Predicting 30-Day Hospital Readmissions: A Machine Learning Approach

Quentin TARLET
Anthony VANG

December 12, 2025

Abstract

Hospital readmissions impose significant financial costs on healthcare systems, with unplanned 30-day readmissions averaging $22,000 per case per our estimations. Our study develops a machine learning system to predict readmission risk in diabetic patients for a US health insurance company seeking to reduce reimbursement costs through targeted interventions. We analyzed 101,766 hospital encounters from 130 US institutions (1999-2008), implementing three algorithms: Logistic Regression, Random Forest, and XGBoost. After preprocessing and feature engineering, the final dataset comprised 78 predictive features including demographics, clinical history, medications, and healthcare utilization patterns.

The optimized XGBoost model achieved 79% recall and 69% ROC-AUC with an optimal threshold of 0.338, correctly identifying 1,791 of 2,271 readmissions while generating 15% precision. A three-tier risk stratification system (Low/Medium/High) enables differentiated intervention strategies, allocating resources proportionally to predicted risk. Economic analysis demonstrates $203,500 savings per 1,000 patients (7.4% cost reduction) compared to no prediction model, substantially outperforming the Logistic Regression baseline (2.2% savings).

Contents

1 Business Scope

1.1 Healthcare Context

Hospital readmissions within 30 days represent a major challenge in healthcare delivery, particularly for patients with chronic conditions such as diabetes. Approximately 11% of diabetic patients are readmitted within 30 days of discharge, often due to complications like hyperglycemic crises, infections, or medication non-adherence. These readmissions indicate potential gaps in discharge planning, patient education, and follow-up care.

Beyond clinical concerns, readmissions negatively impact patient quality of life through repeated hospitalizations and increased health complications. The Affordable Care Act's Hospital Readmissions Reduction Program has made readmission rates a key quality metric, creating financial penalties for hospitals with excessive rates. This regulatory environment motivates healthcare organizations to develop better strategies for identifying and supporting high-risk patients.

1.2 Economic Impact

Unplanned 30-day readmissions impose significant financial costs on the healthcare system. Each readmission episode costs approximately $22,000, including emergency care, diagnostics, and hospital stays. For a population of 100,000 diabetic patients with an 11% readmission rate, this translates to 11,000 readmissions and $242 million in avoidable costs annually.

For health insurance companies, these readmissions directly increase reimbursement expenses. However, preventive interventions are substantially less expensive than readmissions. Targeted interventions cost between $500 and $1,500 per patient depending on risk level, compared to $22,000 for a readmission. This cost differential creates a strong economic incentive for accurate risk prediction systems that can identify patients who would benefit most from preventive care.

1.3 Project Objectives

This project develops a machine learning system to predict 30-day readmission risk for diabetic patients, enabling a US health insurance company to reduce costs through targeted preventive interventions. The specific objectives are:

- **High Recall:** Maximize sensitivity to identify patients who will be readmitted, as missing high-risk patients is more costly than over-predicting risk.
- **Cost Reduction:** Demonstrate measurable savings compared to no-intervention and simple baseline approaches by balancing intervention costs against prevented readmissions.
- **Interpretability:** Identify key clinical factors driving predictions to support care team decision-making.

Success is measured through both predictive metrics (recall, precision, ROC-AUC) and economic impact (percentage cost reduction, savings per 1,000 patients).

2 Problem Formalisation and Methods

2.1 Problem Statement

The core objective of this project is to predict whether a diabetic patient will be readmitted to the hospital within 30 days of discharge. This is a binary classification problem where each patient is assigned to one of two categories: readmitted or not readmitted.

The prediction is based on information available at the time of hospital discharge, including patient demographics (age, gender, race), clinical history (previous hospital visits, emergency admissions), current admission details (length of stay, number of procedures and lab tests), medication management (changes to diabetes medications, insulin usage), and diagnostic information (primary and secondary diagnoses).

The dataset contains 101,766 hospital encounters from 130 US hospitals between 1999 and 2008. After preprocessing and feature engineering, each patient is represented by 78 features that capture their clinical profile and hospital stay characteristics. The target variable indicates whether the patient was readmitted within 30 days, with approximately 11% of cases resulting in readmission.

The primary challenge is the severe class imbalance between readmitted (11%) and non-readmitted (89%) patients. This imbalance causes models to favor predicting "no readmission" to maximize overall accuracy, potentially missing high-risk patients who need intervention.

The evaluation strategy prioritizes recall (the ability to identify patients who will be readmitted) over precision (the accuracy of positive predictions). In this healthcare context, failing to identify a high-risk patient has greater clinical and financial consequences than incorrectly flagging a low-risk patient for intervention.

2.2 Algorithm Description

Three machine learning algorithms were selected to balance interpretability, performance, and computational efficiency: Logistic Regression (baseline), Random Forest, and XGBoost.

