Predictive Modeling of Diabetic Patient Readmission Using Statistical Learning Techniques

Anna Simeone, Marta Zecchini

245744, 246790

June 13, 2025

Contents

1 Introduction

Hospital readmission of diabetic patients remains a widely recognized and pressing healthcare issue. Effective treatment of hyperglycemia in hospitalized individuals has a significant impact on clinical outcomes, influencing both morbidity and mortality rates [1]. In this study, our objective was to develop a machine learning model capable of accurately predicting hospital readmission for diabetic patients, classifying them into three categories: readmission within 30 days (`<30`), readmission after more than 30 days (`>30`), or no readmission (`NO`).

2 Materials and Methods

2.1 Dataset

The dataset spans ten years of clinical care (1999 - 2008) collected from 130 hospitals and integrated delivery networks across the United States. It includes 50 features related to patient demographics, diagnoses, treatments, and hospital outcomes. Data were extracted for inpatient admissions with a diagnosis of diabetes, length of stay between 1 and 14 days, laboratory tests conducted, and medications administered during the hospital stay. The final dataset contains 101,766 entries.

2.2 Data inspection and preprocessing

2.2.1 Data exploration, cleansing and grouping

A comprehensive inspection of the dataset was performed prior to any transformation. For each variable, descriptive statistics were computed, and graphical visualizations were generated to assess distributions, cardinality, and the presence of missing or inconsistent values.

To further assess consistency in the dataset, encounters related to the same patient but showing conflicting demographic information (e.g., same patient with different gender or race) were identified, suggesting possible human errors during data entry. These encounters were removed to reduce noise and improve data quality.

All hospitalizations ending in *death* or *discharge to hospice* were excluded. Although valid from a data perspective, such cases introduce bias, as the model could trivially learn to predict no readmission simply due to patient death.

After harmonization, high-cardinality categorical features were **grouped** into broader, clinically meaningful categories to reduce sparsity and overfitting. Administrative codes (admission type, source, discharge disposition) were reclassified into macrogroups; rare race categories merged into "Other"; age bands aligned with life stages; and medical specialties consolidated into a few clinical domains

Medication features were binarized to indicate drug administration, ignoring dosage changes ("Up," "Down," "Steady").

The *glucose serum* and *A1c* features were binarized to indicate whether each test was performed. Despite substantial missingness, they were retained due to their clinical relevance and the assumption that test absence may reflect physician judgment, thus carrying potential predictive value.

To account for repeated hospitalizations, a binary feature was added to indicate whether an encounter was the patient's first. Additionally, a reduced dataset containing only the first admission per patient was created to assess model sensitivity to repeated observations.

2.2.2 Missingness analysis and imputation

Several variables contained missing values, initially analyzed at the feature level (see Table 1).

Table 1: Percentage of missing values across selected features.

Feature	**Missingness (%)**
Weight	96.85
Medical specialty	48.95
Payer code	38.05
Admission type	10.21
Admission source	6.90
Discharge disposition	4.71
Race	2.11
Diagnosis group 3	1.43
Diagnosis group 2	0.36
Diagnosis group 1	0.02
Gender	∼0

Missingness was also evaluated per encounter by computing the percentage of missing features out of 51 variables, with a 10% threshold (i.e., at least 6 missing features). Only 0.17% of encounters (166

records) exceeded this threshold. Given their minimal impact on prediction and the small fraction affected, these encounters were removed to simplify preprocessing without losing relevant information. The `weight` feature was excluded due to its high missingness (96.85%), which is assumed to be Missing Completely At Random (MCAR).

For variables with over 30% missingness, we investigated the missing data mechanism. In both `payer_code` and `medical_specialty`, missingness varied across observed variables (e.g., race, admission type), suggesting a *Missing At Random* (MAR) pattern. As such, imputing a distinct category (e.g., `Unknown`) was deemed appropriate to preserve potential signal.

For features with less than 30% missingness, we adopted a two-step imputation strategy: categorical variables were imputed with the most frequent category using `SimpleImputer`, while numerical ones were imputed via `KNNImputer` ($k = 5$) to better capture local structure.

2.2.3 Feature Correlation

A correlation matrix was computed to assess multicollinearity among numerical features (Figure 1).

