Predicting Hospital Readmission Risk

Using Explainable Machine Learning on Public Health Data

Binger Yu | Savina Cai | Yansong Jia

COMP 9170 Project Report - December 4, 2025

Problem Statement

The Clinical Challenge

Predicting 30-day readmissions for diabetic patients is difficult.

- Risk is influenced by polypharmacy, comorbidities, and care transitions.
- Relationships between factors are highly non-linear.

Why It Matters

- Improve care transitions and early-stage interventions
- Reduce preventable readmissions
- Avoid HRRP penalties on high readmission rates

Accurate and interpretable predicting model needed

Motivation: The Need for Advanced Prediction

Practice: High readmission rates are costly and harmful:

- Diabetes readmission contributes to rising healthcare cost ($327 billion/year)
- High readmission -> poor outcomes for patients

Research: Factor analysis reveals contributors

Gap Analysis & Our Unique Contribution

Prior Work

- Mostly used simple models (LogReg, basic trees)
- Limited features
- Weak predictive power

What is missing

- Low AUC performance (0.55-0.65)
- Limited Feature Engineering
- Poor interpretability (missing SHAP or LIME)
- fairness rarely evaluated

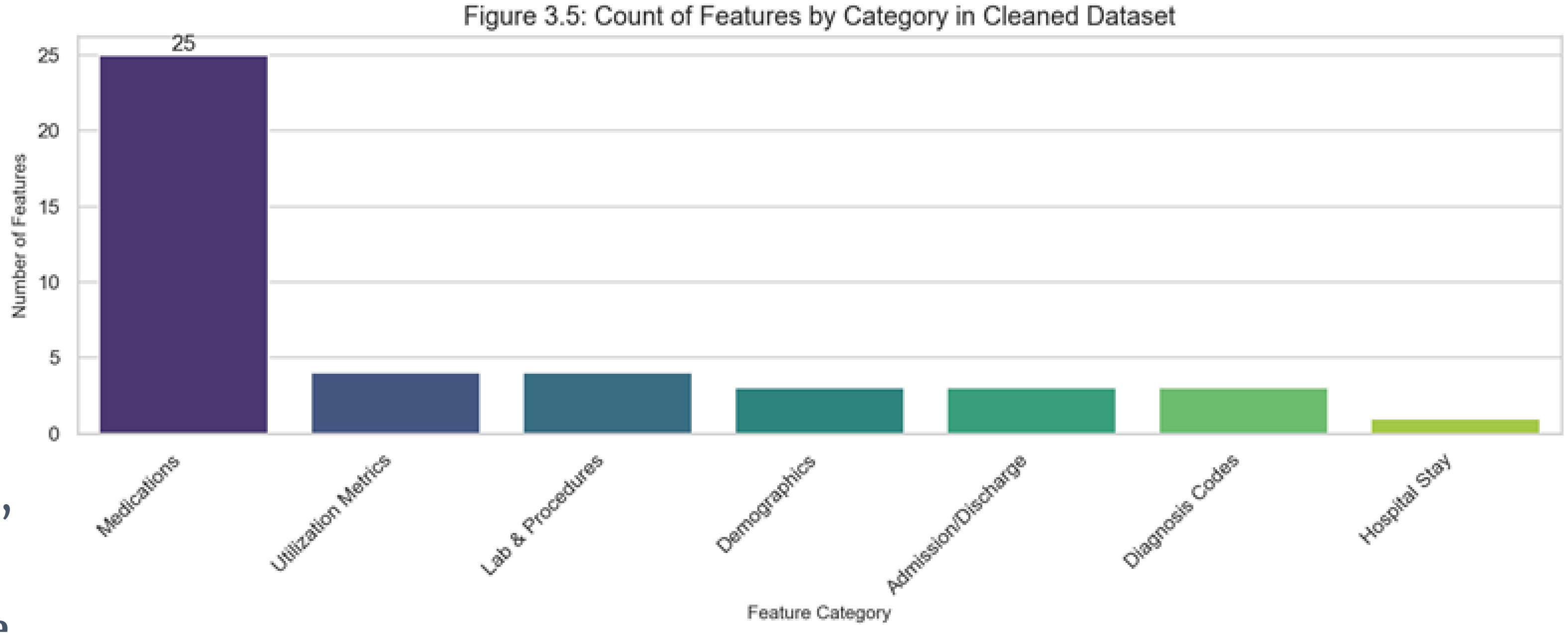
Our Contribution

- Benchmark advanced GBMs (XGBoost, LightGBM)
- Engineer utilization features
- Apply SHAP for explainability
- Conduct a multi-metric fairness audit

