Hospital Readmission Final Modeling Report

Authors: Andrew Castillo, Terry Zhuang, Sarthak Dhanke, Jiayi Wang

Abstract

Hospital readmissions pose significant financial and clinical challenges, particularly among patients with chronic conditions such as diabetes. To address this issue, we developed a predictive modeling framework to assess the likelihood of 30-day hospital readmission, leveraging machine learning techniques on patient-level clinical and demographic data. The modeling approach includes handling class imbalance through SMOTENC-based synthetic oversampling and class weighting, hyperparameter tuning to maximize the F2 score (which prioritizes recall over precision), and model calibration to ensure accurate probability outputs. Key predictive features identified using SHAP (Shapley Additive Explanations) include risk score, inpatient visit history, number of procedures, medication counts, and discharge disposition. We evaluated multiple models, including Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost, comparing their performance under different class balancing techniques. Threshold tuning was conducted to optimize cost-effectiveness, considering the $16,300 average cost per readmission and a $2,500 intervention cost per high-risk patient. Our findings indicate that XGBoost with Class Weights provides the most consistent financial savings, while XGBoost with SMOTE yields the highest savings under peak intervention effectiveness.

By integrating machine learning with cost-benefit analysis, this study provides a scalable, explainable, and financially sustainable solution for reducing preventable hospital readmissions. The proposed framework supports data-driven decision-making in healthcare, improving resource allocation, patient care, and hospital operational efficiency.

Table of Contents

Problem Statement

Problem Defined

Hospital readmissions are a critical issue in the healthcare system, particularly for patients with chronic conditions such as diabetes. The 30-day readmission rate among diabetic patients poses significant financial and clinical challenges for hospitals, insurers, and patients. From a financial perspective, hospitals may face substantial penalties under the Hospital Readmissions Reduction Program (HRRP), with potential fines reaching up to $217,000 per hospital annually. Additionally, frequent readmissions contribute to increased healthcare costs, straining hospital resources and insurance systems.

Beyond financial burdens, readmissions are associated with worsened patient outcomes, including increased morbidity and diminished quality of care. Many factors contribute to the likelihood of readmission, including a patient's clinical history, demographic characteristics, discharge information, and prescribed medications. However, accurately identifying high-risk patients remains a challenge due to the complex interplay of these factors.

Proposed Supporting System

To address this issue, predictive modeling techniques can be employed to develop robust risk assessment tools for hospital readmissions among diabetic patients. By leveraging machine learning and statistical modeling, hospitals can:

- Identify high-risk patients early, allowing healthcare providers to allocate targeted interventions.
- Improve long-term health outcomes by facilitating proactive management strategies, such as enhanced follow-up care, medication adjustments, and patient education programs.
- Reduce hospital readmission rates and associated costs for healthcare institutions, insurers, and patients, ultimately improving the efficiency of healthcare delivery.

Goal

This study aims to develop predictive models that assess the likelihood of hospital readmission among diabetic patients within 30 days of discharge. By analyzing relevant patient data and identifying key risk factors, this research seeks to provide actionable insights that can aid hospitals in implementing data-driven strategies to improve patient care and reduce preventable readmissions.

Data Overview

Dataset Description

The primary dataset used in this study originates from 130 US hospitals and integrated delivery networks, spanning a ten-year period (1999–2008). This dataset specifically focuses on diabetes-related hospital admissions and subsequent 30-day readmissions, which serve as the target variable for predictive modeling. The dataset is publicly available through the UCI Machine Learning Repository: Diabetes 130-US Hospitals Dataset (1999–2008).

Key Features

The dataset includes clinical, demographic, and administrative information related to hospital stays. Some of the critical variables are:

- Clinical Factors: Duration of hospital stay, lab test results, primary and secondary diagnoses, prescribed medications.
- Demographics: Patient attributes such as age, race, and gender.
- Admission & Discharge Details: Admission type, admission source, and discharge disposition.
- Medication History: Use of diabetes-related medications, including insulin, metformin, and other diabetes treatments.
- Target Variable: 30-day hospital readmission, which is modeled as a binary classification problem (readmitted within 30 days or not).

Data Quality

Several data-related challenges and modeling considerations emerge from this dataset:

- Presence of missing values in key features such as race and laboratory tests.
- High dimensionality, with over 50 variables influencing patient outcomes.

Complexity of Readmission Prediction

- Multiple comorbidities and external socioeconomic factors impact readmission likelihood.
- The interplay between patient demographics, prescribed medications, and hospital policies introduces challenges in defining causality.

Class Imbalance

- The dataset exhibits an imbalance in readmission rates, with only ~11% of cases belonging to the positive class (readmitted within 30 days).
- This imbalance necessitates the use of techniques like resampling, cost-sensitive learning, or advanced evaluation metrics to avoid biased model performance.

Modeling Considerations

- The trade-off between model interpretability and predictive accuracy is crucial, given the clinical setting.
- Explainability of predictions is essential for real-world adoption, requiring approaches like logistic regression, decision trees, or SHAP-based feature analysis in complex models.

Exploratory Data Analysis

The dataset used in this study consists of 101,766 patient records collected from 130 US hospitals and integrated delivery networks over a ten-year period (1999–2008). It focuses on diabetes-related hospital admissions and examines 30-day readmissions as the primary outcome variable. The dataset contains 50 attributes covering demographics, clinical factors, admission details, medication history, and discharge information.

The target variable is "readmitted", indicating whether a patient was readmitted within 30 days, after 30 days, or not at all.

Attribute Analysis

Each feature was systematically evaluated concerning its data type, prevalence of missing values, presence of noise, utility in predictive modeling, and statistical distribution.

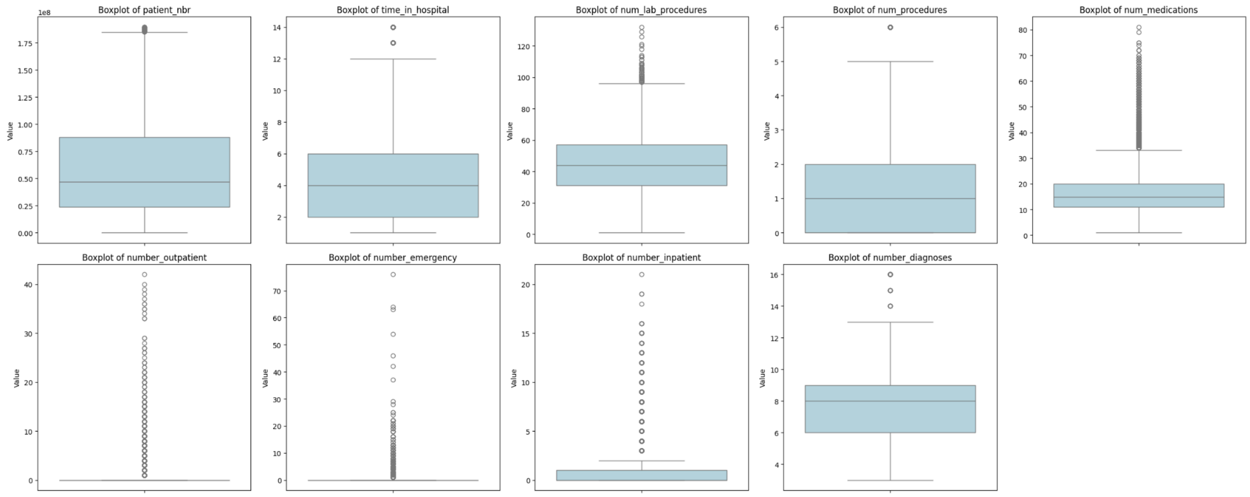
Feature	Type	Missing Values	Issues & Considerations	Predictive Utility	Distribution
`encounter_id`	Integer (ID)	0%	Unique identifier; not predictive	No	Unique ID
`patient_nbr`	Integer (ID)	0%	Redundant patient identifier	No	Unique ID
`race`	Categorical	~2% (? values)	Missing values necessitate imputation or encoding	Yes	Discrete
`gender`	Categorical	0%	Binary classification	Yes	Binary
`age`	Categorical (brackets)	0%	Requires conversion to numerical bins	Yes	Discrete
`weight`	Categorical	~97%	High sparsity; likely uninformative	No	Sparse
`admission_type_id`	Integer (Categorical)	0%	Requires mapping to categorical labels	Yes	Discrete

discharge_disposition_id	Integer (Categorical)	0%	Requires mapping to categorical labels	Yes	Discrete
admission_source_id	Integer (Categorical)	0%	Requires mapping to categorical labels	Yes	Discrete
time_in_hospital	Integer	0%	Key predictor	Yes	Right-skewed
num_lab_procedures	Integer	0%	No inherent noise	Yes	Gaussian-like
num_procedures	Integer	0%	High frequency of zero values	Yes	Skewed
num_medications	Integer	0%	Strong association with readmission	Yes	Right-skewed
number_outpatient	Integer	0%	Many zero values; sparse distribution	Maybe	Skewed
number_emergency	Integer	0%	Sparse, but relevant for high-risk patients	Maybe	Skewed
number_inpatient	Integer	0%	High importance for severe cases	Yes	Skewed
diag_1, diag_2, diag_3	Categorical	0%	ICD-9 coding; requires aggregation	Yes	Discrete

number_diagnoses	Integer	0%	No missing values; useful	Yes	Right-skewed
max_glu_serum	Categorical	~50%	High missing rate; may require imputation	Maybe	Sparse
A1Cresult	Categorical	~80%	Significant sparsity; utility uncertain	Maybe	Sparse
medication columns	Categorical	0%	Ordinal encoding necessary	Yes	Discrete
change	Categorical	0%	Binary (Yes/No)	Yes	Binary
diabetesMed	Categorical	0%	Binary (Yes/No)	Yes	Binary
readmitted	Categorical	0%	Requires binarization (<30 days vs. No)	Yes	Imbalanced

