HOSPITAL READMISSION PREDICTION

COGNIZANT - NURTURE PARTNER NETWORK PROGRAM

CTS MENTOR
Mr. Sauvik De

REC MENTOR
Dr.Vijay K

TEAM MEMBERS

RAHUL R PRAVEEN R PRAVIN A PRIYA E PUGAL M

RAMYA N RATHESHVER S RISHI S RAZEEN BATSHA S

Rajalakshmi Engineering College

Department of Artificial Intelligence and Machine Learning

Contents

Background

- Hospital readmissions indicate gaps in care quality and increase risks for complications, infections, delayed recovery, and even higher mortality. They also contribute to emotional stress for patients and financial strain on families. Addressing this challenge can improve patient outcomes while reducing healthcare costs.
- Early identification of high-risk patients allows targeted interventions, such as enhanced discharge planning, timely follow-up care, and better patient education. This proactive approach helps prevent avoidable readmissions and supports more efficient use of healthcare resources.

Objective

To develop an accurate predictive model that classifies patients as likely or unlikely to be readmitted to the hospital within a specified time frame by:

- **Understand why readmissions happen** – whether due to medical, social, or system-level factors.
- **Help hospitals cut costs** while raising quality.
- **Reduce avoidable readmissions,** improving both patient lives and healthcare sustainability

Dataset features

The columns available in the dataset are:

- **age**: Age bracket of the patient
- **time_in_hospital**: Length of hospital stay (1–14 days)
- **n_procedures**: Number of procedures performed during the stay
- **n_lab_procedures**: Number of lab procedures performed
- **n_medications**: Number of medications administered
- **n_outpatient**: Outpatient visits in the year prior
- **n_inpatient**: Inpatient visits in the year prior
- **n_emergency**: Emergency visits in the year prior
- **medical_specialty**: Admitting physician's specialty
- **diag_1**: Primary diagnosis (Circulatory, Respiratory, Digestive, etc.)
- **diag_2**: Secondary diagnosis
- **diag_3**: Additional secondary diagnosis
- **glucose_test**: Glucose serum test result (high, normal, not performed)
- **A1Ctest**: A1C level result (high, normal, not performed)
- **change**: Change in diabetes medication (yes/no)
- **diabetes_med**: Diabetes medication prescribed (yes/no)
- **readmitted**: Target variable (yes/no)

Dataset Description

Hospital readmissions dataset revolves around predicting hospital readmissions for diabetic patients.It contains 25,000 rows and 17 columns.Out of these, one column acts as the target variable (readmission) while the rest serve as predictor variables.

- **Clean & Balanced Dataset** : Proportion of readmissions is 53%-YES, indicating no high imbalance in the target variable
- **Focused around Diabetes** : Most of the patients in the dataset have diabetes, making it disease-specific and clinically meaningful.
- **Transforming Data into Insights** : We can create derived features such as Medication burden score for more powerful predictions.

Our approach

Data-Driven Storytelling

- Start with in-depth Exploratory Data Analysis (EDA) to uncover patterns and relationships driving hospital readmissions.

End-to-End Machine Learning Pipeline

- EDA-driven preprocessing to clean, encode, and transform data
- Predictive model building with rigorous evaluation using cross-validation
- Hyperparameter tuning to optimize performance

Beyond Accuracy

- Moved beyond simple accuracy to a multi-metric evaluation
- Recall-focused as in clinical settings, false negatives are costly (missing high-risk patients can lead to severe complications and readmissions)
- Added custom cost analysis to capture clinical + economic value of predictions

Architecture diagram

Hospital_readmission.csv

Initial Exploration

- Inspecting the dataset's size & statistics
- Checking missing values and duplicate records
- Class imbalance Checking - Target Variable

