

Predictive Modelling for In-Hospital Mortality: A Comparative Analysis

Introduction

The Intensive Care Unit (ICU) is a specialized medical facility that provides intensive care to individuals in severe conditions. For these critically ill patients, features such as temperature and blood urea nitrogen (BUN) concentration could be critical indicators of mortality.^{1,2} The research question aims to answer whether in-hospital death can be predicted using patient data in the first 24 hours of ICU admission; performance will also be compared among XGBoost, logistic regression, and neural network models. This research question is important because reliable early mortality predictions may serve as complementary tools for medical professionals in terms of patient-centred care, medical resource allocation, strategic decision-making, and family support.

Data Engineering Process

This data set consisted of 91,713 observations of patient visits to the ICU within one year.³ There were 186 variables, including identifiers, demographics, vitals, labs, APACHE covariates, and GOSSIS predictions.^{4,5} For this study, in-hospital death (binary outcome: 0 for survival, 1 for death) and mean death probabilities were studied. To handle missing values, columns with over 75% missing data were dropped, resulting in 27,795 complete rows. Scaling was applied to numerical and encoded categorical variables. Columns with zero variance were removed, and highly correlated features (>0.7) were pruned to prevent instability and avoid multicollinearity. Identifiers, redundant and irrelevant variables were also removed. 80% training and 20% test sets for logistic regression and neural network models were used; 64% training, 16% validation, and 20% test sets for the XGboost model were used. The outcome in the training set was highly imbalanced, with most observations having a death outcome of 0. To further select features, recursive feature elimination with cross-validation (RFECV) was performed.

Analysis

For the XGBoost model, grid search with stratified 10-fold cross-validation was used to tune hyperparameters, including number of trees, maximum depth, minimum samples per leaf, and learning rate. An XGBoost model with the resulting hyperparameters was then trained to predict the binary death outcome, and its performance was evaluated on training and test sets, respectively.

A logistic regression model was performed to predict the binary outcome of hospital death, using RFECV. The model was trained using the training set, and then the test set was used to make predictions of the outcome. The model was evaluated using the confusion matrix and classification report.

The first neural network model used two hidden layers with 32 neurons each to predict the binary outcome. Both layers used the tanh activation function, and the output layer used the sigmoid activation function since the outcome is binary. To calculate the loss over iterations, 200 epochs were used. The model's performance was evaluated using 100 epochs by comparing training data accuracy with validation data accuracy. Recall and f1-score accuracies were also calculated to compare with the other models.

For the second neural network, the data was split into 10 classes, delineating bins from 0 to 1, each representing a specific range of probabilities. It used layers with different neuron counts, batch normalization, and dropout. The output layer had 10 neurons, using softmax to assign probabilities to the classes. Across 500 epochs, it evaluated both probabilities and binary outcomes. For the latter, various thresholds were applied to binarize death events, enabling a nuanced evaluation of model performance in handling different degrees of risk or confidence levels associated with the predictions

Following that, the performances of logistic regression, XGBoost, and neural network models

were compared on the test set. The focus was on the recall scores because the aim was to identify as many true positives as possible in mortality prediction. We also looked at the f1-scores because the data set was extremely imbalanced, with the majority of observations having an outcome level of 0 (no death).

Findings

Table 1: Comparison of accuracy, recall, and F-1 scores for test sets by model

	Average Accuracy	Recall for Death = 1	F1-Score for Death = 1
XGBoost	0.92	0.68	0.43
Logistic Regression	0.92	0.69	0.33
Neural Network 1	0.91	0.15	0.25
Neural Network 2(on Binary Data)	0.91	0.25	0.35
Neural Network 2 (on Categorical Data)	0.75	-	-

For the XGBoost model, the tuned hyperparameters were 100 trees, a maximum depth of 3, a minimum of 50 samples per leaf, and a learning rate of 0.2. A model was fitted, and the test set's overall accuracy was 0.92. However, the performance of this model on the test set was very different for the two levels of outcome, where it yielded 0.31 precision, 0.68 recall, and 0.43 f1-scores for death, while all scores were above 0.9 for no death. This might be due to the unbalanced outcome distribution. To address this, the same model was trained on the up-sampled, balanced training set, but the resulting performance on the test set had an even lower recall of 0.44. This means that the XGBoost model missed a significant number of true positive cases and had many false positive predictions for mortality.

The recursive feature elimination curve indicated that the performance of the model increases with the number of selected features, with an optimal number of 53 features. The f1-score for the training set was 0.30, indicating that the model did not adequately fit the training data. The f1-score for the test set was 0.33, indicating that the trained model did not adequately generalize to the new (test) data. The recall for the test set was 0.69, and the accuracy was 0.92, which was comparable to the XGBoost model. The area under the curve was 0.77, which indicated a moderate discriminatory power. This means the model can moderately discriminate between positive and negative cases of hospital deaths. Therefore, the logistic regression model using RFECV did not make effective predictions regarding hospital deaths.

The first neural network model found that loss over iterations began to stabilize around 150 epochs. Overall, the training data accuracy and validation data accuracy were both high, approximating 0.916 and 0.911 on average, respectively, using 100 epochs. On the other hand, the neural network model had a very low recall of 0.15 and a very low f1-score of 0.25. This may have been caused by the disproportionate outcome of hospital deaths in the dataset.

The second neural network stabilized its loss after about 200 epochs. It achieved 0.75 accuracy in classifying different probability bins and 0.91 accuracy in predicting binary death events. Notably, when adjusting the threshold for death predictions, our model showed decreasing recall for death instances (coded as 1). This decline might be due to the dataset's high number of zeros for death events and probabilities, likely stemming from the imbalanced representation of hospital deaths in our data.

Conclusion


All models showed high accuracy but struggled with recall and f1-scores for predicting mortality. This means they didn't effectively capture actual death instances, posing risks for patient outcomes within the first 24 hours of ICU admission. Precision for mortality was also unsatisfactory, implying potential errors in identifying at-risk patients. Consequently, despite high correctness overall, these models should be used cautiously and might not offer actionable insights for predicting mortality in ICU patients.

Individual Contributions

Each person contributed to different parts of the code, report, and presentation slides.

Code and Presentation

GitHub repository: <https://github.com/Yutong-Lu/Datathon-4>

Google slides:  18-CHL5230-F23

References

1. Erkens R, Wernly B, Masyuk M, Muessig JM, Franz M, Schulze PC, Lichtenauer M, Kelm M, Jung C. Admission Body Temperature in Critically Ill Patients as an Independent Risk Predictor for Overall Outcome. *Med Princ Pract*. 2020;29(4):389-395. Available from: <https://doi.org/10.1159/000505126>.
2. Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig J, Masyuk M, Lauten A, Schulze PC, Hoppe UC, Kelm M, Jung C. Blood Urea Nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU. *PLoS One*. 2018 Jan 25;13(1):e0191697. Available from: <https://doi.org/10.1371/journal.pone.0191697>.
3. Lee M, Raffa J, Ghassemi M, Pollard T, Kalanidhi S, Badawi O, Matthys K, Celi L A. WiDS (Women in Data Science) Datathon 2020: ICU Mortality Prediction (version 1.0.0). *PhysioNet*. 2020. Available from: <https://doi.org/10.13026/vc0e-th79>.
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHEII: a severity of disease classification system. *Crit Care Med*. 1985;13:818-29.
5. GOSSIS: Global Open Source Severity of Illness Score: International Benchmarking for Critical Care [Internet]. Available from: <https://gossis.mit.edu/>.