Predicting In-Hospital Mortality Risks with Explainable AI

Christopher White
CSC 4810, Department of Computer Science
Georgia State University
Atlanta, Georgia, USA
itsme@controlaltnerd.com

Abstract – **Artificial intelligence tends to engender a healthy amount of distrust among non-technical users who do not understand the inner workings of complex machine learning models. This lack of trust is especially noticeable in areas where decisions have profound impacts on health and safety. Explainable AI has the potential to promote trust in and understanding of tools built on complex models. This study looks at the potential for explainable AI to bridge the trust gap in medical settings and how it might look in practice.**

I. Introduction

Explainable AI is an area of artificial intelligence that focuses on enabling transparency into the reasoning behind the output of complex AI models, which are naturally "black boxes" to people who are non-experts in AI domains. Such transparency would certainly be desirable in any situation, but is of special importance in areas where humans are expected to rely on AI when making decisions that affect health and safety. Medicine is a field where its inherent complexity makes it a promising area for AI-based innovation, but proves to be a challenging target due to its impact on human life and the lack of technical expertise possessed by healthcare providers in matters of AI [1].

The overarching goal of healthcare providers when caring for patients is to restore health and quality of life while minimizing risks, especially that of death. The ability to predict patient mortality in-hospital is of great benefit to providers, and length of stay (LOS) is one of the most common metrics used to determine a patient's morbidity and mortality risks. LOS predictions have been well-explored with a variety of machine learning methods, such as decision trees, random forests, and convolutional neural networks [2],[3]. However, the correlation between inpatient complications and LOS is not always clear [2], and LOS is also used to estimate administrative factors that are of more concern to managers than doctors [3]. Indeed, one study of over 12,000 hospital admissions found that when physicians tend to prioritize shorter LOS, a patient's 30-day risk of mortality increases greatly [4].

Thus, focusing more directly on risks to the patient and the factors that drive those risks seems likely to improve health outcomes for patients. This is a task well-suited to explainable AI, which can bridge the gap between innovation and caution with trust, by offering domain-specific insight into the process by which a highly technical tool arrives at a particular output that could affect a patient's well-being.

This research focuses on an approach to mortality risk prediction using explainable AI that allows healthcare providers to examine a given patient's risk based on complex data, and to easily assess the reasoning behind a model's predictions. Explainability in this system is driven by detailing which features are of greatest significance for each prediction, and to what extent they affect the prediction.

II. Methodology

A. System Design

At a high level, the system contains four primary components. First is the data pre-processing pipeline in which data undergoes feature engineering, transformation, and imputation. Next is the predictive layer, a "black box" LightGBM gradient-boosting model. Third is the explainable layer for analyzing the reasoning behind the black-box model's predictions and providing a domain-relevant analysis. Last is the Python application that allows user interaction with the system.

B. Dataset

The dataset used in this study to train and validate the predictive model is the GOSSIS-1-eICU subset of the Global Open Source Severity Illness Score (GOSSIS-1) dataset, obtained from PhysioNet[1]. It contains 131,051 patient records with 216 features, collected from more than 200 hospitals across the United States from 2014-2015. All data was collected within 24 hours of patient admission to an ICU and excludes:

- Patients less than 16 years old
- Patients whose heart rate was not recorded
- Patients who were readmitted to the ICU
- Patients for whom the mortality outcome was not recorded

The GOSSIS-1-eICU set provided a preprocessed dataset that did not require training and testing data to go through the described preprocessing pipeline; however, new data, including records sourced elsewhere and synthesized data, that is used for predictions does require preprocessing. Minor

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

[1] Used with permission of PhysioNet under terms of research usage.

changes were made to the preprocessed dataset to remove metadata categories that had no effect on the target feature. 1,000 records were split from the dataset and reserved for manual testing, and the remaining records were split between training (80%) and evaluation (20%).

C. *Predictive Layer*

Gradient boosting is a technique that combines several weak learners, often decision trees, in an ensemble, where each model learns from the training of the prior model to gradually increase its accuracy. Gradient boosting is commonly applied to classification problems, especially those that rely on tabular datasets.

In this study LightGBM was chosen for gradient boosting because it excels at handling very large datasets and has built-in support for explainability. The model was tuned to focus on recall ability without sacrificing generalization. To control tree complexity, each sub-model was limited to 31 leaves and a maximum depth of 7, while requiring at least 20 samples per split and 500 samples per leaf to avoid overfitting.

The model's learning rate was set to 0.03 across 1,000 boosting rounds with early stopping to maximize convergence. Feature subsampling was used at a rate of 20%, and together with moderate regularization also helped to mitigate overfitting. Class imbalance was handled by applying the ratio of negative to positive samples as a class weight.

