Final Project Report: Predicting Hospital Admissions

Mahi Sheth, Ashwin Ajit, Rebecca Han, Anupam Pradeep

01:960:486:01 Applied Statistical Learning

Dr. Michael LuValle

December 19, 2025

Abstract

Hospital readmissions within 30 days of discharge represent a significant challenge for healthcare systems, contributing to increased costs, hospital penalties, and poorer patient outcomes. This report investigates predictors of hospital readmission among diabetic patients using ten years of clinical data from U.S. hospitals. Through extensive exploratory data analysis, feature engineering, and the application of both linear and nonlinear predictive models, we assess the extent to which patient demographics and healthcare utilization intensity explain readmission risk. Across all models, predictive performance is modest (AUC ≈ 0.61–0.64), suggesting that there are structural limits to predictability from administrative clinical data alone. Our findings highlight utilization intensity as the dominant signal, motivating future work that incorporates social, behavioral, and longitudinal patient data.

1. Introduction

Hospital readmissions, commonly defined as unplanned inpatient admissions occurring within 30 days of discharge, represent a persistent and costly challenge within the United States healthcare system. Readmission rates are frequently used as indicators of hospital quality, continuity of care, and the effectiveness of care coordination. High readmission rates not only signal potential gaps in post-discharge care but also impose substantial financial burdens on both hospitals and patients. In response, policymakers have implemented programs such as the Hospital Readmissions Reduction Program (HRRP), which financially penalizes hospitals with excess readmissions for certain conditions. These policies have intensified the need for reliable methods to identify patients at elevated risk of readmission.

Patients with chronic diseases are particularly vulnerable to readmission due to the complexity of their conditions and the ongoing management required after discharge. Diabetes, one of the most prevalent chronic illnesses in the United States, poses a unique challenge in this regard. Diabetic patients often experience multiple concurrent illnesses, require complex medication regimens, and undergo frequent laboratory testing and clinical monitoring. These factors increase the likelihood of complications, treatment non-adherence, and subsequent hospital utilization. As a result, understanding and predicting readmission risk among diabetic patients is of both clinical and economic importance.

Accurate prediction of hospital readmissions has the potential to significantly improve patient outcomes and healthcare efficiency. From a clinical perspective, the early identification of high-risk patients enables providers to implement targeted interventions, such as enhanced discharge planning, medication reconciliation, follow-up appointments, and outpatient support services. From a systems-level perspective, predictive models can help hospitals allocate limited resources more effectively, prioritize care coordination efforts, and reduce avoidable costs associated with readmissions. However, despite substantial research efforts, predicting readmissions remains a difficult task, with many studies reporting only modest predictive performance.

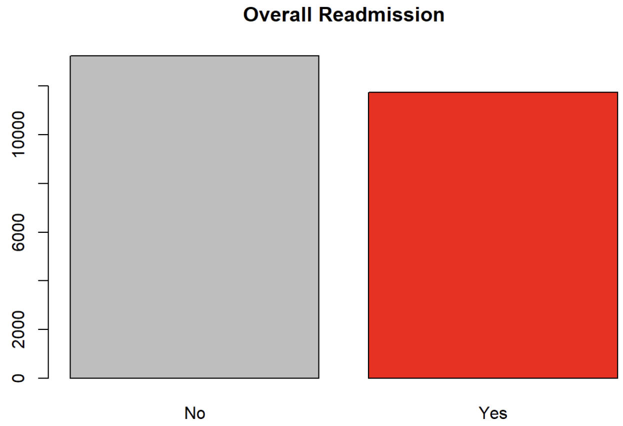
This report addresses the problem of hospital readmission prediction through a modeling framework applied to a large, real-world clinical dataset of diabetic patients. Using ten years of hospital encounter data from 130 U.S. hospitals, we perform extensive exploratory data analysis, feature engineering, and model comparison across a range of statistical and machine learning approaches. Particular emphasis is placed on healthcare utilization intensity, such as the frequency of hospital visits, length of stay, and volume of medications and procedures, as these factors are hypothesized to be strong drivers of readmission risk.

Rather than focusing solely on maximizing predictive accuracy, this study critically evaluates the practical limits of readmission prediction using routinely collected hospital data. By comparing linear, penalized, and nonlinear models within a consistent validation framework, we assess whether increasing model complexity yields meaningful performance gains. The results demonstrate that while all models perform better than random guessing, overall discrimination remains modest, suggesting that structural limitations in the available data constrain predictive power. These findings emphasize the importance of integrating statistical rigor with domain expertise and underscore the need for more comprehensive data sources, including social and behavioral determinants of health, in future readmission prediction efforts.

2. Dataset Description

The dataset used consists of hospital encounter records for patients diagnosed with diabetes, collected over a ten-year period from 1999 to 2008 across 130 hospitals and integrated delivery networks within the United States. Each observation represents a single inpatient hospital stay, capturing detailed information on patient demographics, diagnoses, laboratory testing, medications, and healthcare utilization during the admission. The primary objective of this dataset is to enable the study of short-term hospital readmissions, with a particular focus on identifying factors associated with a patient being readmitted shortly after discharge. After initial preprocessing and quality checks, the final analytic dataset contains approximately 25,000 observations, with no missing values in the selected variables.

The primary outcome of interest is hospital readmission, defined as whether a patient experienced a subsequent inpatient admission within 30 days of discharge from the index hospitalization. This outcome is encoded as a binary variable, where a value of 1 indicates a readmission and 0 indicates no readmission within the specified time window. In the dataset, readmissions occur at a moderately high rate, with approximately 47% of hospital stays followed by a readmission. While the classes are not perfectly balanced, neither class overwhelmingly dominates the other, making the dataset suitable for classification modeling without the need for extreme rebalancing techniques. However, this level of imbalance requires the use of evaluation metrics beyond simple accuracy, including sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC).



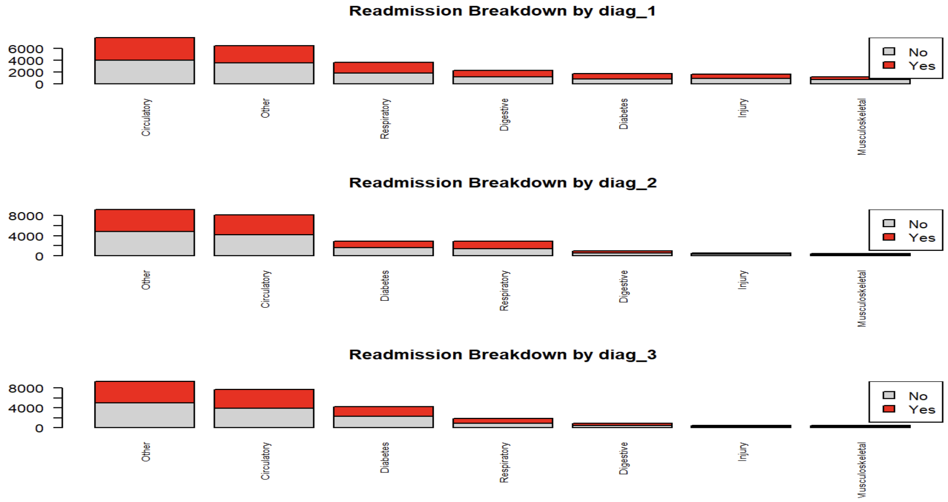
Patient age is recorded as a categorical variable representing ten-year age intervals (e.g., 40–50, 50–60, 60–70). For modeling purposes, age was treated as an ordered factor to preserve the natural ordering of age groups. Older age groups constitute a substantial portion of the dataset, reflecting the higher prevalence of diabetes and hospitalization among older adults. No direct identifiers such as patient names, exact dates of birth, or geographic identifiers are included in the dataset, ensuring patient anonymity. As a result, the analysis focuses on clinical and utilization-based predictors rather than socioeconomic or geographic factors.

