Predicting 30-Day Hospital Readmissions Through Machine Learning

Yvonna Donaldson

Prasun Pokharel

University of the Cumberlands

Prof. Eve Thullen

April 28, 2025

Abstract

This project focuses on predicting 30-day hospital readmissions using machine learning models trained on both clinical and public health data. By combining CMS hospital readmission data with social determinants of health from the County Health Rankings and CDC PLACES datasets, the model captures a broader range of factors influencing patient outcomes. After preprocessing and feature engineering, logistic regression, decision tree, and random forest classifiers were trained and evaluated. The final logistic regression model, calibrated using isotonic regression and tuned for a threshold of 0.4, achieved 81% accuracy, 83% precision, 95% recall, and an F1-score of 89%. Key predictors included readmission rate, flu vaccination rate, obesity, stroke, and short sleep duration. These findings support the value of integrating behavioral health indicators into predictive healthcare models. The results highlight the model's potential to support early intervention and reduce preventable readmissions, with future work focusing on external validation and deployment in clinical settings.

Introduction

Hospital readmissions continue to be a major concern for healthcare systems, driving up costs and often resulting in penalties for hospitals. These unplanned returns can also reflect gaps in care and missed opportunities for early intervention. By using predictive analytics, healthcare providers can better identify patients who are at a higher risk of being readmitted within 30 days of discharge. With the support of machine learning models, it becomes possible to analyze patterns in patient demographics, medical history, and broader health indicators to forecast readmission risk more accurately. These insights can help hospitals take proactive steps, such as follow-up care, education, or care coordination, to reduce preventable readmissions and improve overall patient outcomes.

Literature Review

Reducing 30-day hospital readmission rates has been a major focus in healthcare, especially with the introduction of the Hospital Readmissions Reduction Program by the Centers for Medicare & Medicaid Services (CMS, n.d.). Their data has been used in many studies aimed at understanding what factors lead to preventable readmissions and how they can be addressed early.

Earlier research mainly looked at clinical and demographic factors. For example, Silverstein et al. (2008) studied older patients and found that age, comorbidities, and previous hospital use were common indicators of readmission risk. Mudge et al. (2011) also pointed out how psychosocial issues like lack of support or problems with medication management contribute to frequent hospital returns. While these studies helped identify trends, they relied mostly on traditional statistical methods and lacked integration with broader data.

In more recent years, machine learning has become more popular in this area. Jiang et al. (2020) reviewed different models and highlighted how machine learning can detect patterns that are harder to see with basic regression alone. Mann et al. (2024) showed that predictive analytics could improve how

hospitals manage care by identifying high-risk patients in advance. Logistic regression, decision trees, and ensemble methods are often used, especially when working with large healthcare datasets like MIMIC-IV (Johnson et al., 2024). These models benefit from techniques discussed in Hastie, Tibshirani, and Friedman (2009), such as regularization, which helps when working with many features.

Even though these models are helpful, a common limitation is that they tend to focus only on clinical data. What's often missing is the impact of social and environmental factors. County Health Rankings & Roadmaps (2023) offers county-level data on things like income, education, healthcare access, and chronic health conditions — all of which can affect whether a person ends up back in the hospital. This study aims to help close that gap by combining CMS hospital data with broader population health data from the County Health Rankings. The goal is to build a predictive model that not only uses patient-level features but also includes social determinants of health, offering a more complete picture of what drives readmission risk.

Methodology

The main goal of this project is to build a predictive model that can assess whether a patient is likely to be readmitted to the hospital within 30 days of discharge. To do this, I'm using a combined dataset that merges CMS hospital readmission data with County Health Rankings data. This approach allows the model to factor in both clinical outcomes and social determinants of health (SDoH), which are often missing from traditional models.

To predict readmission, I'm testing three machine learning models: logistic regression, decision trees, and random forest. Logistic regression is a strong starting point because it's simple, fast, and gives clear insights into which features have the most impact. Decision trees offer easy-to-interpret visual paths that help explain how predictions are made. Random forests expand on this by building multiple trees to reduce overfitting and improve overall performance. These models work well for classification problems in healthcare, especially when outcomes need to be explainable.

