A Comparison of Machine Learning Models for Mortality Prediction: National Health and Nutrition Examination Survey (NHANES III)

Roy Williams[1], Miguel Alonso Jr.[2], Prasad Bhoite[3], Emir Veledar[1], Zoran Bursac[1]
[1]Department of Biostatistics
[2]Department of Computing and Information Science
[3]Department of Humanities, Health and Society
Florida International University, 11200 SW 8th St, Miami, FL, 33199

Abstract

Predicting all-cause mortality is a major goal of public health and often medicine in general. Variables such as high blood pressure, advanced age, smoking status, and other factors have been associated with an increased risk in all-cause mortality. The CDC's Third National Health and Nutrition Examination Survey (NHANES III) is a large nationwide probability sample of 39,695 persons. NHANES provides sample information regarding relevant health metrics for example blood pressure or age, in addition to patient vital status during the time period. To retrospectively determine which features are most relevant in predicting mortality, a selection of machine learning models including logistic regression, decision tree classifier, and a random forest classifier were trained on the dataset and compared based on accuracy, precision, F1 score, and subsequently area under a receiver operating characteristic curve. Overall, the random forest classifier seemed to provide the best predictive performance with an accuracy of 0.99, exceeding the 0.95 threshold. Scientists could apply this methodology to guide mortality prediction or other specific outcomes.

Key Words: Machine learning, NHANES, logistic regression, decision trees, random forest

1. Introduction

Previous analyses have been conducted using the NHANES dataset to examine mortality, although these have been primarily survival analyses. A recent analysis from 2019 observed that all-cause mortality was associated with mean blood pressure, age, metabolic syndrome status, and was also influenced modestly by sex (1). A 2013 paper using NHANES in turn found that smoking status is also associated with all-cause mortality (2). This is consistent with previous research findings, and what has become conventional clinical wisdom that smoking can contribute to illness. Notably, neither of these papers applied machine learning techniques to analyze mortality in NHANES (1,2). Where machine learning models have been applied to mortality prediction in medicine, they have been applied prospectively or concurrently in settings such as an intensive care unit (ICU) following an unplanned extubation, as a recent landmark paper published in Nature shows (3). While impressive, an unplanned extubation in an ICU is a very specific situation and leaves a more general question concerning mortality prediction

unanswered. Namely, how well can commonly deployed machine learning algorithms predict mortality in a large national dataset? In this analysis, a logistic regression, random forest classifier, decision tree classifier, Naive Bayesian Classifier, XG Boost, and CatBoost were deployed and compared upon their prediction accuracy.

2. Methods

NHANES itself is unique based on the fact it combines diagnostic information with risk factor data obtained from surveys involving the same participants. This was the primary reason the NHANES dataset was selected. NHANES III contains 15838 rows and 32 columns composed of 32 separate risk factors. The average set of values for each variable for those who lived, and those who dead were summarized in the dataset.

The primary reason a logistic regression, decision tree and random forest models are selected is because we were attempting to solve a binary classification and prediction problem. Since these are typically well-regarded methods for classification and prediction, they allow us to predict mortality in the NHANES dataset. Support vector machines and neural network-based models were not included due to time limitations of the project. It remains likely that models could be expected to preform similarly well, and future analysis may include these procedures.

Prior to analysis, data was balanced by minority oversampling and K-fold cross validation with k=10 used to generate more reliable testing and training sets. Feature selection was performed to determine which features were of the most relevant interest in predicting all-cause mortality. Initially, factors shown to be associated with all-cause mortality such as age and blood pressure were used.

Generally, models with an accuracy of 95% or greater are recommended in prediction and classification. Thus, this was the primary benchmark upon which we evaluate the performance of our models, although comparisons will also be shown regarding F1 score, ROC curve and confusion matrices.

Data preprocessing posed the greatest detriment to the project and thus required the most time. The first issue was the sheer amount of missing data. Missing data was analyzed by each variable. Overall, NHANES was missing a 63% of low-density-lipoprotein (LDL) data. This accounted for roughly 9,500 missing entries. This was dropped from the analysis. Secondly, there was substantial collinearity between the variables, for example, the dataset contained not only continuous data for LDL and age, but also stratified data for these variables. All stratified data variables were also eliminated from the analysis. The next step was to properly encode our binary target and feature variables. Binary encoding was done for mortality status, obesity, smoking status, high glucose, metabolic syndrome, fasting, hypertensive status, and high triglyceride status. Dummy variable encoding was done for diabetes status, where the groups analyzed were diabetic, pre-diabetic, and non-diabetic, and for race/ethnicity, where the groups analyzed were white, black, Hispanic and multi-race.