3. Exploratory Data Analysis

An initial assessment of data quality revealed no missing values among the selected variables, and no duplicate rows were identified. The dataset includes three diagnostic variables corresponding to the primary diagnosis and two secondary diagnoses associated with each hospital admission. Diagnoses are grouped into broad clinical categories, such as circulatory, respiratory, and diabetes. This categorical structure reduces the dimensionality of diagnosis information while retaining clinically meaningful distinctions between major disease groups. To avoid sparsity and unstable parameter estimates in downstream models, entries labeled as "Missing" or "Unknown" were grouped into a single "Other" category. This preprocessing step balances interpretability and statistical stability, ensuring that diagnostic variables contribute meaningful information without inflating model variance.

Binary variables, such as readmission status, medication changes, and diabetes medication use, were encoded numerically to simplify modeling. Categorical variables were converted to factors, and ordered factors were used where natural ordering existed, such as in age categories. These features collectively support a comprehensive analysis of hospital readmissions and enable the exploration of both linear and nonlinear relationships between patient characteristics and readmission risk.

Primary, secondary, and tertiary diagnosis codes were grouped into broad clinical categories to reduce sparsity and improve interpretability. Bar plots comparing readmission rates across diagnosis categories reveal modest variation, with circulatory and respiratory conditions exhibiting slightly higher readmission proportions. However, differences across diagnosis groups are relatively small compared to variation driven by utilization measures.
This finding suggests that while clinical diagnosis contributes to readmission risk, it may act as a proxy for underlying severity rather than a direct driver. Additionally, the overlap among diagnosis categories across primary and secondary fields highlights the complexity of diabetic patients, many of whom present with multiple conditions.

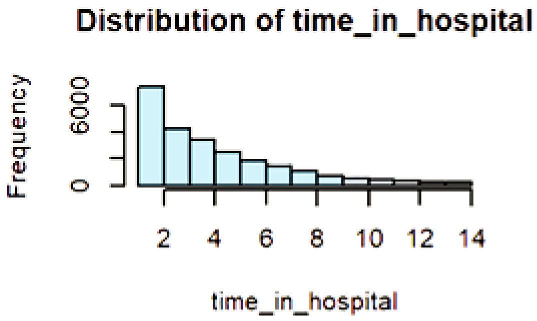


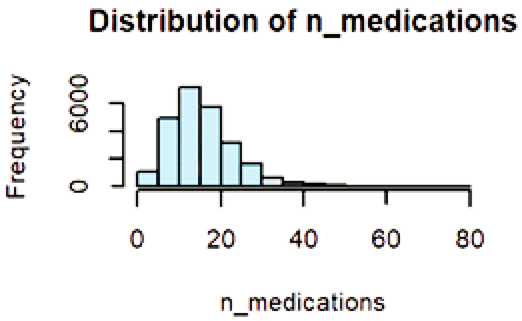
Pairwise correlations among numerical predictors reveal several moderate relationships. The strongest correlations are observed between time in the hospital and the number of medications, procedures, and laboratory procedures. These relationships are expected, as longer hospital stays typically involve more diagnostic tests and treatments. Importantly, most correlations fall within the low to moderate range (approximately 0.1 to 0.4), suggesting that severe multicollinearity is unlikely. Nevertheless, the presence of correlated utilization measures raises concerns about the stability of coefficients in linear models. This insight motivated later variance inflation factor (VIF) testing and the use of penalized regression methods. EDA thus directly informed both model choice and diagnostic evaluation.

The EDA results highlight several important implications for modeling hospital readmissions. First, moderate class imbalance necessitates evaluation metrics beyond accuracy. Second, skewed utilization variables benefit from transformation and normalization. Third, while diagnosis and demographic variables provide contextual information, utilization intensity appears to be the dominant signal in the data. Finally, modest correlations among predictors justify the use of both traditional and penalized regression approaches, as well as nonlinear models. Overall, EDA confirms that the dataset contains meaningful signals for predicting readmissions but also reveals structural limitations that constrain predictive performance. These insights provide the foundation for the feature engineering strategies and modeling approaches described in the following sections.

4. Feature Engineering / False Discovery Rate

To improve the predictive performance across various models, we performed targeted feature engineering based on thorough statistical testing and established background literature. Rather than relying on the raw utilization counts of the broadly encompassing dataset, we constructed features modified to better represent trends through normalization and captured unrealized trends by developing composite features.





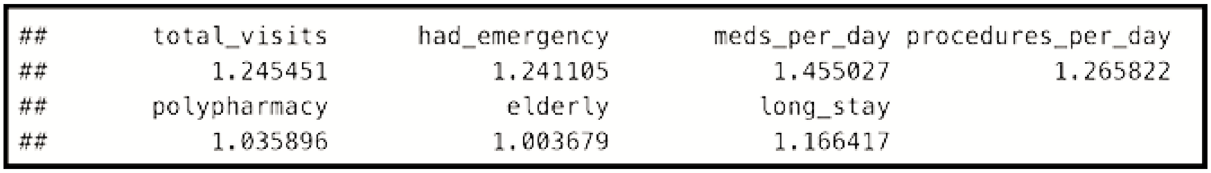
To counter the right-skewed nature of the raw utilization counts, we logarithmically normalized the data to achieve as much of an even distribution as possible. The modified data points were proven to show more statistically significant evidence when tested under univariate logistic regression models. This removed the issue of extreme-case patients which were outliers adding noise to the data. We also combined multiple features together into ratios to give the models context rather than raw count values. This would reflect treatment intensity better than counts since they could have been miscontextualized over any given period of time.