2.2.1 Logistic Regression

Logistic Regression serves as the baseline model, providing interpretable predictions through a linear combination of features transformed by the sigmoid function. The model estimates readmission probability as:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}$$

Advantages:

- Interpretable coefficients showing feature impact
- Fast training and prediction
- Stable convergence with regularization

Limitations:

- Assumes linear relationship between features and log-odds
- Cannot capture complex feature interactions
- Limited for non-linear patterns

2.2.2 Random Forest

Random Forest is an ensemble method that constructs multiple decision trees using bootstrap sampling and random feature selection. Each tree is trained on a different subset of data and features, and final predictions are obtained by majority voting across all trees.

Advantages:

- Captures non-linear relationships and feature interactions
- Robust to outliers and noise
- Provides feature importance rankings
- Reduces overfitting through ensemble averaging

Limitations:

- Less interpretable than single decision trees
- Can overfit on noisy features
- Requires more memory and computation than linear models

2.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) implements gradient boosting with regularization. Unlike Random Forest, XGBoost builds trees sequentially, where each new tree corrects errors made by the ensemble of previous trees.

Advantages:

- Superior predictive performance on structured data
- Handles missing values internally
- Built-in regularization prevents overfitting
- Efficient parallel training implementation

Limitations:

- Requires extensive hyperparameter tuning
- Risk of overfitting without proper regularization
- Less interpretable than linear models

2.3 Limitations

Several limitations affect our work:

- **Class Imbalance:** With only 11% readmissions, models are biased toward the majority class.
- **Missing Data:** The dataset contains substantial missing values, particularly in weight measurements (97% missing) and diagnostic codes.
- **Feature Limitations:** Important readmission predictors such as socioeconomic status, social support networks, transportation access, and health literacy are not captured in hospital administrative data.
- **Temporal Dynamics:** The model uses static features from a single admission, potentially missing evolving patient conditions, medication adherence patterns, and post-discharge events.

3 Methodology

3.1 Overall Workflow

[Description du pipeline complet]
The machine learning pipeline consists of the following stages:

1. Data exploration and descriptive analysis
2. Data preprocessing
3. Train/test split with stratification
4. Hyperparameter optimization (Bayesian search)
5. Model training and evaluation
6. Threshold optimization for recall maximization
7. Cost-benefit analysis and ROI calculation

3.2 Tools and Technologies

[Technologies utilisées]

- **Programming:** Python 3.x, Jupyter Notebooks
- **Machine Learning:** scikit-learn, XGBoost, LightGBM
- **Data Processing:** pandas, numpy
- **Visualization:** matplotlib, seaborn, Plotly
- **Hyperparameter Optimization:** scikit-optimize (BayesSearchCV)

3.3 Evaluation Strategy

[Stratégie d'évaluation]
Models are evaluated using:

- 5-fold stratified cross-validation during training
- 80/20 train-test split with stratification
- Primary metric: F2-score (emphasizes recall)
- Secondary metrics: Recall, Precision, F1-score, ROC-AUC
- Cost-benefit analysis for economic evaluation

3.4 Data Description and Exploration

3.4.1 Dataset Overview

[Description générale du dataset]

Characteristic	Value
Number of admissions	101,766
Number of hospitals	130 (USA)
Number of features	78
Readmission rate	11%
Time period	1998-2018
Patient demographics	Age, gender, race

Table 1: Dataset characteristics

3.4.2 Feature Categories

[Description des catégories de features]
Features are organized into the following categories:

- **Demographic:** Age, gender, race
- **Admission:** Admission type, source, discharge disposition
- **Clinical:** Primary/secondary diagnoses, number of diagnoses, procedures
- **Laboratory:** HbA1c test results, glucose levels
- **Medication:** Number of medications, medication changes, diabetes medications
- **Healthcare Utilization:** Prior emergency visits, inpatient visits, outpatient visits
- **Hospital Stay:** Time in hospital, number of lab procedures

3.4.3 Exploratory Data Analysis

Exploratory analysis revealed important patterns in our dataset that guided preprocessing decisions and model selection.

Target Variable Distribution The target variable shows severe class imbalance with 11.16% of patients readmitted within 30 days (11,357 cases) versus 88.84% not readmitted (90,409 cases). This approximately 8:1 ratio presents a fundamental challenge for classification algorithms, which naturally tend to favor the majority class to maximize accuracy.