During this study an ensemble model was considered as well that would combine the LightGBM model with a neural network designed for analyzing tabular data, and make a final prediction using another gradient-boosting algorithm. However, the ensemble model did not evaluate well despite extensive tuning, so the LightGBM model was chosen to stand alone in the predictive layer.

D. *Explainable Layer*

Several explainability methods were explored for this study, and ultimately SHapley Additive exPlanations (SHAP) was chosen for its simplicity, human readability, and ease of integration with LightGBM. This method uses Shapley values from game theory to provide a significance score to various features that drive a model's prediction. SHAP provides the model with easily readable explanations – features with positive values contribute to a prediction, while features with negative values detract from a prediction.

E. *User Interface*

The application devised for model interaction in this study was built with Python 3.13, using the following packages for model development and deployment: pandas, numpy, lightgbm, shap, sklearn. It provides a semi-automatic way to check individual patient records from the 1,000 records separated from the full dataset. The application loads one record at a time and displays the patient's chart to the user. After displaying each chart, the model generates the patient's predicted mortality risk along with the top features that contributed to the prediction, listed in order of significance with percentages. The program also provides to the user the option to generate SHAP waterfall plots for a visual aid in understanding each prediction.

III. Results

A. *Evaluation*

The predictive model was evaluated with several metrics (precision and recall are for the positive class only, death in this study):

TABLE I. Predictive Model Evaluation

ROC AUC	**PR AUC**	**Precision**	**Recall**
0.90	**0.58**	**0.28**	**0.86**

The most important metric here is recall, for which the model scored 86%. Also known as the true positive rate (TPR), recall measures the model's ability to correctly predict patient deaths relative to all actual patient deaths. The higher recall is, the less likely the model is to miss when a patient has an increased mortality risk. While the low precision means the model will flag many false positives, and there is certainly room for improvement to increase the model's utility and reduce false alarms, recall is of paramount importance for such a high-stakes prediction like risk of death.

B. *Explanation*

Shapley values are output by the explainable model along with the top ten features that most strongly influence the model's prediction, and formatted for readability and interpretability by a domain expert in the Python application. Figure 1 shows an example of significant features and their corresponding importance to a prediction. Red (positive) values point to the right and contribute to the model's prediction, while blue (negative) values point to the left and detract from the model's prediction, making it less likely to predict the positive class (death). The graph can be read from bottom to top, with each value pushing the prediction toward or away from the positive class.

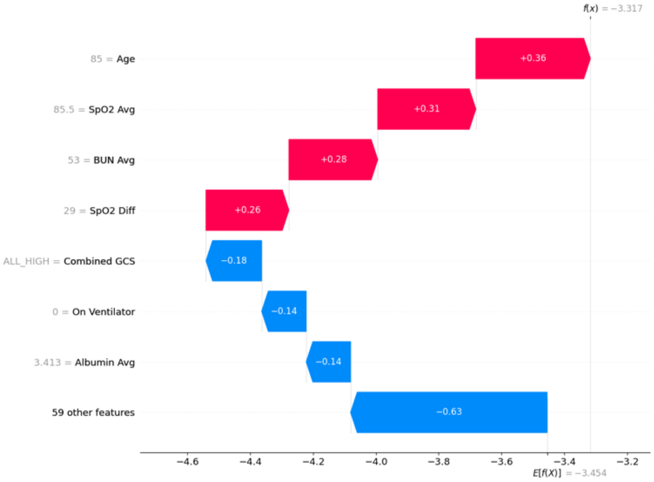
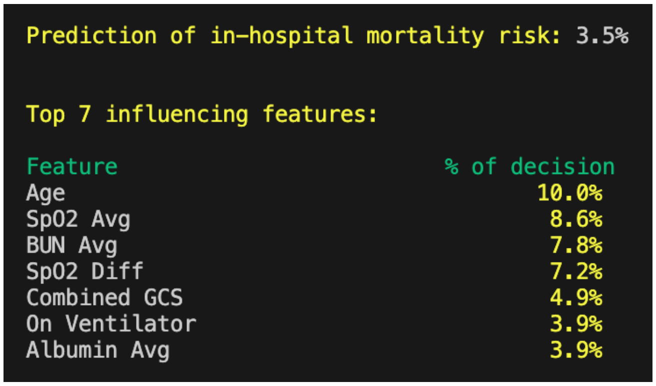


Fig. 1. SHAP waterfall plot showing feature importance.