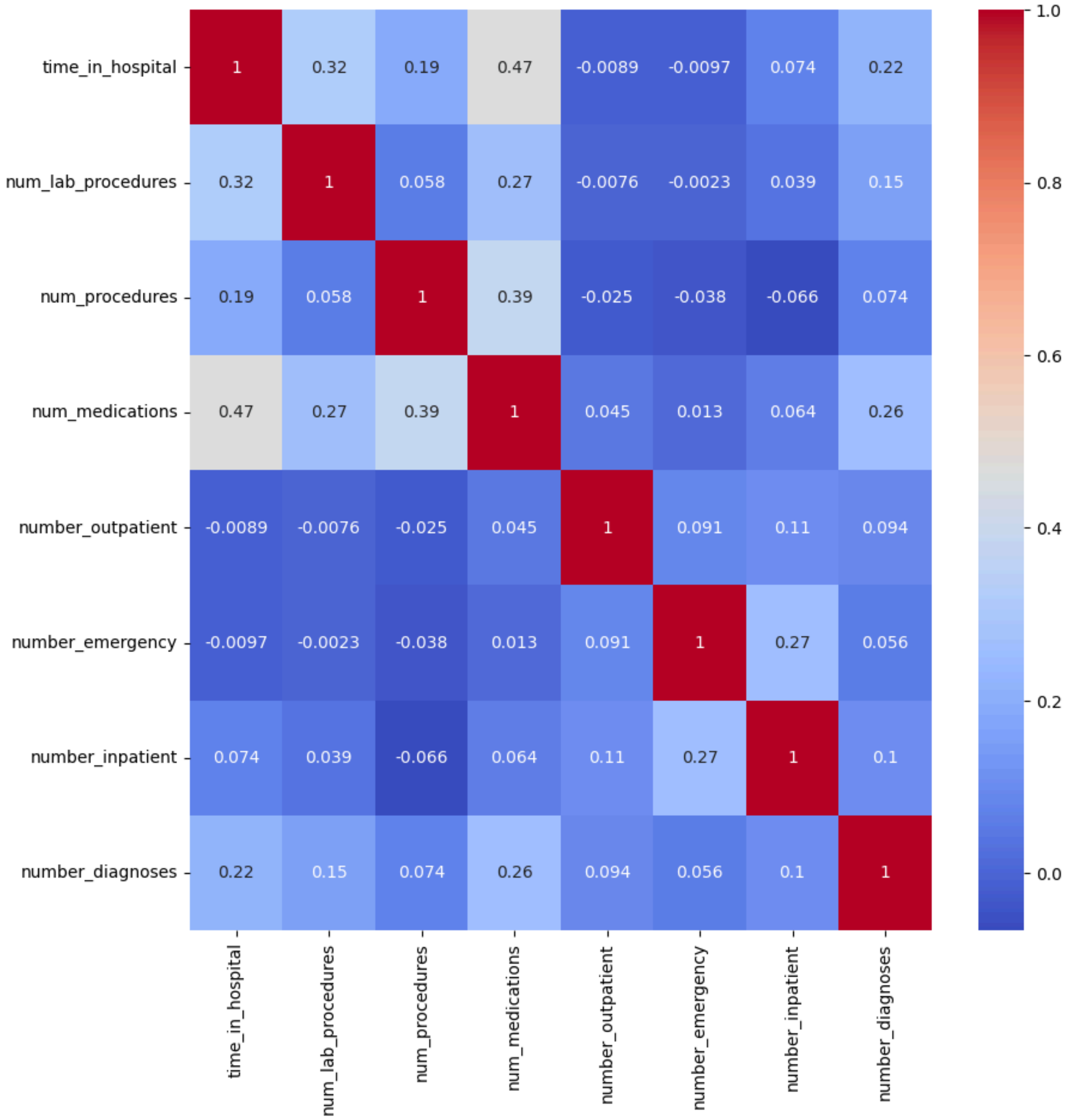
Please see Appendix for visualizations of all the features

Outlier Analysis and Retention



Outlier detection was performed using Interquartile Range (IQR) analysis and visualized through boxplots on numerical features. The boxplots indicate the presence of some extreme values, particularly in the number of outpatient visits, number of emergency visits, and number of inpatient visits. However, given the class imbalance in the dataset—where early readmissions are relatively rare—these outliers may represent anomalies that are critical to the predictive task. In this case, rather than removing outliers, they were retained as they could be valuable indicators of early hospital readmissions. Removing these data points could result in the loss of important predictive information, reducing the model's ability to capture real-world trends.

Correlation Analysis



	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses
time_in_hospital	1	0.32	0.19	0.47	-0.0089	-0.0097	0.074	0.22
num_lab_procedures	0.32	1	0.058	0.27	-0.0076	-0.0023	0.039	0.15
num_procedures	0.19	0.058	1	0.39	-0.025	-0.038	-0.066	0.074
num_medications	0.47	0.27	0.39	1	0.045	0.013	0.064	0.26
number_outpatient	-0.0089	-0.0076	-0.025	0.045	1	0.091	0.11	0.094
number_emergency	-0.0097	-0.0023	-0.038	0.013	0.091	1	0.27	0.056
number_inpatient	0.074	0.039	-0.066	0.064	0.11	0.27	1	0.1
number_diagnoses	0.22	0.15	0.074	0.26	0.094	0.056	0.1	1

A Pearson correlation analysis among numerical variables revealed:
`time_in_hospital` has a moderate correlation with `num_medications` (0.47), which aligns with the expectation that patients requiring longer hospital stays are likely to receive a higher number of medications for treatment and management of their condition. This correlation is particularly relevant when considering models incorporating medication-based predictors, as prolonged hospitalization may indicate severe cases requiring intensive pharmaceutical interventions. Additionally, `time_in_hospital` shows a weaker correlation with `num_lab_procedures` (0.32) and `num_procedures` (0.19), suggesting that while these factors contribute to the overall hospital experience, they are not as strongly tied to the length of stay as medication administration.

Similarly, `num_medications` and `num_procedures` exhibit a notable association (0.39), reflecting the likelihood that patients undergoing more procedures also receive a broader range of medications for post-procedure management and recovery. The correlation between `number_inpatient` and `number_emergency` (0.27) is relatively weak, indicating that while emergency visits may contribute to inpatient admissions, other factors such as planned treatments or elective procedures likely play a more significant role.

Overall, the relatively low correlation coefficients across numerical features suggest that this dataset exhibits high-dimensional and complex relationships with intricate interdependencies. This highlights the need for advanced modeling approaches that account for potential non-linear interactions and multicollinearity, particularly in regression-based analyses where highly correlated predictors may distort model coefficients and lead to biased interpretations.

Data Processing & Engineering

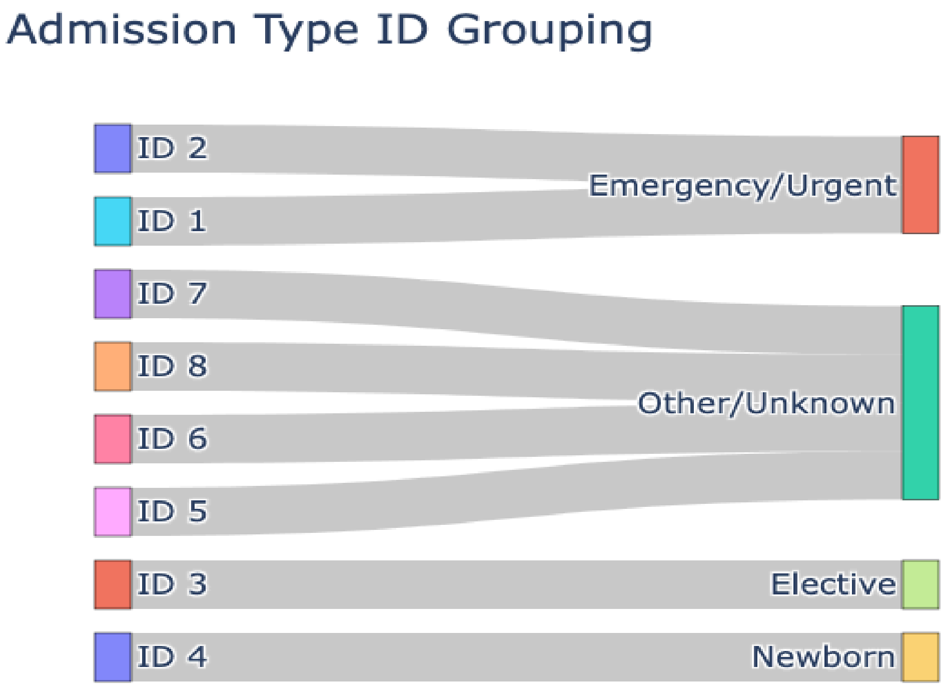
Handling Missing Values

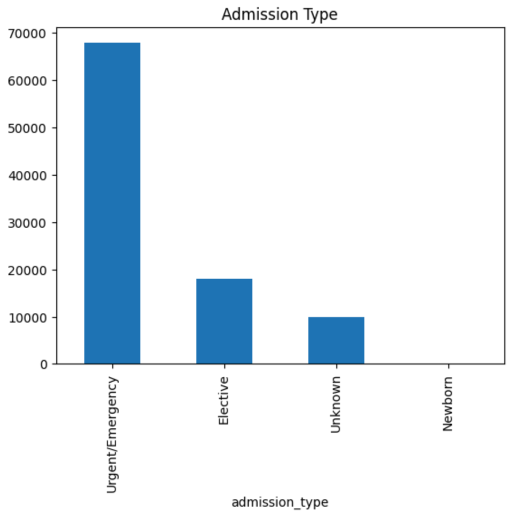
Several features in the dataset contained a significant percentage of missing values. Features with excessively high missing rates, such as `weight` (98% missing), `payer_code` (40% missing), `medical_specialty` (50% missing), `max_glu_serum` (95% missing), and `A1Cresult` (83% missing), were removed from the dataset. The missing values represented a considerable proportion of the data, and imputing them would have introduced uncertainty. Additionally, placeholders such as `?` and `Unknown/Invalid` were identified and replaced with `NaN` values. The remaining missing values were dropped, as most of the affected columns were categorical and did not significantly impact predictive modeling.

Categorical Encoding & Data Type Adjustments

To enhance interpretability, categorical identifier columns, including `admission_type_id`, `discharge_disposition_id`, and `admission_source_id`, were converted to categorical data types. This step ensured that numerical identifiers were not mistakenly treated as continuous variables. Standardization of categorical text values was also performed to maintain consistency across the dataset.