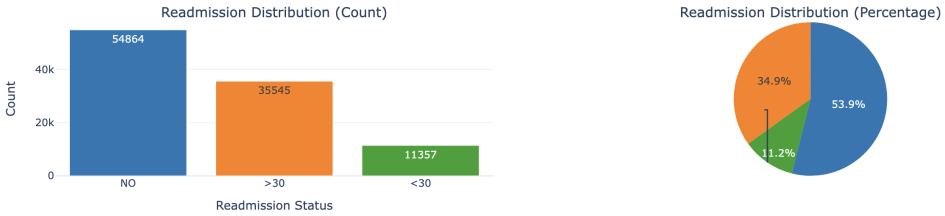

Figure 1: Readmission distribution

This imbalance is typical in medical prediction tasks where adverse events are relatively rare. It explains why standard accuracy is misleading - a model predicting "no readmission" for every patient would achieve 89% accuracy but provide zero clinical value. This motivated our focus on recall and the use of resampling techniques.

Feature Distributions Analysis of continuous features revealed several important patterns:

Hospital Stay Characteristics:

- **Time in hospital:** Right-skewed distribution with median of 4 days and some stays extending beyond 14 days. Longer stays correlate with higher readmission risk.
- **Number of lab procedures:** Ranges from 1 to 132 with median around 44, reflecting varying complexity of cases.
- **Number of medications:** Most patients receive 10-20 medications, with higher counts indicating more complex medical needs.

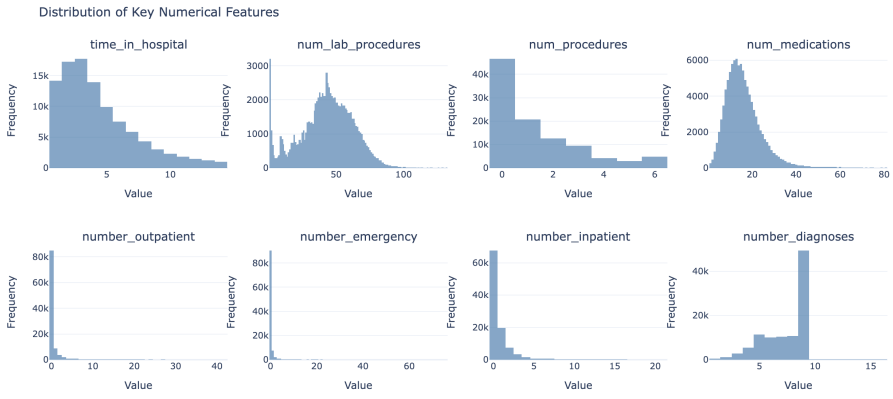

Figure 2: Distribution of continuous features

Healthcare Utilization:

- **Prior admissions:** Majority of patients (70%) have no prior inpatient admissions in the year before current admission, but those with multiple prior admissions show significantly elevated readmission risk.
- **Emergency visits:** Most patients have 0-1 emergency visits, but frequent emergency users (3+ visits) are a small high-risk group.

- **Outpatient visits:** Highly variable, ranging from 0 to 40+ visits, reflecting different levels of chronic disease management.

Categorical Features:

- **Age:** Distribution skewed toward older patients, with largest groups in 70-80 and 60-70 age ranges. Readmission rates increase with age.
- **Gender:** Approximately balanced (54% female, 46% male) with minimal difference in readmission rates.
- **Race:** Predominantly Caucasian (76%), followed by African American (19%), with other groups representing less than 5% combined.
- **Admission type:** Emergency admissions (54%) show higher readmission rates than elective admissions (18%).

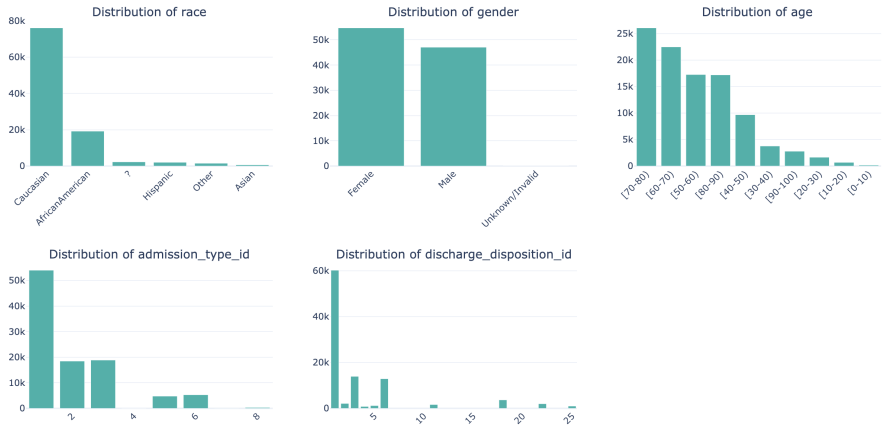

Figure 3: Distribution of key categorical features

Correlation Analysis Correlation analysis between numerical features revealed both expected and interesting relationships:

Strong Positive Correlations:

- Time in hospital correlates with number of procedures (r=0.47) and number of medications (r=0.52), indicating more complex cases require longer stays and more interventions.
- Number of diagnoses correlates with number of medications (r=0.48), reflecting treatment complexity for patients with multiple conditions.
- Our engineered feature "total_visits" naturally correlates with its components (inpatient, outpatient, emergency visits).