Figure 2 shows the output of the Python application corresponding to the waterfall plot, and includes the mortality risk as a percentage.



```
Prediction of in-hospital mortality risk: 3.5%


Top 7 influencing features:

Feature                          % of decision
Age                                      10.0%
SpO2 Avg                                  8.6%
BUN Avg                                   7.8%
SpO2 Diff                                 7.2%
Combined GCS                              4.9%
On Ventilator                             3.9%
Albumin Avg                               3.9%
```

Fig. 2. Python program output showing top features and predicted in-hospital mortality risk.

The program also displays the actual result of a patient's stay in the hospital when manually evaluating test data from the GOSSIS-1-eICU dataset. Manual testing seemed to validate the recall vs. precision imbalance, with predictions tending to be overly cautious in favor of patient care. This was an interesting observation because although the model demonstrated low precision during evaluation, the final output of the model is not a binary classification but rather a percentage. Given both the complexity of medicine as a science and the interventional care received while in-hospital, it would be reasonable to expect that the majority of patients who are admitted to the hospital would survive despite some having a high risk of death.

IV. Discussion

While initial results demonstrate utility to non-technical users, there are several areas where the models could be improved. Precision of the predictive model could be increased with further parameter tuning, or possibly through feature engineering. The explainable model could be evaluated and validated by domain experts to ensure that the SHAP values generated are reasonable and expected given the corresponding patient records. Further research and work could also yield more useful application given feedback from domain experts in real-world settings.

V. Conclusion

As artificial intelligence becomes increasingly prevalent across all areas of life, the need for explainability will grow tremendously. In areas of critical importance such as healthcare, where AI has the potential to cause great harm as well as provide great benefit, explainability is a necessary feature that must be ingrained into the design of AI-based systems to promote trust and ensure safety.

Providing interfaces that interpret not just the results of diagnostics and predictions but also the reasoning behind those results is a key factor in driving adoption of AI in these areas. To that end, this study developed a simple yet effective interface design that clearly conveys the target prediction, risk of mortality, and the significant features influencing that prediction, in both tabular and graphic formats. Healthcare providers are more likely to utilize AI-driven tools such as this when they are provided with interfaces that instill confidence in the tools' abilities through features such as domain-relevant insights.

References

[1] D. Srivastava, H. Pandey, A. K. Agarwal, and R. Sharma, "Opening the Black Box: Explainable Machine Learning for Heart Disease Patients," *2023 International Conference on Advanced Computing Technologies and Applications*, October, 2023, DOI: 10.1109/ICACTA58201.2023.10392874

[2] R. Jain, M. Singh, A. R. Rao, and R. Garg, "Machine Learning Models To Predict Length Of Stay In Hospitals," *2022 IEEE 10th International Conference on Healthcare Informatics*, June, 2022, DOI: 10.1109/ICHI54592.2022.00105

[3] M. A. S. Iskandar, T. Badriyah, and I. Syarif, "Prediction of Length of Stay in Hospital Using Hyperparameter Optimization in the Convolutional Neural Networks Method," *2024 International Electronics Symposium*, August, 2024, DOI: 10.1109/IES63037.2024.10665859

[4] R. W. Krell, M. E. Girotti, and J. B. Dimick, "Extended length of stay after surgery: complications, inefficient practice, or sick patients?", *JAMA Surgery*, August, 2014, DOI: 10.1001/jamasurg.2014.629

[5] A. Clarke, "Why are we trying to reduce length of stay? Evaluation of the costs and benefits of reducing time in hospital must start from the objectives that govern change.", *Quality in Health Care*, pp. 172-178, 1996, DOI: 10.1136/qshc.5.3.172

[6] W. N. Southern and J. H. Arnsten, "Increased Risk of Mortality among Patients Cared for by Physicians with Short Length-of-Stay Tendencies,"*Journal of General Internal Medicine*, January, 2015, DOI: 10.1007/s11606-014-3155-8

[7] J. Raffa, A. Johnson, T. Pollard, and O. Badawi, GOSSIS-1-eICU, the eICU-CRD subset of the Global Open Source Severity of Illness Score (GOSSIS-1) dataset (version 1.0.0), *PhysioNet*, 2022. https://doi.org/10.13026/gbmg-a531

[8] J.D. Raffa, A.E.W. Johnson, Z. O'Brien, T.J. Pollard, R.G. Mark, L.A. Celi, et al, "The Global Open Source Severity of Illness Score (GOSSIS)," *Critical Care Medicine*, 2022, DOI: 10.1097/CCM.0000000000005518

[9] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, and H.E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation [Online], 101 (23), pp. e215–e220