Evaluation of the Effectiveness of Applying Different ML Algorithms on MIMIC-IV Dataset to Reduce Racial Bias in ICU Mortality Prediction

Elijah Chou, Tianqi Bao, Tiantian Li

Emory University

Abstract

Despite wide utilization of severity scoring systems for case-mix determination and benchmarking in intensive care units, the possibility of scoring bias across ethnicities has yet to be thoroughly examined. Recent guidelines on the use of illness severity scores to inform triage decisions for allocation of scarce resources such as mechanical ventilation during the current coronavirus pandemic warrant examination for possible bias in these models. This study will specifically examine the SOFA severity scoring system that is currently one of the most widely used scoring systems across the United States.

In this paper, alternative models are used in predicting the mortality of patients including KNN classifier, decision tree classifier, logistic regression classifier, random forest classifier and naive bayes classifier. Models are then accessed based on accuracy, ROC, and AUROC. After selecting the optimized random forest model based on the assessment, we then calculated AUROCs and 95% confidence intervals for individual races to test for racial bias by comparing the patterns to those found in the AUROCs of the original logistic regression model trained only by SOFA scores. Upon calculation of AUROCs and their respective confidence intervals, there was no significant difference in the racial bias between the two models. There was still an overall increase in AUROC in all studied races after preprocessing and implementation of the random forest model.

Introduction

Severity scoring systems are employed in the intensive care unit (ICU) to perform severity adjustment for the purposes of benchmarking and research (Vincent and Moreno 2010). In particular, the Sequential Organ Failure Assessment (SOFA) score was developed based on expert opinion, incorporating organ function scores from six organ systems to characterize severity state in sepsis but has been repurposed to predict patient outcomes (Vincent et al. 1996). It has generally been assumed that these systems are fair and objective in terms of their use across different ethnic groups. However, while it is known that such models may perform differently among disparate geographic populations or between different centers (Poncet et al. 2017), the assumption of scoring neutrality among ethnic groups within a given population has not been closely examined. Disparities in ICU outcomes may result from pre-admission clinical factors, socioeconomic determinants, the quality of ICU care, and cultural practices (Quindemil et al. 2013; Orlovic et al. 2019). Another possible source of disparity emanates from the use of biased algorithms (Wiens et al. 2020; Obermeyer et al. 2019; McLennan 2020).

The current COVID-19 pandemic raises two intersecting issues that demand a closer evaluation. First, relatively higher mortalities have been observed ethnic populations, specifically African Americans (Ferdinand and Nasser 2020). Second, severity scores have been proposed by professional societies and various policy groups to be incorporated into triage systems for potential scarce resource allocation (VENTILATOR ALLOCATION GUIDELINES 2015; State of Michigan 2012). It is therefore imperative to determine whether biased scoring systems could be adding to existing baseline disparities in healthcare.

Background

Previously, researchers have investigated the performance of three severity scoring systems across ethnic groups in two large ICU databases and found statistical evidence that suggested that illness severity scores did not discriminate for severity of disease, but were poorly calibrated for Blacks and Hispanic patients where these scores over predicted mortality (Sarkar et al. 2021). In this study, a simple logistic regression model was used to predict mortality using only the average scores as a model feature. In another retrospective study, researchers concluded that non-Hispanic Black patients were more likely to be denied medical resources if the SOFA score, a prominently used severity scoring system, is utilized. This further justifies the need for a better machine learning model to be implemented in limited hospital resource allocation (Raschke et al. 2021).

In another study, researchers used patients' three latest SOFA scores recorded during their ICU stay as input for various machine learning models and ensemble learning models and found that an ensemble model of linear and logistic regression achieved the highest AUC compared to other previous works (Aperstein et al. 2019). They also found that by adding additional gastrointestinal failure scores, the ensemble model and others improved in their prediction. We hope that by adding more parameters in our models, we can produce a model with high accuracy and reduced bias in prediction.

