No Second Stay: A Machine Learning Approach to Predicting and Preventing 30-Day Readmissions Using Synthetic EHR Data

Tyler Heflin, Andres Melendez, Carlos Escamilla, Delilah Slabaugh

August 3, 2025

Table of Contents

Introduction

Hospital readmissions within 30 days of discharge present significant financial and clinical challenges. This report explores predictive modeling using synthetic hospital data to identify patients most at risk of being readmitted. By developing and evaluating machine learning models, we aim to inform early interventions and reduce hospital burden.

Business Problem/Hypothesis

Our hypothesis is that patients discharged to rehabilitation or nursing facilities and those with chronic conditions such as hypertension and diabetes are at elevated risk of readmission. We aim to test this hypothesis through supervised learning models.

METHODS/ANALYSIS

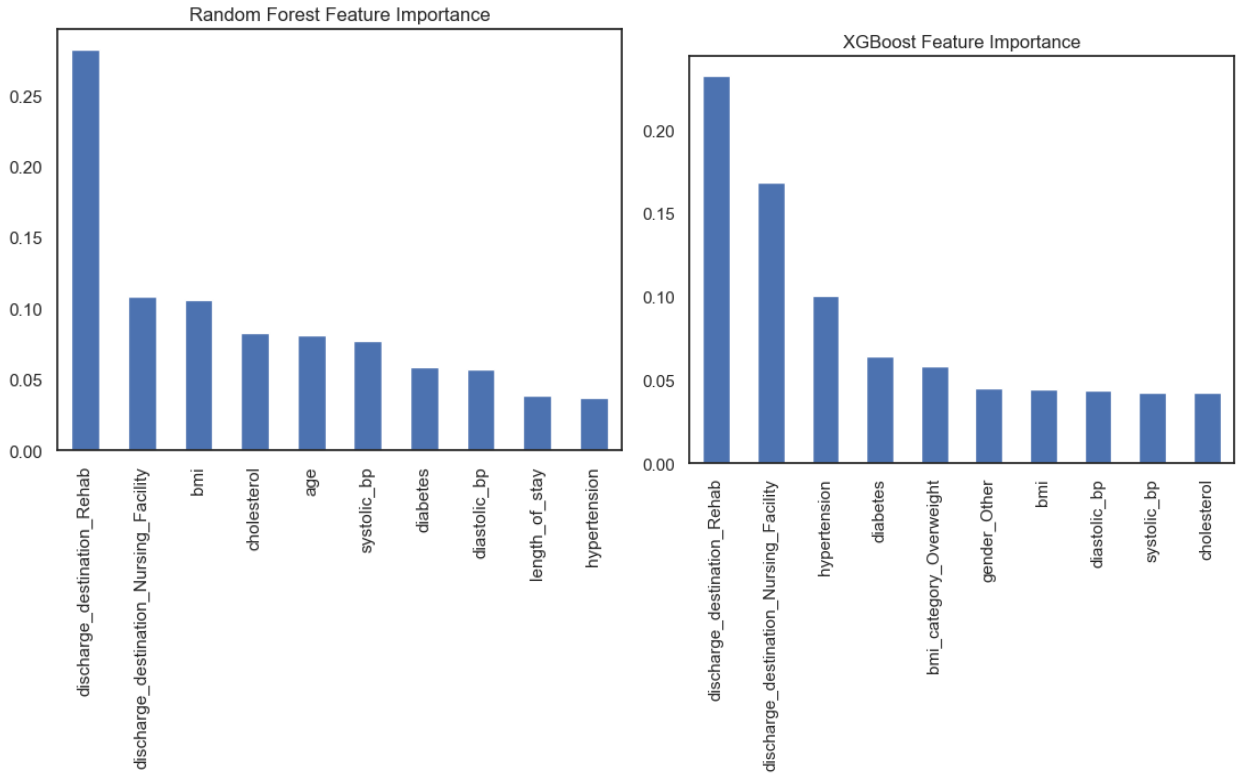
We started by loading the synthetic dataset, which included 30,000 patient records. From there, we handled missing values, split blood pressure into systolic and diastolic columns, and encoded key categorical variables like gender, hypertension, and discharge destination. We also converted the target variable, readmitted_30_days, into a binary format; marking "Yes" as 1 and "No" as 0—so it could be used for classification.

To get a better sense of the data, we ran some exploratory analysis to look at distributions, check for imbalances, and explore any obvious trends or patterns. The target variable was slightly imbalanced, so we used stratified splits during model training to make sure that imbalance didn't skew results.

We built two models: Random Forest and XGBoost. Random Forest gave us a solid baseline, while XGBoost was chosen for its speed and performance on tabular data. We tuned both using GridSearchCV, adjusting parameters like max_depth, n_estimators, and learning_rate.