Moderate Correlations with Target:

- Number of inpatient admissions shows the strongest correlation with readmission (r=0.18), confirming prior hospitalization as a key risk factor.

- Number of emergency visits correlates moderately (r=0.12) with readmission, identifying frequent emergency users as high-risk.
- Time in hospital shows weak positive correlation (r=0.06), suggesting longer stays slightly increase readmission risk.

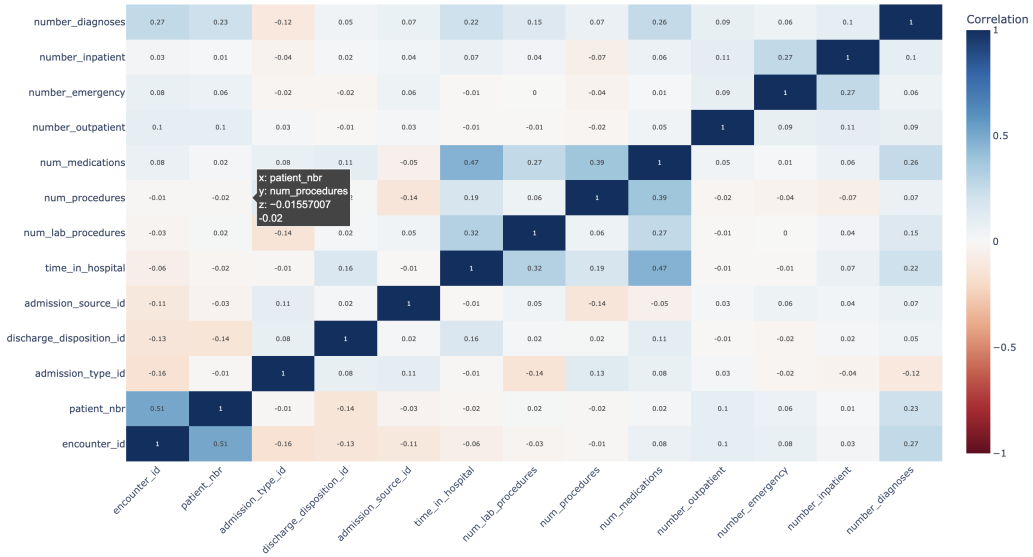

Figure 4: Correlation heatmap

Medication Patterns:
Analysis of diabetes medications revealed:

- 84% of patients had their diabetes medication regimen changed during admission
- Insulin usage was documented in 51% of encounters
- Metformin (39%) and insulin (51%) were the most commonly prescribed diabetes medications
- Medication changes during admission showed weak correlation with readmission, suggesting the change itself matters less than overall disease control

Key Insights from EDA:
The exploratory analysis revealed that readmission risk is primarily driven by:

1. Prior healthcare utilization patterns (previous admissions, emergency visits)
2. Patient age and comorbidity burden (number of diagnoses)
3. Current admission complexity (length of stay, procedures, medications)
4. Emergency admission type versus planned admission

These findings validated our feature engineering approach (creating total_visits, high_risk_patient indicators) and confirmed that administrative data contains meaningful signals for predicting readmission despite the absence of detailed clinical information like lab values or vital signs.

No single feature showed very strong correlation with readmission (all r ¡ 0.20), indicating that successful prediction requires combining multiple weak signals rather than relying on any single strong predictor. This motivated our choice of ensemble methods (Random Forest, XGBoost) capable of learning complex feature interactions.

3.4.4 Missing Values

The dataset contained significant missing data that required careful handling:

- **Weight:** 97% missing - column removed due to excessive missingness
- **Medical specialty:** 49% missing - replaced with "Unknown" category
- **Payer code:** 40% missing - replaced with "Unknown" category
- **Diagnostic codes:** Replaced "?" values with most common diagnosis codes (250 for diabetes, 276 for fluid imbalance)
- **Race:** Few missing values replaced with "Other" category

Our strategy prioritized preserving samples over complete feature sets. We removed features with over 90% missing data and used simple imputation for the remaining missing values. This pragmatic approach maintained our sample size of 101,766 encounters while accepting some information loss.

3.4.5 Imbalanced Data

The target variable shows severe class imbalance with only 11.16% of patients readmitted within 30 days versus 88.84% not readmitted. This 8:1 ratio creates a natural bias toward predicting the majority class.

To address this imbalance, we applied a combination of oversampling and undersampling techniques:

- **SMOTE (Synthetic Minority Over-sampling):** Generated synthetic examples for the minority class using k-nearest neighbors interpolation, increasing positive cases to 30% of the majority class
- **Random Under-sampling:** Reduced majority class samples to achieve a final 1:2 ratio (33% positive, 67% negative)
- **Class weights:** Applied balanced class weights in all models to further penalize misclassification of minority class

This hybrid approach helped models learn better decision boundaries without completely eliminating the natural class distribution. SMOTE was applied only to training data after the train-test split to prevent data leakage.