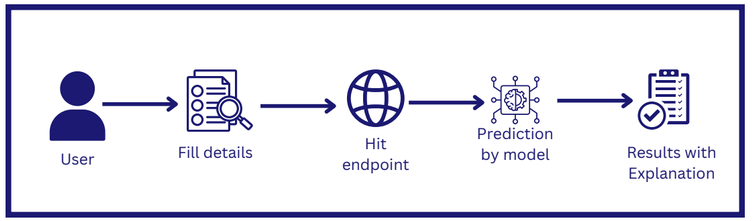
Exploratory Data Analysis

- Analyse the dataset's features
- Visualize the records
- Collinearity Checking

Original Dataset Splitting

- Training Dataset- 75%
- Evaluation Dataset- 15%
- Testing Dataset -10%

Process Flow Diagram



Sklearn Pipeline - Feature Engineering & Modeling

Feature Added

- Number of visits
- Procedure Medication Ratio

Encoding

- One - Hot Encoding - Numerical Columns
- Label Encoding - Categorical Columns

Modeling

Baseline Models

- Logistic regression
- Random forest
- Gradient boosting
- Bagging KNN
- MLP-Neural network
- Naive bayes
- XGboost

Training | **Evaluation** | **Model selection** | **Tuning**

Randomized Search CV

1. Random Forest
2. Gradient Boosting
3. XG Boost

Chosen Best Model

Random Forest (Fine Tuned)

Recall - 86.13 %
Accuracy - 55.72 %
ROC - AUC - 65.67 %

.pkl

Pickle file

WEB UI

Explanaibility

Reasoning of why we get the result and report the basis of the result using SHAP.

Predictions

Readmission probability is predicted based on the input given , i.e., if the patient is likely to be readmitted or not