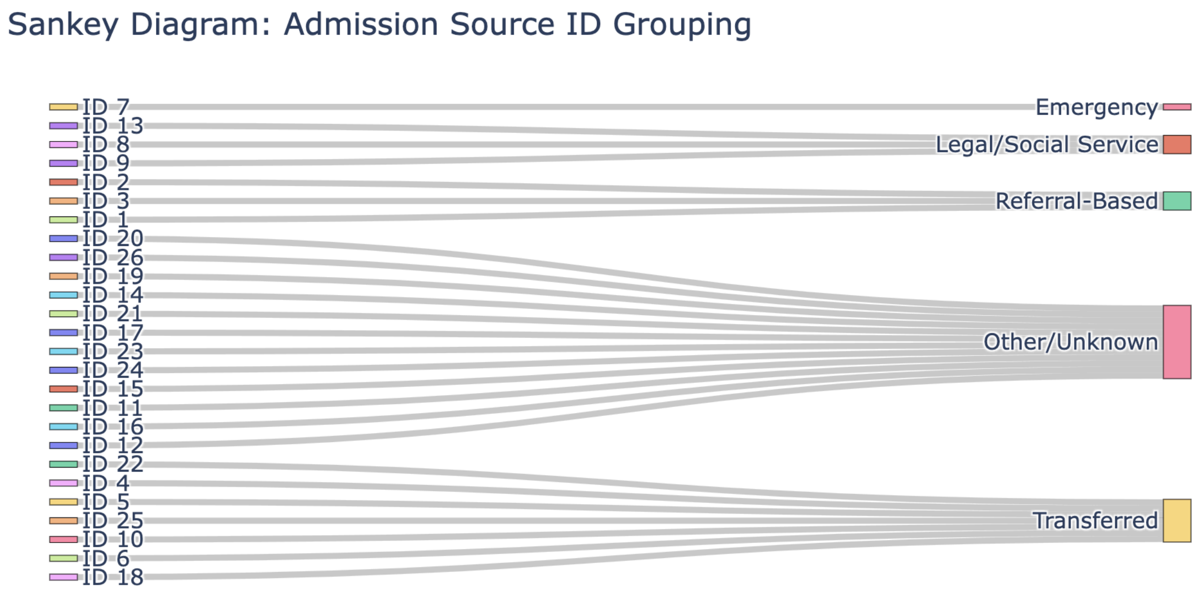
Simplifying Admission Type Categories

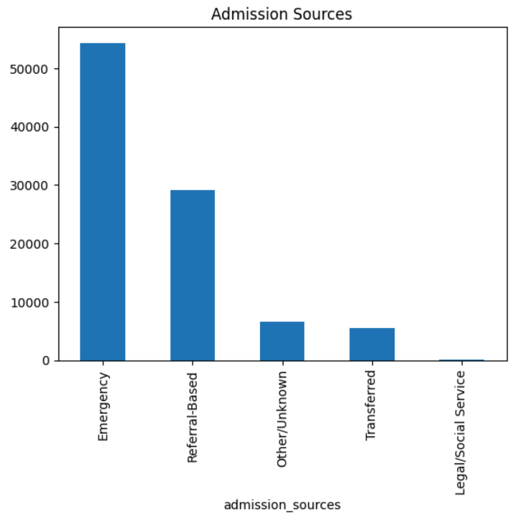




The original dataset included eight different admission types. To enhance model interpretability, these categories were consolidated into four meaningful groups. Admissions categorized as Emergency, Urgent, and Trauma Center were combined into a single category called Emergency/Urgent. Elective admissions were maintained as a separate category, as they follow distinct medical protocols. Newborn admissions were also preserved as an independent category, as they hold unique clinical significance. Finally, admissions labeled as Not Available, Not Mapped, and NaN were grouped under the Unknown category to simplify analysis.

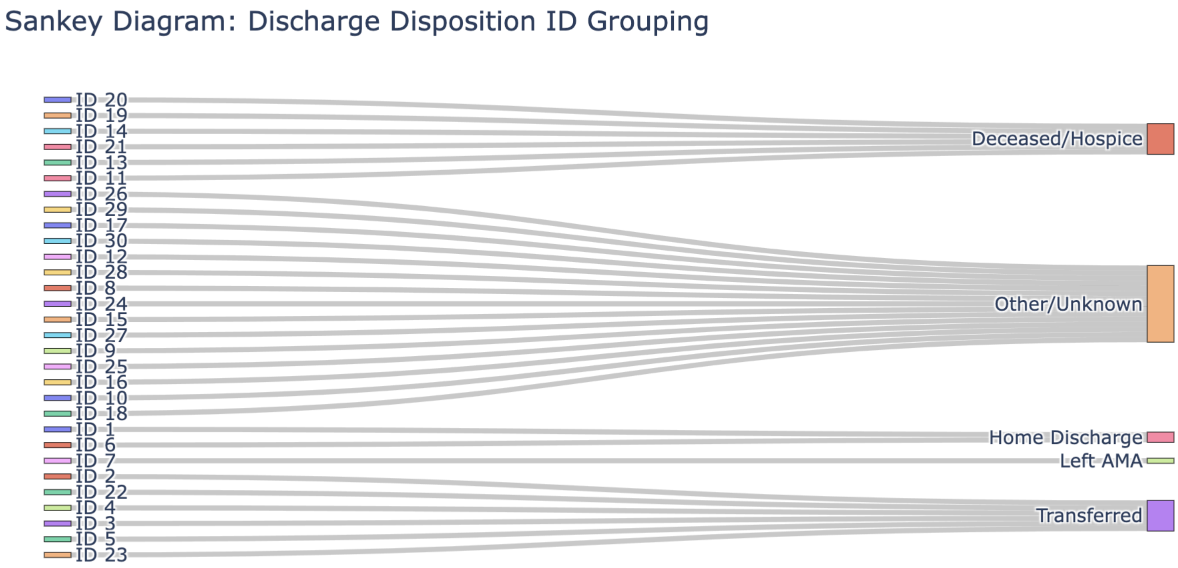
Mapping Admission Sources

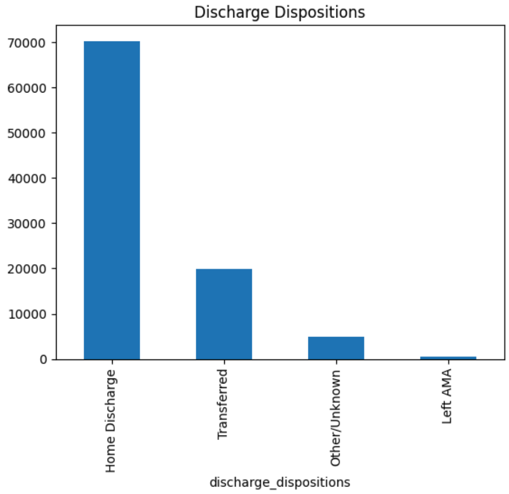




Similarly, the dataset contained 25 different admission source categories. These were grouped into five main categories: Emergency admissions included all patients admitted through the emergency room. Referral-Based admissions encompassed physician, clinic, and HMO referrals. Transferred admissions included patients transferred from hospitals, skilled nursing facilities, or other healthcare settings. Legal/Social Service admissions were categorized separately to account for cases involving court-ordered hospitalization or social service interventions. The remaining categories were labeled as Other/Unknown to handle miscellaneous cases.

Mapping Discharge Dispositions

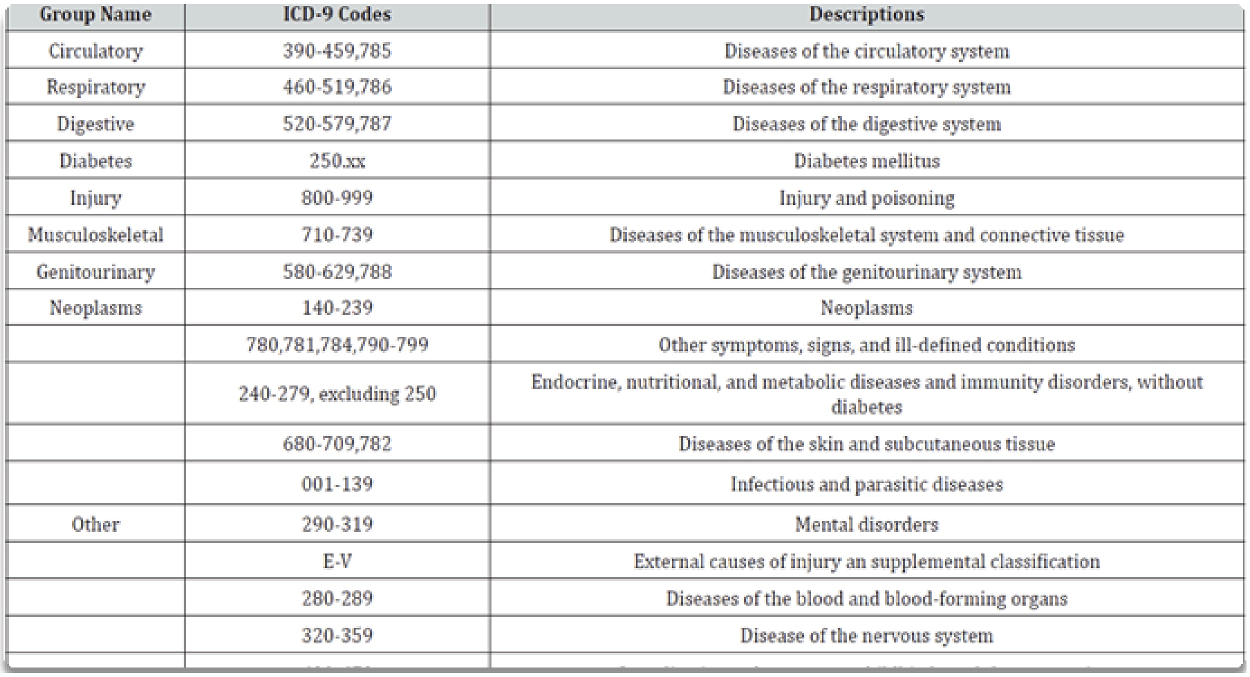




Discharge disposition refers to the patient's status at the time of discharge. The dataset originally contained 30 different categories for discharge disposition, which were condensed into four primary groups. Home Discharge included patients who were discharged to their residence or received home healthcare services. Transferred patients included those moved to other healthcare facilities such as skilled nursing facilities or rehabilitation centers. Left Against Medical Advice (AMA) cases were categorized separately, as these patients chose to leave before completing treatment. All other miscellaneous cases were grouped under Other/Unknown.

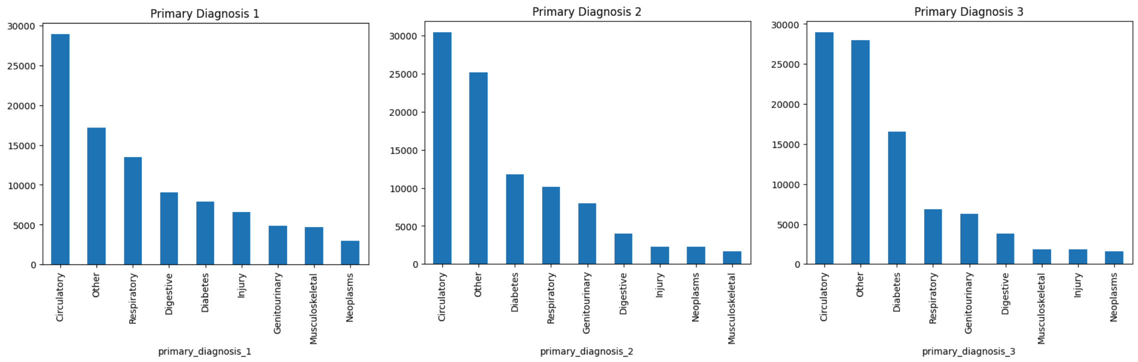
Additionally, basing on domain knowledge, since the objective is to predict readmissions, those patients who died during this hospital admission were excluded. Encounters with “Discharge disposition” values of 11, 13, 14, 19, 20, or 21 are related to death or hospice which means these patients cannot be readmitted.