Dataset Overview: UCI Diabetes 130-US Hospitals

Key Characteristics

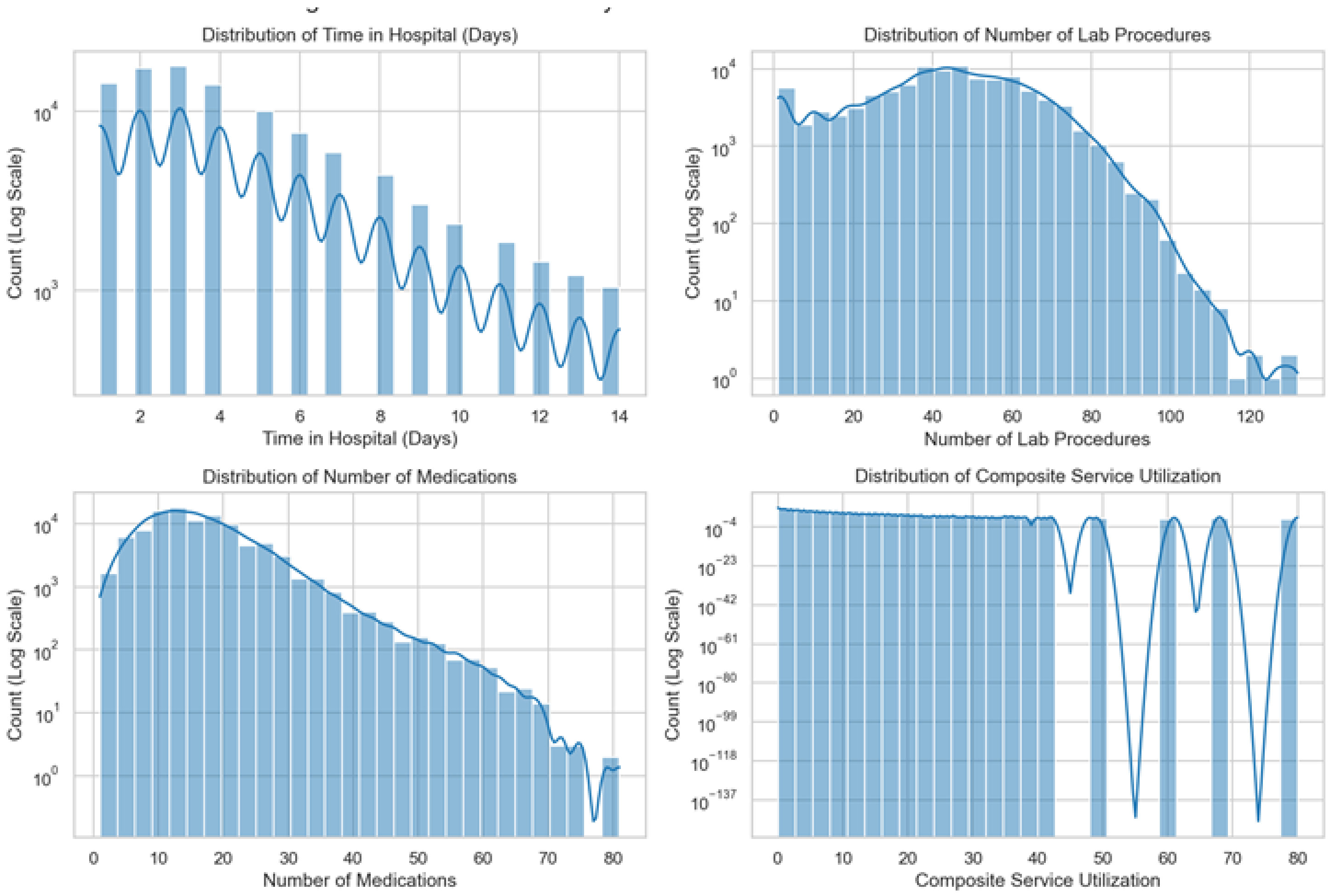
- 101 k+ hospital encounters (1999-2008)
- Data source: UCI Machine Learning Repository (130 US hospitals)
- Binary target: 30-days readmission
- 46 input features: demographics, labs, medication, utilization
- Class imbalance: 11.16% positive readmissions



