Predictive Modeling for Blood Donation Propensity Using Machine Learning

Author: Akshat Banga
Institution: Unified Mentor – Data Science Internship

Keywords: Blood donation, Gradient Boosting, Calibration, Threshold tuning, Model evaluation

Abstract

This report presents an analytical study of predicting blood donation behavior using machine learning models. The approach focuses on interpretability, calibration, and threshold optimization. Gradient Boosting emerged as the best-performing model with an F1 score of 0.57 and Recall of 0.77 at an optimized threshold of 0.20. Reliability and threshold sweep plots justify the choice of the model and its operating conditions, ensuring balanced precision and recall suitable for real-world donor outreach applications.

1. Introduction

Blood donation forecasting is vital for sustaining healthcare supply chains. This project, undertaken as part of the Unified Mentor internship, explores machine learning-based prediction of donor return likelihood. Instead of relying solely on default classifier thresholds, this work integrates threshold optimization and reliability analysis to ensure decisions are both statistically sound and operationally meaningful. The language and explanation style adopted are in an indirect Indian English tone for clarity and contextual understanding.

2. Dataset Description and Exploratory Analysis

The dataset 'blood.csv' includes attributes such as Recency, Frequency, Monetary value, Time since first donation, and the target class representing donor activity. Exploratory Data Analysis indicated that Frequency and Monetary are strongly correlated, and active donors typically show lower Recency values and higher Frequency counts. These insights motivated ratio-based feature engineering.

3. Methodology

The pipeline comprised preprocessing, feature transformation, model training, calibration, and threshold tuning. Cross-validation ensured reliability while preventing overfitting. Models compared included Logistic Regression, Random Forest, Gradient Boosting, and a soft-voting ensemble. Calibration was conducted through sigmoid and isotonic scaling methods, evaluated via reliability

curves.

4. Results and Discussion

Among all models, Gradient Boosting displayed the most consistent and interpretable behavior. The threshold sweep demonstrated that F1 score peaked at a threshold of 0.20, where recall significantly improved without a large drop in precision. This operational point was ideal for maximizing donor identification. Reliability curves further showed mild underconfidence at low probabilities, corrected partially through isotonic calibration. The SoftVote ensemble achieved marginally higher ROC-AUC but was less stable under calibration tests.

Model	Best Threshold	Precision	Recall	F1 (pos)
Gradient Boosting	0.20	0.45	0.77	0.57
Sigmoid Calibrated	0.25	0.48	0.68	0.57
Isotonic Calibrated	0.30	0.50	0.59	0.54
SoftVote Ensemble	0.35	0.47	0.68	0.56

Threshold and reliability visualizations further validated the statistical performance. The threshold sweep plot highlighted the precision-recall trade-off curve where the model attained maximum F1. The reliability curve demonstrated good calibration with probabilities aligning closely with the ideal diagonal, indicating that the predicted donation likelihoods were practically trustworthy.

5. Conclusion

In conclusion, this internship project demonstrates that optimized Gradient Boosting with probability calibration and tuned thresholds can accurately forecast potential donors. The inclusion of reliability curves and threshold sweeps provided a justified analytical foundation for deployment readiness. The model achieves a meaningful balance between accuracy and interpretability, aligning with healthcare data ethics and operational goals of donor management systems.

References

1. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12. 2. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. Proceedings of ICML. 3. Zadrozny, B., & Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. KDD.