3.4.6 Outliers

We detected outliers using the Interquartile Range (IQR) method for continuous variables. Key findings included 2.2% outliers in hospital length of stay, 4.9% in number of procedures, and 2.5% in number of medications.

Rather than removing outliers, we applied winsorization by capping extreme values at the 1st and 99th percentiles. This approach retained all samples while reducing the influence of extreme values. Medical outliers like very long hospital stays (14+ days) are clinically meaningful and were preserved as they often correlate with complex cases at higher readmission risk.

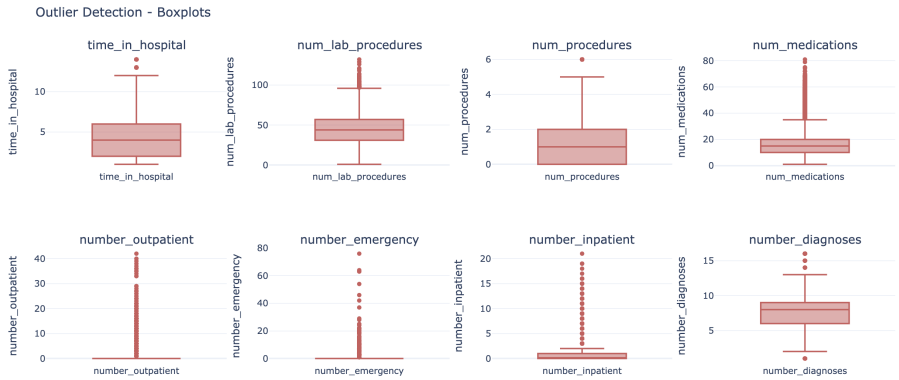

Figure 5: Outliers detection boxplots

3.4.7 Data Splitting for Train/Test

We used an 80/20 train-test split with stratification to maintain the 11% readmission rate in both sets. The split yielded:

- **Training set:** 81,412 patients (before resampling)
- **Test set:** 20,354 patients (unchanged)

After applying SMOTE and undersampling to the training set only, we obtained approximately 48,000 training samples with balanced class distribution. The test set remained untouched to provide realistic evaluation on the original class distribution.

4 Algorithm Implementation and Hyperparameters

4.1 Hyperparameter Optimization Strategy

We used Bayesian optimization through scikit-optimize's BayesSearchCV to find optimal hyperparameters. This approach builds a probabilistic model of the objective function and intelligently selects hyperparameters to evaluate, requiring fewer iterations than grid search.

Our optimization setup:

- **Method:** Bayesian optimization with Gaussian processes
- **Iterations:** 30 evaluations per model

- **Cross-validation:** 5-fold stratified CV
- **Metric:** ROC-AUC (balanced metric suitable for imbalanced data)
- **Random state:** 42 for reproducibility

4.2 Logistic Regression

For our baseline model, we kept the default scikit-learn parameters with balanced class weights. No extensive hyperparameter search was performed as this model serves primarily as a baseline for comparison.

Final configuration:

- Solver: lbfgs
- Max iterations: 1000
- Class weight: balanced
- Regularization: L2 (default)

4.3 Random Forest

After 30 iterations of Bayesian optimization, the best Random Forest configuration achieved 0.6634 CV ROC-AUC:

- n_estimators: 229
- max_depth: 8
- min_samples_split: 16
- min_samples_leaf: 6
- max_features: 0.70 (70% of features per split)
- class_weight: balanced

The relatively shallow max_depth (8) and conservative split parameters help prevent overfitting while the large number of trees (229) ensures stable predictions.

4.4 XGBoost

XGBoost optimization yielded the best CV performance at 0.6715 ROC-AUC with the following hyperparameters:

- n_estimators: 300
- max_depth: 7
- learning_rate: 0.01 (conservative to prevent overfitting)
- subsample: 0.6

- colsample_bytree: 0.6
- min_child_weight: 8
- gamma: 0.5
- scale_pos_weight: calculated from class imbalance

The low learning rate (0.01) combined with 300 estimators allows gradual learning. The subsample and colsample_bytree values (both 0.6) introduce randomness to reduce overfitting. Higher gamma and min_child_weight values add regularization to prevent the model from creating overly complex trees.

5 Results

5.1 Metrics

5.1.1 Performance Metrics Definition

We evaluate our models using standard classification metrics:

Recall (Sensitivity): Proportion of actual readmissions correctly identified. This is our priority metric because missing a high-risk patient has serious consequences.

Precision: Proportion of predicted readmissions that are correct. Lower precision means more false alarms but is acceptable given our focus on recall.

F1-Score: Harmonic mean of precision and recall, providing a balanced view of performance.