Encoding Primary Diagnoses



Group Name	ICD-9 Codes	Descriptions
Circulatory	390-459,785	Diseases of the circulatory system
Respiratory	460-519,786	Diseases of the respiratory system
Digestive	520-579,787	Diseases of the digestive system
Diabetes	250.xx	Diabetes mellitus
Injury	800-999	Injury and poisoning
Musculoskeletal	710-739	Diseases of the musculoskeletal system and connective tissue
Genitourinary	580-629,788	Diseases of the genitourinary system
Neoplasms	140-239	Neoplasms
Other	780,781,784,790-799	Other symptoms, signs, and ill-defined conditions
	240-279, excluding 250	Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
	680-709,782	Diseases of the skin and subcutaneous tissue
	001-139	Infectious and parasitic diseases
	290-319	Mental disorders
	E-V	External causes of injury an supplemental classification
	280-289	Diseases of the blood and blood-forming organs
	320-359	Disease of the nervous system

The dataset contained three diagnosis fields (`diag_1`, `diag_2`, and `diag_3`), each representing different aspects of a patient's condition using ICD-9 codes. To improve interpretability and reduce dimensionality, the primary diagnosis was mapped into nine broader disease categories: Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Other. This transformation allowed for a more structured analysis of disease types affecting readmission likelihood.



Additional Features

To enhance risk signaling for hospital readmission, we engineered additional features by leveraging domain knowledge. These features integrate **medication management**, **hospital utilization**, **patient risk profiles**, and **admission characteristics**, allowing the model to capture nuanced patterns associated with readmission risk.

Medication Management Features

To enhance risk signaling, new medication-related features were engineered. A binary indicator (`on_multiple_meds`) was created to flag patients prescribed three or more medications. The total number of medication changes (`med_changes_total`) was calculated by summing instances where medications were either increased or decreased. Additionally, an `insulin_binary` variable was introduced to indicate whether a patient was using insulin.

Hospital Utilization Features

Hospital utilization patterns play a crucial role in predicting readmission risk. The `total_visits` feature was derived by summing the total number of outpatient, emergency, and inpatient visits. A `repeated_emergency` flag was introduced for patients with multiple emergency visits, as frequent emergency care utilization is often associated with a higher likelihood of readmission. The `all_usage` feature was created to identify patients who utilized all three care settings: outpatient, emergency, and inpatient.

Patient Risk Profile Features

$$risk\_score \;=\; Age\_R_{w} + (2 \times Impatien\_V_{w}) + (1.5 \times Freq\_Emgcy\_V_{w}) + (1.2 \times Med\_Ct_{w})$$

A composite risk score (`risk_score`) was developed to capture patient risk levels. The score incorporates multiple factors, including **age risk**, **inpatient visit history** (weighted x2), **emergency visit frequency** (weighted x1.5), and **medication count** (weighted x1.2). Additionally, `age_risk` was mapped based on age group categories, assigning higher risk scores to older patients, as they are generally more susceptible to complications and readmissions.

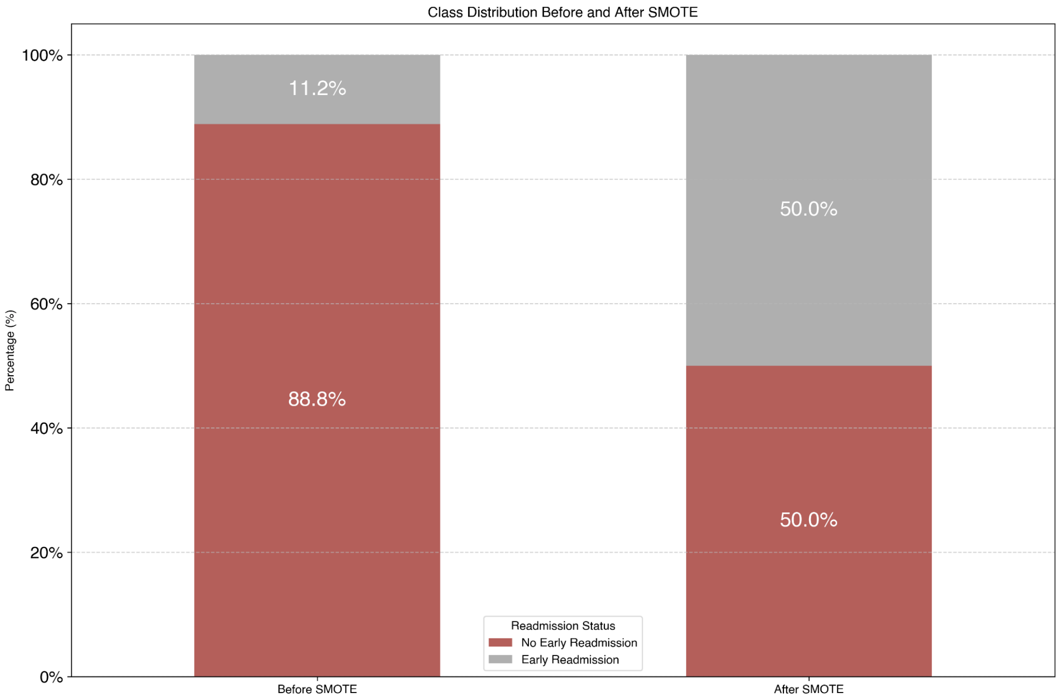
Admission Characteristics

To identify critical patterns in admissions, a binary indicator (`is_emergency`) was introduced to flag admissions classified as urgent or emergency-related. This feature helps distinguish between planned admissions and those requiring immediate medical attention.

Class Imbalance in Readmission Data

The dataset exhibited significant class imbalance, with only 11.5% of patients readmitted within 30 days. **Two different approaches** were implemented to handle this issue.

1. The first approach involved using SMOTENC (Synthetic Minority Over-sampling Technique for Categorical Data) to generate synthetic samples for the minority class, thereby transforming the class distribution to a balanced 50-50 ratio. This technique ensures that the predictive model is trained on a more balanced dataset, reducing bias toward the majority class.



2. The second approach maintained the natural class imbalance but applied class weighting in the model training phase. By adjusting class weights, the model penalizes misclassifications of the minority class more heavily, ensuring that it learns to recognize patterns associated with readmission even in the presence of an imbalanced dataset. These dual approaches provided flexibility in model training, allowing for comparisons to determine the most effective strategy for improving predictive performance.

Model Selection

To assess the effectiveness of different machine learning models in predicting hospital readmissions, we tested a diverse set of algorithms while addressing class imbalance using SMOTENC-based synthetic oversampling and class-weighted learning. The initial models selected for evaluation include:

- Logistic Regression
    - A simple and interpretable baseline model that serves as a foundational comparison for more complex methods. It is particularly useful for understanding feature importance and evaluating the impact of class imbalance when using techniques like class weighting.
- Random Forest
    - A bagging-based ensemble model that is highly effective in handling class imbalance. By aggregating multiple decision trees, it reduces variance and improves generalization, making it robust against overfitting. Additionally, its ability to handle both categorical and numerical variables makes it suitable for hospital readmission prediction.
- XGBoost
    - A gradient boosting model that iteratively corrects errors in predictions by focusing more on difficult-to-classify cases. This makes it particularly effective for class imbalance, as it adjusts the learning process to improve the recall of the minority class. Its ability to capture complex relationships between features enhances predictive performance.
- LightGBM
    - A boosting algorithm optimized for speed and efficiency, making it well-suited for large datasets. It employs histogram-based learning, which reduces memory consumption and accelerates training while maintaining high accuracy. Additionally, it incorporates built-in handling for imbalance through custom loss functions.
- CatBoost
    - A boosting method specifically designed for categorical data, reducing the need for extensive preprocessing. It handles categorical features efficiently and reduces the risk of overfitting. Its ability to perform well with minimal hyperparameter tuning makes it an excellent choice for datasets with a high proportion of categorical variables, such as hospital admission data.

Each of these models brings unique strengths to the predictive task, and their performance will be systematically evaluated to determine the best approach for predicting early hospital readmissions.

Model	**Purpose**	**Strengths**
Logistic Regression	Baseline model	Simple, interpretable
Random Forest	Uses bagging to reduce variance	Effective for class imbalance, improves generalization

XGBoost	Corrects errors iteratively	Captures complex relationships, improves minority class performance
LightGBM	Gradient boosting optimized for speed	Handles large datasets efficiently, reduces overfitting
CatBoost	Boosting method optimized for categorical features	Works well with categorical data, less hyperparameter tuning needed

Assessing Model Performance Before Hyperparameter Tuning

Bootstrapping vs. Boosting

Bootstrapping and boosting are two distinct ensemble learning techniques that approach predictive modeling differently. Bootstrapping-based methods, such as Random Forest that we have selected as one of our models, work by generating multiple subsets of the training data through random sampling with replacement. Each decision tree is trained independently, and the final prediction is obtained by aggregating the results, typically through majority voting (for classification). This technique helps reduce variance and prevent overfitting, making it particularly effective when dealing with noisy or imbalanced data. However, because each tree is trained independently, it may not always prioritize difficult-to-classify cases, which can be crucial for our imbalanced readmission dataset.

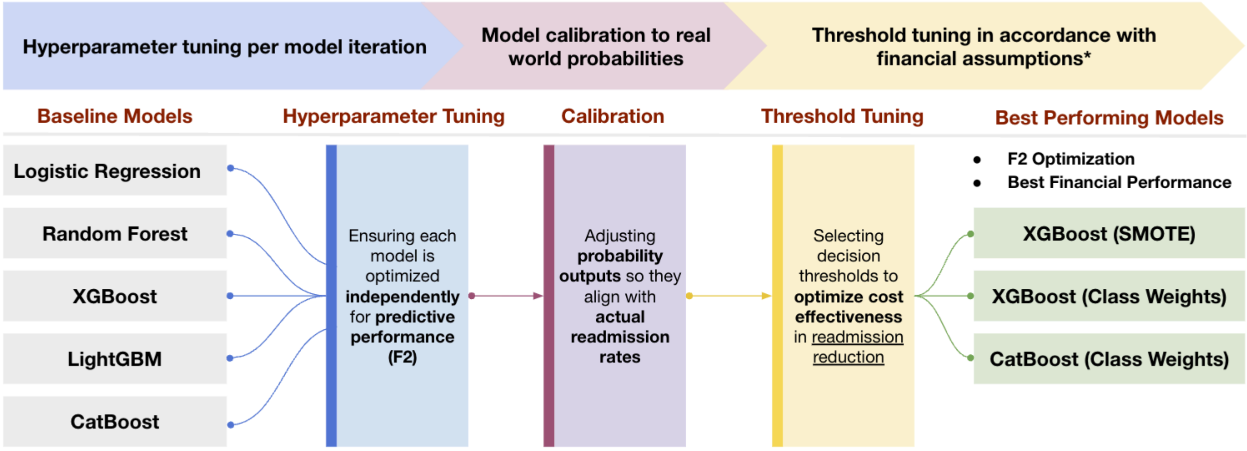
On the other hand, boosting methods, such as XGBoost, LightGBM, and CatBoost, build models sequentially, where each new tree focuses on correcting the errors of the previous ones. This iterative process gives more weight to misclassified samples, making boosting methods particularly effective at improving the recall of minority classes. However, boosting is also more prone to overfitting, especially on smaller datasets, since it aggressively learns from patterns, including noise.