Feature Pool:

- Log_visits (logarithmic transformation of # of visits)
- Log_meds (logarithmic transformation of # of medications taken)
- Long_stay (binary predictor for patient with a hospital stay with a length over one week)
- Polypharmacy (binary predictor for patient that takes at least five medications per day)
- Age_num (age converted into numeric form from bins)
- Procedures_per_day
- Had_emergency (binary predictor for whether the patient had an emergency visit)

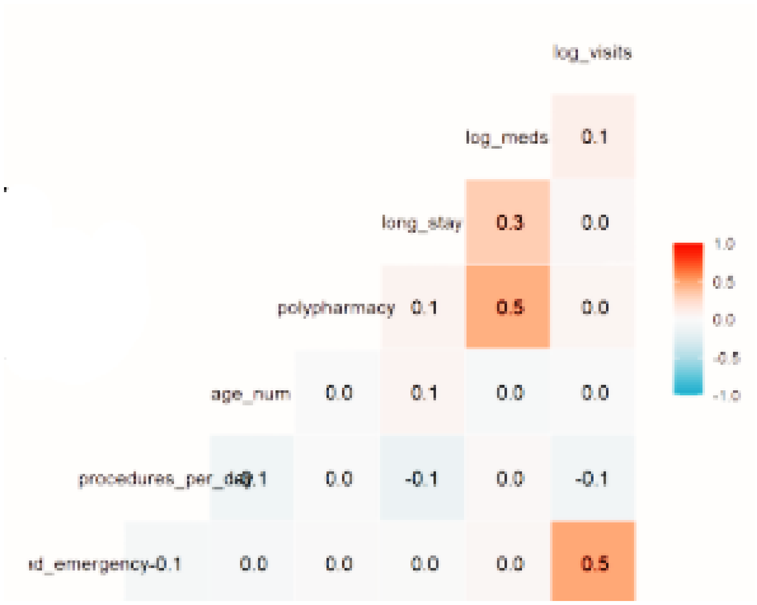
In order to guarantee that only statistically meaningful features were used in our modeling process we implemented the Benjamini-Hochberg False Discovery Rate test on our feature pool. This would ensure that the features we considered as statistically significant were not false positives under the assumption of 99% confidence. Since we were testing for multiple features, this allowed us to identify robust predictors of readmission while limiting false positives. The test sorted p-values from smallest to largest before adjustments based on the number of tests provided and their rank. If the adjusted p-values were under our threshold of 0.01, the feature would be deemed viable.

VIF Test (Variance Inflation Factor)



```
##        total_visits       had_emergency        meds_per_day procedures_per_day
##            1.245451            1.241105            1.455027           1.265822
##        polypharmacy             elderly           long_stay
##            1.035896            1.003679            1.166417
```

Variance Inflation Factor (VIF) measures how much a regression coefficient’s uncertainty is increased because that predictor is correlated with other predictors. This is calculated by temporarily treating each feature as the outcome and analyzing the returned R^2 value. This is converted to the VIF through the formula $VIF=1/(1-R^2)$. Since our features had relatively low VIF values, close to 1, we determined that there was mild to no collinearity amongst our features.



The collinearity matrix is another visual representation of the correlation between features. To reduce feature redundancy that would add noise to the models and overfit certain data, we set a threshold of 0.5 which would describe moderate to mild collinearity. This further narrowed down our feature pool until we had a set of seven features that were conceptually sound at a high level in relation to hospital readmission rates and were also statistically proven to be significant without excessive redundancy and the chance of false positives.

5. Modeling

To evaluate the ability to predict hospital readmissions, we implemented and compared a diverse set of classification models. These include:

- Logistic Regression
- Penalized Logistic Regression
- Random Forest
- Generalized Additive Model
- Support Vector Machine
- XGBoost
- Tuned XGBoost

These models were selected to explore the tradeoff between interpretability and predictive flexibility. Essentially, it highlights how a model can adapt to complicated patterns while still knowing which variables mattered. All models were evaluated using consistent training, validation, and test splits, and performance was assessed using accuracy, sensitivity, specificity, and area under the ROC curve (AUC).

a) Logistic Regression

Logistic regression serves as a natural baseline for binary classification problems such as hospital readmissions. THe model estimates the probability that a patient will be readmitted based on a linear combination of predictors including healthcare utilization, medication burden, procedural intensity, and demographic indicators. While it is the simplest of the models, it still performs well. The formula for logistic regression is:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots b_n X_{ni})}}$$

where:

- $P(Y = 1)$ indicates hospital readmissions
- X_i are the values of the predictors (e.g., total visits)
- β_0 is the intercept
- β□ tells us how much this characteristic changes the chance of readmission

- n is number of predictors

Results:

- **Accuracy (~0.60)** indicates acceptable overall classification.
- **Low sensitivity (~0.34)** shows that many readmitted patients are missed.
- **High specificity (~0.82)** reflects strong performance in identifying non readmitted patients.
- **AUC (~0.63)** indicates only moderate discrimination.

Conclusion:
Logistic regression establishes a baseline but lacks sufficient flexibility to model the complexity of hospital readmissions.

b) Penalized Logistic Regression

Penalized logistic regression extends the standard logistic framework by adding regularization terms to the likelihood function. LASSO encourages sparsity, while Ridge shrinks correlated coefficients. These approaches aim to reduce overfitting and improve generalization. The penalties take the form:

Ridge: $$\sum_{j=1}^{p}(\beta_j)^2 < c.$$

LASSO: $$\sum_{j=1}^{p}|\beta_j| < c.$$

where:

- p: total number of predictors included in the model
- j: index for predictors, where j=1,2,…,p□
- β□: regression coefficient associated with the j-th predictor
- c > 0: regularization constraint that controls the strength of shrinkage
  - Smaller c -> stronger regularization
  - Larger c -> weaker regularization

Results (LASSO):

- **Accuracy (~0.60)** indicates acceptable overall classification performance.
- **Low sensitivity (~0.33)** shows that many readmitted patients are missed by the model.
- **High specificity (~0.84)** reflects strong performance in identifying non readmitted patients.
- **AUC (~0.63)** indicates only moderate discriminatory ability.

Results (Ridge):

- **Accuracy (~0.60)** indicates acceptable overall classification performance.
- **Low sensitivity (~0.35)** shows that many readmitted patients are missed by the model.
- **High specificity (~0.83)** reflects strong performance in identifying non readmitted patients.
- **AUC (~0.62–0.63)** indicates only moderate discriminatory ability.

Both penalized models produced performance extremely similar to standard logistic regression. This happens because LASSO and Ridge mainly help by shrinking coefficients and reducing variance, but they still assume a linear relationship. Since readmission risk likely depends on nonlinear effects and interactions, regularization alone does not dramatically improve predictive power. The main practical effect here is slightly more stable generalization across validation and test sets.