Exploratory Data Analysis Insights

Trends & Patterns

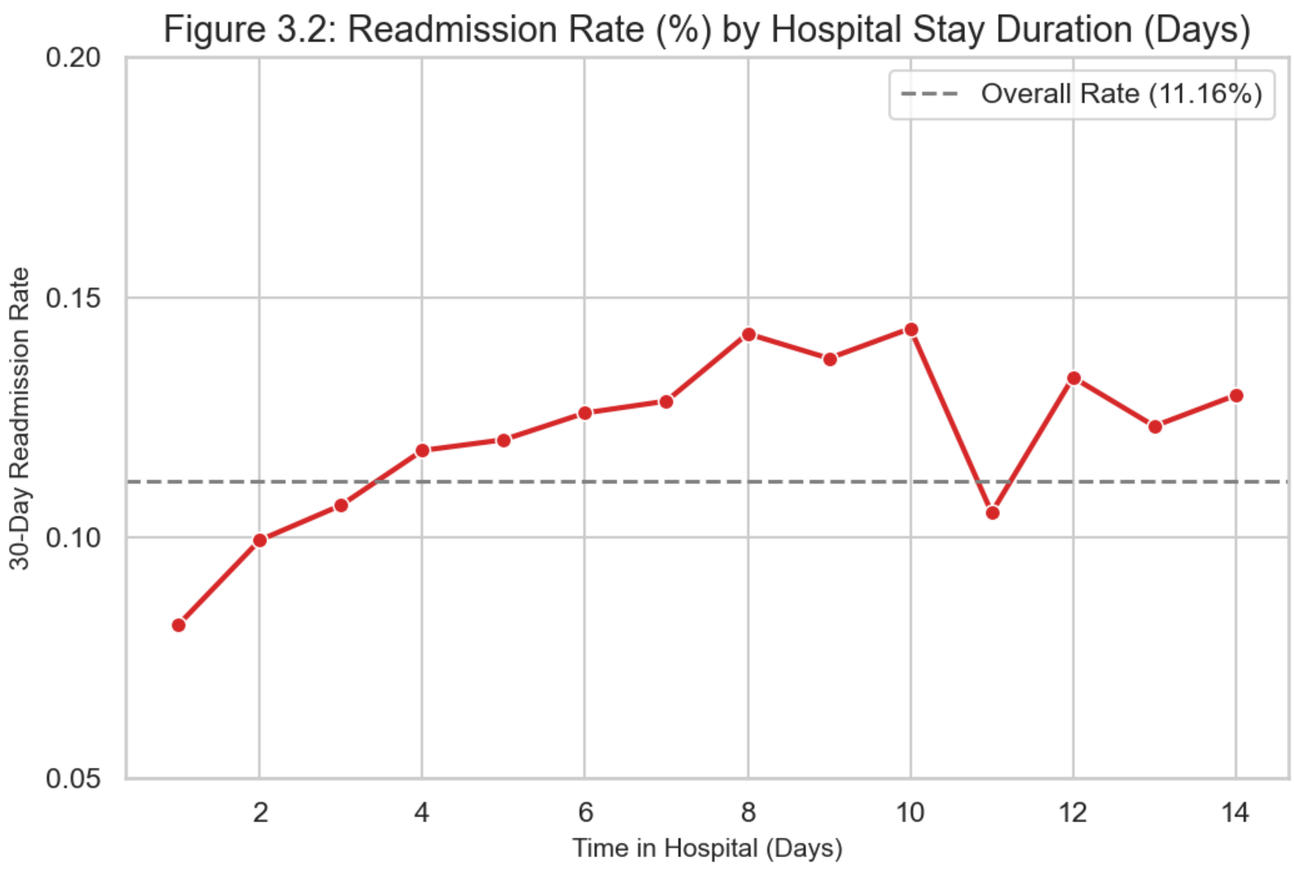
- Right-skewed distributions in meds, labs, and utilization.