Given that our dataset is relatively small (95,672 records) and imbalanced (only 11.5% readmitted cases), boosting methods may provide higher predictive accuracy, particularly in capturing the minority class patterns. However, they will require careful tuning to avoid overfitting. Bootstrapping methods like Random Forest remain a strong option due to their robustness and ability to generalize well without extensive tuning.

In our case, boosting may be more effective in identifying early readmissions due to its ability to focus on harder-to-classify cases. However, we should complement this approach with regularization techniques (such as early stopping in boosting models) and compare results with Random Forest to determine whether the added complexity truly improves predictive performance.

Hyperparameter Tuning

Framework for Developing Models for Cost-Optimal Readmission Prevention


Hyperparameter tuning per model iteration
Model calibration to real world probabilities
Threshold tuning in accordance with financial assumptions*
Baseline Models
Hyperparameter Tuning
Calibration
Threshold Tuning
Best Performing Models
Logistic Regression
Random Forest
XGBoost
LightGBM
CatBoost
Ensuring each model is optimized independently for predictive performance (F2)
Adjusting probability outputs so they align with actual readmission rates
Selecting decision thresholds to optimize cost effectiveness in readmission reduction
F2 Optimization
Best Financial Performance
XGBoost (SMOTE)
XGBoost (Class Weights)
CatBoost (Class Weights)

We have developed a Framework for Cost-Optimal Readmission Prevention to systematically optimize our predictive models while incorporating real-world readmission patterns and financial considerations. This framework ensures that our modeling approach not only enhances predictive performance but also aligns with cost-effectiveness objectives in reducing hospital readmissions.

The framework consists of four key stages. Baseline model selection begins with evaluating multiple machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost, and train them with data that were upsampled through SMOTENC or with class weight considerations, to determine the most effective techniques for handling the complexities of hospital readmission prediction. Following model selection, we implement hyperparameter tuning, where each model is optimized independently to maximize its F2 score, prioritizing recall to better identify high-risk patients.

Once models are tuned for predictive accuracy, we perform probability calibration, ensuring that model outputs align with actual readmission rates. This step enhances model interpretability and facilitates better decision-making in clinical applications. The final stage involves threshold tuning, where decision thresholds are strategically adjusted to optimize cost-effectiveness in readmission reduction. By calibrating these thresholds, we ensure that our predictive models not only improve classification performance but also drive actionable insights that support hospital resource allocation and financial planning.

Through this framework, we have identified XGBoost with SMOTE, XGBoost with Class Weights, and CatBoost with Class Weights as the most effective models for readmission prediction, balancing recall, precision, and cost considerations. Our approach emphasizes iterative refinement, allowing for adjustments in response to evolving financial assumptions and healthcare priorities. By leveraging this structured methodology, we enhance the clinical and economic impact of predictive modeling in hospital readmission prevention.

Stage 1: Hyperparameter Tuning

In our modeling framework, hyperparameter tuning plays a significant role in optimizing model performance, particularly in the context of imbalanced hospital readmission data. Given the rarity of early readmissions in our dataset, selecting the right hyperparameters ensures that our models effectively capture high-risk patients while minimizing false negatives.

Hyperparameter Tuning Strategy

Hyperparameter tuning focuses on three key parameters that significantly impact model performance:

1. **Learning Rate**: This parameter controls how quickly the model adapts to errors. A lower learning rate ensures stable convergence but requires more training iterations, while a higher learning rate speeds up training but increases the risk of overshooting optimal solutions.
2. **Tree Depth**: This parameter controls the complexity of individual trees in tree-based models. Deeper trees capture more complex patterns but can lead to overfitting, especially on small datasets. We optimize tree depth to strike a balance between capturing important relationships and maintaining generalization.
3. **Regularization Parameters**: Regularization helps prevent overfitting by penalizing overly complex models. Techniques such as L1 (Lasso) and L2 (Ridge) regularization are applied in models like Logistic Regression, while tree-based models use L2 leaf regularization and min_child_weight to control overfitting.

By systematically adjusting these hyperparameters using grid search and randomized search techniques, we ensure that each model is tuned for maximum predictive effectiveness.

Why We Chose F2 Score as Our Evaluation Metric

In highly imbalanced datasets, traditional evaluation metrics such as accuracy can be misleading. Given that approximately 88% of patients in our dataset are not readmitted, a naive model could achieve high accuracy simply by predicting "No Readmission" for all cases. However, this approach would fail to identify high-risk patients who require intervention.

The F1 score, a commonly used metric for classification, balances precision and recall equally. However, in the context of hospital readmissions, Type II errors (false negatives) are far more costly than Type I errors (false positives). A false negative means a high-risk patient was incorrectly classified as low risk, leading to missed interventions and potential complications. To address this, we selected the F2 score, which places greater emphasis on recall than precision. The formula for F2 is:

$$F2 = 5 \times Precision \times Recall\ /\ 4 \times Precision + Recall$$

This metric ensures that our models prioritize correctly identifying high-risk patients, even at the expense of some false positives. While false positives may lead to additional patient monitoring, they are a less severe consequence than failing to flag a patient at risk of readmission. By focusing on hyperparameter tuning and selecting F2 as our primary evaluation metric, our framework ensures that predictive models are both effective in identifying at-risk patients and aligned with real-world healthcare priorities.

Stage 2: Calibration

Model calibration is a crucial step in our predictive framework, ensuring that the probabilities generated by our models align with actual readmission rates. Without proper calibration, models may overestimate or underestimate risk, leading to misinformed clinical decisions. In hospital readmission settings, where the cost of false negatives is particularly high, well-calibrated probability outputs improve decision-making by providing realistic estimates of patient risk.

Why Calibration is Necessary

Probabilistic models, particularly boosting methods like XGBoost and CatBoost, often produce overconfident predictions. This means that if a model predicts a 70% chance of readmission, the actual likelihood may be significantly lower or higher. Such miscalibration can lead to incorrect risk assessments, causing hospitals to over-allocate or under-allocate resources for patient care. In clinical settings, calibrated probability estimates are essential for aligning predictions with real-world outcomes, improving resource planning, and ensuring that high-risk patients receive the necessary interventions.

Calibration Techniques Used

For model calibration, we applied Isotonic Regression, a non-parametric technique that learns a monotonic function to map raw model outputs to calibrated probabilities. This technique is particularly effective for models that exhibit non-linear miscalibration, such as gradient boosting algorithms.

- Isotonic Regression adjusts the predicted probabilities without assuming a fixed functional form, making it highly flexible.
- It ensures that probability estimates remain monotonically increasing, meaning that if one patient has a higher predicted risk than another, this relationship is preserved after calibration.
- In our framework, we utilized CalibratedClassifierCV for most models but required a custom implementation for CatBoost, given its unique handling of categorical data.

Special Considerations for Calibrating CatBoost

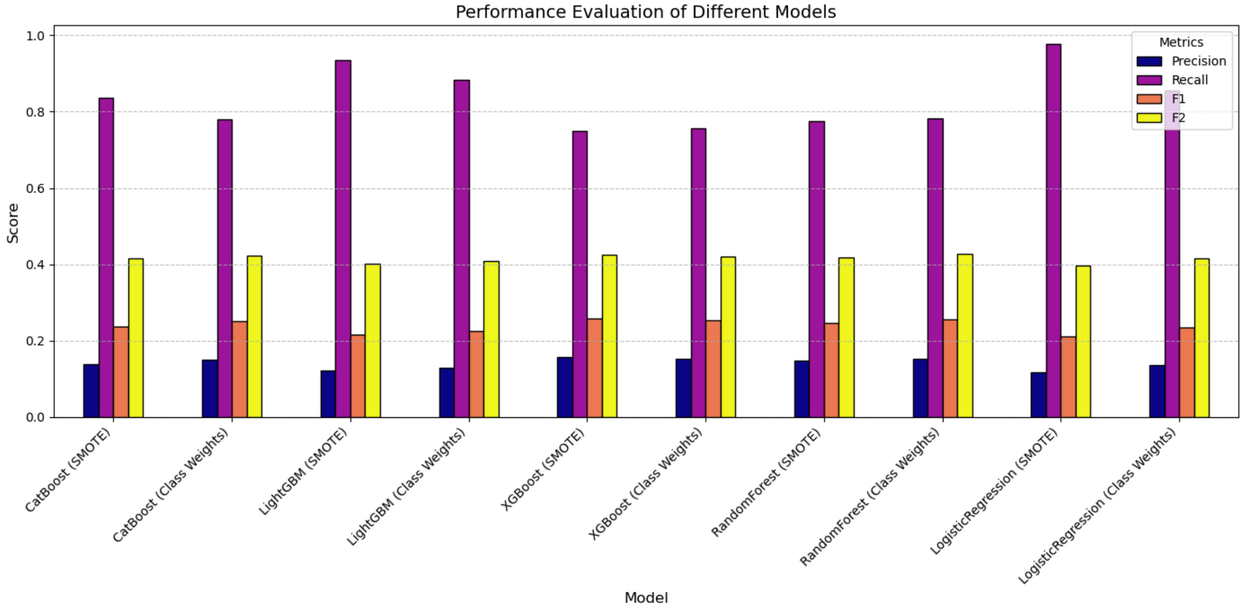
CatBoost differs from other boosting models in how it processes categorical features. Instead of one-hot encoding or label encoding, CatBoost internally converts categorical variables using ordered statistics, reducing overfitting and improving generalization. However, this internal encoding creates challenges when applying external calibration techniques like Isotonic Regression. Unlike XGBoost or LightGBM, which output raw probability scores that can be directly calibrated, CatBoost's categorical handling introduces additional transformations that require custom calibration adjustments.

To account for this, we implemented a custom CatBoost calibration pipeline:

1. Train CatBoost on SMOTE-balanced data while retaining its built-in categorical feature handling.
2. Extract raw probability scores from an independent calibration set to avoid overfitting.
3. Apply Isotonic Regression to adjust these probabilities, ensuring a well-calibrated mapping to real-world readmission likelihood.
4. Wrap the calibrated model into a new CalibratedCatBoost class that transforms probability outputs before making predictions.

This approach ensures that CatBoost's unique categorical encoding remains intact, while its probability outputs are corrected for overconfidence. By applying this structured calibration method, we enhance the reliability of our risk predictions, reducing misclassification errors and improving patient triage decisions.