Conclusion:
Penalized logistic regression improves coefficient stability and helps control overfitting, but it does not substantially improve predictive performance compared to standard logistic regression for this dataset. This suggests that the limitation is more about the linear structure of logistic regression than instability in coefficient estimates, motivating the use of more flexible nonlinear models.

c) Random Forest

Random Forests are ensemble learning methods that combine predictions from a large number of decision trees. Each tree is trained on a bootstrap sample of the training data, and at each split only a random subset of predictors is considered. This randomness reduces correlation between trees and helps control overfitting. Unlike logistic regression, Random Forests automatically capture nonlinear relationships and interactions among predictors without requiring them to be explicitly specified. The general Random Forest take the form:

$$\bar{T} = \frac{1}{n} \sum_{i=1}^{n} T_i.$$

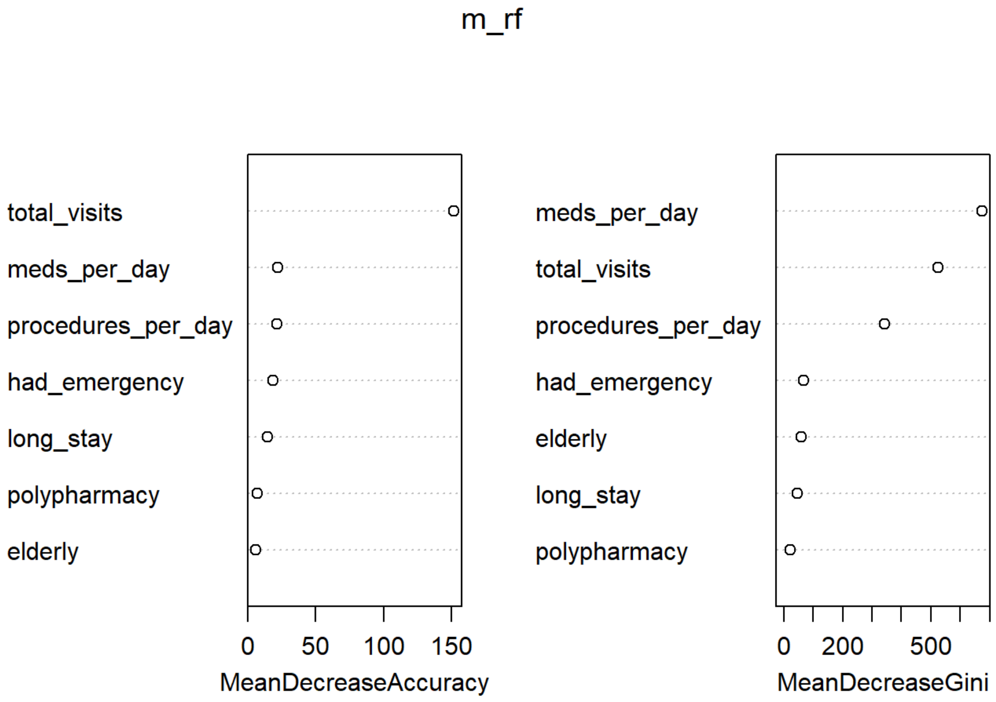
where:

- $\overline{T}$: final Random Forest prediction (average predicted probability of readmission)
- n: total number of trees in the Random Forest
- i: index of an individual tree, where $i = 1 , 2 , \dots , n$
- T_i: predicted probability of readmission produced by the iii-th decision tree

Results:

- **Accuracy (~0.60)** is comparable to logistic and penalized logistic regression.
- **Sensitivity (~0.45)** improves substantially, indicating better identification of readmitted patients.
- **Specificity (~0.73)** decreases, reflecting an increase in false positives.
- **AUC (~0.61)** indicates moderate discriminatory ability.

The Random Forest model differs from the linear and penalized logistic regression models because it can capture nonlinear relationships and interactions among predictors automatically. By averaging predictions across many decision trees, the model is better able to identify patients at higher risk of readmission. This increased flexibility improves detection of readmissions but also leads the model to flag more patients overall. As a result, Random Forests tend to emphasize sensitivity over specificity. Overall, the model shows that allowing nonlinear structure improves performance, though gains in overall discrimination remain limited.



The Random Forest variable importance plot shows how much each predictor contributes to the model using Mean Decrease in Accuracy and Mean Decrease in Gini. The chart indicates that total_visits is the most important predictor, followed by meds_per_day and procedures_per_day, highlighting the strong role of healthcare utilization. Variables such as polypharmacy, elderly, and long_stay contribute less to the model's predictions.

Conclusion:
The Random Forest model demonstrates that incorporating nonlinear relationships improves the identification of patients at risk of readmission, but the overall gains remain modest, indicating that more refined ensemble methods may be needed for further improvement.

d) Generalized Additive Model (GAM)

The Generalized Additive Model (GAM) extends logistic regression by allowing certain predictors to have **nonlinear effects** on the probability of hospital readmission while still maintaining interpretability. Instead of assuming a strictly linear relationship between predictors, GAMs model selected predictors using smooth functions. This makes GAMs particularly well suited for healthcare data, where risk often increases in a nonlinear manner as utilization measures grow.

The GAM used in this analysis includes smooth terms for total visits, medications per day, and procedures per day, while binary indicators such as emergency visits, polypharmacy, elderly status, and long hospital stays are included as linear terms. The formula is:

$$g\left(E[y|\mathbf{x}]\right) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

where:

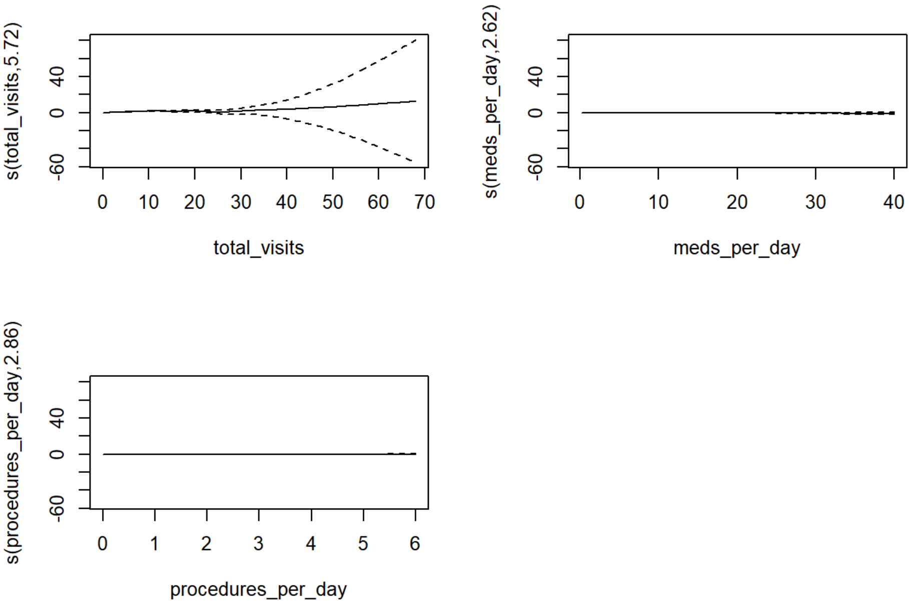
- Y: binary response variable
- X: (x_1 to x□) is vector of predictor variables
- g(): link function that connects the mean of the response to the linear predictor
- β_0 is the intercept
- p is total number of predictors

Results:

- **Accuracy (~0.61)** indicates strong overall classification performance and is slightly higher than logistic and penalized logistic regression.
- **Sensitivity (~0.47)** improves substantially, reflecting better identification of readmitted patients compared to linear models.
- **Specificity (~0.72)** decreases slightly, indicating an increase in false positives relative to logistic regression.
- **AUC (~0.63–0.64)** indicates good discriminatory ability and an improvement over linear and penalized models

Compared to logistic and penalized logistic regression, the GAM shows higher sensitivity and AUC. This improvement occurs because the model can capture nonlinear relationships between healthcare utilization variables and readmission risk. Instead of assuming that each additional visit or procedure has the same effect, the GAM allows the risk of readmission to increase more sharply at higher levels of healthcare use. This behavior is consistent with clinical expectations. At the same time, the model remains stable and avoids the high variability sometimes seen in more complex machine learning models.

