and numerous lab measurements and vital sign measurements were highly correlated, we used a logistic regression model with the LASSO $L1$ penalty for variable selection and kept all predictor variables with non-zero coefficients from LASSO. Using LASSO, we can identify some important variables associated with the likelihood of patient mortality. LASSO selected 68% of the variables. In the following sections, we go into more details about the modeling algorithms.

2.3. Logistic Regression

Logistic regression is a widely known statistical model that estimates the probability of an event occurring for a binary outcome. Logistic regression typically does not make as many assumptions about the data as linear regression

[1] `https://eicu-crd.mit.edu/`.

does. Logistic regression generally requires that the relationship between the independent variables and the log odds be linear. However, this assumption rarely holds. An advantage of this learner is that it is interpretable in terms of the coefficients which quantify the contribution of each feature to the risk estimate of the outcome. It also tends to be more calibrated compared to other learners in regards to prediction probabilities.

2.4. Random Forest

Random Forest is an ensemble algorithm which combines multiple de-correlated decision trees each based on a subset of the data samples. Random forests typically generate reasonable predictions and can be fairly accurate across a wide scope of problems. Unlike logistic regression, random forests are robust to outliers, can model interactions without manually specifying them, do not require feature scaling and can handle a large number of input variables without overfitting. It also provides feature importance by assessing how much each variable contributes to the predictions relative to other variables. With this model, we tune the number of trees (`n_estimators`), the maximum depth of the trees (`max_depth`), and the class weights using a grid search with 5-fold cross validation. The best parameter setting that is chosen is the one that maximizes the area under the precision-recall curve (AUPRC) metric.

2.5. XG-Boost

The XG-Boost algorithm is an improved algorithm based on gradient boosting decision trees in terms of scalability. With gradient boosting, the model is sequential in that each weak learner learns from the previous learner's errors and tries to correct them. Algorithmic enhancements of this model include regularization with the lasso and ridge penalty and the ability to handle sparse data. Given its superior prediction performance and speed, it is the algorithm of choice for many winning teams in data science competitions. With this model, we tune the number of estimators (`n_estimators`), the maximum depth of the trees (`max_depth`), and the learning rate. The hyper-parameters are selected using a grid search with 5-fold cross validation and are chosen based on the AUPRC similar to random forest.

2.6. Long Short-Term Memory (LSTM)

Multilayer Perceptrons (MLP) are a class of learning algorithms which have exploded in popularity due to their highly flexible nature and role as universal function approximators. LSTMs are a form of recurrent neural networks that naturally lend themselves to sequential or time series data. The model is trained by optimizing weights between each layer in a series of optimization steps that propagate back and forth throughout the network. In tuning our model, we varied the optimizer, the sequence of layers, number of nodes in each layer, and nonlinear activation function applied in each layer. Ultimately, our model with the best performance was a single LSTM layer with 24 nodes and the hyperbolic tangent function, followed by a dense output layer using the sigmoid activation function. The model was trained using the Adam optimizer.

3. Evaluation and Results

3.1. Evaluation

To evaluate our learning algorithms, we divide the data set into a training and testing set with an 80%-20% split using stratified sampling by hospital region and teaching status of the hospital that the patient is in. We utilize the stratified sampling technique to ensure that the range of hospitals that patients are being cared in are equivalent in the training and testing set. All learning algorithms are evaluated based on area under the precision-recall curve (AUPRC), precision, recall and the F-β score where $\beta = 2$. We use these evaluation metrics because the classification problem is imbalanced with 12% of the patients deceased and 87% of the patients alive at the end of their ICU

Method	AUPRC	Precision	Recall	F2 score
Dummy Classifier	0	0	0	0
Logistic Regression	0.35	27.8%	68.6%	0.53
Random Forest	0.40	28%	73.1%	0.55
XG-Boost	0.46	30.6%	78%	0.60
LSTM	**0.86**	**58.5%**	**83.8%**	**0.77**

Table 1: Evaluation results of proposed learning algorithms to predict patient mortality of respiratory patients on the test data set.

stay. In addition, false negatives, or failing to predict death of a patient, is more costly so we aim to use recall as the defining metric. Thus, the final model selected is the model that maximizes the recall score. Recall is the proportion of actual positive instances that are correctly classified. Since the classification problem is imbalanced, we decide that using a probability threshold of 0.5 might not be the most suitable in hard classifying the labels. Thus, we tune the threshold for each learner and select the threshold that maximizes the F-β score where $\beta = 2$. The F-β measure is a weighted mean of the precision and recall where $\beta > 1$ places more weight on recall. Lastly, we evaluate the calibration of the learner's predicted probabilities using a calibration plot.