Exploratory Data Analysis Insights

Trends & Patterns

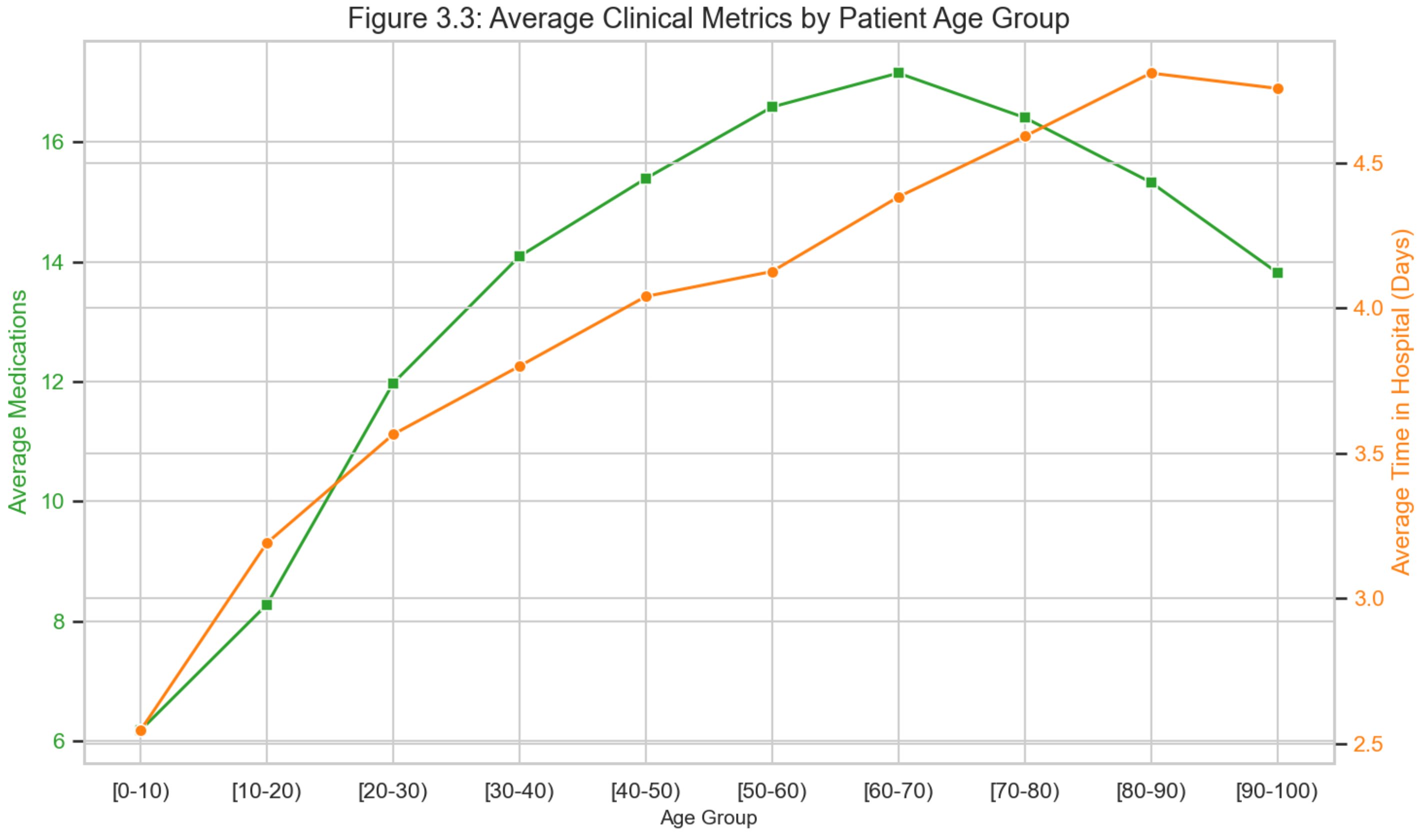
- Right-skewed distributions in meds, labs, and utilization.
- Longer stays (>4 days) -> higher readmission risk



Exploratory Data Analysis Insights

Trends & Patterns

- Right-skewed distributions in meds, labs, and utilization.
- Longer stays (>4 days) -> higher readmission risk
- Older patients (70-80) -> more complex conditions and resource use.



