

Explainable Machine Learning for 30-Day Hospital Readmission Risk Prediction in Patients with Diabetes: A Comparative Analysis of Model Performance and Algorithmic Fairness

Zisheng Liang

liangzisheng6@gmail.com

University of Pennsylvania

Research Article

Keywords: Machine Learning, Predictive Analytics, Scikit-learn, Statistical Data Analysis, Feature Engineering, Logistic Regression, Random Forest , Model Evaluation, Explainable AI (XAI)

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11517784/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Explainable Machine Learning for 30-Day Hospital Readmission Risk Prediction in Patients with Diabetes: A Comparative Analysis of Model Performance and Algorithmic Fairness

Zisheng Liang

University of Pennsylvania

1. Introduction

Hospital readmission within 30 days is a costly but preventable problem in the U.S. health system. In 2018, there were about 3.8 million readmissions, at an average cost of \$15,200 each (Weiss & Jiang, 2021). The 30-day readmission rate for diabetic patients reaches 22.7%, nearly double the general rate. Common attributes include low income, racial or ethnic minority status, health conditions, and weak discharge support (Rubin, 2015). The federal Hospital Readmissions Reduction Program (HRRP) cuts a hospital's Medicare payments by up to 3% when readmission rates stay high (Centers for Medicare and Medicaid Services, 2025). However, many hospitals still have no way to identify high-risk patients before discharge.

This project uses machine learning to predict which diabetic inpatients are most likely to be readmitted within 30 days. The goal is to promote a risk score system to hospitals that identifies high risk patients who may need extra follow-up, medication review, or discharge counseling. This system should serve as an assistive tool and not replace clinical judgment.

2. ML Problem Framing

This project's main goal is framed as a binary classification problem. The model answers a yes-or-no question: will this patient be readmitted within 30 days? In deployment, we use the predicted probability as a risk score for the hospitals to follow up with patients. Machine learning is a good fit for our goal for two reasons. First, readmissions depend on many features. In our case, it depends on over 190 features at once, which is challenging for clinicians to manually weighting at the point of discharge. Second, nonlinearly may exist between features, and machine learning is a good fit to capture these patterns. This framing fits the real policy use case. A hospital does not need perfect prediction. What it needs is a tool that can rank patients according to their risks before they are discharged from the hospital.

3. Data Collection

We use the "Diabetes 130-US Hospitals for Years 1999-2008" dataset from the UCI Machine Learning Repository (Clore et al., 2014). The raw dataset contains 101,766 diabetic inpatient encounters from 130 U.S. hospitals. It includes demographics, prior hospital visits, diagnoses, lab tests, medications, and discharge information. Race, gender, and age are also recorded. After cleaning, 99,340 encounters remain from 69,987 patients. About 11.4% of these (11,314 encounters) ended in a 30-day readmission. The data is imbalanced, so accuracy is not a useful metric. A model that predicts "no readmission" for everyone would achieve 88.6% accuracy and miss all high-risk patients. The raw data also has substantial missingness; the weight column is 97% missing, the medical specialty column is 49% missing, and the payer code column is 40% missing.

4. Data Preprocessing

4.1 Imputation Strategies

Data cleaning decisions included dropping the *weight* feature because about 97% of the column was missing. We also dropped 15 drug columns because most rows had a "No" value. Three rows with invalid gender codes were dropped. Patients who passed away in the hospital or transferred to hospice were removed. The "None" values in *max_glu_serum* and *A1Cresult* columns indicate that the doctor did not order the test. We recoded "None" to "NotMeasured" so the model could treat it as clinical information. We filled in the missing values with "Unknown" for *medical_specialty* and *payer_code*, because the non-missing values still contained useful information. We filled a small portion of missing *race* values with "Other".

4.2 Feature Engineering

We created new features to make the data easier for the models to use. The original diagnosis columns used ICD-9 codes, which are difficult to understand. We grouped these codes into broader clinical categories, such as circulatory, respiratory, diabetes and injury. This made the features more interpretable (Strack et al., 2014). We also created *num_active_meds*, which counts how many diabetes medications a patient was taking. *Any_med_change* column was created to record whether any medication dose changed. We created a *total_prior_visits* column by aggregating prior outpatient, emergency, and inpatient visits. Age groups buckets such as [50-60) are converted to ordered numeric values from 0 to 9.

