Automated versus Manual Machine Learning with Small Data for Predictions of Six-Month Outcomes Among Patients in the Intensive Care Unit

This manuscript (permalink) was automatically generated from gweissman/icu_ml_ms@468f06f on March 27, 2020.

Authors

- **Gary E. Weissman**
  0000-0001-9588-3819 · gweissman · garyweissman
  Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; Division of Pulmonary, Allergy, and Critical Care Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

- **Trang T. Le**
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **William La Cava**
  0000-0002-1332-2960 · lacava · w_la_cava
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Marzana Chowdhury**
  Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

- **Steven Brooks**
  · stevegbrooks
  Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

- **Elizabeth L. Cooney**
  Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

- **Francisca Oredeko**
  Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

- **Trishya Srinivasan**
  · TrishyaS
  School of Medicine, Wayne State University, Detroit, MI 48201

- **Stephanie Szymanski**

Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

- **Michael E. Detsky**
  0000-0001-8305-2137 · michael_detsky
  Department of Medicine, Sinai Health System, Toronto, Ontario, Canada; Interdepartmental Division of Critical Care Medicine and Department of Medicine, University of Toronto, Toronto, Ontario, Canada

- **Jason H. Moore**
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Scott D. Halpern**
  0000-0002-3603-4769 · ScottHalpernMD
  Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; Division of Pulmonary, Allergy, and Critical Care Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Traditional statistical modeling approaches are usually compared to machine learning methods using large, retrospective datasets. Their relative performance across manual and automated methods, how approaches to missing data, use of repeated measures across time, and split-sampling approaches, when applied to small, prospectively collected datasets is unknown. We sought to address these questions using a small ($N = Z$), prospectively collected data from patients admitted to an intensive care unit. Therefore, we compared multivariable logistic regression, penalized logistic regression, XGBoost, TPOT-1, TPOT-2, and Feat approaches to model tuning and selection. Each model received three different datasets with varying strategies for handling missing data, including prior imputation, inclusion of missing fields, and complete cases analysis. Each model variably received data from the first and second day of inlusion in the cohort. Each model also received 80/20 and 50/50 split samples or training and testing. We found that…

Introduction

A surge of interest in predictive modeling techniques has paralleled the increasing availability of large data sets and open source software packages that allow nearly out-of-the-box model development. The popularity of and potential for data science methods is particularly relevant to the health care setting where decision making under uncertainty with large and varied data inputs are the daily norm. However, many advanced modeling approaches have failed to yield evidence for their superiority over traditional statistical methods.[1,2] Comparisons between statistical and machine learning methods have focused on relatively large datasets, or “big data.” However, prospectively collected, clinically rich datasets of cohorts with relevant, patient-centered outcomes are more rare. With less noise in the cohort selection and training labels, these prospective cohorts, albeit typically smaller due to the expense of constructing them, offer an opportunity to better isolate the effects of different modeling approaches.

However, small prospectively collected datasets present additional unique and unexplored questions. First, how much data are wasted in using a split-sampling approach for internal validation?[3] With an extremely large dataset with millions of observations, the difference between a testing sample of 20% or 25% may not matter. But if the data set has only a few hundred observations, a careful consideration of sufficient sample size in the training set to fit a model is balanced against the need for sufficient sample size in the test set to construct a clinically meaningful confidence interval. Second, the tradeoffs in approaches to missing data — common in clinical datasets — for such small datasets used for prediction is unknown.[4,5] The removal of complete cases is relatively costly given the small number of observations while imputation may introduce or reinforce bias. Third, with a small dataset, does incorporating repeated measures across a patient’s trajectory improve predictive performance?[6] Finally, all of these decisions could be guided by statistical expertise and clinical insight into the problem at hand, or could be left to purely automated methods -— called automated machine learning” to use the data itself to guide analytic choices around model selection and imputation.[7]

Therefore, using a small, prospectively collected clinical dataset with six-month outcomes, we sought to compare different approaches to split sampling, handling of missing values, use of repeated measures across time, and model selection across two long-term outcomes in patients with critical illness.

Methods

We compared the predictive performance of manual versus automated modeling strategies across different approaches for split-sampling, handling of missing data, and the use of repeated measures over time. Individual investigators were responsible for implementing the manual (GW), automated (TL), and automated with temporal features (WL) modeling approaches in a competition-style format. Only one investigator (SB) had access to the outcomes in the testing dataset for evaluation until after all models had been trained.