3.2. Results

Table 1 shows the results of our proposed learning algorithms in predicting mortality of patients. From Table 1, we note that the long short-term memory network achieves the best performance in all metrics. Given that this model obtains the highest recall rate, we conclude that this model is the most accurate in classifying mortality of the patients with respiratory illnesses followed by XG-Boost. The performances of the models seem to follow what we expected. We posit that LSTM performed the best out of all classifiers given that the vital signs were measured for each patient periodically and because the measurements were taken frequently in time this contributed to over 2 million instances within the data in which a neural network architecture such as this could thrive given the amount of data. Also given that the data used to train this model was inherently time series, the LSTM is structured specifically to model time series data which could also attribute to its performance. Likewise, the other models were trained using aggregated data so they were mostly likely not as accurate because of the loss of information in aggregating the vital sign measurement. Additionally, logistic regression performed the worst compared to the black-box models suggesting that the data is most likely non-linear.

3.3. Model calibration

While we focus on how well the models do on maximizing recall, we also want to make sure that the predicted probability of the positive label for each patient is calibrated properly especially if doctors wanted an estimated probability of death for a patient instead of a hard coded label result. Figure 1 shows the calibration curves for all of the classifiers. All of them are not perfectly calibrated, however, we note that overall if we wanted to use predicted probabilities to determine occurrence of death for a patient then logistic regression might be the best model for that task. For all models, the predicted probabilities are well calibrated for lower probabilities from 0 to 0.4, however, for probabilities greater than 0.4, all of the predicted probabilities diverge from the ideally calibrated line suggesting that most of the models struggle in calibration for higher predicted probabilities. Random Forest appears to be the worst in terms of calibration of the predicted probabilities.

4. Related work

Predicting patient mortality risks has been widely studied, and is the premise for placing patients in the ICU in the first place. Multiple systems for standardized triage have been developed, such as Assessment and Chronic Health Evaluation (APACHE) II and III and the Simplified Acute Physiology Score (SAPS) II and III, which convert patient history and vitals to a calibrated risk score for the patient. While effective for standardizing risk assessment and evaluating a hospital's resource allocation, these scores are not without their limitations. APACHE was created only using data from patients admitted to the ICU, and is meant to assess mortality risk on a population level; if a patient has a rare condition or compromised immune system, then their mortality risk could be drastically underestimated.

With the vast improvements in machine learning methods in the past two decades, ICU outcomes have been a natural area of focus for research. Approaches have spanned countless different algorithms from SVMs, regression variants, decision trees, and neural networks. Frequently, ensemble methods were particularly effective at combining the best of these approaches into one coherent model. [5]

More recently, research has expanded beyond merely predicting patient outcome but also a patient's length of stay in the ICU. One notable example is that of [4], whose recent work in predicting patient stay was used to pre-process the eICU time series data for training our neural network. They note that while neural networks are often applied to physiological time series data, it is unusual for approaches to incorporate more sparse data from the eICU database such as medications administered and lab results. These sparse features are particularly important for detecting rare disease patterns, and are a promising opportunity for future work.

5. Discussion and Conclusion

Given the constantly evolving nature of both machine learning and healthcare, it should come as no surprise that the most recently refined practices of recurrent neural networks and medicine offer such complementary benefits. In all fairness, the precursors to ML-based patient outcomes like APACHE II were created using multivariate logistic regression, which makes our baseline measure especially appropriate given the context. The significance of our project is in an assessment of the performance of the status quo (linear methods) in comparison with less traditional approaches. While the empirical results in terms of the metrics speak for themselves, consideration for interpretability and fairness are particularly important in a medical context. While there is always a potential for unfairness in machine learning methods, it is almost certainly more apparent in linear methods which tend to be more interpretable. Another takeaway from our results is the old adage ”less is more”; when training the LSTM, a simple architecture of just two layers proved most effective at mitigating the risk of overfitting.