Flask app

Handle endpoints and import pickle model

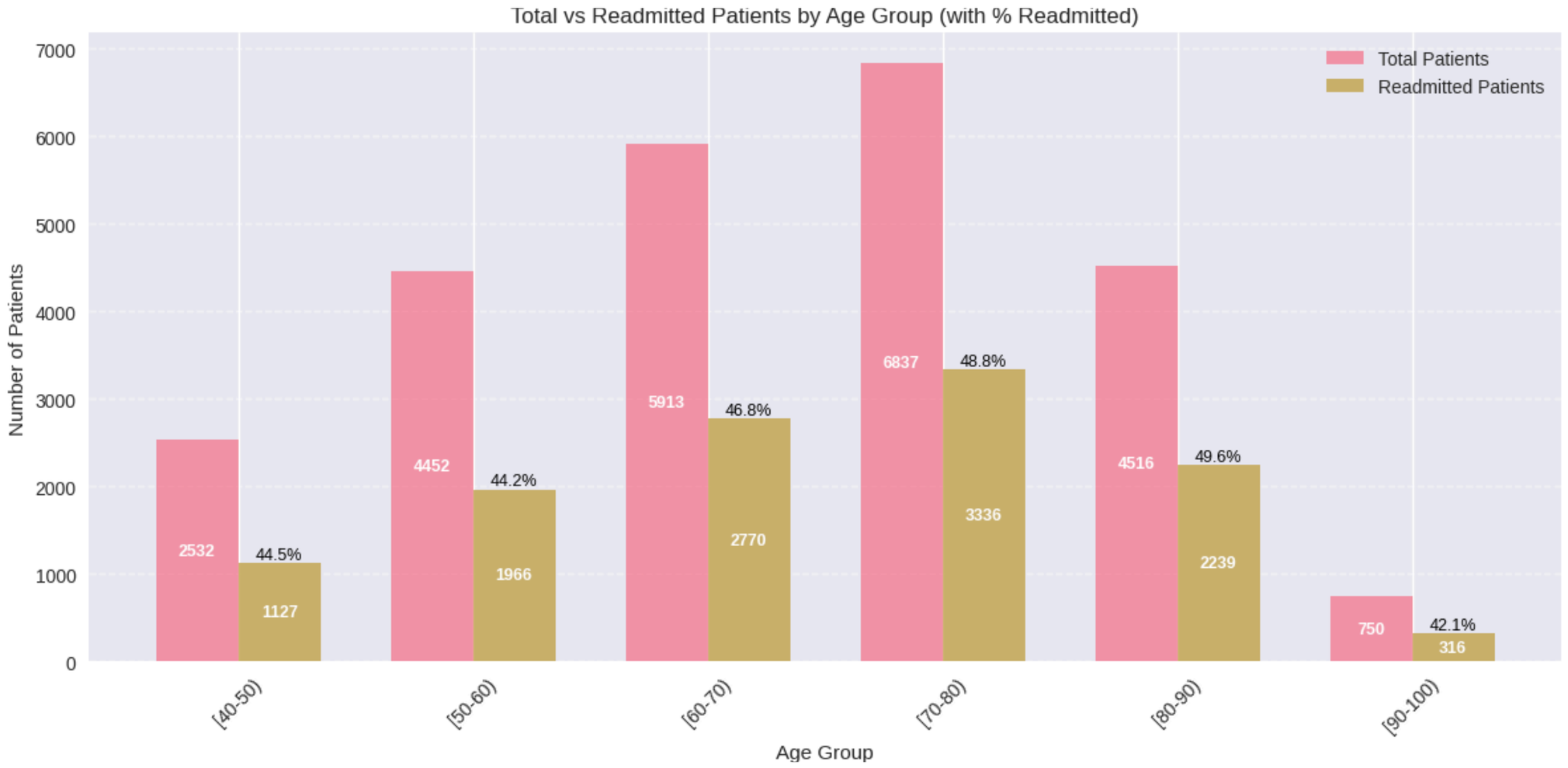
User Interface

Handle the inputs from user & display the results

Exploratory Data Analysis

EDA STEPS	DESCRIPTION
Data Overview	- Checked dataset shape, column types, and summary statistics. - Inspected missing values and unique values in categorical features.
Target Variable Analysis	- Visualized distribution of readmitted vs. non-readmitted patients
Univariate Analysis	- Analyzed frequency of categorical variables (e.g., admission type, gender, age groups). - Plotted boxplots for numerical features like time in hospital, number of lab procedures.
Bivariate Analysis	- Explored relationship between readmission and predictors . diagnosis(primary), age .
Correlation Analysis	- Generated a heatmap for numerical variables to identify multicollinearity.
Outlier Detection & Cleaning	- Identified and assessed outliers in features such as inpatient, outpatient & emergency visits

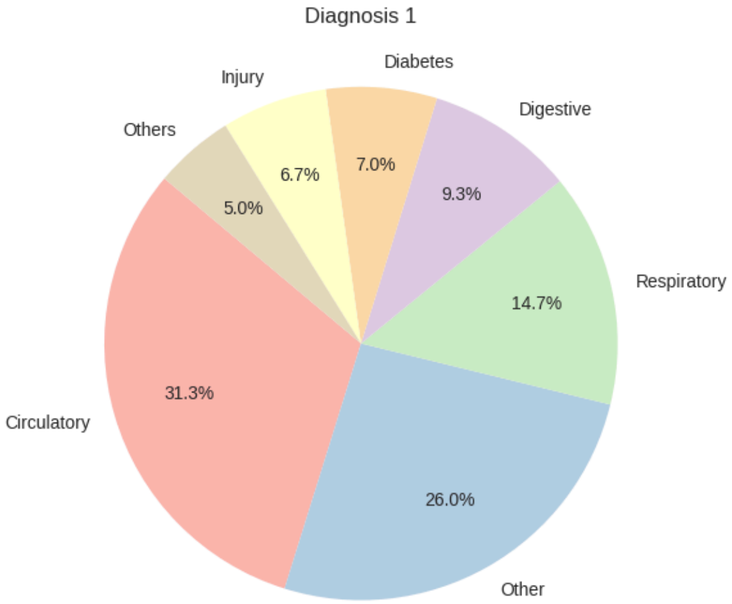
Distribution of Readmission Rate by Age Group