Methods

Step 1: Model Selection

Since the output label for our dataset is binary, and the features are numerical values, we considered the following models:

1. **KNN classifier:** predicting the output label by k-number of nearest neighbors of the selected item, choosing the mode of two labels and using it as the predicted result.
2. **Decision tree classifier:** building the tree by selecting the best split points through the calculation of gini index/entropy. Predicting the result by inputting features of the selected item through the route of the tree branches.
3. **Logistic regression classifier:** building the model by updating coefficients of the model. While we have the final model, using it to predict by inputting all the features of the selected item. The output would be a float number, which is the probability that the result is 1. Thus, if the output ≥ 0.5, we predict that to be 1. Otherwise, we predict that to be 0.
4. **Random forest classifier:** building clusters of decision tree classifiers and outputting the average result as the final prediction result.
5. **Gaussian naive bayes classifier:** predicting based on Bayes theorem formula.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

6. **Ensemble method:** predict each record by taking into account the prediction done on that sample by all previous classifiers. Predicted result comes from taking the mode of previous predictions.

Step 2: Hyper parameter tuning: select the best parameters for each algorithm

For each algorithm, using k-fold cross validation to estimate the performance of different parameters, and then using misclassification error(accuracy) for evaluation, choosing the parameters that have the lowest MSE or highest accuracy for prediction. Since we only

have parameters in KNN classifier, decision tree classifier, and random forest classifier, we will test on these three while leaving out the logistic regression classifier and gaussian naive Bayes classifier in this step.

1. KNN:

   Tested on:

   - k= 5, 10, 50, 100, 500, 1000

   Best parameter: k = 10.

2. Decision tree:

   Tested on:

   - criterion = ["gini", "entropy”]
   - max_depth = [5, 10, 15, 20, 30]
   - min_samples_leaf = [10, 50, 100, 500, 1000]

   Best parameter: [‘entropy’, 10, 100]

3. Random forest

   Tested on:

   - number of trees = [10, 50, 100, 200, 400]
   - criterion = ["gini", "entropy”]
   - maximum depth = [5, 10, 15, 20, 30]
   - minimum leaf = [10, 50, 100, 500, 1000]
   - max number of features = looping through all features
   - bootstrap or not = [True, False]
   - using out-of-bag samples or not = [True, False]

   Best parameter: [400, ‘entropy’, 20, 50, 17, True, True]

   (note: while testing on if using out-of-bag samples or not, the premise is to set “bootstrap or not” to be True.

Experiments/Results

Original data description and data cleaning steps taken

The original dataset contains over 40,000 samples with 17 features including information about the patients such as ID, marital status, ethnicity, etc., along with statistics regarding patients' status at admission, and the length of the recorded hospital stay. The target label is the ‘hospital_expire_flag’ column, in which 1 represents the death of the patient while 0 represents the patient recovered. There are a total of 5 time-stamp features, 7 categorical features, 2 continuous numerical features, and 3 features regarding patients’ identity. In order for the data to be more interpretable, we adopted several data cleaning steps.

1. Removing irrelevant features: To reduce the unnecessary model complexity, we removed irrelevant features including identification number, anonymized timestamps, death time (if applicable), and language. Since data went through an anonymization process, we do not

know if the time-stamp features still preserve valid information. Hence, we decided that it would be safer to leave the time-stamp features out of consideration.

2. Converting categorical data: Some models require data to be numerical, so we did the following to convert all categorical data to numerical:
    - 'Insurance': categorical data with two levels:

      ['Medicare', 'other']

      to binary data with 0 representing other and 1 representing Medicare.

    - 'admission_location': categorical data with 11 levels:

      ['TRANSFER FROM HOSPITAL', 'EMERGENCY ROOM', 'PHYSICIAN REFERRAL', 'PROCEDURE SITE', 'PACU', 'WALK-IN/SELF REFERRAL', 'CLINIC REFERRAL', 'INTERNAL TRANSFER TO OR FROM PSYCH', 'AMBULATORY SURGERY TRANSFER', 'TRANSFER FROM SKILLED NURSING FACILITY', 'INFORMATION NOT AVAILABLE']

      to binary data with 0 representing non-emergency room and 1 representing emergency room.