.



The smooth plots show that total visits have a strong nonlinear effect on readmission risk. The risk increases sharply as the number of visits becomes large. In contrast, medications per day and procedures per day have mostly flat effects, indicating a weaker influence once other variables are controlled for. Overall, the GAM improves performance by capturing the nonlinear relationship between frequent healthcare use and readmission risk.

Conclusion:
The Generalized Additive Model improves on logistic regression, LASSO, and Ridge by capturing nonlinear patterns in healthcare use while remaining easy to interpret. Its higher sensitivity and AUC show better identification of high risk patients. Compared to Random Forest, the GAM provides more stable results and clearer insight into how predictors affect readmission risk, making it a strong overall model for this analysis.

e) Support Vector Machine (SVM)

Support Vector Machines (SVMs) are classification models that separate observations by constructing a decision boundary with the largest possible margin between classes. To allow for nonlinear relationships among predictors, an SVM with a **radial basis function (RBF) kernel** was used. The RBF kernel enables the model to capture complex patterns in the data that cannot be represented by linear decision boundaries. Since SVMs rely on distance based calculations, all predictors were standardized using the training set mean and standard deviation prior to model fitting. The model is written as:

$$f(\mathbf{x}) = \sum_{i}^{N} \underset{\uparrow}{\alpha_i} y_i k(\underset{\nwarrow}{\mathbf{x}_i}, \mathbf{x}) + b$$

where:

- f(x): decision value for observation x
- N: number of support vectors
- α_i: learned weight for the i-th support vector
- y_i: class label of the i-th support vector (−1 or +1)
- x_i: i-th support vector
- x: input observation being classified
- $k(x_i,x)$: kernel function evaluating similarity between x_i and x
- b: intercept (bias) term

And the RBF kernel:

$$k(x, y) = e^{\left(-\gamma \| x - y \|^2\right)}$$

where:

- k(x,y): kernel similarity between observations x and y
- x: first input vector
- y: second input vector
- γ: kernel bandwidth parameter controlling smoothness
- // x−y // ^2: squared Euclidean distance between x and y

Results:

- **Accuracy (~0.60)** is comparable to logistic regression methods.
- **Sensitivity (~0.52)** is substantially higher than linear and penalized logistic regression models.

- **Specificity (~0.67)** is lower, indicating an increased rate of false positives.
- **AUC (~0.62)** suggests moderate discriminative ability.

Compared to linear models such as logistic regression, LASSO, and Ridge, nonlinear models show better sensitivity, meaning they are more effective at identifying patients who are likely to be readmitted. This improvement occurs because nonlinear models can capture more complex relationships between predictors and readmission risk. However, higher sensitivity comes with lower specificity, since more patients are classified as high risk even when they are not readmitted. When compared to other nonlinear approaches like Random Forests and Generalized Additive Models, overall performance remains moderate, with only small improvements in AUC. This indicates that although nonlinear patterns exist in the data, the separation between readmitted and non readmitted patients is limited, and the performance gains over linear models are modest rather than substantial.

Conclusion:
The SVM improves sensitivity compared to linear models, meaning it identifies more patients who are likely to be readmitted. However, this comes with lower specificity, as more non readmitted patients are incorrectly classified as high risk. Overall, the SVM performs similarly to other nonlinear models such as Random Forests and Generalized Additive Models and does not show a clear advantage over them.

f) XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful learning method that builds decision trees sequentially, where each new tree is trained to correct the mistakes made by the previous trees. Unlike Random Forests, which build trees independently, XGBoost focuses more heavily on observations that are difficult to classify, such as patients who are repeatedly readmitted. This makes XGBoost well suited for complex healthcare data where nonlinear relationships and interactions between variables are expected. The formula is:

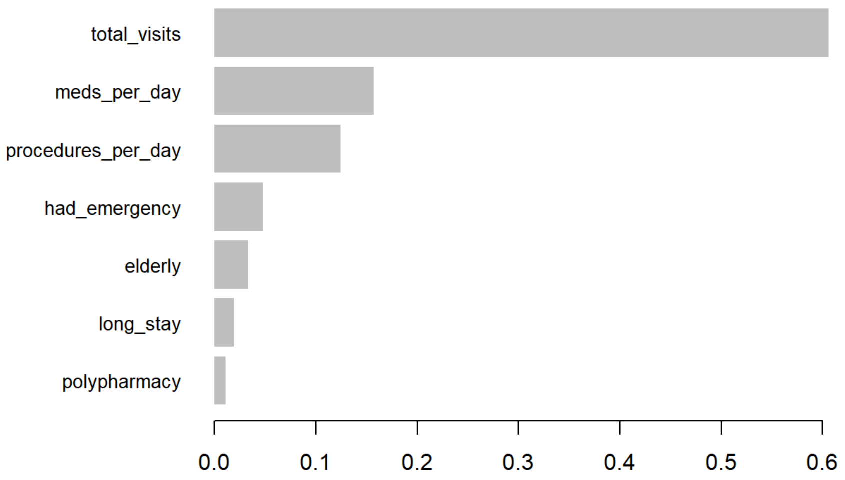
$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

where:

- $\mathcal{L}^{(t)}$: loss function at boosting iteration t
- n: number of observations
- i: index for observations, i=1,…,n
- y_i: observed outcome for observation i

- $\hat{y}_i^{(t-1)}$: predicted value for observation iii from the model after t−1 trees
- f□(x_i): prediction from the decision tree added at iteration t for observation i
- x_i: vector of predictor values for observation i
- l(): loss function
- Ω(f□): regularization term that penalizes the complexity of the tree added at iteration t
- t: boosting iteration index

Compared to the other models tested, XGBoost shows the strongest overall performance. Logistic regression and penalized logistic regression had similar accuracy but low sensitivity, meaning many readmitted patients were missed. Random Forest, SVM, and GAM improved sensitivity but either reduced specificity or showed only moderate gains in AUC. XGBoost achieved the best balance, with higher sensitivity and competitive AUC while maintaining reasonable specificity, making it the most effective model among those evaluated.



This feature importance plot from the XGBoost model shows that total_visits is by far the most important predictor of hospital readmission. It contributes much more to the model than any other variable. meds_per_day and procedures_per_day are the next most influential features, indicating that treatment intensity also plays an important role. Variables such as had_emergency, elderly, long_stay, and polypharmacy have smaller contributions, suggesting they provide additional but less dominant information in predicting readmission risk.