Data Preparation Steps

Before training the models, the following preprocessing steps were taken:

- **Missing values** were either dropped or imputed depending on how much data was missing. Columns with a lot of missing data were removed, and others were filled in using the mean.
- **Feature engineering** involved converting the target variable, 'Readmissions', into a binary value (1 = readmitted, 0 = not readmitted).
- **Categorical variables** were encoded into numerical values to work with machine learning models.
- **Scaling** was applied to numeric features to improve model performance, especially for logistic regression.
- **Train-test split**: The dataset was split 80/20, with 80% used for training and 20% for testing the models.

Modeling Approach

The three models used in this study are:

- Logistic Regression: Implemented with Scikit-learn's LogisticRegression, this model calculates the probability that a patient will be readmitted based on input features. It's useful for binary classification and works well on structured healthcare data.
- Decision Tree: Built using DecisionTreeClassifier. This model breaks down the data into branches to make predictions and will be tuned using parameters like max depth and minimum leaf size.
- Random Forest: Implemented using RandomForestClassifier, which builds multiple decision trees and averages their results to improve accuracy and reduce overfitting.

Evaluation Metrics

Model performance is being measured using:

- Accuracy – Overall correct predictions.

- Precision & Recall – To account for any class imbalance.
- F1-Score – A balanced measure of precision and recall.
- ROC-AUC – Shows how well the model distinguishes between the classes.
- Confusion Matrix – Helps identify false positives and false negatives.