    - 'Marital_status': categorical data with five levels:

      ['WIDOWED', 'DIVORCED', 'MARRIED', nan, 'SINGLE']

      to binary data with 0 representing others and 1 representing married.

    - Created dummy variable columns for 'ethnicity'.

      ['WHITE', 'ASIAN', 'HISPANIC/LATINO', 'BLACK/AFRICAN AMERICAN']

    - Created dummy variable columns for 'admission_type'.

      ['OBSERVATION ADMIT', 'URGENT', 'EW EMER.', 'DIRECT OBSERVATION', 'SURGICAL SAME DAY ADMISSION', 'AMBULATORY OBSERVATION', 'EU OBSERVATION', 'DIRECT EMER.', 'ELECTIVE']

3. Removing highly correlated features: To reduce redundancy in the dataset that could lead to unnecessarily complex models, we attempted to remove highly correlated features by mapping the correlation between all pairs of features on a heat map. We plotted the correlation between all pairs of features, including labels, using heatmap (see fig. 1). None of the features has a high correlation (>0.8) with another feature, so we did not remove any feature.
4. Normalizing the data: Normalized data using sklearn.standardscaler. Each feature in the resulting dataset would have a mean of 0 and a standard deviation of 1.

   #Details of feature selection and extraction can be found in feautre_extract.py

5. Train/Test split: Splitted the dataset into training set and test set, with test size = 0.3*size of original dataset. #Details of code can be found in train_test_splitting.py

Empirical results

After finding the optimal parameters for each of the model proposed above, we predicted the test data using the models and collected accuracy scores along with false positive rate (FPR) and true

positive rate (TPR). Then, we used the ROC curve (see Fig. 2) to visually compare the performance of each model on our dataset. Each point along an ROC curve represents a trade-off between sensitivity (or TPR) and specificity (or 1-FPR). We prefer a model that results in a ROC curve that approaches the top left corner more, or in other words, has a larger area under the ROC curve. We can visually observe that the area under the ROC curve for KNN is significantly smaller than that for other models, indicating that KNN is the least reliable model for predicting the dataset. We also used AUROC and accuracy scores to compare the models (see Table 1). AUROC stands for Area Under ROC Curve. As mentioned earlier, a model that generates a larger AUROC is a more reliable model.

	Accuracy	AUROC
Logistic Regression classifier	0.9328	0.8471
Decision Tree classifier	0.9319	0.8352
Random Forest classifier	0.9328	0.8520
Gaussian NB classifier	0.5493	0.8267
Ensemble Model	0.9327	N/A

Table 1

From the above table, we can see that both logistic regression classifier and random forest classifier achieve the highest accuracy, while random forest achieves the most optimal AUROC. Thus, we conclude that Random Forest is the best model to predict our dataset. #Details of code can be found in model_assessment.py