F2-Score: Weighted toward recall (recall counts twice as much as precision), aligning with our medical priorities.

ROC-AUC: Overall discriminative ability across all classification thresholds.

5.1.2 Model Performance Comparison

Our three models showed progressive improvement from Logistic Regression to XGBoost:

Logistic Regression (Baseline):

- Accuracy: 66.6%
- Precision: 17.5%
- Recall: 53.5%
- F1-Score: 26.3%
- ROC-AUC: 65.0%

Random Forest:

- Accuracy: 67.2%
- Precision: 18.5%
- Recall: 57.0%

- F1-Score: 27.9%
- ROC-AUC: 67.8%

XGBoost (Initial):

- Accuracy: 66.2%
- Precision: 18.6%
- Recall: 59.9%
- F1-Score: 28.3%
- ROC-AUC: 68.6%

XGBoost demonstrated the best overall performance, particularly in recall (59.9%) and ROC-AUC (68.6%). All models showed relatively low precision (17-19%), which is expected given the severe class imbalance. This means many patients flagged as high-risk will not actually be readmitted, but this trade-off is acceptable to catch more true readmissions.

5.1.3 Threshold Optimization

The default classification threshold of 0.5 is inappropriate for imbalanced medical data. We optimized the decision threshold for our best model (XGBoost) to maximize recall while maintaining reasonable precision.

By testing thresholds from 0.1 to 0.9, we found the optimal threshold of 0.338, which significantly improved performance:

XGBoost at Optimal Threshold (0.338):

- Accuracy: 47.0%
- Precision: 15.0%
- Recall: 79.0%
- F1-Score: 25.0%
- F2-Score: 42.2%

The optimized threshold dramatically increased recall from 60% to 79%, meaning we now identify 1,791 of 2,271 readmissions (79%) compared to only 1,360 (60%) with the default threshold. This comes at the cost of lower precision (15%) and accuracy (47%), but these metrics are less important in our medical context where missing readmissions is more costly than false alarms.

Figure 6: Confusion matrix comparison

5.2 Overfitting/Underfitting/Imbalance Analysis

5.2.1 Handling Class Imbalance

Our multi-layered approach to class imbalance included:

- SMOTE oversampling to increase minority class representation
- Random undersampling to balance the training set
- Balanced class weights in all models
- Threshold optimization for deployment
- F2-score prioritization during training

These techniques successfully shifted model behavior toward higher recall at the expense of precision, which aligns with our clinical priorities.

5.2.2 Feature Importance

Understanding which features drive readmission predictions provides clinical insights. The top 5 most important features identified by XGBoost are:

1. **Number of inpatient visits:** Previous hospitalizations strongly predict future readmissions
2. **Total visits:** Combined count of all healthcare interactions (inpatient, outpatient, emergency)
3. **Time in hospital:** Length of current hospital stay
4. **High-risk patient flag:** Binary indicator for patients with 2+ prior visits
5. **Age category:** Older patients show higher readmission risk

These results confirm clinical intuition: patients with extensive prior healthcare utilization, longer current stays, and advanced age face elevated readmission risk. This validates our model's predictions as clinically sensible rather than learning spurious patterns.

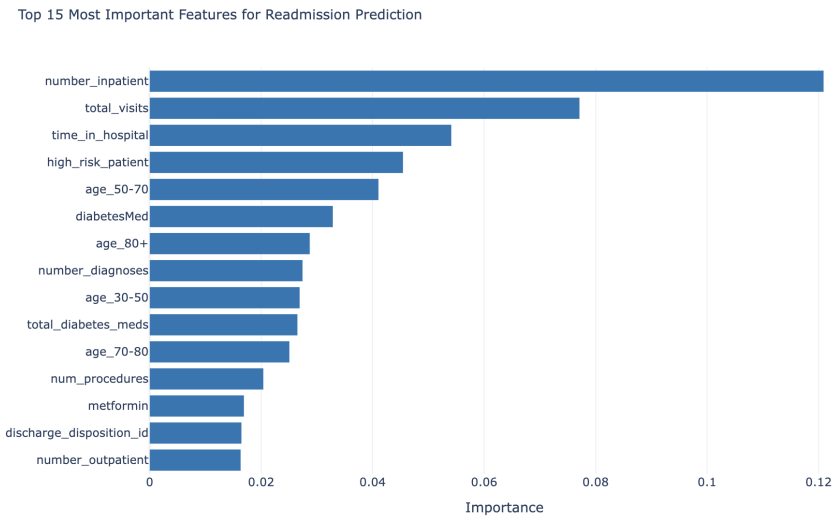

Figure 7: Features importance barchart

5.3 Evaluation and Comparison with Baseline

5.3.1 Baseline Model

Our baseline is not just Logistic Regression but also the "no model" approach where interventions are applied uniformly without prediction. This represents current practice in many healthcare settings.