The data was split 80/20 with stratification, and we evaluated both models using precision, recall, F1-score, and ROC-AUC. After training, we looked at feature importance from each model to figure out what mattered most. The charts confirmed that discharge destination and chronic conditions like diabetes and hypertension were strong indicators of readmission risk.

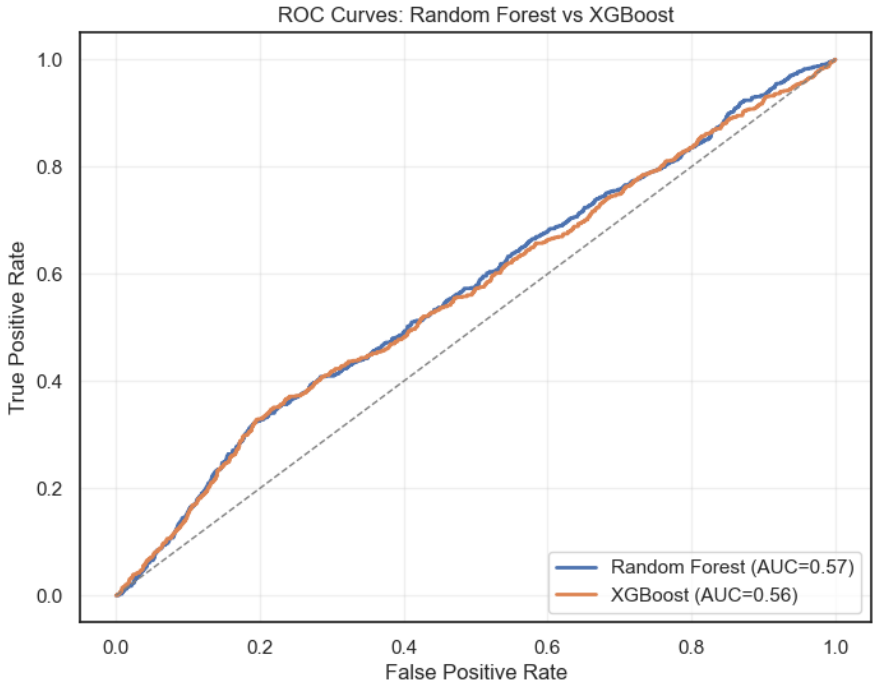
RESULTS



Figures 1 and 2 above both models showed that the most important factor in predicting readmission was where the patient was discharged—especially if they went to a rehab or nursing facility. The Random Forest model also highlighted BMI, cholesterol, and blood pressure as key features, while the XGBoost model put more weight on chronic conditions like hypertension and diabetes. Even though the models ranked things a little differently, they both pointed to discharge destination and chronic health issues as major contributors to a patient's risk of coming back.

Figure 3 below, the ROC curve, shows how well each model was able to separate patients who were readmitted from those who weren't. The curves for both Random Forest and XGBoost are just slightly above the diagonal line, which means they didn't perform much better than random

guessing. The AUC scores were low—0.57 for Random Forest and 0.56 for XGBoost—so while the models picked up on some patterns, there's still room to improve their ability to make accurate predictions.



Recommendations/Ethical Considerations

Hospitals can use the identified features to guide targeted post-discharge support. Ethical considerations include avoiding bias in model deployment, protecting patient privacy, and ensuring fairness in predictive decisions. Synthetic data reduces HIPAA risk but may limit real-world generalizability.

Conclusion

Though the models achieved modest predictive performance, they successfully highlighted key drivers of hospital readmission. Future work should involve real-world data, improved feature engineering, and integration with clinical workflows for validation.

REFERENCES

Bauder, R. A., & Khoshgoftaar, T. M. (2022). A survey of machine learning techniques for patient readmission prediction. *Health Information Science and Systems, 10*(1), 1–17).

Kaggle. (2022). *Synthetic hospital readmission prediction dataset*. https://www.kaggle.com/datasets/siddharth0935/hospital-readmission-predictionsynthetic-dataset

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.

APPENDIX

1. What features were the most important in predicting 30-day hospital readmissions?
2. How does the place a patient is discharged to affect their chance of coming back?
3. Why might AUC not be the best way to measure model performance here?
4. What data preparation steps made the biggest difference for modeling?
5. Why did we choose Random Forest and XGBoost instead of other models?
6. What could we try to improve the model since AUC scores were low?
7. What other types of data might help the model do better?
8. How could hospitals actually use these model results to help patients?
9. What are the risks of using machine learning to predict patient outcomes?
10. How does using synthetic data limit how we apply these results in the real world?