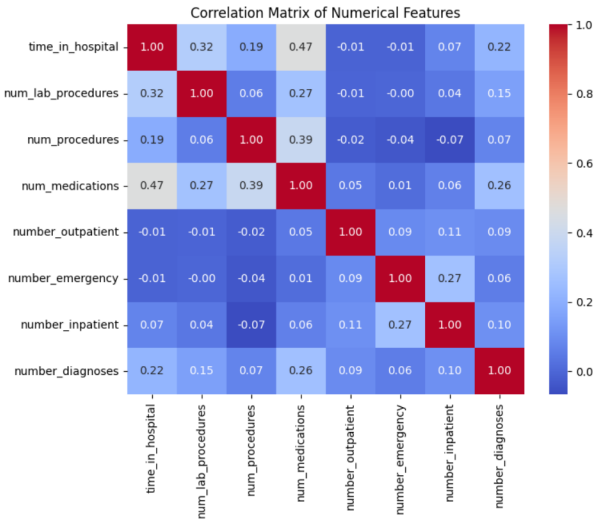


Figure 1: Correlation matrix of the numerical features.

Moderate positive correlations were observed (e.g., between `num_medications`, `num_procedures`, and `time_in_hospital`), reflecting clinical overlap. However, correlation levels were not high enough to warrant feature removal, and preliminary tests showed no performance gain from exclusion. All features were therefore retained.

2.2.4 Outlier Detection

To enhance data quality, we applied both rule-based and multivariate outlier detection. Rule-based filters flagged short hospital stays (≤ 1 day) with unusually high medication or lab counts, mostly acute, high-intensity treatments, which were retained for clinical plausibility. Concurrently, an Isolation Forest identified 2,968 (3%) multivariate outliers representing atypical but valid cases, such as frequent low-intensity admissions linked to chronic or social care; these were also kept due to lack of data errors.

2.2.5 Skewness

Several count-based numerical features showed right-skewed distributions. To reduce skewness and stabilize variance—particularly for models sensitive to feature distribution, like logistic regression—we applied a log transformation. This compressed large values, improved symmetry, and enhanced model compatibility (Figure 2).

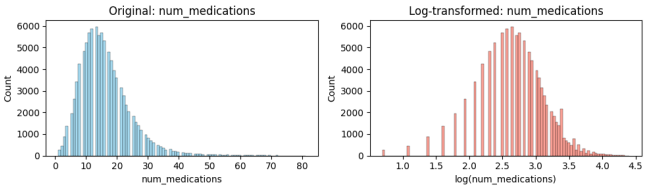


Figure 2: Distributions of original continuous features before and after transformation.

2.2.6 Data Preparation

After preprocessing, the final dataset consisted of 98,913 patient encounters and 50 features. The target variable `readmitted` was recoded into three classes: within 30 days, after 30 days, and not readmitted. Categorical features were one-hot encoded to avoid spurious ordinal effects.

A stratified split preserved class proportions in both sets: the training set included 52.7% `NO`, 35.8% `>30`, and 11.5% `<30`; the test set showed similar distributions. To prevent data leakage, all encounters from the same patient were assigned to only one subset.

Non-informative IDs (`patient_nbr` and `encounter_id`) were dropped. Selected numerical

features were standardized via z-score normalization (fit on training data only) to ensure zero mean and unit variance, improving model convergence and stability.

2.3 Models and Training

Model selection and hyperparameter tuning were performed on the training set using 5-fold stratified **cross-validation**, ensuring stable results across different data splits.

Final performance was evaluated on the held-out test set. To estimate variability and provide confidence intervals for each metric, we applied **bootstrap** resampling with 100 iterations.

The combination of cross-validation (for model tuning) and bootstrapping (for uncertainty estimation) yields a robust evaluation pipeline and more reliable generalization estimates.

Handling Class Imbalance To address the pronounced class imbalance, we employed:

- **Class weighting**: Applied during model training to penalize errors on minority classes.
- **SMOTE**: Synthetic oversampling of the minority class was performed using feature-space similarity.

Models We evaluated multiple classification algorithms offering different trade-offs between interpretability, scalability, and robustness to class imbalance:

- **Logistic Regression**: interpretable baseline for linear modeling.
- **Random Forest**: ensemble method capturing nonlinear interactions and robust to imbalanced data.
- **XGBoost**: gradient boosting framework optimized for accuracy and flexibility.

Additional models such as **SGDClassifier** and **Decision Trees** were explored during development but excluded from final evaluation due to lower performance.