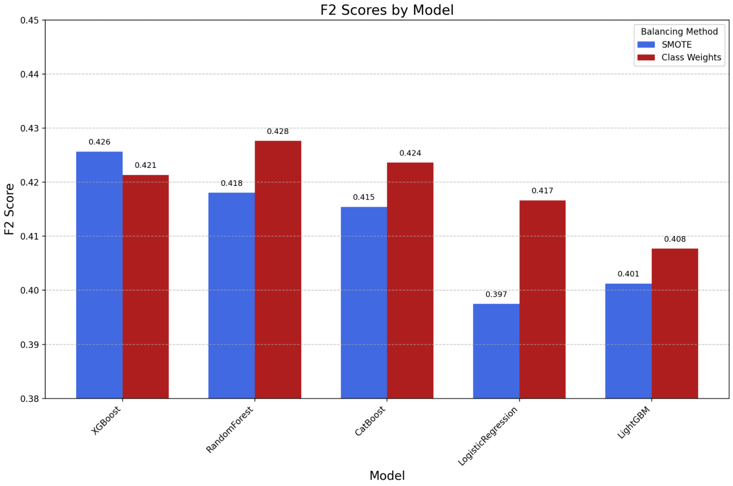
Modeling Performance Evaluation



Model performance was evaluated based on the F2 score, which prioritizes recall over precision to better identify high-risk patients for hospital readmission. Given the class imbalance in the dataset, we tested two different strategies to address the imbalance: SMOTENC (Synthetic Minority Over-sampling Technique for Categorical Data) and Class Weights. The goal was to determine which approach optimally balances predictive accuracy with clinical applicability.

Key Findings from Model Evaluation

1. Class Weights Outperform SMOTE on Average



   a. Class weighting yielded better F2 scores across 4 out of 5 models, including Random Forest, CatBoost, Logistic Regression, and LightGBM.
   b. This suggests that re-weighting the loss function during training is more effective than synthetic oversampling in ensuring models correctly classify minority (readmitted) cases.
2. Random Forest Achieves the Highest F2 Scores
   a. Among all models, Random Forest with class weighting delivered the best F2 score, demonstrating its robustness in handling class imbalance.
   b. However, SMOTE negatively impacted Random Forest performance, likely due to overfitting on synthetic samples rather than learning generalized patterns.
3. XGBoost and CatBoost Performed Well in Both Settings
   a. These gradient boosting models were relatively resilient to the choice of balancing method, though class weighting still led to slightly higher F2 scores.
   b. This suggests that boosting algorithms already incorporate mechanisms to address imbalance, making external resampling methods like SMOTE less impactful.

Aspect	**Class Weights**	**SMOTE**
Approach	Adjusts loss function to penalize misclassification of minority class	Synthesizes new minority class samples to balance distribution
Effect on Model Training	Retains original data distribution, better generalization	Increases dataset size but may introduce noise
Impact on Performance	Consistently improves F2 score, better recall	Can lead to overfitting, especially in tree-based models
Best Models with Approach	Random Forest, CatBoost, XGBoost	XGBoost, LightGBM (some improvement)

Clinical Interpretation

The optimal probability threshold for classification was identified between 0.08 and 0.10, ensuring that the models effectively capture at-risk patients while minimizing excessive false positives. Given that early readmission detection is crucial for intervention, class weighting provides a more stable and generalizable solution, as SMOTE may cause models to overfit on synthetic patterns rather than actual clinical variations.

Ultimately, class weighting emerged as the preferred approach, particularly for models like Random Forest, CatBoost, and Logistic Regression, as it enhanced predictive reliability without distorting the dataset. However, boosting models like XGBoost still performed well under both balancing strategies, reinforcing their robustness in imbalanced classification tasks.

Stage 3: Threshold Tuning

Threshold tuning is also a critical component of our modeling framework, ensuring that predictive models not only optimize recall but also balance the financial impact of interventions. While our initial evaluation prioritized maximizing the F2 score to correctly identify high-risk patients, threshold tuning refines these predictions by adjusting the probability cutoff at which a patient is classified as high risk. This step ensures that interventions are not only cost-effective but also tailored to improve patient welfare, encompassing a broader view of care beyond hospitalization.

Predictive models generate probability scores indicating the likelihood of a patient being readmitted. However, without proper threshold tuning, models may either misclassify patients at risk of readmission or flag too many low-risk patients for intervention, leading to increased healthcare costs. Lowering the threshold increases the sensitivity of the model, capturing more true positive cases but also increasing the number of false positives, which results in unnecessary patient interventions. Conversely, raising the threshold prioritizes precision, reducing false positives but increasing the likelihood of missing high-risk patients who require intervention. To address this trade-off, we tested multiple probability thresholds ranging from 0.08 to 0.10 across different models to determine which settings maximize financial savings while maintaining clinical efficacy.

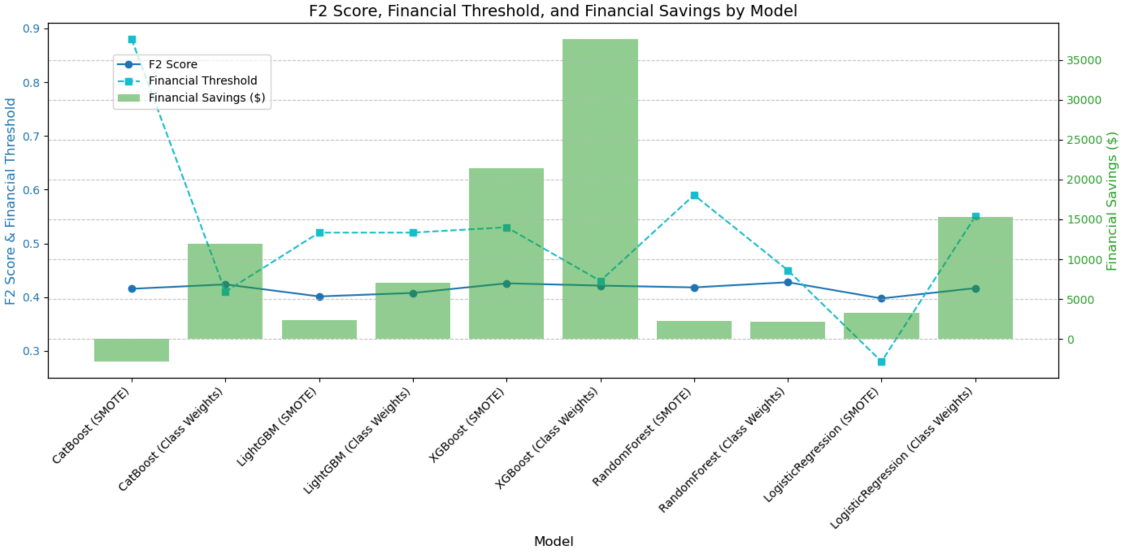
In our cost-benefit analysis, we considered an intervention cost of $2,500 per patient, representing a maximized cost estimate that accounts for a wide range of potential post-discharge care strategies. While this cost could reflect extended hospital stays, it is not limited to in-hospital treatments. It also includes hospice follow-up care, remote patient monitoring, additional nurse check-ins, enhanced outpatient services, and coordinated aftercare plans aimed at preventing readmission. These interventions are designed to proactively support high-risk patients post-discharge, ensuring they receive the necessary medical attention outside of the hospital setting. By investing in preventative care, hospitals can mitigate avoidable complications, reduce the burden of emergency readmissions, and enhance overall patient outcomes.

The financial impact of threshold tuning was assessed by incorporating this holistic intervention approach into our savings calculation. With an average readmission cost of $16,300 per patient, early intervention has the potential to generate substantial cost savings by preventing costly hospital returns. The effectiveness of interventions was modeled at 30%, 40%, and 50% reductions in readmission risk, allowing us to evaluate the trade-offs between precision, recall, and financial viability. Our formula for healthcare savings,

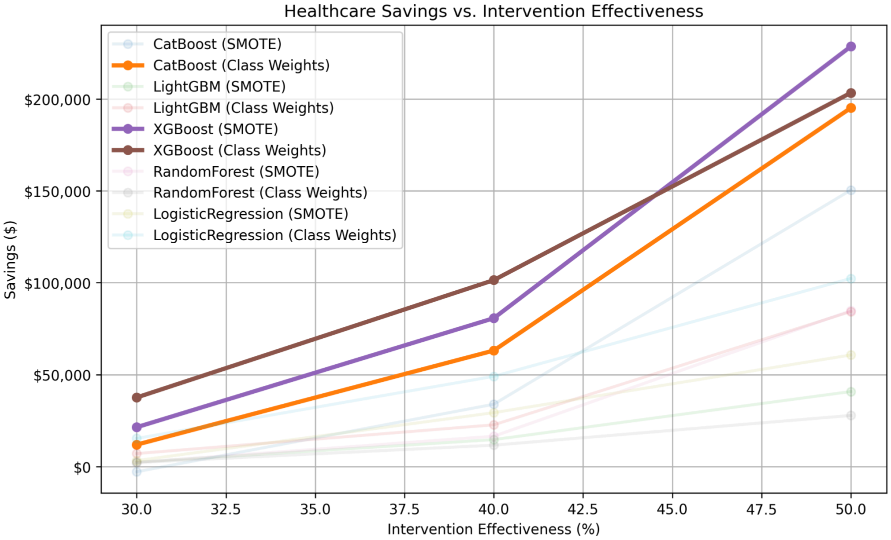
$$Savings = (TP \times IE \times 16,300) - ((TP + FP) \times 2,500)$$

captures both direct cost reductions and the broader benefits of improved patient care.

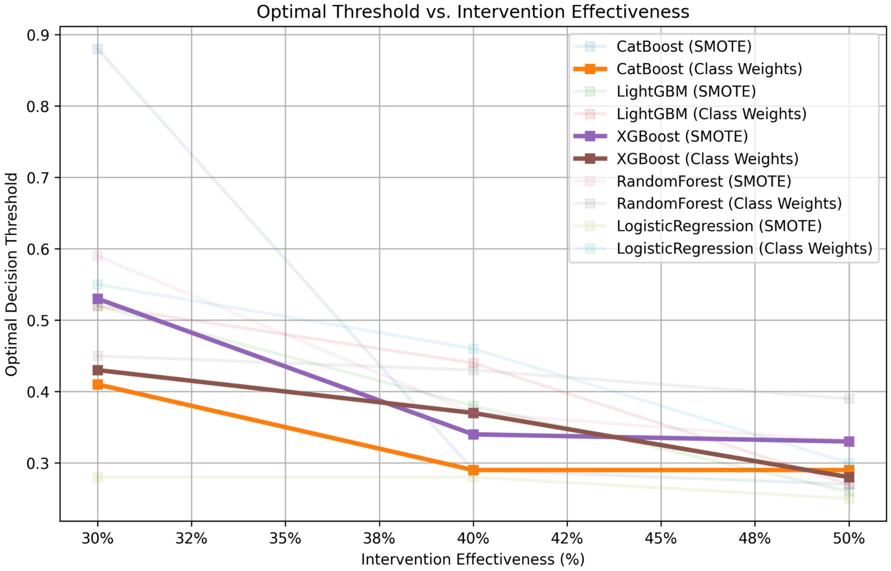
We assessed how different models and probability thresholds balance savings with effective intervention strategies.