4.3 Train-Test Split

A patient-aware and time-aware train-test split is engineered. Single-visit patients were split with a stratified 80/20 split. For patients with multiple visits, a standard random split could result in the same patient's earlier and later visits being assigned to the wrong side. Therefore, we routed their latest encounter to the test set and all earlier encounters to the training set. This mimics real deployment, where the model always predicts the future from the past. This split

also reduces leakage. The final split produced 72,269 training encounters and 27,071 test encounters.

4.4 Scaling

Numeric variables were scaled using “StandardScaler”. The scaler was first fitted to the training data and then applied to the test data. This step can prevent information from the test set from entering the training process. Scaling is important because some models are sensitive to the size of numbers.

4.5 OneHotEncoding

The model one-hot encoded categorical variables such as race, gender, admission type, discharge disposition, admission source, lab results, payer code, medical specialty, and diagnosis categories. One-hot encoding converts each of these categories into new 0/1 binary columns.

5. Experimentation

5.1 Models Used and Why

We tested baseline Logistic Regression, tuned Logistic Regression, tuned Decision Tree, and tuned Random Forest in this project. The tuned logistic regression adds regularization to check how much the model performance increases from the baseline logistic regression. The decision tree generates outcomes from a set of simple rules, and they are simple for a clinician to audit. The random forest averages across many trees and captures more nonlinearity than a single tree. It can handle the mixture of numeric, binary, and one-hot encoded features well; therefore, it is our preferred choice in terms of tools that focus on recall rate. All models used `class_weight = 'balanced'` since only 11.4% of encounters in the cleaned data ended in a 30-day readmission. This setting prevents models from predicting "no readmission" for every patient.

5.2 Hyperparameters and Why

We adjusted the "C" parameter of the logistic regression model, which determines the penalty intensity of the model for larger coefficients. A smaller C value indicates a stronger regularization effect, while a larger C value indicates a weaker regularization effect. Ultimately, a "C" value of 0.1 yielded the best results.

For the Decision Tree, we tuned “max_depth” to determine how many splits the tree can make. A shallow tree is easier to explain, but it may miss important patterns. On the other hand, an overly deep tree may memorize the training data and perform worse on new patients. The best tuning choice is “max_depth = 3.”

For the Random Forest, we tuned “n_estimators” and “max_depth” to control the number of trees and tree complexity. More trees could yield more reliable predictions, but also increase computing time. `n_estimators = 200` and `max_depth = 10` are chosen as the final setting.

5.3 Hyperparameter Tuning Strategy

This project uses 5-fold cross-validation on the training set to split data into smaller portions and tests on 5 different folds. This result is more trustworthy because each time a different part of the data is used as the validation set, and the rest is used for training. We chose the F1 score as the tuning metric because it balances recall and precision with an unbalanced dataset.

5.4 Evaluation Metrics

The model uses recall, precision, F1, ROC-AUC, and PR-AUC as evaluation metrics. Accuracy is not taken into consideration because with 11.4% positive cases, a model could predict "no readmission" for every patient and still achieves 88.6% accuracy. Recall shows how many true readmission cases the model detects. It is the most important metric because missing a high-risk patient is the most serious outcome in our case. Precision shows of all the patients the model warned about, how many get readmitted. Low precision means hospitals waste time and resources on patients who would not have returned. F1 balances recall and precision, and its high value needs both scores to be high. ROC-AUC measures how well the model ranks readmitted patients above non-readmitted ones across all thresholds. PR-AUC focuses more on the positive class, making it more useful than ROC-AUC for imbalanced data.

6. Results

6.1 Model Performance Summary

Table 1 below shows the two Logistic Regressions produced the highest precision, ROC-AUC, and PR-AUC. The tuned Random Forest achieved the highest recall of 0.710, which means it correctly identified 71% of the readmitted patients in the test set. Overall, all models produced moderate performance. The results show that 30-day readmission is challenging to predict using the structured hospital data.