Conclusion:

Overall, this feature importance analysis helps explain why the XGBoost model performs well, as it effectively prioritizes the most informative predictors related to healthcare utilization. The strong influence of total visits and treatment intensity aligns with clinical expectations and supports the model's improved sensitivity. Building on these results, the next step is to further refine XGBoost through hyperparameter tuning to assess whether additional performance gains can be achieved.

g) XGBoost Tuned

Building on the baseline XGBoost model, a tuned version of XGBoost was developed to improve performance through hyperparameter optimization rather than changes to the model structure. A randomized search was used to select values for parameters controlling learning rate, tree depth, subsampling, and regularization, with early stopping based on validation AUC to reduce overfitting.

The tuned model follows the same boosting framework as the original XGBoost model but achieves better generalization by limiting excessive tree complexity. By applying stronger regularization and more conservative tree growth, the model better balances flexibility and stability, leading to more consistent performance on unseen data.

Results:

- **Accuracy (~0.61)** is comparable to the baseline XGBoost model and slightly higher than most linear and ensemble alternatives.
- **Sensitivity (~0.43)** remains substantially higher than logistic and penalized logistic regression, indicating improved identification of patients at risk of readmission.
- **Specificity (~0.76)** improves relative to the untuned XGBoost model, showing fewer false positives.
- **AUC (~0.64 on validation, ~0.63 on test)** represents the strongest overall discriminative performance among all models evaluated.

The performance improvement of the tuned XGBoost model comes from using stronger regularization and more conservative tree settings. A smaller learning rate and stricter splitting rules prevent the model from fitting noise in the training data while subsampling helps reduce variability across trees. As a result, the tuned model generalizes better to new data and produces more stable predictions. Compared to the baseline XGBoost model, tuning slightly improves specificity while maintaining strong sensitivity, meaning the model becomes more selective without missing substantially more readmissions. Overall, the tuned XGBoost achieves a better balance between sensitivity and specificity than Random Forests, GAMs, and SVMs, leading to stronger overall discrimination.

Conclusion:

Hyperparameter tuning improves the XGBoost model by helping it perform more consistently on new data. Although the overall structure of the model does not change, adjusting key settings leads to better overall performance. The tuned XGBoost model achieves the highest AUC and a good balance between sensitivity and specificity, making it the strongest model for predicting hospital readmission risk in this study.

6. Results

Validation Performance Metrics:

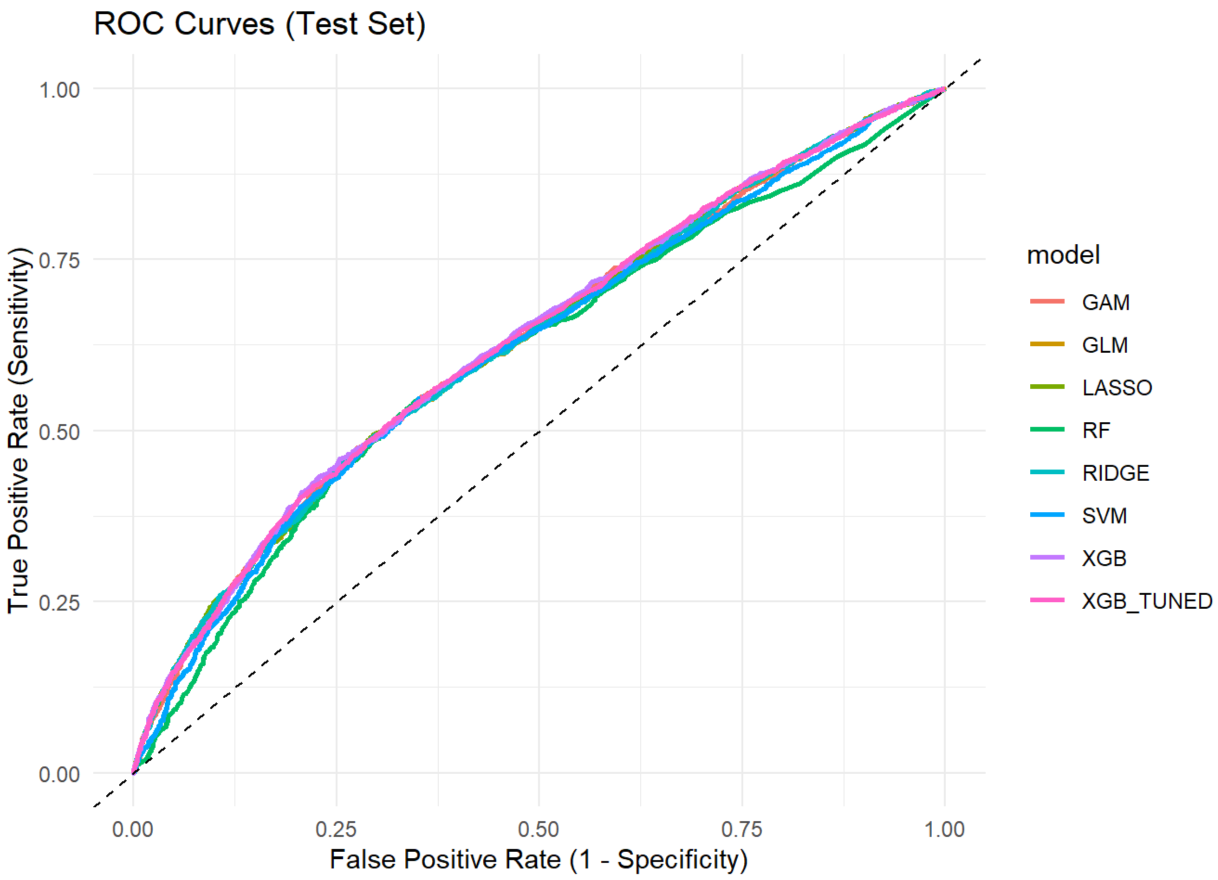
Model	Accuracy	Sensitivity	Specificity	AUC
GLM	0.6072	0.3556	0.8280	0.6347
LASSO	0.6060	0.3406	0.8389	0.6360
RIDGE	0.6048	0.3513	0.8273	0.6340
RF	0.6034	0.4536	0.7349	0.6185
XGB	0.6100	0.5208	0.6883	0.6396
XGB_TUNED	0.6072	0.4288	0.7638	0.6427
GAM	0.6108	0.4728	0.7319	0.6405
SVM	0.6052	0.5220	0.6782	0.6241

Validation Test Metrics:

Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
GLM	0.5994	0.3395	0.8240	0.6241
LASSO	0.6022	0.3305	0.8371	0.6250
RIDGE	0.6028	0.3430	0.8274	0.6241
RF	0.6014	0.4573	0.7260	0.6060
XGB	0.6002	0.5198	0.6696	0.6298
XGB_TUNED	0.6084	0.4297	0.7629	0.6285
GAM	0.6084	0.4745	0.7241	0.6259
SVM	0.6000	0.5242	0.6655	0.6169