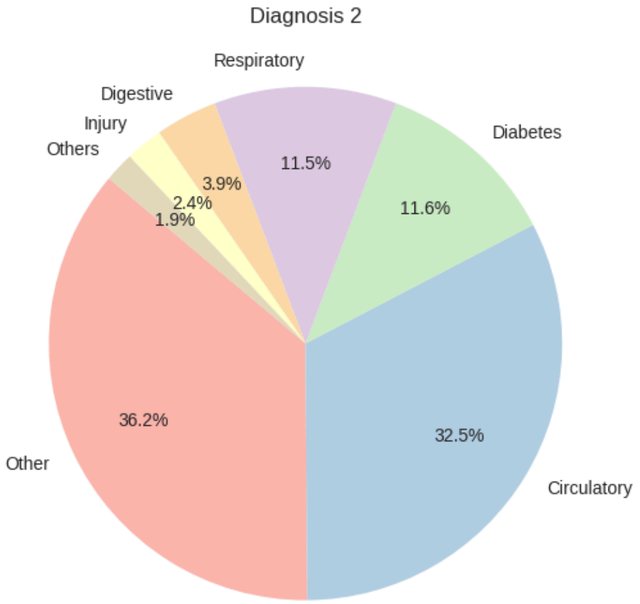


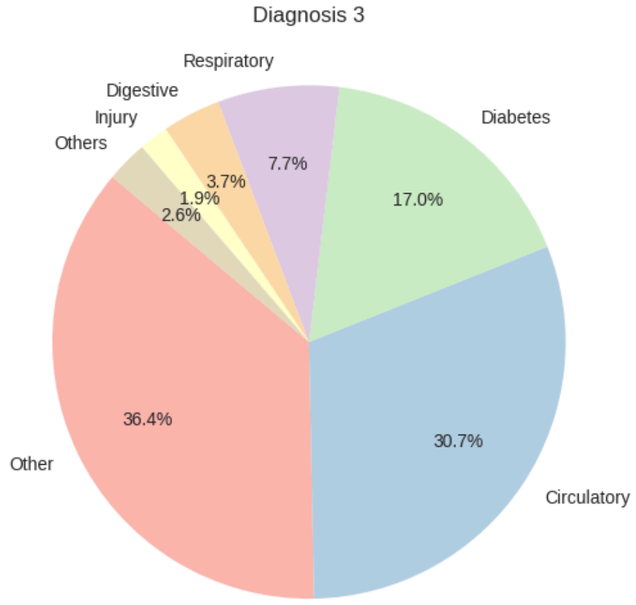
Older Age = Higher Risk: Readmission % is highest in 70–90 age group (≈49%), showing greater vulnerability.

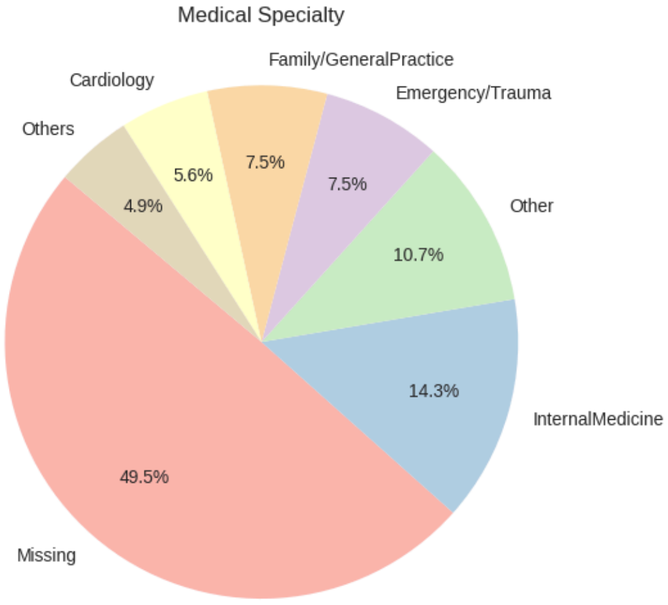
Key Action: Focus follow-up care on older patients (70–90) to reduce preventable readmissions.