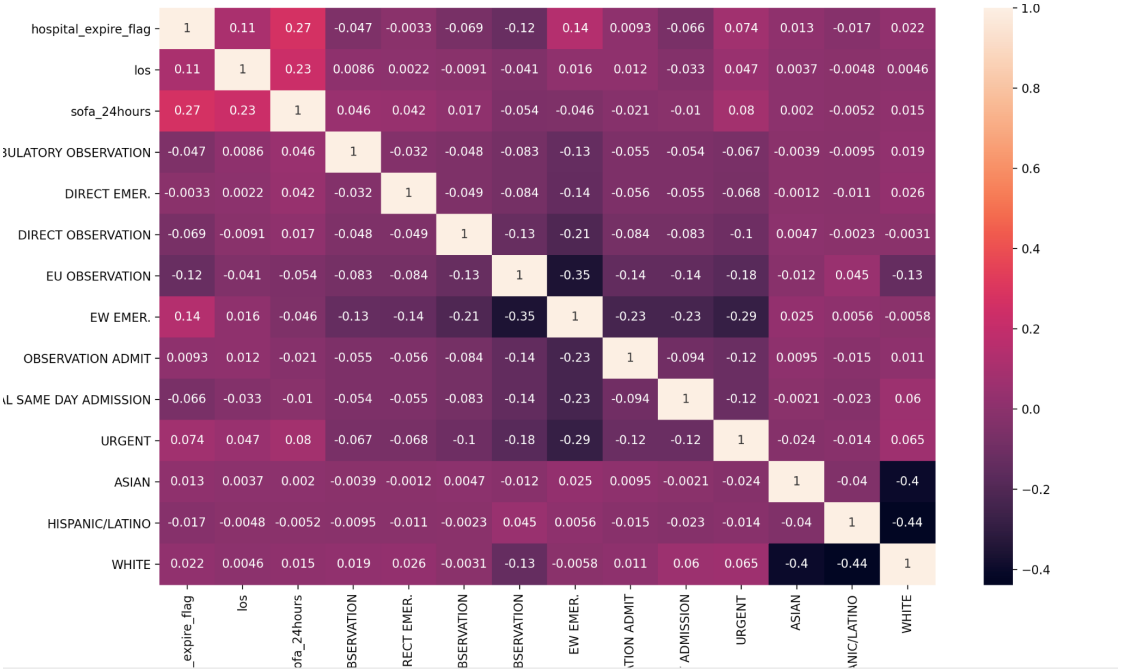

Fig. 1: Heatmap representing correlation between each pair of features.

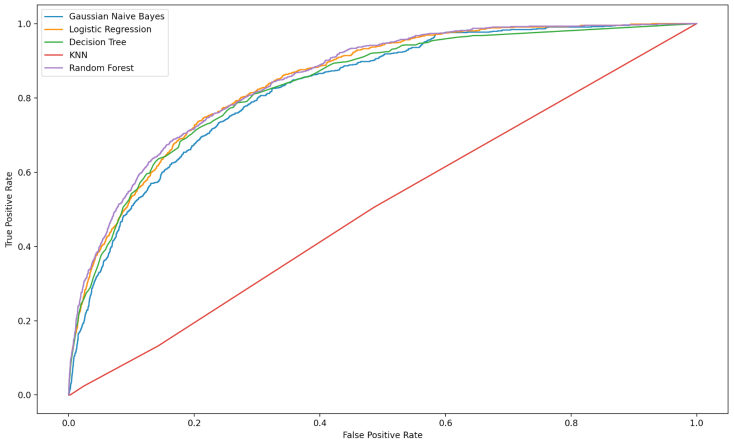

Fig. 2: ROC curves of ML models used in experiments.

For the racial bias analysis, we generated the following forest plots as a test of the model's discrimination. The AUROCs and 95% confidence intervals were calculated with the "AUROC Confidence Intervals.ipynb" Jupyter notebook in the project directory, and the forest plots were generated with Microsoft Excel.

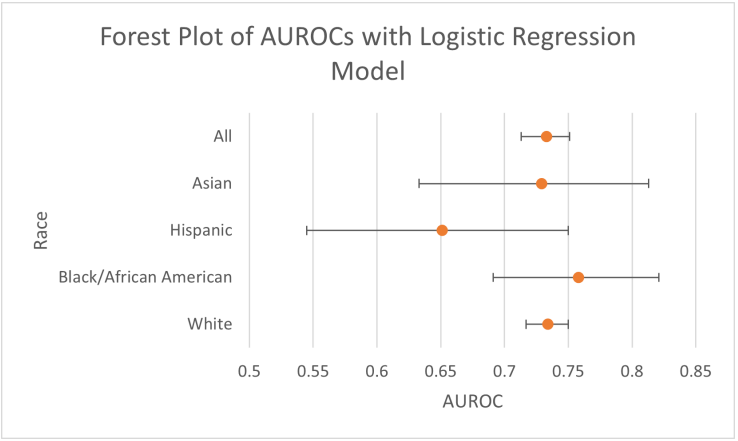

Fig. 3: Forest plot of AUROCs for each race in testing set with Logistic Regression Model trained with only SOFA scores.