The next step was balancing data for our target variable, mortality status. The initial split between “dead” and “alive” was 2162 dead and 9244 alive. This initially posed significant problems when the analysis was run without balancing and without cross validation. For

example, the logistic regression AUC was initially 0.73, compared with the rectified 0.82. In order to rectify the balance issue, minority oversampling was used. Minority over sampling resulted in an even split with 9244 alive and 9244 classified as dead. The next step was model selection where a full model involving diabetes status, age, SES, cholesterol, triglycerides, HDL, BMI, serum albumin, diastolic blood pressure, sex and smoking status was selected based on previous research and public health domain expertise. The same model was applied using all three techniques.

3. Results

Following figures show the results of various methods we applied in order to predict and classify mortality.

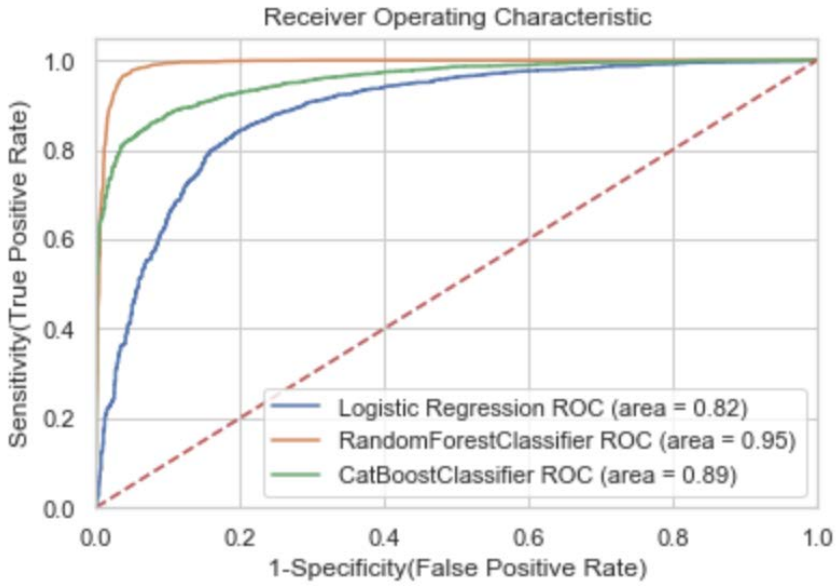

Figure 1: In comparing a Logistic Regression with Random Forest Classifier and CatBoost Classifier, the Random Forest Classifier had the greatest AUC at 0.95. CatBoost Classifier had an AUC of 0.89, while the Logistic Regression had an AUC of 0.82. The Random Forest had the highest AUC of all models tested.

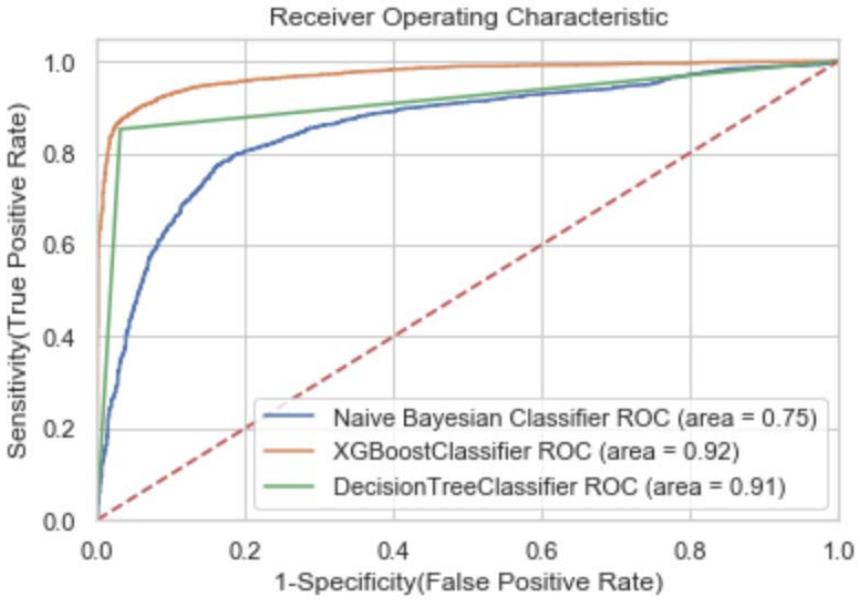

Figure 2: In comparing a Naive Bayesian Classifier, XGBoost Classifier and a Decision Tree Classifier, the XGBoost classifier had the greatest AUC with an AUC of 0.92. The Decision Tree Classifier was a close second with an AUC of 0.91. The Naive Bayesian Classifier was the weakest performing model of all tested with an AUC of 0.75.