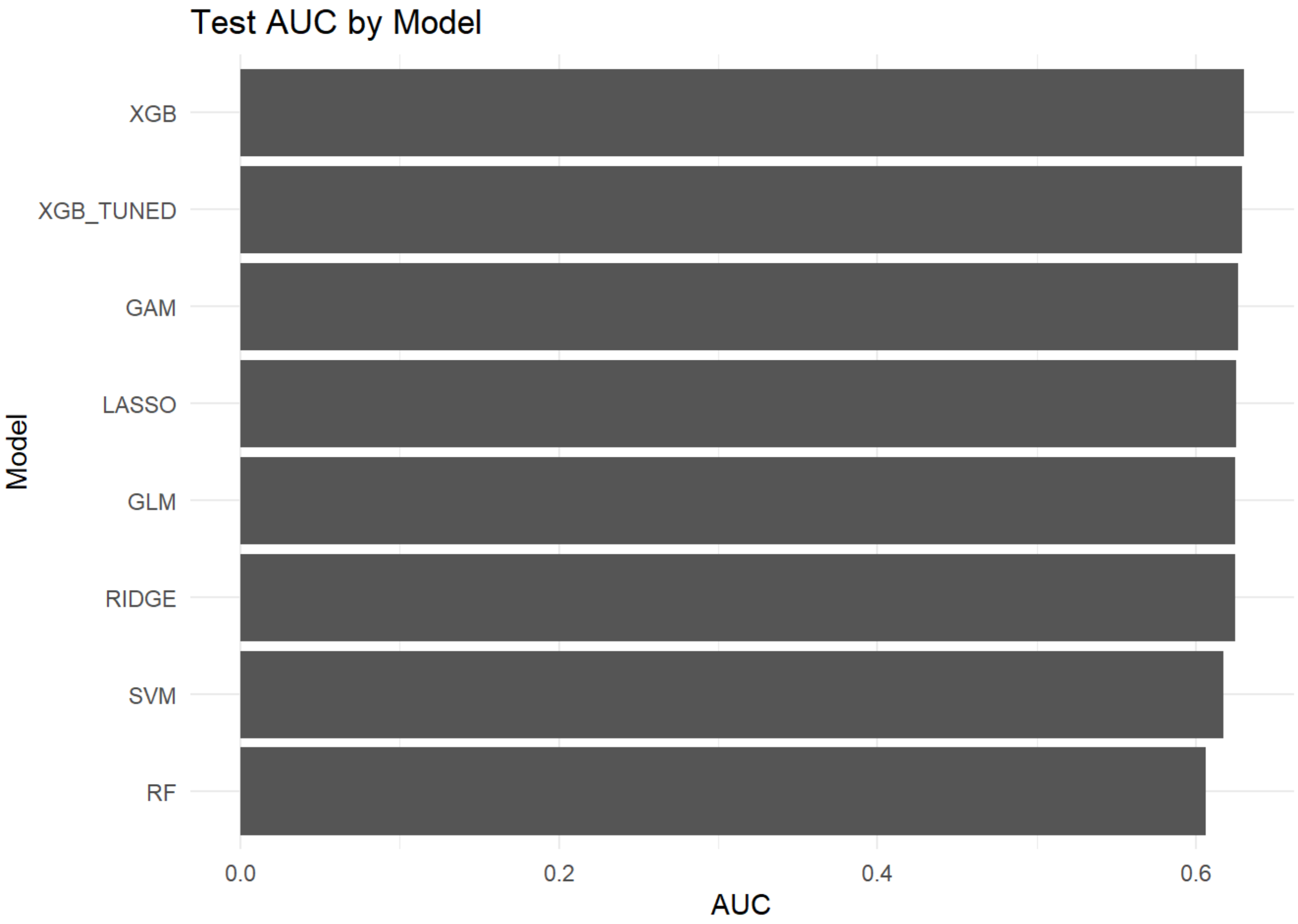
Across both the validation and test sets, overall accuracy remains similar across all models, clustering around 0.60. This suggests that accuracy alone is not sufficient to distinguish model performance in this setting. Larger differences emerge when comparing sensitivity, specificity, and AUC. The linear models GLM, LASSO, Ridge consistently show high specificity but low sensitivity, indicating that they correctly identify non readmitted patients but fail to detect a large portion of actual readmissions. In contrast, nonlinear models such as Random Forest, GAM, SVM, and XGBoost substantially improve sensitivity, capturing more high risk patients at the cost of increased false positives. Among all models, XGBoost and tuned XGBoost achieve the strongest balance, with relatively high sensitivity and the highest AUC values on both validation and test sets. These results indicate that nonlinear and boosted models provide better overall discrimination than linear approaches.

ROC Curves (Test Set)



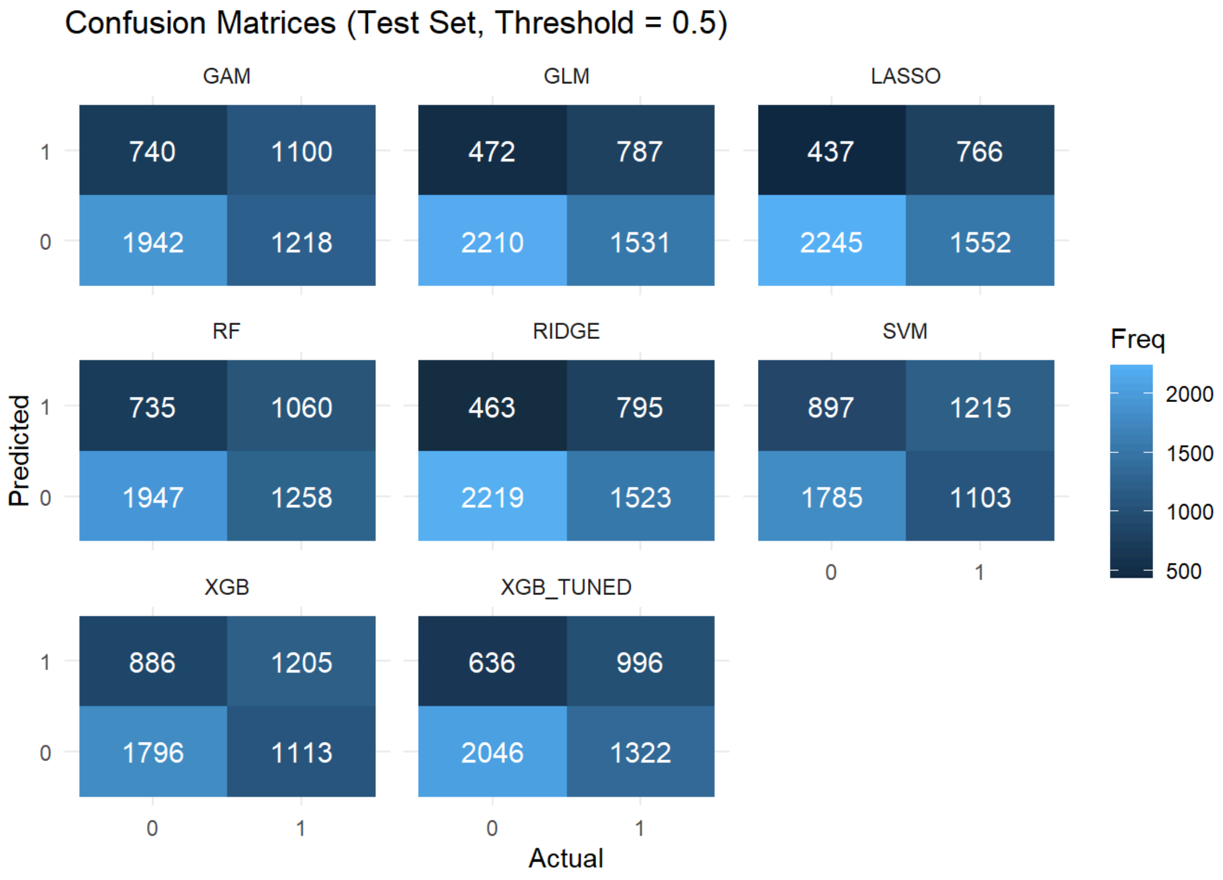
The ROC curves for all models look very similar, which matches the fact that their AUC values are close to each other. This means that all models have a similar ability to rank patients by readmission risk. Some nonlinear models, such as GAM, SVM, and XGBoost, perform slightly better than the linear models in parts of the curve, but the differences are small and hard to see visually. The tuned XGBoost model shows a small improvement, which is clearer in its AUC value than in the shape of the ROC curve. Overall, the ROC curves suggest small performance differences rather than major improvements across models.