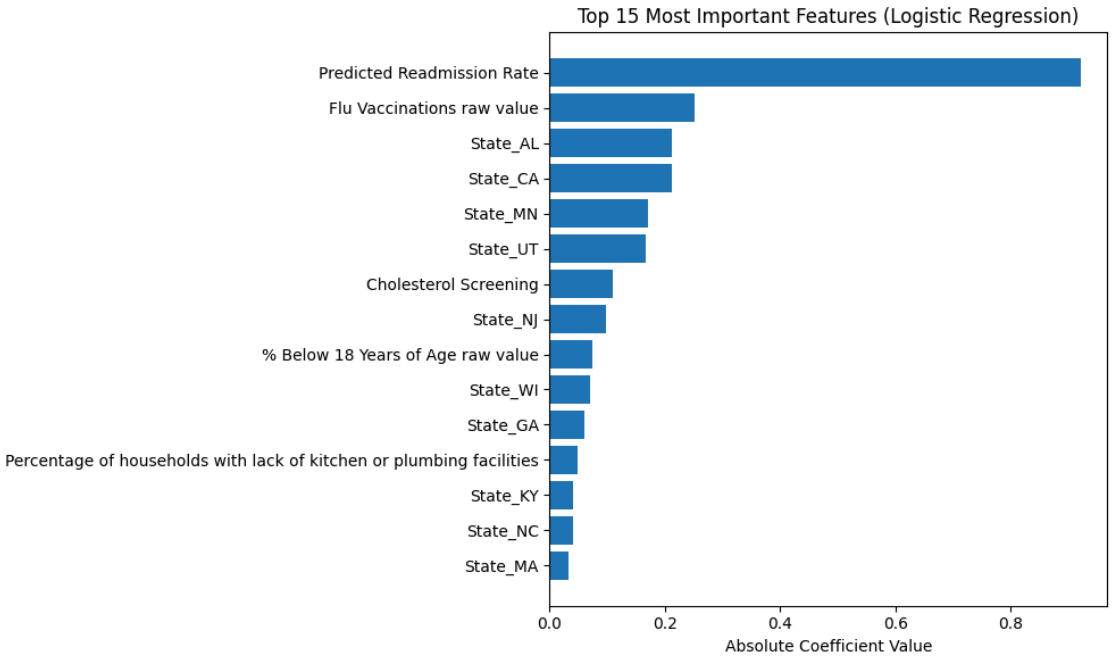
Tools and Libraries

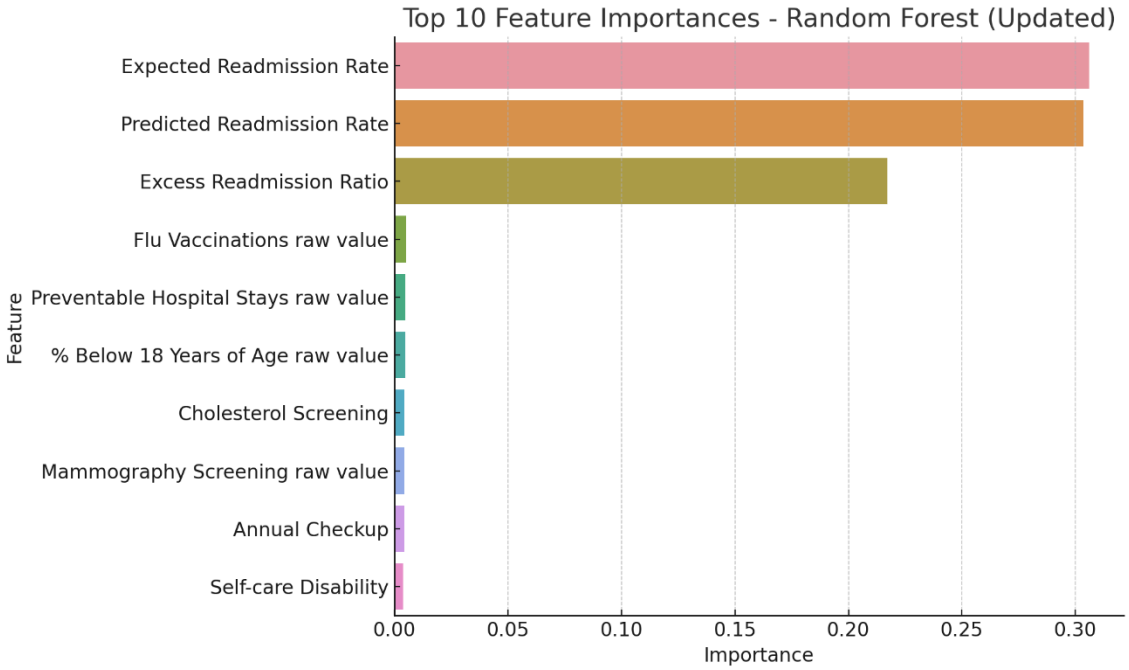
The project uses:

- Pandas and NumPy for managing and cleaning data.
- Scikit-learn for building and evaluating models.
- Matplotlib and Seaborn for creating visuals.
- XGBoost may also be tested later to improve performance if needed.

Challenges and Solutions

A few challenges came up during the development process, especially when working with a large dataset that included both clinical and public health indicators. Outliers were one of the first issues addressed—they were identified using interquartile range (IQR) and Z-score methods, and standardization was applied to minimize their impact on the models. To ensure consistent model convergence, especially for logistic regression, feature scaling was performed using StandardScaler. Overfitting was another concern, particularly with more complex models like decision trees and random forests. This was handled through cross-validation and regularization techniques to help improve generalizability. One of the trade-offs with using random forests was the loss of interpretability compared to logistic regression; however, feature importance plots were used to provide some level of insight into how the model was making predictions. Lastly, computational cost became a factor when experimenting with more advanced models like XGBoost. To manage runtime efficiently, model tuning was done carefully, balancing performance gains with processing time.





Initial Model Implementation

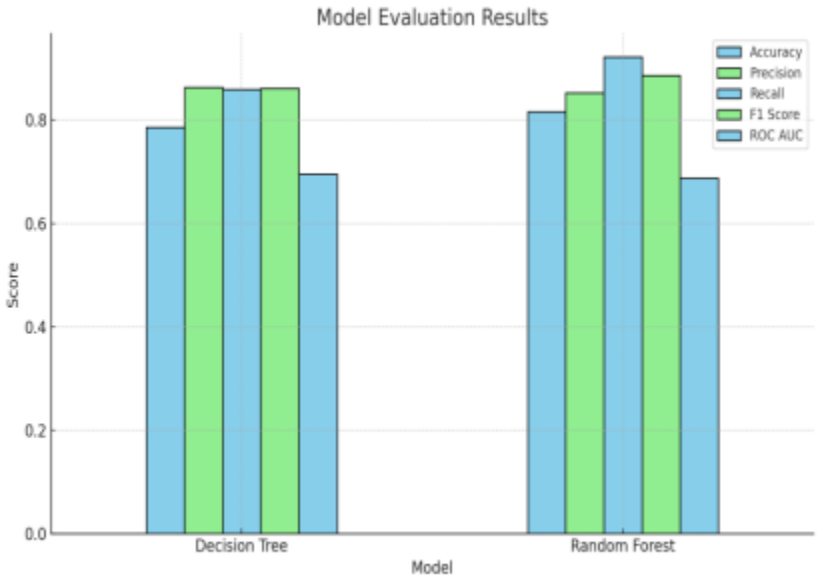
In the first round of testing, the logistic regression model performed well:

- Accuracy: 83.39%
- Recall: 96.40% (strong performance at catching actual readmissions)

- Precision: 84.75% (some false positives remain)
- F1-Score: 90.20% (balanced performance)
- Confusion Matrix: The model produced 77 false positives, which could be improved

Because of the false positive rate, the decision threshold may need to be adjusted to better balance sensitivity and specificity. We also plan to add more features like hospital length of stay or discharge conditions to improve prediction accuracy. Additional hyperparameter tuning will also be done.