Exploratory Data Analysis Insights

Outliers Identified

- 10.44% -> unusually emergency visits
- 15.44% -> excessive outpatient visits.
- 6.68% -> frequent prior hospitalizations.
- These subgroups indicate high-risk, vulnerable patients.

Challenges Discovered

- Heavy class imbalance (only 11.16% readmitted).
- Many categorical features requiring encoding.
- Right-tail outliers may distort model performance.
- Non-linear relationships -> simple models perform poorly.

Methodology (Approach / Model / Architecture)

Problem & Strategy

- Binary classification task: predict 30-day hospital readmission
- Applied a 5-model comparison strategy to benchmark improvement over simple baselines.

Model Selection

- Baselines : Logistic Regression, Decision Tree (depth=10)
- Advanced Models : Random Forest, XGBoost, LightGBM

Why this Architecture

- Gradient boosting captures non-linear interactions better than linear models.
- Performance gain (~+5% AUC over baseline) shows true pattern learning, not overfitting.
- Using multiple models prevents algorithm selection bias.

Methodology (Pre-processing Pipeline)

Data Split (80/20)

- Only 11% of samples are positive (readmissions).
- Used a stratified split to maintain class balance across train/test sets.)

Numeric Features

- Applied median imputation (robust to outliers).
- Standardized using StandardScaler to ensure fair contribution across features

Categorical Features & Imbalance

- Used constant imputation + OneHotEncoder for categorical variables.
- Applied class weighting (scale_pos_weight = 10) to address severe imbalance - balancing 89% negative vs. 11% positive class distribution.

Methodology (Baseline Used & Design Justification)

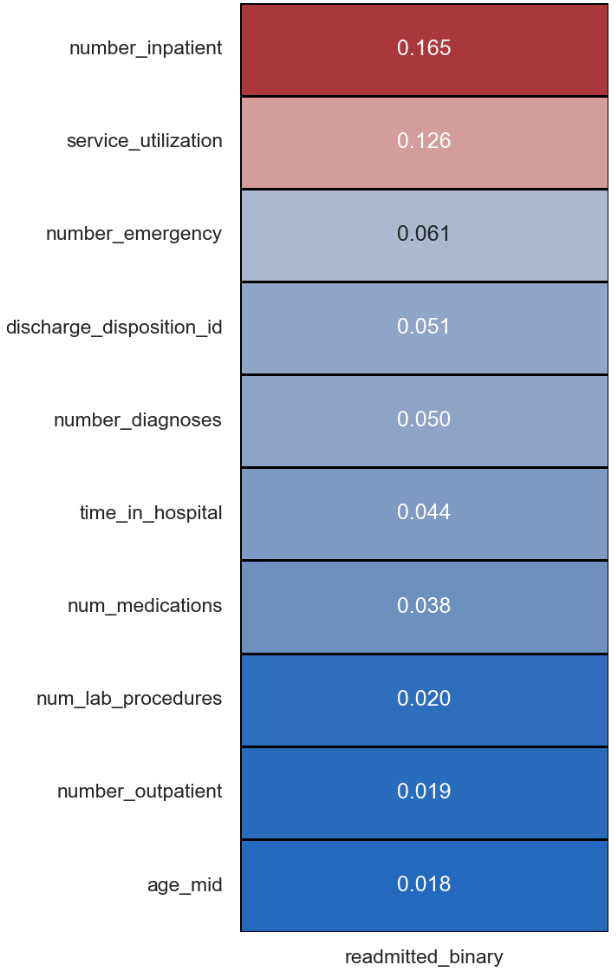
Two Baselines

- Logistic Regression: AUC=0.64 -> establishes minimum acceptable performance.
- Decision Tree (depth=10): AUC=0.65 -> prevents overfitting; aligns with clinical interpretability

Why This Design?

- Class weighting: Reduces cost of missing high-risk readmissions.
- Gradient boosting: Captures non-linear relationships (comorbidities, medication interactions)
- Stratified split: Ensures reliable evaluation on imbalance data.
- 5-model comparison: Highlights true improvements beyond simple baselines.