Clearly, evaluating model performance must go far beyond accuracy alone when dealing with a domain as sensitive and nuanced as healthcare. In our efforts to design a machine learning model that performs well on the most relevant metrics, we find that the sparse, non-linear nature of medical time series data readily lends itself to non-linear methods, particularly long short term memory models. It is also worth noting that given LSTM is not an interpretable model, we use variable importance of the second most accurate classifier (XG-Boost) and determine that duration of the ICU stay, mean blood pressure and heart rate measured for the APACHE score, if the patient is on a ventilator, and patients who are aged 70-79 are the top 5 most important features relative to the others in predicting mortality of the patients. Two of these features are related to vital signs from APACHE which indicate that vital signs are

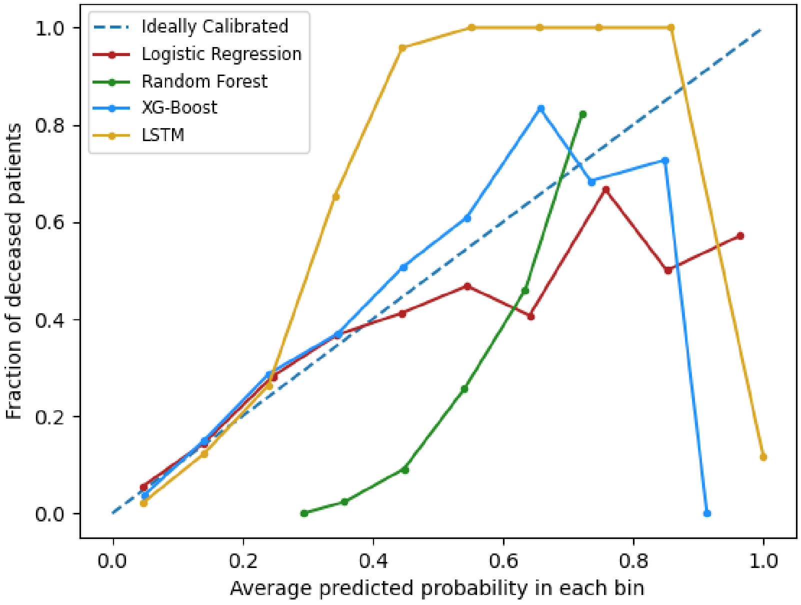

Figure 1: Calibration plot of proposed learning algorithms

key indicators in predicting mortality.

If we had more time to work on this project, some possible future work that we might consider include doing an anomaly detection on the vital sign measurements and lab results to remove patient anomalies from the data set in order to improve our models. Researching the appropriate ranges of these measurements often rely on medical domain knowledge in which we could consult with experts to determine the anomalies within the data. Another potential task that we could also consider would be using a cluster analysis to determine representative groupings of the diagnoses for the respiratory diseases and see if the clusters help in improving the predictions of our models.

6. Access Project Github Repository

The data and code for our project can be accessed from the following URL link in the footnote [2].

[2] https://github.com/elishuas/SI-670-Final-Project.

References

[1] "An Evaluation of Outcome from Intensive Care in Major Medical Centers". In: *Annals of Internal Medicine* 104.3 (1986). PMID: 3946981, pp. 410–418. DOI: 10.7326/0003-4819-104-3-410. eprint: https://www.acpjournals.org/doi/pdf/10.7326/0003-4819-104-3-410. URL: https://www.acpjournals.org/doi/abs/10.7326/0003-4819-104-3-410.

[2] Min Hyuk Choi et al. "Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records". In: *Scientific Reports* 12.1 (2022), p. 7180. DOI: 10.1038/s41598-022-11226-4. URL: https://doi.org/10.1038/s41598-022-11226-4.

[3] Tom J Pollard et al. "The eICU Collaborative Research Database, a freely available multi-center database for critical care re-

search". In: *Scientific data* 5.1 (2018), pp. 1–13.

[4] Emma Rocheteau et al. "Predicting patient outcomes with graph representation learning". In: (Jan. 2021). arXiv: `2101.03940 [cs.LG]`.

[5] Ikaro Silva et al. "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012". In: *2012 Computing in Cardiology*. IEEE. 2012, pp. 245–248.