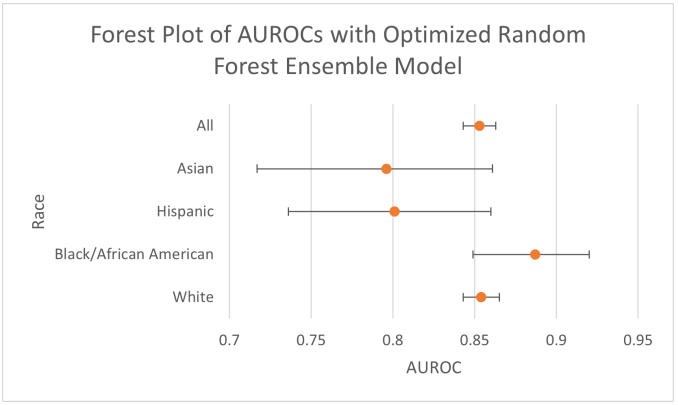

Fig. 4: Forest plot of AUROCs for each race in testing set with the Optimized Random Forest Model.

In order to determine whether our experiments improved on the studied logistic regression model found in the main paper this project is based on, we compared the above forest plots (Figures 3 and 4) to see if there was a qualitative difference in the performance of the models based on the AUROCs of each model using both individual and all races in the test data. According to the analysis of the aforementioned paper, racial bias is detected if any of the individual race 95% confidence intervals are out of range from the other intervals, which would indicate a significant difference in the model's predictive performance for that particular race (Sarkar et al. 2021). When examining Figure 3, it can be noted that there is no observable outlier in confidence intervals for the original logistic regression model trained by only the SOFA scores. Figure 4 also shows a similar pattern upon observation. However, there is a decrease in the relative performance of the random forest's model on the Asian test set compared to the relative performance of that in the logistic regression model, which is worth noting despite it being a statistically insignificant difference. Even so, it can also be seen that the overall performance of the model increased after using the optimized random forest model when comparing the AUROCs from Figure 4 to those in Figure 3.

Discussion

A key takeaway from our experimental results is that including more features other than the SOFA score and processing the data with feature selection and extraction techniques can achieve a very high accuracy rate and a reasonably high AUROC. Even just comparing the overall AUROC of the logistic regression model trained only by SOFA scores with that of the logistic regression model trained with more features, there is an increase in the performance of the model.

While we do not see a significant difference in any racial bias in the predictive modeling of ICU hospital mortality, we do see an overall increase in the performance of the model in terms of AUROCs. Another main takeaway from our experiments is that, while there does not seem to be any significant improvement or worsening of racial bias in predictive modeling when utilizing preprocessing techniques and a more robust machine learning algorithm, there will still be an increase in the overall performance of the model after implementing such measures. This can be both a good and bad thing: the finding is beneficial in the sense that we would be able to implement more accurate predictions in the ICUs in hospitals for possibly better outcomes and more effective resource allocation in these units without introducing more racial bias in these predictions, but it is also not favorable since these experimental methods do not seem to help reduce racial bias in predictions. Not only should these experiments be replicated with more models or techniques to support or reject the findings in this study, but other ways of improving racial bias should be explored as well.

One possible next step from this project would be to examine the other available datasets stored within the MIMIC-IV database for more useful features that may help in improving the predictive capabilities of machine learning models. There are still many other variables we did not include in this study that can be found in the dataset, and the addition or removal of some variables may prove to be advantageous in improving not only the accuracy of the predictive model but also possibly reducing racial bias in predictions. This could also be done by using more ICU datasets that other hospital systems have. For example, these experiments could be replicated on Emory ICU datasets if the data could be accessed with permission.