The results of our threshold tuning analysis revealed key insights into model performance. XGBoost with Class Weights emerged as the most consistent performer, delivering stable financial savings across all intervention levels. This model effectively balanced early intervention with cost-effectiveness, making it the preferred choice for clinical deployment. Meanwhile, XGBoost with SMOTE achieved the highest savings at peak intervention effectiveness, reaching $200,000 in savings at 50% intervention effectiveness. However, class-weighted models proved to be more reliable in optimizing financial outcomes, as they prioritized recall while maintaining a more generalized understanding of patient risk compared to SMOTE-based oversampling techniques.

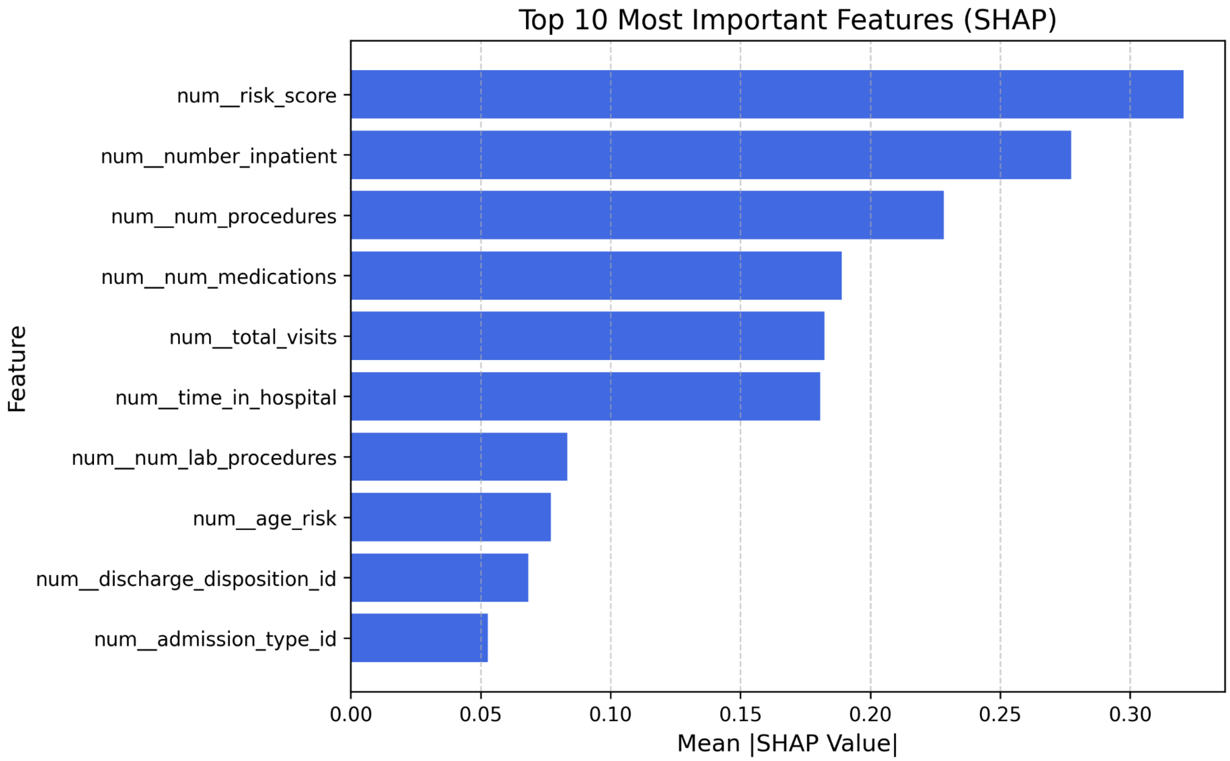


Our approach to threshold tuning also accounted for the evolving nature of intervention strategies. At 30% intervention effectiveness, models maintained higher decision thresholds (~70%), ensuring that interventions were focused on the most severe cases to prevent unnecessary costs. As intervention effectiveness increased to 50%, thresholds were lowered, allowing a broader group of high-risk patients to receive proactive care while still maintaining cost efficiency. This adaptive approach reflects a strategic shift in balancing financial constraints with the goal of improving long-term patient well-being.



Ultimately, XGBoost with Class Weights was identified as the optimal model, effectively adjusting decision thresholds in response to increasing intervention effectiveness while maintaining financial efficiency. Our threshold tuning framework goes beyond cost savings—it integrates a patient-centric approach that prioritizes improved healthcare outcomes. By considering not just hospitalization costs but also outpatient support, hospice care, and continued medical monitoring, our model ensures that interventions align with both financial sustainability and enhanced patient welfare. The ability to adjust decision thresholds dynamically based on intervention effectiveness allows for scalable, cost-effective, and clinically meaningful improvements in hospital readmission prevention.

Feature Importance Analysis



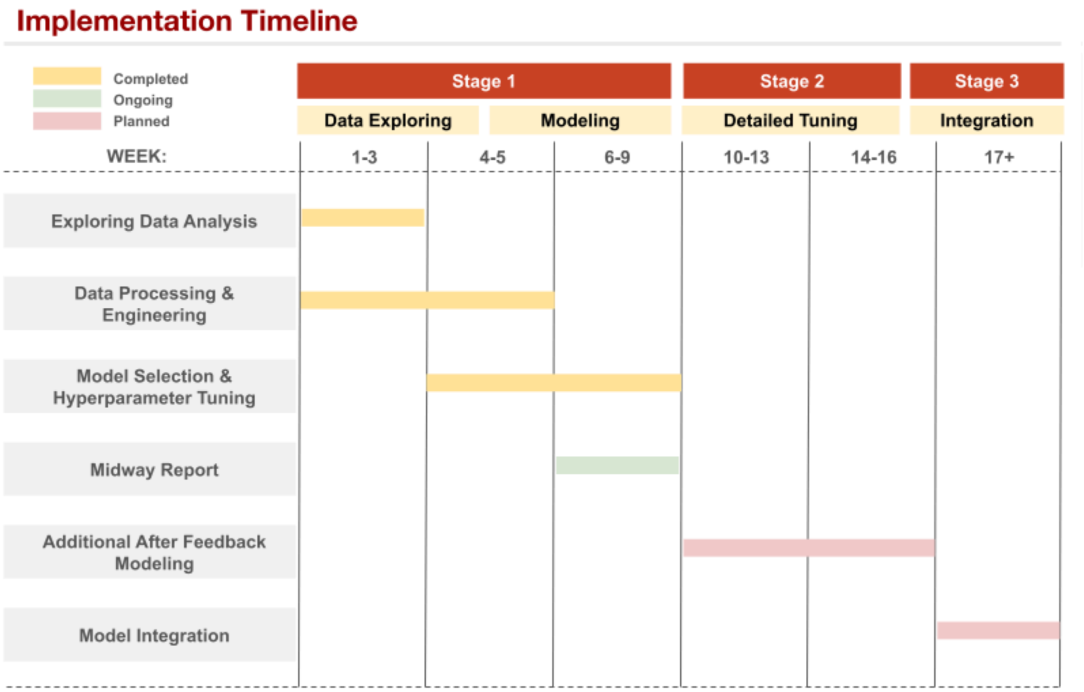
To gain deeper insights into the key drivers of hospital readmission risk, we employed SHAP (Shapley Additive Explanations) to evaluate feature importance within our predictive model. SHAP values measure the marginal contribution of each feature to the model's predictions, helping us understand how individual variables influence the likelihood of readmission. Higher SHAP values indicate stronger predictive power, meaning that these features have a greater impact on model decisions.

The analysis revealed that the risk score was the most influential feature in predicting readmissions. Given that this score is a composite measure incorporating factors such as age, inpatient visit history, emergency visit frequency, and medication count, its prominence highlights the importance of holistic patient risk assessment. Following the risk score, the number of inpatient visits emerged as the second most critical predictor, emphasizing that prior hospitalizations play a significant role in determining readmission likelihood. Patients with a higher frequency of inpatient stays are more likely to require continued medical attention, increasing their risk of being readmitted shortly after discharge.

Additional key predictors included the number of procedures and number of medications prescribed during hospitalization. These variables suggest that patients undergoing multiple procedures or receiving extensive pharmacological treatment may have more complex medical conditions, which inherently elevate their readmission risk. Similarly, the total number of visits, including outpatient, emergency, and inpatient encounters, was identified as a crucial factor, reinforcing the idea that frequent healthcare interactions signal underlying health instability.

Other notable features contributing to the model's predictions included time spent in the hospital, lab procedures, and discharge disposition. The length of hospitalization often correlates with disease severity, while the type and frequency of lab tests provide insights into a patient's overall clinical condition. Meanwhile, discharge disposition, which reflects where a patient is sent post-discharge (e.g., home, rehabilitation, or hospice), plays an essential role in determining the likelihood of a successful recovery versus the need for rehospitalization.

Implementation Plan



The implementation plan for integrating predictive modeling into hospital readmission prevention follows a structured timeline that ensures seamless model deployment while aligning with clinical workflows. The plan is divided into three key stages: data exploration, model development, and detailed tuning, culminating in full-scale integration.

During the data exploration phase, we focused on analyzing the dataset, identifying key predictive features, and addressing class imbalance through techniques such as SMOTENC and class weighting. The model development phase involved selecting the best-performing algorithms, conducting hyperparameter tuning, and optimizing models based on the F2 score to maximize recall while ensuring precision. This phase also included performance evaluations to balance predictive accuracy with financial and clinical outcomes. The final phase, detailed tuning and integration, ensures that models are properly calibrated, decision thresholds are optimized for cost-effectiveness, and region-specific financial constraints are taken into account. This stage also includes identifying potential disparities in model performance across demographic and patient groups, ensuring that predictive insights are equitable across diverse populations.

Following model refinement, the implementation plan transitions into system integration and deployment. The first step involves automating data ingestion from hospital systems such as Epic and Cerner, enabling real-time updates to enhance prediction accuracy. A real-time prediction API will then be deployed, allowing healthcare professionals to assess readmission risk at the point of discharge, with models providing explainable outputs for decision support. To ensure that predictive insights translate into actionable interventions, clinical workflow integration will embed these predictions into hospital discharge planning systems, allowing healthcare providers to make informed decisions about patient follow-up care. The final step in implementation focuses on continuous monitoring and improvement, where model drift and performance metrics will be evaluated periodically, ensuring that predictions remain accurate and aligned with evolving healthcare policies and patient needs.