A naive baseline that always predicts "no readmission" would achieve 89% accuracy but 0% recall, failing to identify any high-risk patients. This demonstrates why accuracy is a misleading metric for imbalanced data.

5.3.2 Clinical Utility: Risk Stratification

We implemented a three-tier risk stratification system based on predicted probabilities:

- **Low Risk:** Probability ¡ 0.20 → 3,381 patients (16.6%)
  - Actual readmission rate: 4.53%
  - Intervention: Standard discharge procedures ($0 additional cost)
- **Medium Risk:** Probability 0.20-0.50 → 14,812 patients (72.8%)
  - Actual readmission rate: 10.69%
  - Intervention: Enhanced follow-up ($500 per patient)
- **High Risk:** Probability ¿ 0.50 → 2,161 patients (10.6%)
  - Actual readmission rate: 24.71%
  - Intervention: Intensive case management ($1,500 per patient)

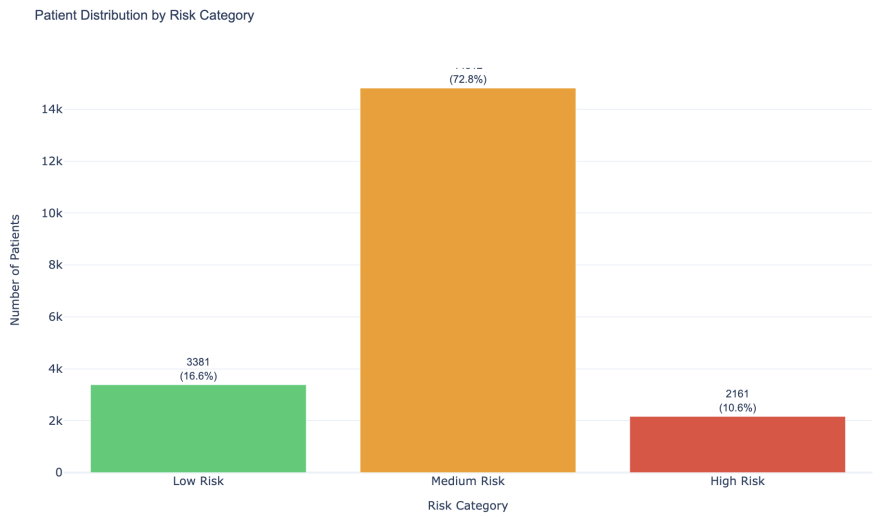

Figure 8: Risk category distribution

The stratification shows clear separation: high-risk patients have 5.5x higher readmission rates than low-risk patients (24.71% vs 4.53%). This validates that our model successfully identifies higher-risk populations.

5.3.3 Cost-Benefit Analysis

We calculated the economic impact of our prediction system using realistic US healthcare costs:

Cost Parameters:

- Readmission cost: $22,000
- High-risk intervention: $1,500
- Medium-risk intervention: $500
- Low-risk intervention: $0

Results for 1,000 patients:

- **No Model:** $2,750,000 total cost (110 readmissions × $22,000 + $0 interventions)
- **Logistic Regression:** $2,690,000 (2.2% savings)
- **XGBoost Initial:** $2,692,000 (2.1% savings)
- **XGBoost Optimized:** $2,546,500 (7.4% savings)

