Balanced vs. Real-World Distributions: A Machine Learning Analysis on ICU Outcome Prediction

Mariah Noelle Cornelio

Data Science Division, University of Texas at Arlington, mnc3287@mavs.uta.edu

Introduction and Background

- ICU mortality and readmission prediction is difficult due to **severe class imbalance**.
- Readmission rates are approximately 4–14% (Lai et al., 2012), while ICU mortality rates are around 30–40% (Armstrong et al., 2021).
- Machine learning models must decide whether **real-world (unbalanced)** or **class-balanced training** leads to better identification of high-risk ICU patients.

Objective

- Evaluate multiple model classifiers under two scenarios: unbalanced (real ICU distribution) and balanced training to determine which approach better supports minority-class detection.

Data

- **Size:** 2520 x 44
- **Target:** bad_outcome
- **Source:** eICU Research Database
- **Features:** labs/vitals, demographics

Table 1. Baseline logistic regression model

Logistic Regression Baseline Model			
	Precision	**Recall**	**F1**
0	0.89	0.99	0.94
1	0.64	0.15	0.24
Macro Avg	0.77	0.57	0.59

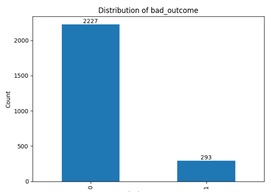


Figure 1. Distribution of target variable bad_outcome (mortality/readmission)

Methodology

6 Models Evaluated for Each Scenario

- Logistic Regression (linear)
- Random Forest (tree-based nonlinear)
- XGBoost (boosted tree nonlinear)
- LightGBM (boosted nonlinear)
- SVM (kernel-based nonlinear)
- Custom Stacked Model
  - LR + XGB as base models
  - LR as meta-learner

Training Procedure

- Stratified k-fold cross-validation for all models
- 80-20 train/test split
- Default threshold = 0.5

Evaluation Metrics

- Recall > Precision > F1
- PR-AUC and ROC-AUC

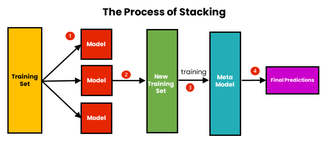


Figure 2. How a stacked model is created

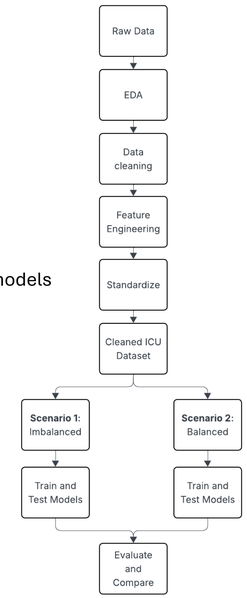


Figure 3. Flowchart of study design

Results

Assessment of Candidate Models

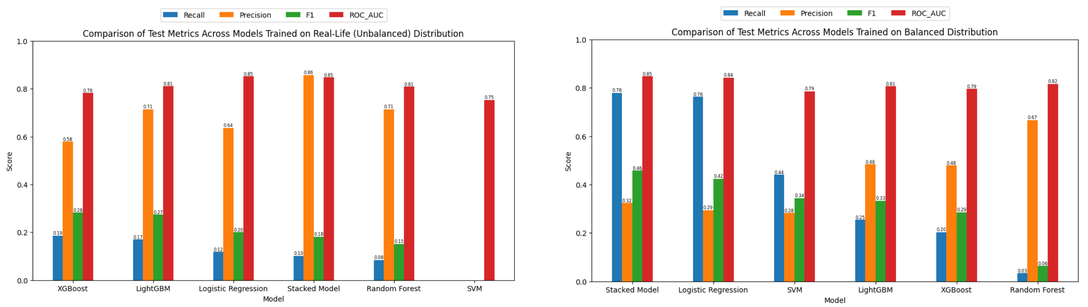


Figure 4. Comparison of test metrics for all models trained on Scenario 1: unbalanced (left) vs. Scenario 2: balanced (right) distributions

Performance Analysis of Best Models from Each Training Scenario

Table 2. Classification report comparison of best-performing models (XGBoost Unbalanced and Stacked Model Balanced)

XGBoost Model (Unbalanced) Classification Report			
	Precision	**Recall**	**F1**
0	0.90	0.98	0.94
1	0.58	0.19	0.28
Macro Avg	0.74	0.58	0.61

Stacked Model (Balanced) Classification Report			
	Precision	**Recall**	**F1**
0	0.96	0.78	0.86
1	0.32	0.78	0.46
Macro Avg	0.64	0.78	0.66

- Balanced (stacked) model is fairer, it detects class 1 minority class better
- Unbalanced (XGBoost) is more conservative
- Macro averages show overall advantage for the balanced scenario in detecting rare cases