Population and Data Collection

We used a dataset derived from a prospective cohort study that was conducted from 2013 to 2014 among patients who spent at least three days in an intensive care unit (ICU).[8] Among 303 patients in the original cohort, 301 (99.3%) had sufficient identifiers to be linked to their original chart in the electronic health record (EHR) to query detailed clinical data. Quality of life and mortality after six-month following discharge were determined in the original study using phone interviews and review of the EHR.

Outcomes

For the primary analysis, each modeling approach was used to predict mortality after six months from hospital discharge. In a secondary analysis, each modeling approach was used to predict the patient's quality of life, defined as a binary variable of whether the quality of life was at least as good as it was prior to the ICU admission.

Clinical variables

Each model had access to the following variables for each patient: age (years), gender (man or woman), race (XXX), diagnosis (XXX), ICU type (medical or non-medical), the presence of any Elixhauser comorbidity categories (see Table XXX), and the Apache score on admission to the ICU. Other variables recorded on the first and second days of the ICU admission included glucose (highest), white blood cell count (WBC; highest), hematocrit (%; lowest), serum sodium (lowest), blood urea nitrogen (highest), total bilirubin (highest), albumin (lowest), pH (lowest), PaCO2 (highest), PaO2 (lowest), temperature (F; highest), heart rate (bpm; highest), respiratory rate (highest), systolic blood pressure (mm Hg; lowest), Glasgow coma scale (GCS; lowest), urine output in the past 24 hours (mL), and the fraction of inspired oxygen (FiO2; %).

Model Types

Manual machine learning

We used the scikit-learn software package in Python to train a traditional multivariable logistic regression model, a penalized regression model (L1 and L2 penalties), and an XGBoost classification model.[9] Because of the small sample sizes and relatively large number of features, each model was trained using the first 20 principle components of each training dataset. In all cases, the same decomposition derived in the training dataset was also used for the testing dataset. Tuning parameters for the penalized regression and the XGBoost model were determined by grid search with 5-fold cross validation.

Automated machine learning

To test the performance of models developed through an automated machine learning (autoML) approach, we used the Tree-based Pipeline Optimization Tool (TPOT).[7] Designed for supervised learning problems, TPOT is a user-friendly autoML library that recommends an optimal series of data processing, feature engineering and classification/regression operators. Using tree-based representations of the pipelines, TPOT explores its search space with genetic programming to arrive at a final pipeline that produces the most accurate predict the outcome in cross-validation. Recently, a new option in TPOT called Template was developed to allow the user to define a desired pipeline structure for TPOT to optimize, trading pipeline flexibility for simplicity and reduced computation time.[10]

In this study, we applied both the standard TPOT approach (TPOT Standard) and TPOT with the *Transformer → Classifier* template (TPOT Template) to predict the six-month outcomes and compare their performance to the other methods. We designate 100 generations at maximum for each TPOT run, each generation with the population size of 100.

Feat

To test the performance of an automated machine learning pipeline[11] ... TODO(Bill)

Missing Data

A manual chart review of the EHR confirmed that none of the missing data elements were due to an error in the dataset or in the database query, but rather were due to data not entered into the EHR. We employed three different approaches to handling missing data to understand their effects on model performance in small clinical datasets and to test how they were related to performance using different modeling approaches.

First, we left all data as missing and allowed each modeling approach to deal with the data differently. For the manually trained models, missingness indicator variables were generated for SBP, pH, albumin, and FiO2, considering that their absence would be informative based on clinical experience caring for patients in the ICU. The remaining missing data were imputed using a k-nearest neighbors procedure. For the TPOT models, median imputation was used for all missing variables. For the Feat models, ... TODO(Bill).

Second, we pre-imputed all missing data so that all modeling approaches used the same imputed dataset. Imputation in this case was performed with the `mice` package in R using Bayesian linear regression.[12]

Third, we performed a complete case analysis by entirely excluding the 8 most missing variables then removing observations that had any missingness among the remaining variables.

Split Sampling

The data were divided into training and testing samples using two different strategies. The first strategy used 80% and 20% splits for training and testing, respectively. The second used 50% and 50% splits. Split sampling for both strategies was performed with balanced stratification on ICU type (medical and non-medical) and by quartile of the APACHE score. The observations were sampled such that the 20% test set is a subset of the 50% test set.

Days of data

Numerous ICU mortality prediction models use data from the first 24 hours following admission. Therefore, we aggregated available laboratory values and vital signs from the first 24 hours of the ICU admission in which each patient was enrolled in the initial study.

However, the trajectory of a patient's illness is sometimes not identifiable within the first 24 hours. It is unknown to what degree such temporal data, if at all, improves predictions of long-term outcomes. Therefore, we included an additional set of models with data from both the first and second 24-hour periods of the ICU admission.