Deep Data Analysis on Diagnosis & Medical SpecialityVariable









1) **Primary Diagnosis:** Most readmissions are linked to Circulatory, Respiratory, and Diabetes-related issues across all diagnosis stages.

2) **Diagnosis Progression:** Share of Diabetes increases in later diagnoses, highlighting chronic condition impact.

3) **Medical Specialty:** Nearly 50% specialty data missing, but among recorded cases, Internal Medicine & Cardiology dominate.

Correlation matrix

Moderate Correlation:

Time in hospital ↔ Medications (0.45) & Lab procedures (0.33) → longer stay = more treatments/tests.

Weak Correlation: Most other features show very low or no correlation.

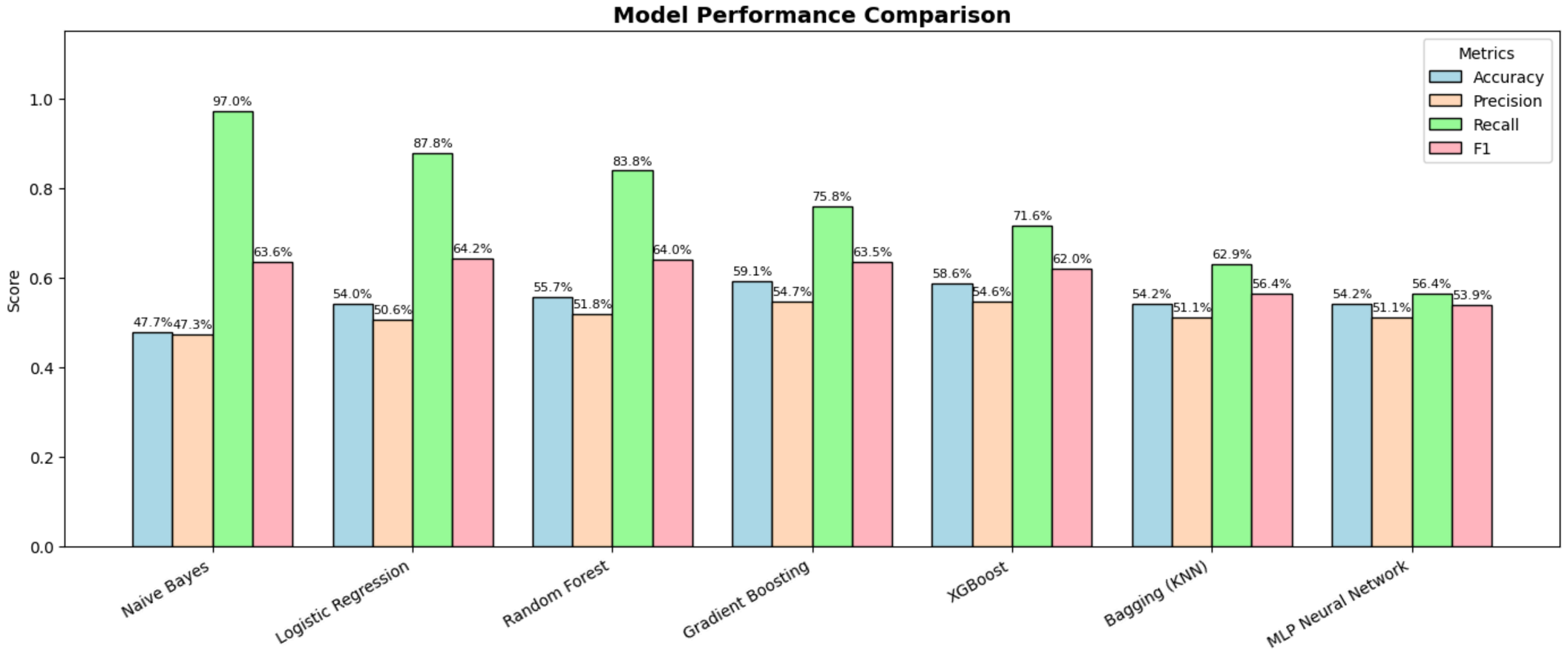
Negative Relation: Procedures ↔ Inpatient visits (-0.08) → more procedures don't necessarily mean more admissions.