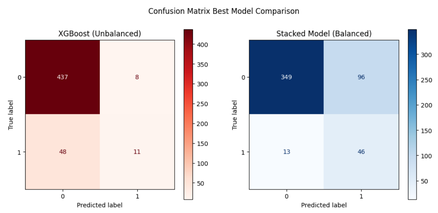


Figure 5. Confusion matrices for the best-performing models: XGBoost (unbalanced, left) and Stacked (balanced, right)

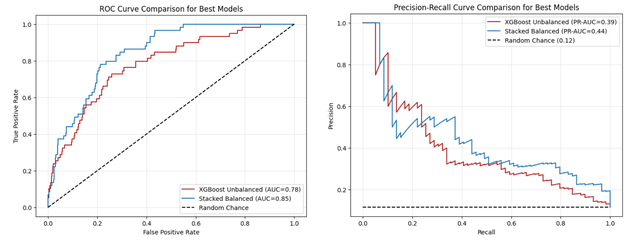


Figure 6. ROC-AUC and PRC-AUC (Precision-Recall Curve) curve comparisons for best-performing models

Best Model Interpretability and Demonstration

Table 3. Predicted vs. actual outcomes for 10 simulated ICU patients using XGBoost unbalanced model and Stacked balanced model

XGBoost Model				Stacked Model			
PatientID	**Predicted**	**Actual**	**Correct**	**PatientID**	**Predicted**	**Actual**	**Correct**
976535	0	0	Yes	976535	0	0	Yes
3133874	0	0	Yes	3133874	0	0	Yes
674731	0	0	Yes	674731	0	0	Yes
219980	0	0	Yes	219980	0	0	Yes
2729266	0	0	Yes	2729266	0	0	Yes
2056365	0	1	No	2056365	1	1	Yes
1856821	0	1	No	1856821	1	1	Yes
1090567	0	1	No	1090567	1	1	Yes
2860660	0	1	No	2860660	0	1	No
3233306	1	1	Yes	3233306	1	1	Yes

- Both models identify physiological severity and organ function markers as strongest predictors, supporting clinical relevance
- XGBoost highlights categorical splits; Stacked model highlights standardized numerical features

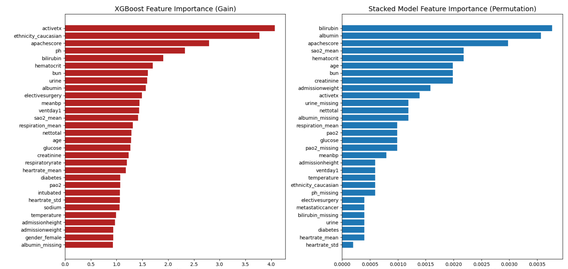


Figure 7. Feature importance for XGBoost (left) using gain and Stacked Model (right) using permutation

Discussion

What I Learned

Summary of Findings

- **Unbalanced training**: high accuracy, poor sensitivity for bad_outcome
- **Balanced training**: improved recall, ROC-AUC, and PR-AUC score; confusion matrices show fewer false negatives (higher recall)
- **Interpretability:** both models highlight severity markers
- **Simulation: Stacked model with balanced training** is more reliable in detecting high-risk patients

Why not just always train on class-balanced data?

- Pros
  - Model learns minority-class patterns
  - Improves discrimination for rare but clinically important outcomes
- Cons
  - Breaks real-world prevalence (model thinks bad outcomes happen 50% of the time, which is untrue)
  - Inflates false positives (false alarms, resource strain)
  - Not ideal when real-world probabilities are needed (dashboards/risk scores)

Which one is better? Balanced or unbalanced training?

- It depends entirely on the goal
  - If early detection/screening→ balanced training
  - If probability estimation → unbalanced training
  - If hybrid → unbalanced training + threshold tuning

Can the unbalanced model be improved?

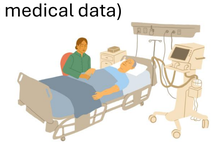
- Yes, through threshold tuning; lowering threshold increases recall for minority class without altering prevalence

Others

- XGBoost tree-based nonlinear model handles class imbalance very well

Future Work

- Threshold optimization for unbalanced training
- Model fine-tuning
- Multiple iteration (common technique used in medical data)
- Improve stacked model architecture
- Expand feature engineering
- Dimensionality reduction (PCA)
- Create a simulation dashboard



References

Armstrong, R. A., Kane, A. D., Kursumovic, E., Oglesby, F. C., & Cook, T. M. (2021). Mortality in patients admitted to intensive care with COVID-19: an updated systematic review and meta-analysis of observational studies. *Anaesthesia*, *76*(4), 537–548. https://doi.org/10.1111/anae.15425

Lai, J. I., Lin, H. Y., Lai, Y. C., Lin, P. C., Chang, S. C., & Tang, G. J. (2012). Readmission to the intensive care unit: a population-based approach. *Journal of the Formosan Medical Association = Taiwan yi zhi*, *111*(9), 504–509. https://doi.org/10.1016/j.jfma.2011.06.012