For the manually created models, the difference between the two time periods was calculated for SBP, WBCs, FiO2, and UOP. These variables were chosen based on clinical experience as potentially relevant for determining a patient's trajectory. For TPOT and Feat, other features were … TODO(Trang), TODO(Bill)

Model Performance

We evaluated the predictive performance of each model using the scaled Brier Score (BS_s) as a measure that captures both discrimination and calibration.[13] Over N predicted probabilities p for some binary outcome y, the Brier Score is defined as

$$BS = \frac{1}{N}\sum_{i}^{N}(y_i - p_i)^2.$$

However, a useful prediction model should do better than just guessing the baseline event rate as a probability, and so scaling the Brier Score to this uninformed guess motivates using the scaled Brier Score such that

$$BS_s = 1 - \frac{\frac{1}{N}\sum_{i}^{N}(y_i - p_i)^2}{\frac{1}{N}\sum_{i}^{N}(y_i - \bar{y})^2}.$$

A Scaled Brier Score of zero indicates that the model is equivalent to guessing the baseline event rate for each observation, a negative score indicates that the model is worse than this, and a positive score indicates that the model is better.

We generated confidence intervals around each performance estimate by calculating the Scaled Brier Score from 1,000 bootstrapped replicates of the predictions and observations for each model. Differences in performance between models were calculated by estimating the boostrapped differences using 1,000 replicates.

Computational resources

We calculated the time it took to train all models of each type using desktop hardware and utilizing parallel processing when available.

Results

For the primary analysis, testing a six-month mortality prediction using one day of data, an 80/20 split, and an individualized imputation strategy, the traditional logistic regression model yielded the highest point estimate of performance, although did not differ statistically ($p > 0.08$ for all comparisons; Figure 1).