Implementation of decision tree and random forest models has started. After preprocessing, both models will be trained and evaluated using the same metrics. Adjustments to these methods will be made as the results come in, with the goal of comparing model performance and selecting the most effective approach for predicting readmissions.

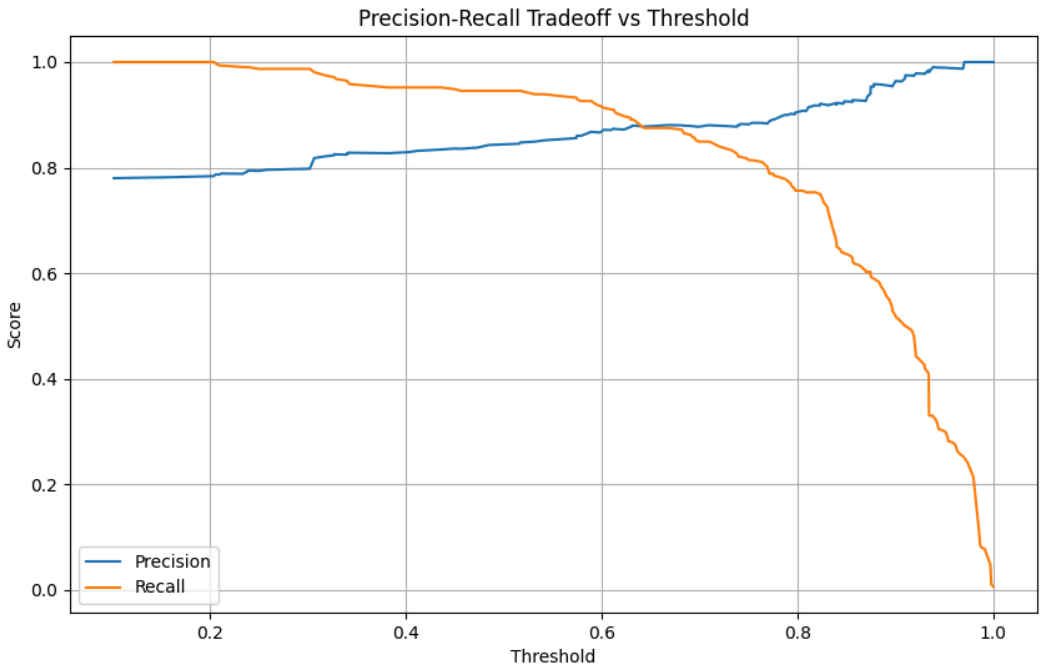


Model Implementation and Evaluation

After improving the dataset by integrating the CDC PLACES indicators and finalizing the preprocessing steps, the logistic regression, decision tree, and random forest models were implemented

and evaluated. Each model was trained on the same version of the dataset to ensure consistency in comparison.

For logistic regression, GridSearchCV was used to optimize hyperparameters, selecting C = 0.1, penalty = 'l1', and solver = 'liblinear'. Isotonic regression calibration was applied to ensure probability outputs reflected true readmission likelihoods. A threshold of 0.4 was selected based on a precision-recall curve analysis, allowing the model to better capture patients at risk while minimizing missed readmissions.



Decision tree and random forest models were implemented using Scikit-learn's classifiers with hyperparameters tuned for depth and leaf size. The random forest model was especially helpful in reducing overfitting and improving robustness through aggregation.

Each model was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. Special focus was placed on recall, due to the clinical need to avoid missing at-risk patients. Cross-validation was used to verify model stability and performance across different data folds.