Action Point: Features are largely independent, reducing multicollinearity risk for modeling.

Strongest Positive Correlations (excluding diagonal):

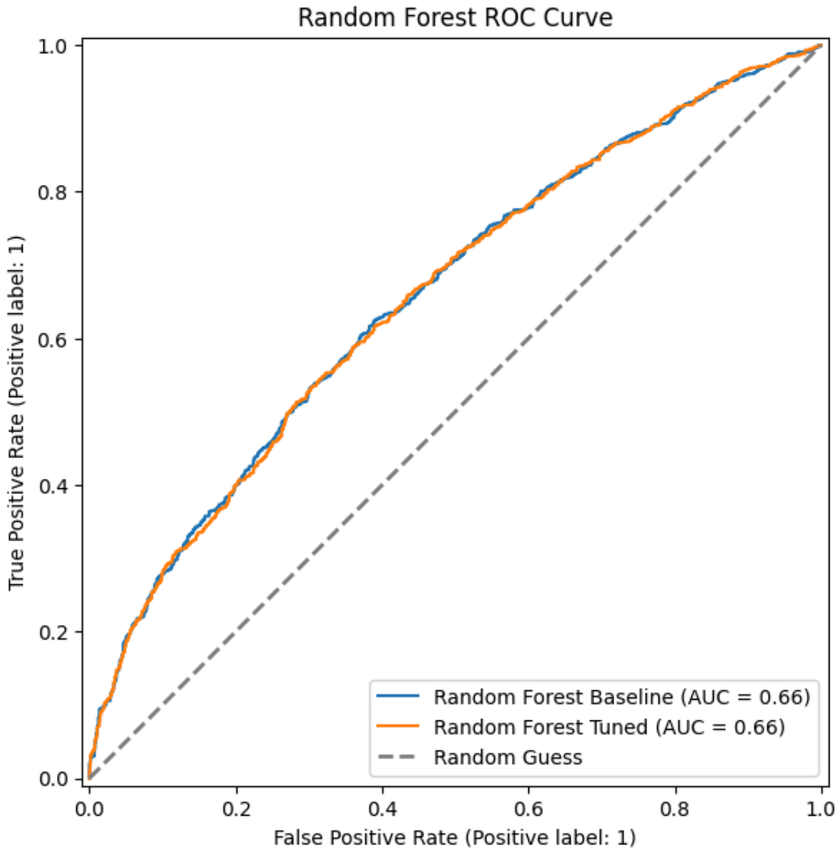
- time_in_hospital ↔ n_medications: 0.449
- n_procedures ↔ n_medications: 0.369
- time_in_hospital ↔ n_lab_procedures: 0.328
- n_lab_procedures ↔ n_medications: 0.272
- time_in_hospital ↔ n_procedures: 0.179