Contributions

Tianqi Bao:

Responsible for the method section including setting up various models for testing and using k-fold to test on different parameters on each model, comparing their accuracies in order to find the best parameters. Typesetting for the report. Discussing with other parts of the project. Responsible for coding, presenting, and reporting the above topics.

Tiantian Li:

Replicating previous approach done on the dataset using a simple logistic regression model with one feature (SOFA score) and separating the prediction result by ethnicities. Data cleaning, feature extraction, selection, and normalization. Train-test splitting and generating appropriate csv files. Model assessment using ROC, AUROC, and accuracy scores. Interpreting model assessment results from graphs and tables. Responsible for coding, presenting, and reporting the above topics.

Elijah Chou:

Conducting background research on the topic and retrieving access to the MIMIC-IV database with the assistance of his principal investigator, Dr. Judy W. Gichoya, from the Emory Healthcare Innovation and Translational Informatics Laboratory of the Emory School of Medicine. In charge of compiling necessary data features from available data to facilitate the preprocessing stage of this project. Responsible for presenting the topic, coding for initial data cleaning/organization and racial bias analysis, and reporting the results shown above.

Code

The following link leads to the Google Drive folder containing all code and datasets that were used throughout the duration of the project:
https://drive.google.com/drive/folders/12PUJXMa2NbR4viv3wEXSxhMbE6QwKOSQ?usp=sharing

References

Aperstein, Y., L. Cohen, I. Bendavid, J. Cohen, E. Grozovsky, T. Rotem, P. Singer, 2019 Improved ICU mortality prediction based on SOFA scores and gastrointestinal parameters. PLoS One. 14(9): e0222599.

Ferdinand, K., and S. Nasser, 2020 African American COVID-19 Mortality A Sentinel Event. J Am Coll Cardiol. 75: 2746-2748.

McLennan, S., M. Lee, A. Fiske, L. Celi, 2020 AI ethics is not a panacea. Am J Bioeth.

Obermeyer, Z., B. Powers, C. Vogeli, S. Mullainathan, 2019 Dissecting bias in data-driven algorithms for healthcare. Science. 366: 447-53.

Orlovic, M., K. Smith, E. Mossialos, 2019 Racial and ethnic differences in end-of-life care in the United States: Evidence from the Health and Retirement Study (HRS). SSM – Popul Heal. 7: 100331.

Poncet, A., T. V. Perneger, P. Merlani, M. Capuzzo, C. Combescure, 2017 Determinants of the calibration of SAPS II and SAPS 3 mortality scores in intensive care: A European multicenter study. Crit Care. 21: 1–10.

Quindemil, K., M. Nagl-Cupal, K. Anderson, H. Mayer, 2013 Migrant and minority family members in the intensive care unit. A review of the literature. HeilberufeScience. 4:128–135.

Raschke, R.A., S. Agarwal, P. Rangan, 2021 Discriminant Accuracy of the SOFA Score for Determining the Probable Mortality of Patients With COVID-19 Pneumonia Requiring Mechanical Ventilation. JAMA. 325(14):1469-1470.

Sarkar, R., C. Martin, H. Mattie, J. W. Gichoya, D. J. Stone, L. A. Celi, 2021 Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. The Lancet Digital Health. 3(4): E241-E249.

State of Michigan, Department of Health O of PHP. Guidelines for Ethical Allocation of Scarce Medical Resources and Services During Public Health Emergencies in Michigan. 2012.

VENTILATOR ALLOCATION GUIDELINES New York State Task Force on Life & the Law New York State Department of Health. 0215; : 11.

Vincent, J. L., R. Moreno, 2010 Clinical review: Scoring systems in the critically ill. Crit Care 14: 1-9.

Vincent, J. L., R. Moreno, J. Takala, 1996 The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 22: 707-10.

Wiens, J., W. N. Price, M. W. Sjoding, 2020, Diagnosing bias in data-driven algorithms for healthcare. Nat Med 26: 2