Table 1. Test-set performance comparison across all four models.

Model	Precision	Recall	F1	ROC-AUC	PR-AUC
Baseline LR	0.085	0.650	0.150	0.650	0.128
Tuned LR	0.085	0.654	0.151	0.651	0.127
Tuned Decision Tree	0.074	0.660	0.133	0.619	0.097
Tuned Random Forest	0.077	0.710	0.138	0.629	0.109

6.2 Confusion Matrix

Figure 1 and Figure 2 show the evaluation dashboards for the tuned logistic regression and tuned random forest. Both dashboards include the confusion matrix, ROC curve, and precision-recall curve.

Figure 1. Tuned Logistic Regression evaluation dashboard

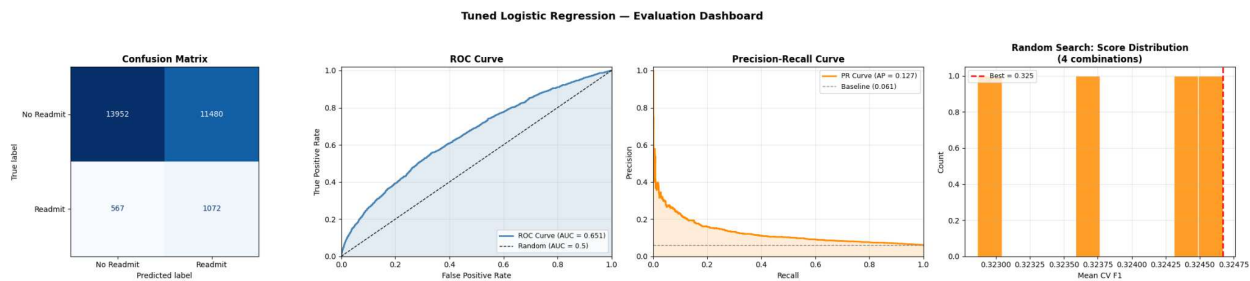
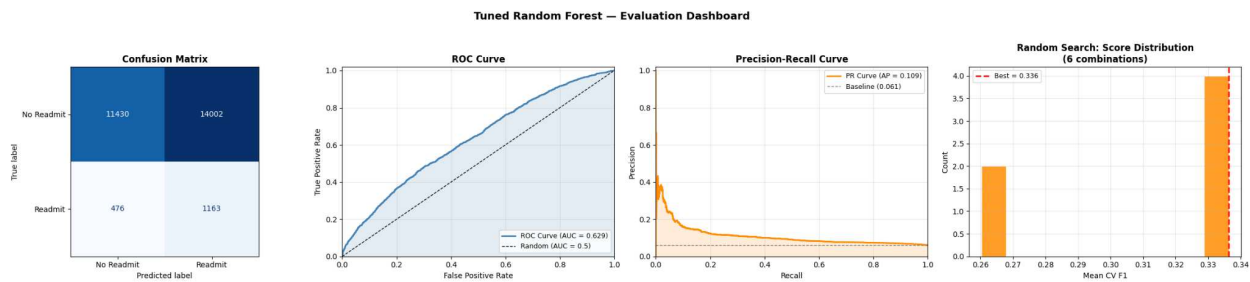


Figure 2. Tuned Random Forest evaluation dashboard



6.3 Graphs and Plots

SHAP values were calculated for the Random Forest to determine the most powerful readmission attributes. Figure 3 shows the SHAP value of the top 20 features, and Figure 4 shows the Beeswarm plot that explains how these features drive the outcome direction.