AUC bar chart



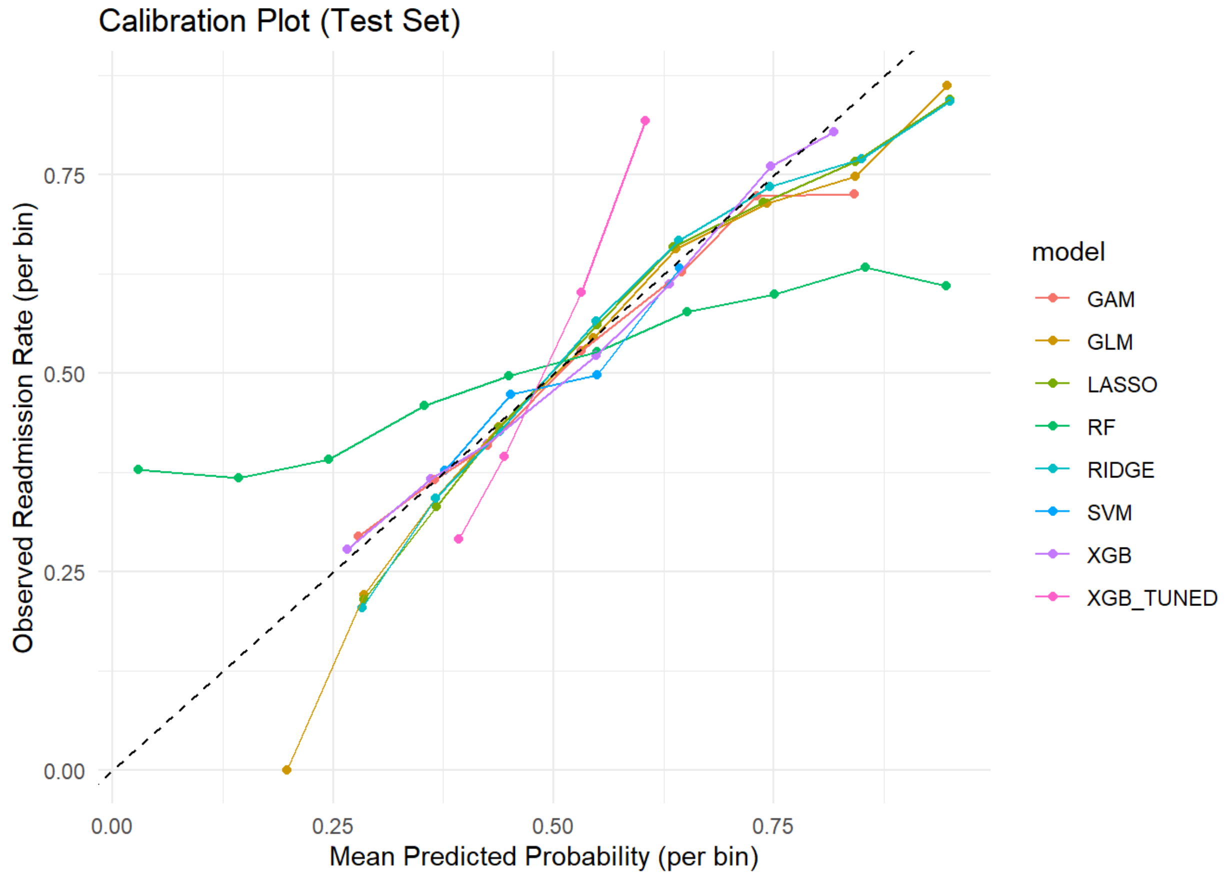
The AUC bar chart shows that all models perform at a similar level, with only small differences in overall performance. Random Forest has the lowest AUC, indicating that it is slightly worse at distinguishing between readmitted and non readmitted patients. The linear models perform somewhat better, but their AUC values still suggest only moderate discrimination. GAM and XGBoost achieve slightly higher AUC values, with the tuned XGBoost model performing best overall. Although these improvements are modest, the results show that XGBoost based models consistently provide a small but reliable advantage over the other approaches.

Confusion matrix heatmaps (threshold= 0.5)



The confusion matrices show that the linear models miss many patients who are actually readmitted. This means they do not catch enough high risk patients. Models like Random Forest, GAM, SVM, and XGBoost identify more readmitted patients, but they also incorrectly flag more patients as high risk. The tuned XGBoost model shows a better balance, catching more readmissions without creating as many false alarms. Overall, these results show a clear tradeoff between missing readmissions and falsely predicting them across models.

Calibration plot



The AUC bar chart shows that all models perform at a similar level, with only small differences between them. Random Forest has the lowest AUC, meaning it struggles the most to separate readmitted and non readmitted patients. The linear models perform slightly better, but still only at a moderate level. GAM and XGBoost, including the tuned XGBoost model, have the highest AUC values, but the improvement is small. In addition, the calibration plot shows that the tuned XGBoost model sometimes overestimates readmission risk, especially in the middle range of predicted probabilities. Overall, even the tuned XGBoost model does not perform especially well, highlighting that predicting hospital readmission remains a challenging task with the available data.

Based on the combined evaluation of AUC, sensitivity, and calibration, the standard XGBoost model provides the best overall balance among the models tested. While tuning slightly improves AUC, it does not consistently improve probability reliability, making the untuned XGBoost model the preferred choice in this analysis.