Hyperparameter Tuning Hyperparameters were optimized through a combination of manual tuning and grid search. For logistic regression, we explored different regularization strategies (`L1`, `L2`), solvers (`liblinear`, `saga`), and penalty strengths (`C`). For tree-based models (Random Forest, Decision Tree, and XGBoost), we adjusted parameters such as `max_depth`, `min_samples_split`, and the number of estimators to balance generalization and overfitting.

Two-Stage Modeling with Gating Mechanism To improve detection of early readmissions (<30 days), we implemented a two-stage *gating architecture* aimed at increasing sensitivity toward this critical and underrepresented class.

The first stage (gate) performs binary classification to identify <30-day readmissions, while the second stage distinguishes between "NO" and >30-day cases, conditioned on the gate's output. This design reflects clinical screening logic—prioritizing recall for high-risk cases, even at the cost of reduced specificity in the first stage.

Logistic Regression, Random Forest, and XGBoost were evaluated in both stages using stratified cross-validation. While all models achieved comparable performance, Logistic Regression was ultimately preferred for its optimal balance between accuracy, interpretability, and computational efficiency.

Other Attempts Several experimental strategies were tested but excluded from the final pipeline due to marginal or inconsistent improvements.

Feature selection: `SelectKBest`, `RFECV`, and L1 regularization slightly improved minority-class recall but added complexity with minimal overall gain.

Ensemble methods: A soft-voting classifier (LogReg, SVM, RF) showed no advantage over simpler models.

Resampling: Random undersampling and NearMiss (V1, V3) increased minority recall but drastically reduced overall performance (e.g., F1 $\approx$ 0.31).

Data filtering: Using only first encounters or removing outliers did not improve performance compared to the full dataset.

Model	Accuracy (%)	F1 weighted (%)	Recall <30 (%)	Recall >30 (%)	Recall NO (%)
LogReg baseline	56.9 (56.3–57.7)	51.2 (50.5–52.1)	0.04 (0–0.1)	34.7 (33.8–35.8)	**85.3** (84.5–85.9)
LogReg balanced	50.4 (49.6–51.1)	51.4 (50.5–52.0)	34.4 (32.2–36.3)	38.2 (37.1–40.0)	62.5 (61.6–63.4)
RF balanced	**54.6** (53.8–55.4)	**54.2** (53.4–55.0)	22.4 (21.1–23.9)	46.4 (45.1–47.7)	67.5 (66.6–68.3)
RF + SMOTE	54.3 (53.5–55.0)	53.8 (53.0–54.5)	17.3 (15.9–18.9)	**49.1** (47.8–50.3)	66.2 (65.4–66.9)
XGBoost balanced	48.3 (47.5–49.3)	49.7 (48.9–50.7)	47.7 (45.7–49.6)	32.4 (31.2–33.4)	59.5 (58.5–60.3)
Gating model	41.9 (41.2–42.5)	44.5 (43.8–45.1)	**59.0** (57.1–61.1)	29.8 (28.9–30.7)	46.4 (45.5–47.4)

Table 2: Comparison of selected models on the test set. Each cell reports the mean value and 95% confidence interval from 100 bootstrap iterations. Boldface indicates the best model(s) for each metric.

3 Results and discussion

To account for class imbalance, model performance was primarily evaluated using the **F1-score**, which being the harmonic mean of precision and recall offers a more reliable assessment than accuracy. Key results are summarized in Table 2.

3.1 Model Performance and Comparison

Table 2 summarizes the performance of all evaluated models. The unbalanced **Logistic Regression** exhibited strong bias toward the majority class (`NO`), achieving high accuracy but failing almost entirely to detect early readmissions (`<30` recall: 0.04%). Introducing class weights significantly improved minority class sensitivity (34.4%) without degrading overall F1-score, confirming the utility of balancing strategies in skewed classification settings.

Among tree-based approaches, the **Random Forest with balanced training** achieved the best overall F1-score (54.2%) and maintained consistent recall across all classes, offering robust generalization. Combining SMOTE with Random Forest further boosted performance on the `>30` class but decreased recall for early readmissions, suggesting a trade-off in class focus when applying synthetic oversampling.

XGBoost, while less stable overall, reached the second-highest recall for `<30` (47.7%), indicating its capacity to prioritize minority class detection but with a notable drop in global performance and interpretability.

