Project

ETC3250 - ETC5250 - Introduction to machine learning - S1

Hoang Gia Phat - 35325453
Monash University

Table of contents

1. Introduction:

The goal of this project is to predict in-hospital mortality of ICU patients using data from the MIMIC-III dataset. This task is approached using various machine learning models, including Logistic Regression, Decision Trees, Neural Networks, and Gradient Boosted Trees (XGBoost). Each model is trained on engineered features derived from the original dataset.

2. Visualization and Feature Importance

From the data visualization to explore the variables distribution that I did, I can observe that age appeared to be a strong indicator of mortality, with older patients it also showing a higher risk. Vital signs such as HeartRate_Mean, SysBP_Mean, SpO2_Mean, and Glucose_Mean also displayed much more differences between the patients who survived and those who did not. For example, patients who died tended to have a higher heart rate with lower oxygen saturation. As a result, I created the Shock Index, and features like MAP, Pulse Pressure, and Risk Score showed clear separation in mortality outcomes. Additionally, the Mortality also varied by ICU unit, thus indicating that the type of care unit would influence patient outcomes.

3. Data Visualization & Feature Insights

To understand the relationship between clinical variables and mortality, I visualized key indicators such as Age, HeartRate_Mean, SysBP_Mean, SpO2_Mean, and Shock_Index across expired vs. non-expired patients. The plots revealed clear patterns:

- Age: Mortality risk increases with age, especially above 70.
- Shock Index: Higher shock index distributions were associated with higher death rates.
- ICU Unit Type: Some units had noticeably higher mortality rates.
- Vital Signs: SpO2 and glucose levels showed clear separability between outcomes.
- Correlation Plot: Engineered features such as Shock_Index and Risk_score had strong mutual correlations and clinical relevance.

These insights helped guide feature selection and confirmed the utility of engineered and clinical variables in the modeling process.

4. Preprocessing Strategy

I applied several preprocessing steps to ensure data quality and model readiness:

- To ensure a robust and meaningful model input, I started by cleaning the unreasonable ages by using the DOB and ADMITTED times to calculate and cleaned the age values making it into a column.
- Vital signs and lab measurements were used to compute clinically significant features such as Shock Index, Mean Arterial Pressure (MAP), Pulse Pressure, and others.

Composite risk scores were engineered to combine multiple risk indicators like hypotension and hypoxia, into a single feature. Outlier flags were created to capture extreme physiological values.

- We check the ICD9 diagnostic codes from the IMIC_metadata_diagnose.csv with the ICD9 diagnostic from the columns to find rare or malformed ICD9 diagnostic codes that were drop. The rest is grouped and generated dummy variables to encode them as binary features retaining ICD9 groups with ≥10 occurrences.
- Categorical variables like age group and ICU admission type were encoded as dummies. Missing values were handled by imputing medians, and numeric features were scaled and standardized.
- Standardized numeric variables via centering and scaling.
- Used PCA (20 components) for Neural Network due to weight constraints.

5. Model Development & Evaluation

So I train the model performance on the Sample with the output result being:

Model	AUC Score
Logistic Regression	0.845
Decision Tree (rpart)	0.586
Neural Network (nnet +PCA)	0.801
XGBoost	0.916

Logistic Regression

The logistic Regression model was trained on processed data with a binomial link. It performs really well, achieving an AUC of 0.845, indicating a great balance of sensitivity and specificity. Which served as an reliable baseline.

Decision Tree

The Decision Tree on the other hand yielded an AUC of 0.586. This result highlighted that eh tree's tendency to overfit or underfit the dataset due to its limited structure without any pruning or boosting.

XGBoost Forest

The XGBoost model demonstrated the highest AUC of 0.916, outperforming all others. It benefited from regularization, handling of missing data, and efficient boosting of weak learners. Feature importance analysis confirmed that engineered clinical features significantly contributed to predictive performance.

Neural Network

The Neural Network which trained using the PCA-reduced input to overcome the curse of dimensionality. The model achieved a good AUC of 0.801. Despite the lower performance thang XGBoost, it captured the non-linear interaction better than the decision tree

6. Conclusion

Among all tested models, the XGBoosted Forest provide the best out-of-sample performance due to its robustness and the ability to model complex features interactions. Its superior AUC also justified it as the best selection for the final prediction

7. Generative AI

Generative AI, specifically ChatGPT, was used extensively during the coding phase of this assignment. It assisted with:

- Engineering code for features such as shock index and composite risk indicators.
- Selecting and tuning appropriate machine learning models (Logistic Regression, Decision Tree, Neural Network with PCA, and XGBoost).
- Writing clean and reproducible R code to evaluate and compare models using ROC-AUC.
- Debugging and interpreting error messages during the model training and prediction stages.
- Tune XGBoost hyperparameters to get the hyperparameters for the Boosted Forest model to get the highest accuracy on the test-set.

8. Preference

- OpenAI (2025) ChatGPT
- ETC3250 Lecture and Tutorial Slide and Code, Monash University