	precision	recall	f1-score	support
0	0.92	0.97	0.95	2809
1	0.97	0.92	0.94	2738
accuracy			0.95	5547
macro avg	0.95	0.95	0.95	5547
weighted avg	0.95	0.95	0.95	5547

	precision	recall	f1-score	support
0	0.99	0.99	0.99	924
1	0.99	0.99	0.99	924
accuracy			0.99	1848
macro avg	0.99	0.99	0.99	1848
weighted avg	0.99	0.99	0.99	1848

Figure 3: A comparison of the two top performing models, XGBoost (top) and Random Forest Classifier (bottom). Overall, the Random Forest Classifier provided the most accuracy with an accuracy of 0.99.

```
[(0.47865909333864204, 'age'),
 (0.09384175438291069, 'uralb'),
 (0.07107384116251834, 'ses'),
 (0.0682203641315275, 'chol'),
 (0.0642660893932913, 'bmi'),
 (0.06261431378424906, 'trig'),
 (0.0574211042963317, 'dbp'),
 (0.05479569237771014, 'hdl'),
 (0.019690082992269757, 'dm_diabetes'),
 (0.017465132172917064, 'smokehx'),
 (0.01195253196763259, 'sex')]
```

Figure 4: Feature importance of the top performing Random Forest Classifier model. Overall, age was the most significant feature.

4. Discussion

To the best of our knowledge, this is the first attempt at applying machine learning models to predict mortality in the NHANES dataset. Of the three methods utilized, the random forest classifier produced the best accuracy, of over 99%. This is similar to the results obtained using ICU data after an unplanned extubation, where the random forest preformed the best of the models used (3). This was compared with 93% accuracy for the decision tree classifier and 81% for the simple logistic regression. The only model that preformed with higher accuracy than the sought after 95% accuracy threshold was the random forest classifier. The ROC AUC was also highest for the random forest classifier with a 0.99%. This was compared with 93% and 82% for the decision tree model and logistic regression model, respectively. Overall, age was the best predictor of all-cause mortality.

Regarding feature importance, there are a few notable results that require further analysis in subsequent studies. Smoking status strikingly offered little predictive capacity (Figure 4). There are many potential reasons for this. Foremost, it is possible individuals near death are too sick to smoke and thus are “non-smokers”. We do not have information concerning their past tobacco usage, or if they used alternate tobacco products. Secondly, due to the fact this is a binary variable, we do not have information concerning the dosage or length of smoking for smokers. Clearly, smoking a cigarette a day vs a pack a day would produce different effects. Finally, there are strict longitudinal limitations involved in this data set. For example, we do not know when individuals are dying, only that their death occurred somewhere between the years 1988 and 1994. It is possible smokers on average die sooner, just not in significantly larger numbers to detect. This is also not concordant with a previous analysis of NHANES (2).

Regarding the longitudinal limitations and the uncertain importance of smoking status, future studies could include a survival analysis and focus primarily on smoking status as a predictive

variable. Furthermore, ablation studies could be useful in this context. It remains to be seen what removing certain features will do to smoking status. This is fertile ground for future analyses. Future studies can also include neural networks and SVMs, which were left out due to time constraints.

References

(1.) Yu, W. W., Randhawa, A. K., Blair, S. N., Sui, X., & Kuk, J. L. (2019). Age- and sex-specific all-cause mortality risk greatest in metabolic syndrome combinations with elevated blood pressure from 7 U.S. cohorts. *PloS one*, *14*(6), e0218307. doi:10.1371/journal.pone.02183

(2.) Patel, C. J., Rehkopf, D. H., Leppert, J. T., Bortz, W. M., Cullen, M. R., Chertow, G. M., & Ioannidis, J. P. (2013). Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States national health and nutrition examination survey. *International journal of epidemiology*, *42*(6), 1795–1810. doi:10.1093/ije/dyt208

(3.) Hsieh, M.H., Hsieh, M.J., Chen, C. *et al.* Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. *Sci Rep* **8,** 17116 (2018) doi:10.1038/s41598-018-35582-2

(4.) Parikh RB, Manz C, Chivers C, et al. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Netw Open.* 2019;2(10):e1915997. doi:10.1001/jamanetworkopen.2019.15997