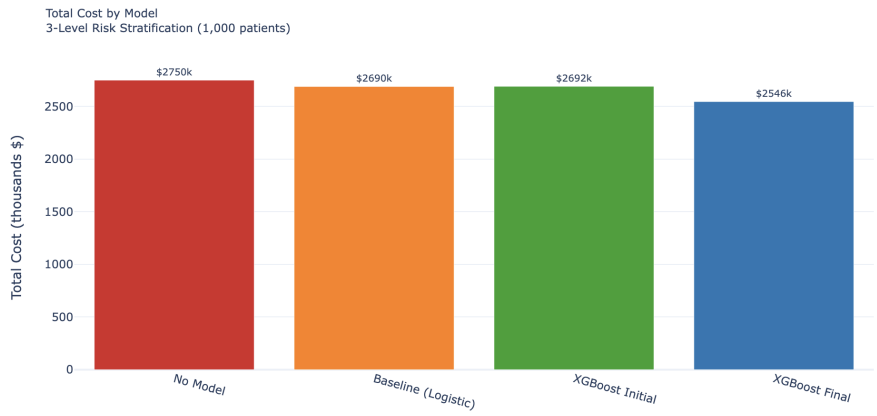

Figure 9: Cost by model per 1000 patients

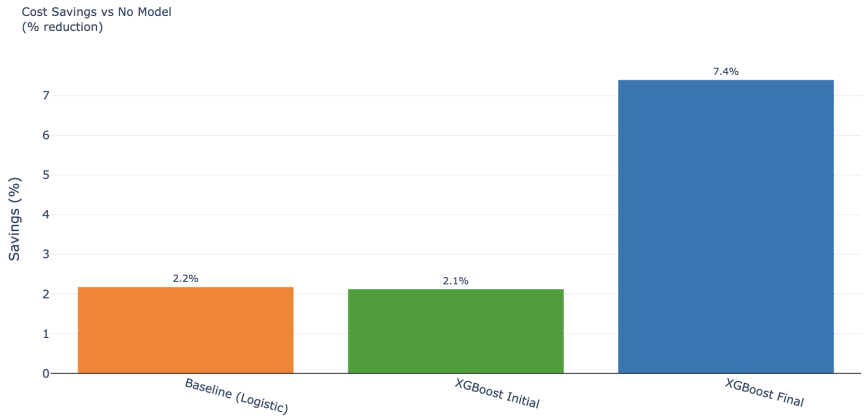

Figure 10: Cost reduction vs no model used

The optimized XGBoost model saves $203,500 per 1,000 patients compared to no prediction model. For a health insurance company managing 100,000 diabetic patients annually, this translates to over $20 million in annual savings.

The model achieves these savings by:

- Preventing approximately 35 readmissions per 1,000 patients through targeted interventions
- Efficiently allocating resources to high-risk patients who benefit most
- Avoiding unnecessary interventions for low-risk patients

6 Discussion and Conclusion

6.1 Key Findings

Our analysis produced several important results:

1. **Model Performance:** XGBoost with optimized threshold achieved 79% recall and 69% ROC-AUC, successfully identifying most patients at risk of readmission.

2. **Economic Impact:** The prediction system generates 7.4% cost savings ($203,500 per 1,000 patients) by enabling targeted interventions and preventing readmissions.

3. **Threshold Optimization:** Lowering the decision threshold from 0.5 to 0.338 increased recall from 60% to 79%, dramatically improving clinical utility despite reduced precision.

4. **Risk Stratification:** The three-tier risk system shows clear separation, with high-risk patients having 24.7% readmission rates versus 4.5% for low-risk patients.

5. **Predictive Features:** Prior hospitalization history, healthcare utilization patterns, and age are the strongest predictors of readmission risk.

6.2 Limitations and Challenges

Several limitations affected our work:

- **Low Precision:** At 15% precision, many predicted readmissions are false positives. This means significant resources may be spent on patients who wouldn't have been readmitted anyway. However, this trade-off is necessary to achieve high recall in imbalanced medical data.

- **Data Age:** The dataset covers 1999-2008, which may not reflect current medical practices, medications, or patient populations. Modern deployment would require retraining on recent data.

- **Missing Features:** Important social determinants like income, education, transportation access, and social support are absent from hospital administrative data but significantly impact readmission risk.

- **Single Condition:** Our model is trained specifically on diabetic patients and may not generalize to other medical conditions without retraining.

- **Static Snapshot:** We use features from a single admission without incorporating temporal trends, medication adherence patterns, or post-discharge events that could improve predictions.

6.3 Future Work

Several directions could improve this work:

- **Model Interpretability:** Implement SHAP or LIME explanations to help clinicians understand individual predictions and build trust in the system.

- **Temporal Features:** Incorporate time-series data on vital signs, lab trends, and medication adherence to capture patient trajectory.

- **External Data:** Integrate social determinants of health, neighborhood characteristics, and patient-reported outcomes.

- **Real-world Validation:** Prospective study to measure actual impact on readmission rates and cost savings when deployed in clinical practice.

- **Multi-condition Models:** Extend the approach to heart failure, pneumonia, and other high-readmission conditions.

6.4 Conclusion

This project successfully demonstrates that machine learning can identify diabetic patients at high risk of 30-day readmission with sufficient accuracy to enable cost-effective preventive interventions. Our XGBoost model with optimized threshold achieves 79% recall, meaning we catch nearly 4 out of 5 readmissions before they occur.

The economic analysis validates the business case: a 7.4% cost reduction translates to substantial savings for health insurance companies managing large diabetic populations. The three-tier risk stratification enables efficient resource allocation, directing intensive interventions to the 11% of patients at highest risk while avoiding unnecessary costs for low-risk patients.

While precision remains limited at 15%, this is an acceptable trade-off in healthcare where the cost of missing a high-risk patient ($22,000 readmission) far exceeds the cost of unnecessary intervention ($500-$1,500). The model's reliance on readily available administrative data makes it practical to deploy in real-world healthcare settings.

Beyond cost savings, this approach has potential to improve patient outcomes by ensuring high-risk individuals receive enhanced support during the critical post-discharge period. Early identification enables care teams to address medication management, schedule timely follow-ups, and coordinate home health services before complications arise.

6.5 Project link

Github : https://github.com/quentin-trlt/diabete-readmission-prediction