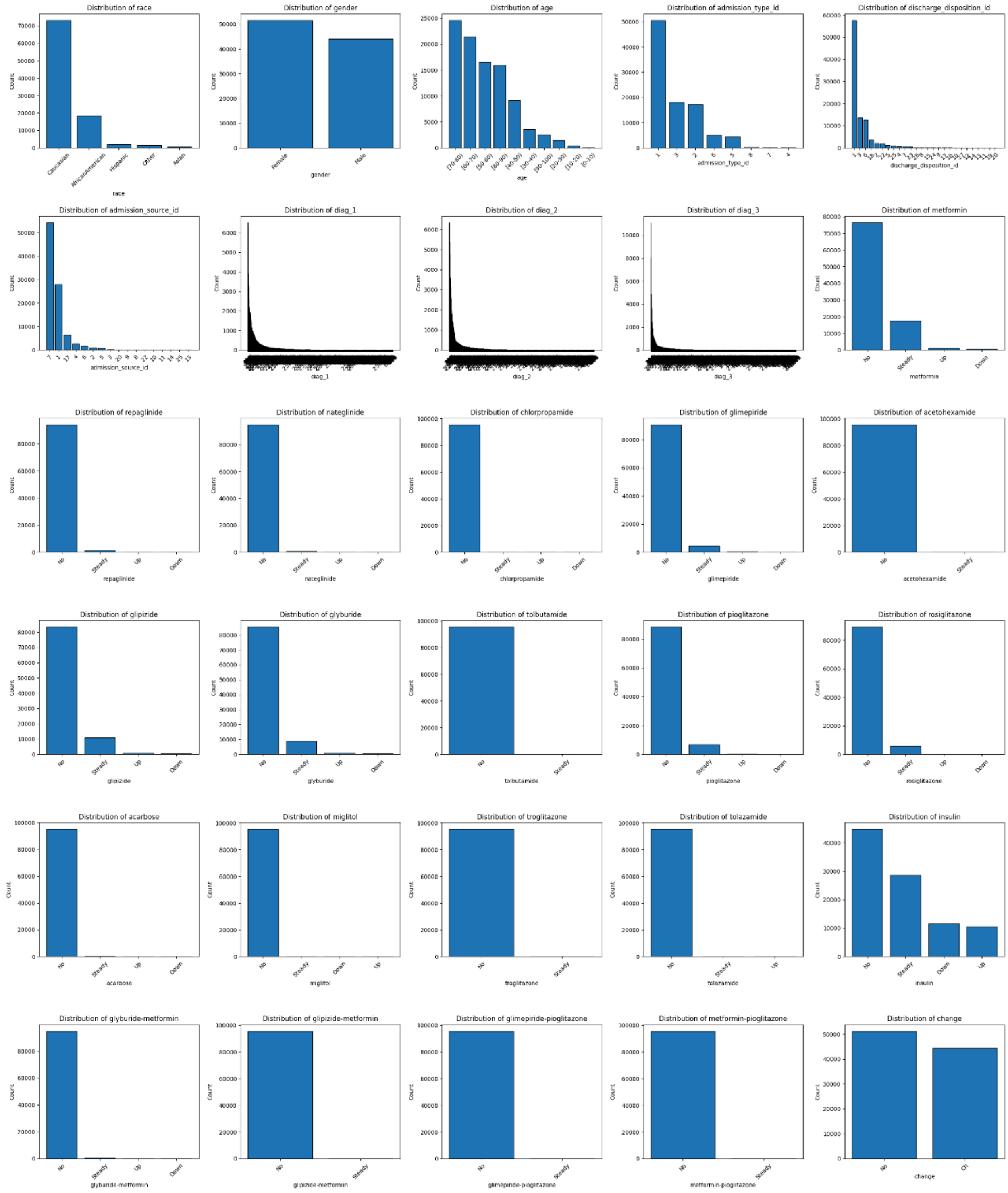
Beyond the implementation plan, we have identified key modeling considerations that influence the robustness and interpretability of predictions. One of the primary challenges is the black-box nature of complex machine learning models, which can obscure explainability in high-dimensional datasets. To address this, we recommend using advanced visualization techniques, such as parallel coordinates and Andrews curves, to provide interpretable insights. Another consideration is the inconsistency in feature importance across different algorithms, which may introduce ambiguity in decision-making. To mitigate this, complementary rule-based models can be developed alongside black-box models to improve transparency.

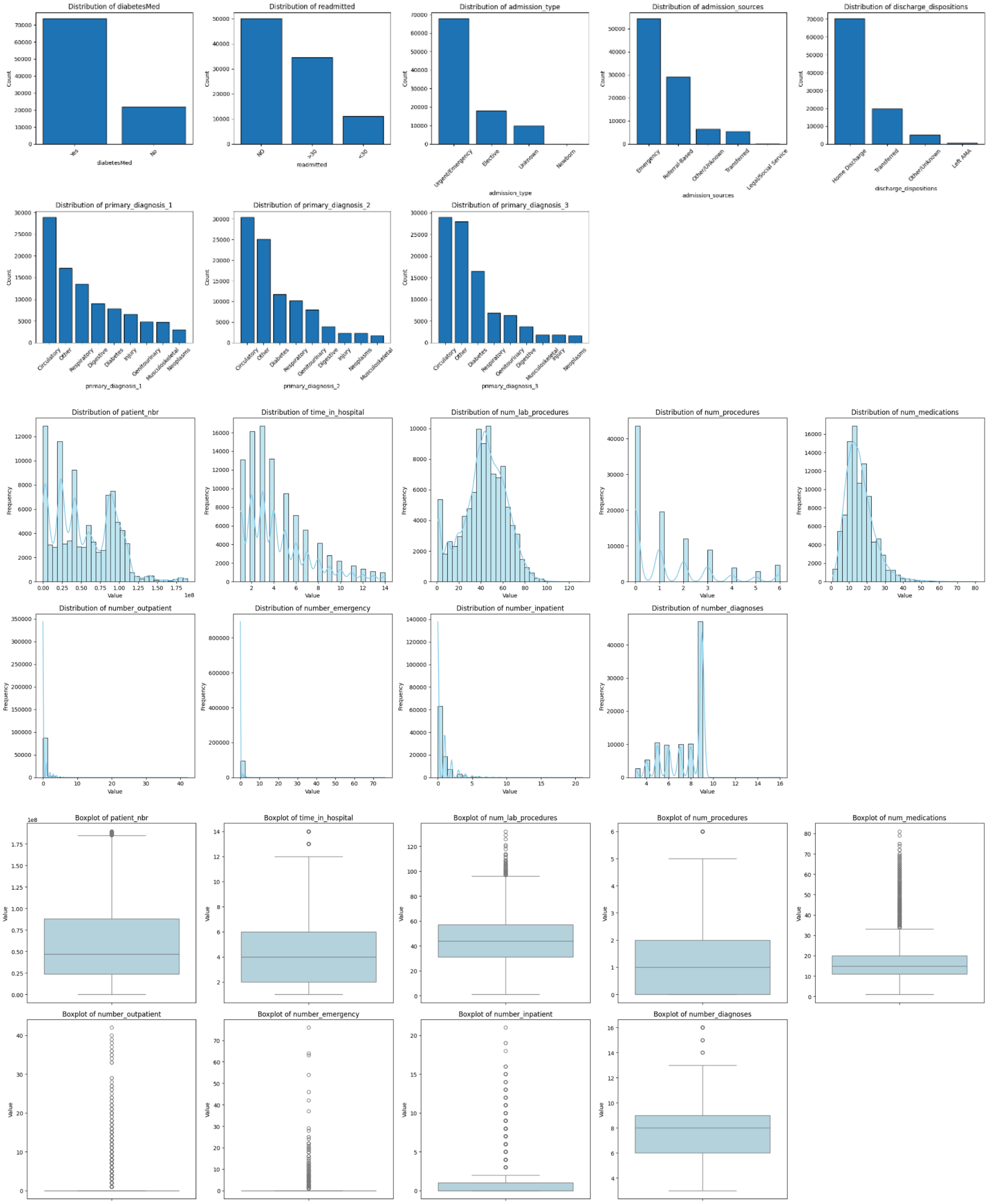
Moreover, while global class weighting strategies effectively address class imbalance, they may overlook intra-group disparities across age, gender, and other demographic factors. To refine this approach, we propose feature-aware class weighting, which prioritizes key categorical variables when adjusting model sensitivity. Additionally, handling missing data presents a challenge, as naïve imputation techniques may fail to capture important domain-specific nuances. A more robust strategy involves using Bayesian imputation with domain-inspired priors, which integrates expert knowledge into the imputation process. Finally, raw probability outputs can be misinterpreted due to over- or under-confidence issues, leading to inaccurate risk assessments. To counter this, we incorporate probability recalibration techniques, such as isotonic regression, and leverage SHAP (Shapley Additive Explanations) to enhance model interpretability.

By implementing a structured integration plan and addressing these key modeling considerations, we ensure that predictive models for hospital readmission prevention are both clinically effective and operationally sustainable. This holistic approach enhances patient care, optimizes resource allocation, and supports data-driven decision-making in hospital settings.

Appendix

Features Visualization




Distribution of diabetesMed
Count
diabetesMed
Yes
No
Distribution of readmitted
readmitted
NO
>30
<30
Distribution of admission_type
admission_type
Urgent/Emergency
Elective
Unknown
Newborn
Distribution of admission_sources
admission_sources
Emergency
Referral-Based
Other/Unknown
Transferred
Legal/Social Service
Distribution of discharge_dispositions
discharge_dispositions
Home Discharge
Transferred
Other/Unknown
Left AMA
Distribution of primary_diagnosis_1
primary_diagnosis_1
Circulatory
Other
Respiratory
Digestive
Diabetes
Injury
Genitourinary
Musculoskeletal
Neoplasms
Distribution of primary_diagnosis_2
primary_diagnosis_2
Circulatory
Other
Diabetes
Respiratory
Genitourinary
Digestive
Injury
Neoplasms
Musculoskeletal
Distribution of primary_diagnosis_3
primary_diagnosis_3
Circulatory
Other
Diabetes
Respiratory
Genitourinary
Digestive
Musculoskeletal
Injury
Neoplasms
Distribution of patient_nbr
Frequency
Value
1e8
Distribution of time_in_hospital
Distribution of num_lab_procedures
Distribution of num_procedures
Distribution of num_medications
Distribution of number_outpatient
Distribution of number_emergency
Distribution of number_inpatient
Distribution of number_diagnoses
Boxplot of patient_nbr
1e8
Value
Boxplot of time_in_hospital
Boxplot of num_lab_procedures
Boxplot of num_procedures
Boxplot of num_medications
Boxplot of number_outpatient
Boxplot of number_emergency
Boxplot of number_inpatient
Boxplot of number_diagnoses