Figure 3: SHAP Feature Importance

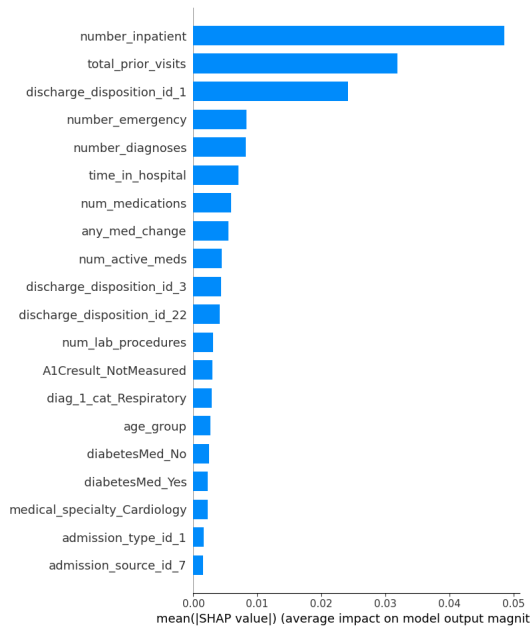
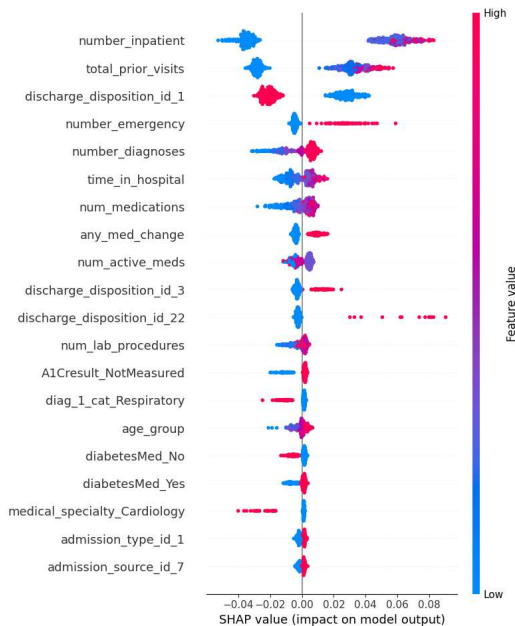


Figure 4. SHAP Beeswarm



The top three SHAP importance features are *number_inpatient*, *total_prior_visits*, and *discharge_disposition_id_1*. These results are consistent with the risk factors in the existing clinical literature (Rubin, 2015).

6.4 Error Analysis

False positives and false negatives are the two main types of errors in the models. False positives are patients who were flagged as high-risk but were not readmitted within 30 days. It would waste hospital extra resources and time. Nevertheless, false positives do not bring the worst type of harm. False negatives are significantly more harmful since they cause the hospital to miss the patient who needs real support.

7. Explainability and Algorithmic Bias

7.1 Explainability

We used SHAP values to indicate feature explainability for the tuned Random Forest model. Figure 3 shows the top 20 features ranked by their SHAP values. The most important features were prior inpatient visits, total prior visits, discharge disposition, length of stay, number of diagnoses, and medication burden. Figure 4 shows whether each feature increased or decreased the prediction of readmission risk. The Beeswarm plot shows the same pattern as the feature importance results.

7.2 Fairness Audit

We audited model performance across race, gender, and age groups because these are sensitive attributes directly related to fairness. The main concern is that the model may produce different outcomes across different races, genders and age groups.

Table 2. Fairness Audit Across Models by Recall Gap

Model	Race gap	Gender gap	Age gap
Base LR	0.270	0.033	0.597
Tuned LR	0.270	0.033	0.597
Tuned DT	0.276	0.020	0.291
Tuned RF	0.341	0.007	0.314

Table 2 shows recall gaps across race, gender, and age groups. Tuned Random Forest has the smallest gender gap and a relatively small age gap, but it has the largest race gap. Tuned Logistic Regression has a smaller race gap, but a much larger age gap. This shows that no model is equally fair across every subgroup. We still choose a tuned Random Forest after the fairness audit because it catches the most 30-day readmissions.

8. Real-World Use

8.1 Final Model Selection

The tuned Random Forest will be the final model chosen for deployment after evaluating the four models. It produces a recall value of 0.710, which is higher than that of the tuned Logistic Regression; however, both its F1 score and ROC-AUC are lower than those of the tuned Logistic Regression. A high recall rate means that most patients, who are truly high-risk, are returned to the health system. False negatives have a more serious impact on readmission, as there is no opportunity to implement necessary interventions. In contrast, false positives primarily lead to unnecessary follow-up rather than being a clinically harmful event. On the other hand, the choice has a trade-off. The Random Forest Race Fairness Gap is larger than that of the Logistic Regression when compared. Thus, the model's recall is aimed at being as high as possible, while fairness concerns are addressed.