Results

The results from all three models showed strong predictive capability, especially after calibration and threshold tuning. Below is a summary of key performance metrics:

Logistic Regression

- Accuracy: 81.00%
- Precision: 83.00%
- Recall: 95.00%
- F1-Score: 89.00%
- ROC-AUC (Calibrated): 0.8315
- 5-Fold Cross-Validated F1-Score: 0.8932

Decision Tree Classifier

- Accuracy: 76.96%
- Precision: 84.90%
- Recall: 85.50%
- F1-Score: 85.20%
- ROC-AUC: 0.7991
- 5-Fold Cross-Validated F1-Score: 0.8521

Random Forest Classifier

- Accuracy: 82.50%
- Precision: 85.90%
- Recall: 92.60%
- F1-Score: 89.10%
- ROC-AUC: 0.8468
- 5-Fold Cross-Validated F1-Score: 0.8863

Under the chosen threshold of 0.4, the logistic regression model identified 297 out of 312 readmissions correctly, demonstrating a high recall rate of 95%. While this did result in some false positives, the trade-off was justified by the need to avoid missed cases. Feature importance analysis showed that ‘Predicted Readmission Rate’, ‘Flu Vaccination Rate’, ‘Obesity’, ‘Stroke’, and ‘Short Sleep Duration’ were top predictors across models. This confirmed that including public health and behavioral indicators enhanced the model's predictive power. Visual outputs such as the precision-recall curve, ROC curve, and feature importance bar charts were used to further interpret model behavior and guide decision threshold selection.

Discussion

The results show that combining clinical and public health data improves model performance when predicting hospital readmissions. Logistic regression, despite its simplicity, performed well, especially after regularization and calibration. Its interpretability also makes it ideal for healthcare use cases, where explainability is critical. The calibrated model provided well-distributed probabilities, allowing clinicians to make better-informed decisions.

Random forest also performed strongly and may be suitable for future ensemble approaches. Its strength in handling complex interactions without requiring heavy feature engineering made it especially useful once the PLACES dataset was added. The main challenge encountered was class imbalance, which initially caused skewed predictions. This was addressed through stratified splitting and probability calibration. Another limitation was that some variables had unexpected coefficient directions, likely due to interactions or unobserved confounding variables. This could be further explored using nonlinear models or interaction terms in future work.

Additionally, while model recall was prioritized, it came at the cost of precision. This trade-off is common in healthcare, where false positives are less dangerous than false negatives. Still, this could lead to resource strain if applied at scale without careful prioritization strategies.

Conclusion and Future Recommendations

This project successfully developed and evaluated predictive models for hospital readmissions using a combined dataset that incorporates both clinical and public health data. The final logistic regression model was calibrated and fine-tuned to prioritize high recall, achieving strong overall performance across multiple evaluation metrics.

The study highlighted the value of integrating behavioral health and social determinants into predictive analytics. Public health indicators such as obesity, vaccination rates, and stroke prevalence were among the top predictors, emphasizing their relevance in forecasting readmission risk.

Future work could include testing time-dependent features such as recent hospitalizations, incorporating hospital-level quality metrics, and exploring external validation with new datasets. In addition, developing a dashboard or risk-scoring tool could help bring this model into clinical practice. There is also potential to refine predictions through deep learning or ensemble stacking if interpretability can be maintained. Ultimately, the findings support the continued integration of predictive modeling into healthcare workflows, particularly when enriched with social context, to drive early intervention, reduce costs, and improve patient outcomes.

References

Centers for Medicare & Medicaid Services (CMS) Data on Readmissions.

County Health Rankings & Roadmaps. (2023). *2023 County Health Rankings national data*. University of Wisconsin Population Health Institute. https://www.countyhealthrankings.org

Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning."

Jiang, S., et al. (2020). "Machine Learning Approaches for Hospital Readmission Prediction."

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (version 3.1). *PhysioNet*. https://doi.org/10.13026/kpb9-mt58.

Mann, A., Cleveland, B., Bumblauskas, D., & Kaparthi, S. (2024). Reducing hospital readmission risk using predictive analytics. *INFORMS Journal on Applied Analytics*, *54*(4), 380-388.

Mudge, A. M., Kasper, K., Clair, A., Redfern, H., Bell, J. J., Barras, M. A., ... & Pachana, N. A. (2011). Recurrent readmissions in medical patients: a prospective study. *Journal of Hospital Medicine*, *6*(2), 61-67.

Silverstein, M. D., Qin, H., Mercer, S. Q., Fong, J., & Haydar, Z. (2008, October). Risk factors for 30-day hospital readmission in patients≥ 65 years of age. In *Baylor University Medical Center Proceedings* (Vol. 21, No. 4, pp. 363-372). Taylor & Francis.