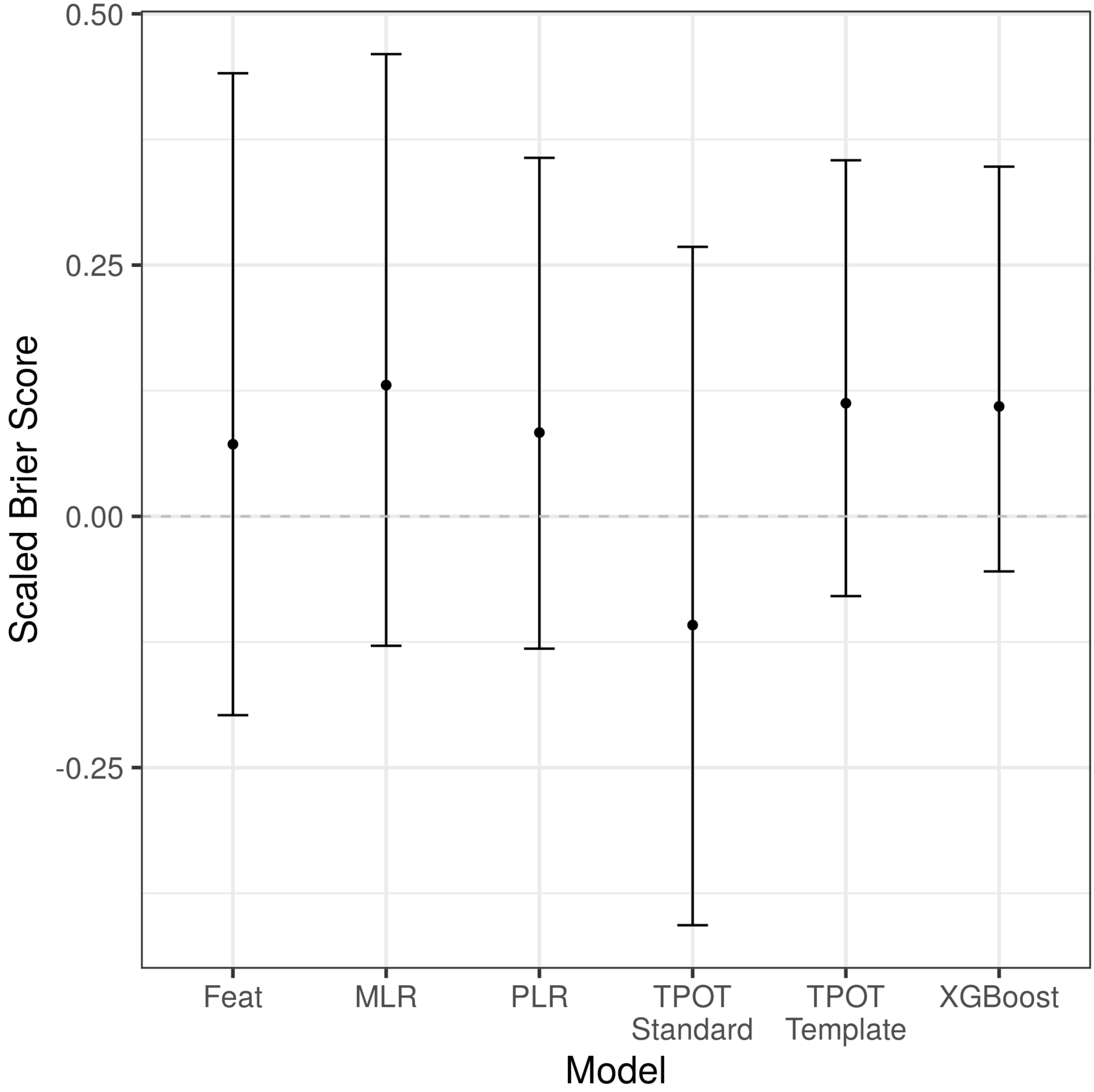

Figure 1: Scaled Brier Score of each modeling approach to predict six-month mortality in the hold-test test set using one day of data, an individualized imputation strategy, and an 80/20 training/testing split. Abbreviations: MLR = multivariable logistic regression, PLR = penalized logistic regression, TPOT = Tree-based Pipeline Optimization Tool.

Computational resources

The (Table 1).

Time required to train all models of each type.

Model type	Time (seconds)	Time (hours)	Relative Time	Hardware
MLR	2.00	<0.001	x1	MacBook Pro (2018), 2.9 GHz Intel Core i9
PLR	5.16	0.001	x2.6	MacBook Pro (2018), 2.9 GHz Intel Core i9
XGBoost	5,439.6	1.5	x2,720	MacBook Pro (2018), 2.9 GHz Intel Core i9
Feat	95,244.82	26.5	x47,613	

Model type	Time (seconds)	Time (hours)	Relative Time	Hardware
TPOT - Template	184,444.6	51.2	x92,222	
TPOT - Standard	258,331.3	71.8	x129,166	

Discussion

These findings reinforce prior work demonstrating no benefit in predictive performance to using machine learning models compared to traditional regression approaches.[14]

References

1. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

2. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. Hernandez Montoya AR, ed. *PLoS ONE*. 2018;13(3):e0194889. doi:10.1371/journal.pone.0194889

3. Steyerberg EW. Validation in prediction research: the waste by data splitting. *Journal of Clinical Epidemiology*. 2018;103:131-133. doi:10.1016/j.jclinepi.2018.07.010

4. Engerström L, Nolin T, Mårdh C, et al. Impact of Missing Physiologic Data on Performance of the Simplified Acute Physiology Score 3 Risk-Prediction Model*. *Critical Care Medicine*. 2017;45(12):2006-2013. doi:10.1097/ccm.0000000000002706

5. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*. 2013;58(1):63-72. doi:10.1016/j.artmed.2013.01.003

6. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*. 2016;102:1-5. doi:10.1016/j.resuscitation.2016.02.005

7. Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Kidd LC, Moore JH. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In: *Applications of Evolutionary Computation*. Springer International Publishing; 2016:123-137. doi:10.1007/978-3-319-31204-0_9

8. Detsky ME, Harhay MO, Bayard DF, et al. Discriminative Accuracy of Physician and Nurse Predictions for Survival and Functional Outcomes 6 Months After an ICU Admission. *JAMA*. 2017;317(21):2187. doi:10.1001/jama.2017.4078

9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. October 2011. https://hal.inria.fr/hal-00650905. Accessed March 27, 2020.

10. Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Kelso J, ed. *Bioinformatics*. 2019;36(1):250-256. doi:10.1093/bioinformatics/btz470

11. La Cava W, Singh TR, Taggart J, Suri S, Moore JH. Learning concise representations for regression by evolving networks of trees. *arXiv:180700981 [cs]*. March 2019. http://arxiv.org/abs/1807.00981. Accessed March 27, 2020.

12. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations inR. *J Stat Soft*. 2011;45(3). doi:10.18637/jss.v045.i03

13. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/ede.0b013e3181c30fb2

14. Gravesteijn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of Clinical Epidemiology*.

March 2020. doi:10.1016/j.jclinepi.2020.03.005