8.2 Real-World Implementation

In clinical practice, the model will be implemented within the Electronic Health Record (EHR). During discharge from an inpatient visit, the EHR will pull the relevant encounters and discharge information, and the model will output a predicted probability of 30-day readmission. This probability will serve as the baseline readmission risk score against which physicians and care teams can compare.

Instead of taking actions for the patient, the model should generate a list of patients in priority order and enable the doctors to have the most senior patients in front of them to decide on tailored intervention. For example, early follow-up appointments, medication reconciliation, discharge education, and so on.

8.3 Users and Benefits

The model helps physicians, nurses, discharge planners, and case managers discharge patients appropriately. Patients experience in-home and outpatient post-acute care before being readmitted to the hospital. The hospital limits the amount of healthcare revenue that HRRP penalties pay out to Medicare through its targeted release.

8.4 Additional Considerations for Real-World Use

Each prediction needs a SHAP justification of the most important features. The doctor needs to understand the reason behind the flag, not just that it occurred. It is a Decision Support system, not a decision maker.

9. Conclusion and Limitations

9.1 Conclusion

Implications of Predicting 30 Readmissions for Inpatients with Diabetes through Machine Learning. It is challenging to predict 30 readmissions for inpatients with diabetes mellitus using machine learning methods. Moreover, surprisingly, none of the four algorithms performs well. Recall varies from 0.65 to 0.71, and the precision is below 0.1 for all models. However, they are still good at filtering the noise. Previous readmissions, duration of stay, discharge location, and overall disease burden are the main predictors and are consistent with the clinical literature. Random forests, which are optimized most appropriately, should be used as an assistive tool. It captures around 71% of the real thirty-day readmissions; thus, it is the most important factor to consider when deciding when patients should be screened and when they should not.

9.2 Limitations

This study is, without a doubt, a narrowing experience. The original data were from 1999 onward, and for this research, current clinical practices are not applicable. Other important predictors are missing. There is a complete absence of disclosure regarding housing, transport, and adherence to medications. The dataset is highly imbalanced: around 11% of the entire class, 13.4% of the training dataset, and 6% of the test dataset after the train/test split belong to the positive class. There is still a fairness issue by race, and a fairness issue by age group. The model worked, but it was not validated with the latest hospital data; therefore, it should not be implemented yet.

9.3 Future Directions

There are three future work considerations. The first is to supplement the existing data with the remaining fields, such as social data (e.g., housing, transport) and outpatient follow-up report data. Then test the model on real data in the future and see whether flags are present for the same patients, whether those patients receive better follow-up care, and whether they have lower readmission rates.

References

- Centers for Medicare and Medicaid Services. (2025). Hospital Readmissions Reduction Program (HRRP).
<https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp>
- Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US hospitals for years 1999-2008 [Dataset]. UCI Machine Learning Repository.
<https://doi.org/10.24432/C5230J>
- Rubin, D. J. (2015). Hospital readmission of patients with diabetes. *Current Diabetes Reports*, 15(4), 17. <https://doi.org/10.1007/s11892-015-0584-7>
- Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, Article 781670.
<https://doi.org/10.1155/2014/781670>
- Weiss, A. J., & Jiang, H. J. (2021). Overview of clinical conditions with frequent and costly hospital readmissions by payer, 2018 (HCUP Statistical Brief No. 278). Agency for Healthcare Research and Quality.
<https://hcup-us.ahrq.gov/reports/statbriefs/sb278-Conditions-Frequent-Readmissions-By-Payer-2018.jsp>

Additional Disclosure

Though not cited, Emi-Johnson, O. G., & Nkrumah, K. J. (2025). Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data. *Cureus*, 17(4), e82437. <https://doi.org/10.7759/cureus.82437>, provides us great insights for the project.