The **Gating Model** delivered the highest sensitivity to early readmissions (recall: 59.0%), clearly outperforming all others on this critical target. However, this came at the cost of accuracy and F1-score, reflecting its screening-oriented nature that favors sensitivity over precision and global balance.

Based on these results, three models were selected for further analysis, each representing a distinct trade-off relevant to deployment scenarios:

- **Balanced Logistic Regression** offers a competitive balance of interpretability, efficiency, and minority class recall. It is well-suited for clinical environments requiring explainable decisions and rapid integration.
- **Balanced Random Forest** delivers the highest overall performance with strong class-wise balance, making it an excellent general-purpose model when predictive accuracy is the main priority.
- The **Gating Model**, by design, optimizes for high sensitivity on early readmissions, making it ideal in high-risk clinical workflows where missing a critical case carries high consequences. It supports a screening logic where recall is prioritized over precision.

Ultimately, the selection of the optimal model should reflect the intended application context—whether interpretability, balanced performance, or high-risk case detection is prioritized.

3.2 Cross-Validation and Model Robustness

To assess generalization and prevent overfitting, 5-fold stratified cross-validation was performed (Table 3). **Random Forest** achieved the highest average F1-score, confirming its strong predictive capacity, but showed higher variability across folds. In contrast, **Logistic Regression** delivered more consistent performance with slightly lower scores. Interestingly, the **Gating Model** exhibited the lowest variance, suggesting stable behavior across data splits.

Table 3: 5-fold cross-validation performance for the selected models. Metrics are reported as mean ± standard deviation.

Model	Accuracy (%)	F1-weighted (%)
LogReg bal	48.0 ± 2.5	51.2 ± 0.3
RF bal	52.4 ± 2.1	52.2 ± 1.9
Gating	41.9 ± 0.3	44.7 ± 0.3

3.2.1 Model Interpretation

To improve interpretability and identify key drivers of readmission, we analyzed both intrinsic and model-agnostic feature importance.

For `Balanced Logistic Regression`, the most influential features (Figure 3) included prior inpatient history, discharge disposition, and payer type. Frequent past hospitalizations strongly correlated with early readmission, while administrative attributes—like transfers or insurance categories—reflected systemic factors influencing follow-up care and access.

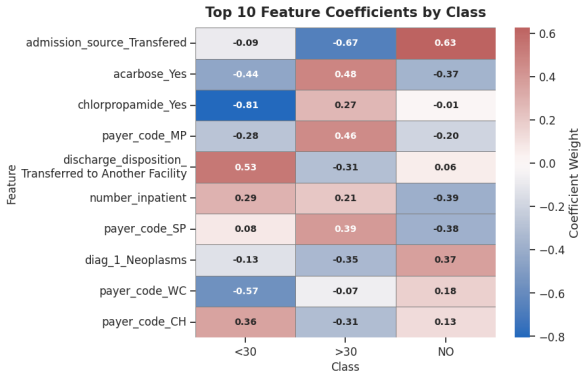


Figure 3: Top 10 feature coefficients by class for the `Balanced Logistic Regression` model.

These results highlight how both clinical and organizational factors contribute to readmission risk, offering interpretable insights to support targeted interventions.

Similarly, the `Random Forest` model emphasized prior admissions, length of stay, and discharge status (Figure 4), reinforcing the importance of care intensity and transition quality.

Overall, both models rely on clinically coherent signals, validating their relevance for real-world clinical decision support.

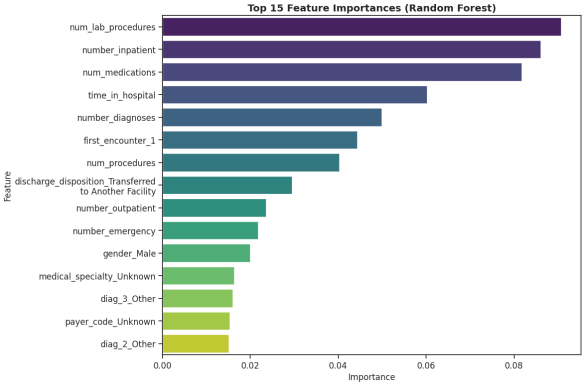


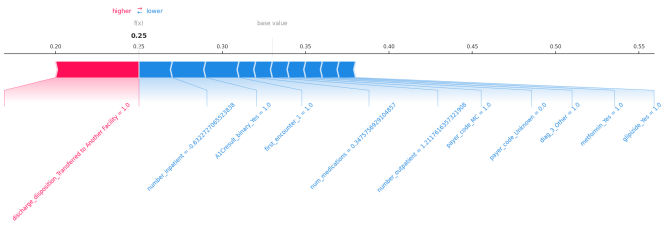
Figure 4: Top 15 most important features for the `Random Forest with SMOTE` model, based on impurity-based importance.