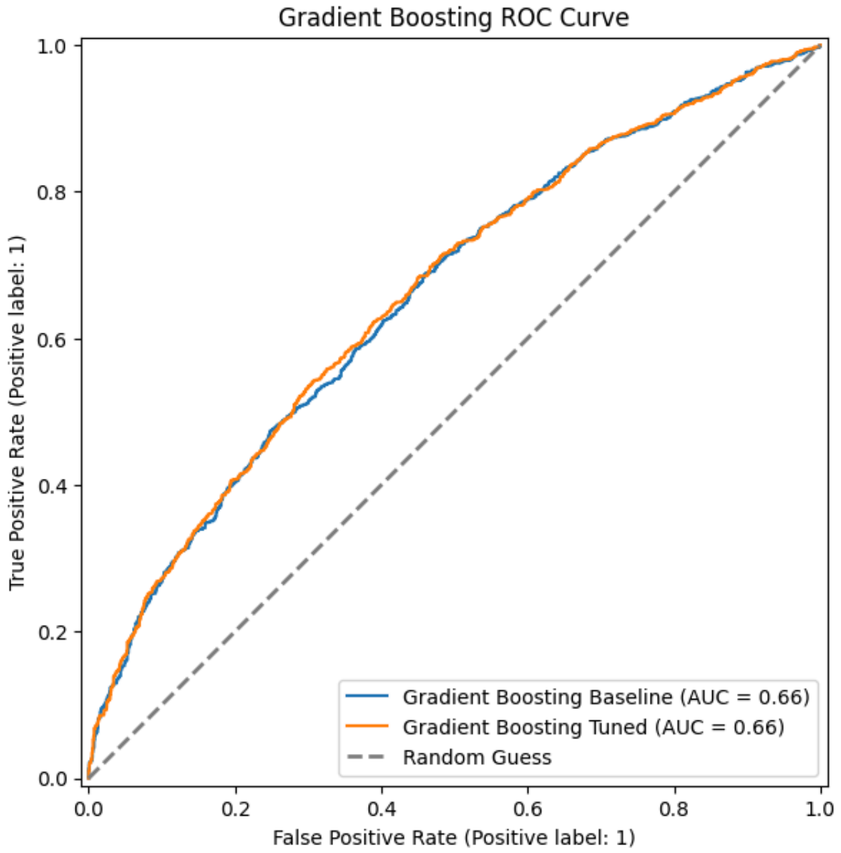
Baseline Models Comparison

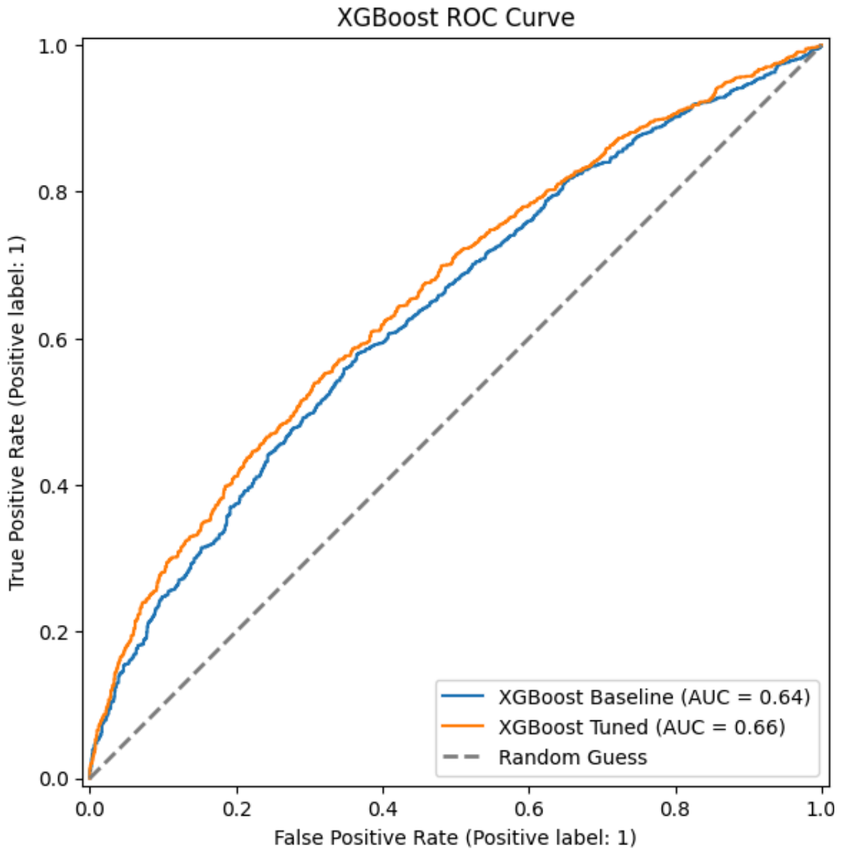


- Naïve Bayes: Highest recall (97%) but poor overall balance → prone to false positives.
- Logistic Regression : Moderate accuracy (~54–56%) with strong F1 (~64%) → more balanced.
- Gradient Boosting , XGBoost,Random Forest: Best trade-off; relatively higher accuracy (~59%) and F1 (~63%).
- Bagging & MLP Neural Network: Lower performance overall (~54% accuracy, ~54–56% F1).

