

Challenging the Complexity Paradigm: A Comparative Study of Logistic Regression and Random Forest for Mortality Prediction in Small Clinical Datasets

Muhammad Huzaifa Saqib
dept. of Software Engineering.

Air University

Islamabad, Pakistan
2502872@students.au.edu.pk
0009-0007-2902-6991

Abstract—While ensemble methods like Random Forests are often recommended for medical classification tasks, this arose the question of how better?. Motivated by a need to validate this common assumption, this research performs a comparative analysis of Logistic Regression and Random Forest models using the Heart Failure Clinical Records dataset. By systematically addressing model convergence through feature scaling and managing class imbalance via cost-sensitive learning, we evaluated the models specifically for "Early Warning" sensitivity. Our results indicate that a Balanced Logistic Regression model significantly outperformed the Random Forest in catching critical events, achieving a recall of 0.64 compared to 0.48. This study suggests that simpler, well-tuned linear models may offer superior clinical utility in high-stakes monitoring where sensitivity is the primary priority.

I. INTRODUCTION

Modern data science literature frequently suggests that complex ensemble models like Random Forest (RF) inherently outperform traditional statistical models. However, in the context of clinical mortality data—where the cost of a "False Negative" is significantly higher than a "False Positive"—the reliability of these complex architectures must be examined. While working on a project, I noticed many AI tools automatically suggested Random Forest, which led me to investigate if that was truly the best choice for medical safety. By comparing a baseline Logistic Regression (LR) against an RF architecture, this study seeks to empirically determine the most effective logic for critical event generation.

II. METHODOLOGY

The study utilized the Heart Failure Clinical Records dataset ($n=299$). The following steps were implemented to ensure a rigorous comparison:

- **Preprocessing:** We identified significant variance in feature scales (e.g., platelets vs. age). We implemented **StandardScaler** to normalize the features, which resolved the initial `lbfgs` convergence failures in the LR model.

- **Addressing Imbalance:** Analysis revealed a 3:2 ratio of survivors to death events. To prioritize safety, we applied **cost-sensitive learning** using the `class_weight='balanced'` parameter to penalize missed critical events more strongly than false alarms.
- **Comparison Framework:** We trained a Baseline LR, a Random Forest (100 estimators), and an Optimized LR to benchmark performance metrics specifically for Class 1 (Death Event).

III. RESULTS AND DISCUSSION

Contrary to the common assumption that ensemble models are superior, the Random Forest model exhibited significant "sensitivity lag," achieving a recall of only 0.48. The Optimized Logistic Regression achieved the highest balanced performance with an accuracy of 0.80 and a recall of 0.64.

Model	Accuracy	Recall (Class 1)	Precision (Class 1)
Unscaled Logistic	0.78	0.60	0.83
Random Forest	0.75	0.48	0.86
Scaled Logistic	0.80	0.56	0.93
Balanced Logistic	0.80	0.64	0.84

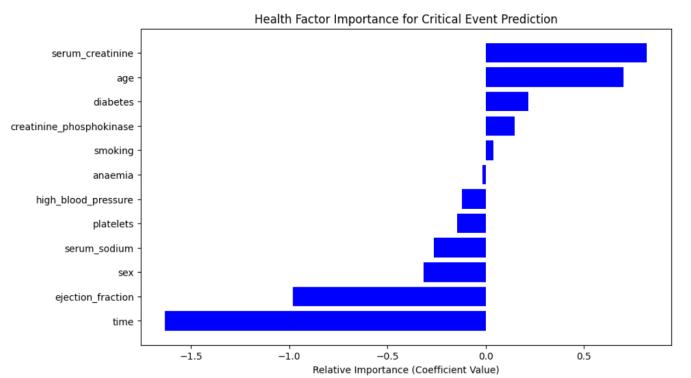


Fig. 1. Relative Importance of Physiological Features in the Balanced Logistic Model

Feature Interpretation: Using coefficient analysis, we determined that `serum_creatinine` and `ejection_fraction` were the most influential physiological predictors. This confirms that the model's mathematical decisions align with established medical understanding, providing an interpretable "Early Warning" mechanism rather than a "black-box" prediction.

IV. CONCLUSION

This research demonstrates that, for small-scale clinical datasets, model simplicity and proper tuning are more effective than algorithmic complexity. For systems requiring high sensitivity to critical health shifts, the Balanced Logistic Regression provides a more reliable foundation than a Random Forest.