SHAP Analysis. To enhance interpretability, we used SHAP (SHapley Additive exPlanations) to analyze individual predictions from top-performing models. SHAP highlights how each feature contributes to a prediction, both globally and locally.

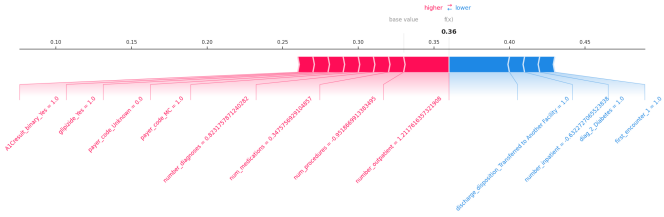
Figure 5 shows force plots for a patient misclassified as `NO` instead of `>30`. Clinical complexity (e.g., multiple diagnoses, outpatient visits) supported the correct `>30` label but was outweighed by signals linked to lower risk, like limited inpatient history. The incorrect `NO` prediction was driven by payer type, medication usage, and prior admissions features that obscured the true risk.

The `<30` class had weak influence, with discharge disposition as the main signal, but outweighed by low care intensity.

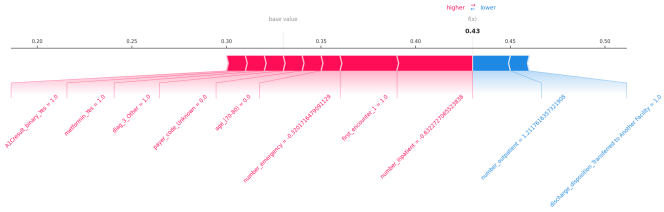
This case illustrates how SHAP clarifies the interplay of competing signals, helping to interpret errors and improve model transparency in clinical contexts.



(a) Prediction for class `<30`



(b) Prediction for class `>30`



(c) Prediction for class `NO`

Figure 5: SHAP force plots show feature contributions for an instance truly labeled `>30` but misclassified as `NO`, with red increasing and blue decreasing class probabilities.

3.3 Ablation Study

An ablation study on a simplified `Balanced Logistic Regression` model using only eight standardized numerical features showed minimal performance loss: 49.8% accuracy (95% CI: 49.0–50.5) and 48.8% weighted F1-score (95% CI: 48.0–49.7), close to the full model's 50.4% accuracy and 51.4% F1. Minority class recall ("`<30` days") remained comparable (34.2% [95% CI: 32.5–35.8] vs. 34.4%).

This suggests numerical features reflecting care intensity drive most predictive power, while categorical variables add limited accuracy, supporting low-dimensional models for efficient, interpretable readmission prediction.

4 Conclusions

This study tackled hospital readmission prediction, emphasizing early readmissions within 30 days due to their clinical and economic importance.

No single model dominated; each offered trade-offs aligned with different deployment needs. The **Balanced Logistic Regression** provided stable, interpretable performance with moderate early readmission recall, ideal for settings prioritizing transparency. The **Random Forest with SMOTE** balanced sensitivity and specificity well, improving recall across classes while relying on clinically meaningful predictors. The **Gating Model** excelled at early detection but with lower overall accuracy, suiting screening-focused applications with tunable thresholds.

Interpretability analyses, including SHAP, confirmed that models depend on key clinical features like prior admissions and discharge disposition, revealing how confounding factors affect misclassifications.

Ultimately, model choice should reflect clinical priorities, be it interpretability, accuracy, or early detection, while addressing class imbalance and ensuring trustworthiness for actionable predictions.

Future Directions Several extensions could improve the framework, including modeling temporal patterns with LSTMs [2], validating on external datasets to ensure generalizability, applying causal inference to clarify clinical factor effects, and integrating with explainable AI in clinical decision support for timely interventions [3].

References

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:1–11, 2014.

[2] Abolfazl Zarghani. Comparative analysis of lstm neural networks and traditional machine learning models for predicting diabetes patient readmission. 2024.

[3] et al. Liu. Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes. *Journal of Medical Artificial Intelligence*, 2024.