ROC AUC Curve - Baseline VS Fine Tuned Models







AUC Scores: Random Forest = 0.66, Gradient Boosting = 0.66, XGBoost = 0.66 (tuned), showing similar discriminative power.

Tuning Impact: Minimal improvement after tuning → models already near optimal.

Overall: All three models perform moderately better than random guess (AUC > 0.5).

Final Selected Model

Final Model Performance on Test Set:

Model	Accuracy	Precision	Recall	F1	ROC-AUC	TP	TN	FP	FN	Cost
RandomForest_Tuned	55.72	51.74	86.13	64.64	65.67	1012	381	944	163	2574
Random Forest_Baseline	56.00	51.96	84.68	64.40	65.78	995	405	920	180	2720
XGBoost_Tuned	58.28	53.91	77.53	63.60	65.80	911	546	779	264	3419
Gradient Boosting_Baseline	58.88	54.41	77.19	63.83	65.79	907	565	760	268	3440
GradientBoosting_Tuned	58.64	54.23	76.94	63.62	66.10	904	562	763	271	3473
XGBoost_Baseline	57.60	53.63	72.34	61.59	63.69	850	590	735	325	3985

Final Model Selection – Random Forest

Chosen Model: Random Forest (Tuned)

Reason for Selection:

Achieves highest Recall (86.13%), ensuring fewer false negatives (critical in medical/healthcare context).

Balanced Precision (51.74%) and F1-score (64.64%) compared to alternatives.

Competitive ROC-AUC (65.67%), similar to Gradient Boosting/XGBoost.

Lowest Cost (2574) among all models tested.

Web User Interface

ReadmitAI

Home Predict About

AI-Powered Hospital Readmission Prediction

Identify at-risk patients before discharge, optimize follow-up care, and reduce avoidable readmissions.

Try Single Prediction

58.68%	0.635	0.45	25,000+
Model Accuracy	ROC-AUC Score	Cost-Optimal Threshold	Patients Analyzed

ReadmitAI

Home Predict About

Predict Hospital Readmission

Age Group: Select

Time in Hospital:

Lab Procedures:

Procedures:

Medications:

Outpatient Visits:

Inpatient Visits:

Emergency Visits:

Medical Specialty: Select

Diagnosis 1: Select

Diagnosis 2: Select

Diagnosis 3: Select

Readmission Likely

Probability: 40.0%

(Threshold: 40.0%)

Why this prediction?

Prediction Summary

Readmission Prediction: YES

Probability: 40.0% | **Risk Level:** MODERATE

Assessment: This patient is just above the threshold for readmission risk.

How the Model Made This Decision

Baseline Population Risk: 47.0% - This represents the average readmission rate for similar patients in the training data.

Key Contributing Factors:

- **n_lab_procedures:** ⬆️ INCREASES risk by 0.021
  Significantly influential - This factor significantly increases readmission risk
  Clinical Note: Clinical significance varies based on patient context
- **time_in_hospital:** ⬇️ DECREASES risk by 0.021
  Significantly influential - This factor significantly decreases readmission risk
  Clinical Note: Clinical significance varies based on patient context

Clinical Insights:

Clinical Insights:

- • **n_lab_procedures** is a major risk factor for this patient
- • **time_in_hospital** is providing significant protection against readmission
- • The baseline readmission risk for similar patients is moderate (47.0%)

Model Confidence: The model is *20.0% certain* about this prediction (distance from neutral 50%).

Feature Importance

THANK YOU