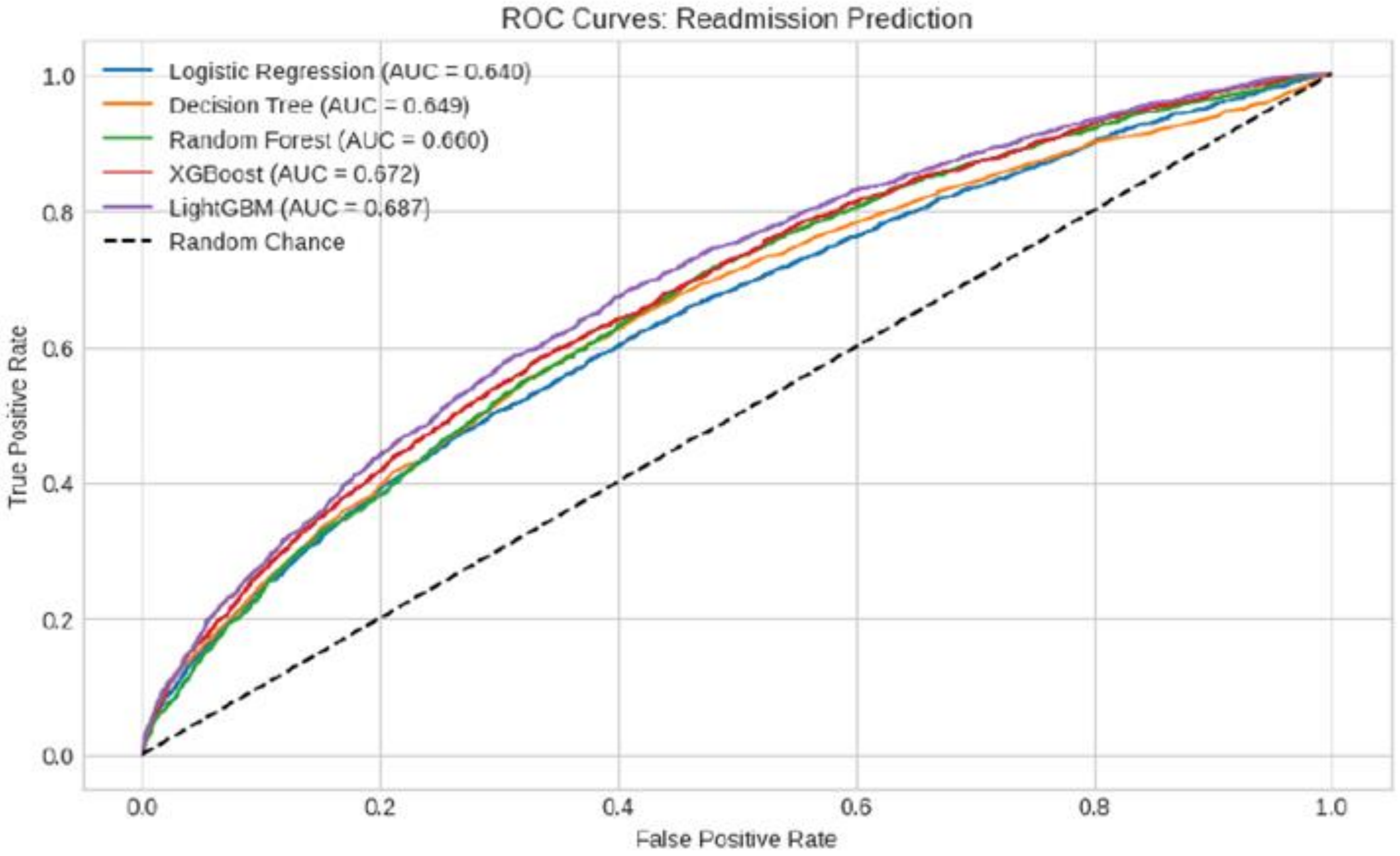
Figure 3.4: Top 10 Feature Correlations with Readmission Target



Key Results - Model Performance Comparison

Model Performance Summary :

Model	Accuracy	AUC-ROC	F1 Score
Logistic Regression (Baseline)	0.64	0.64	0.25
Random Forest	0.68	0.66	0.27
LightGBM (Top Performer)	0.66	0.69	0.28
XGBoost	0.57	0.67	0.26



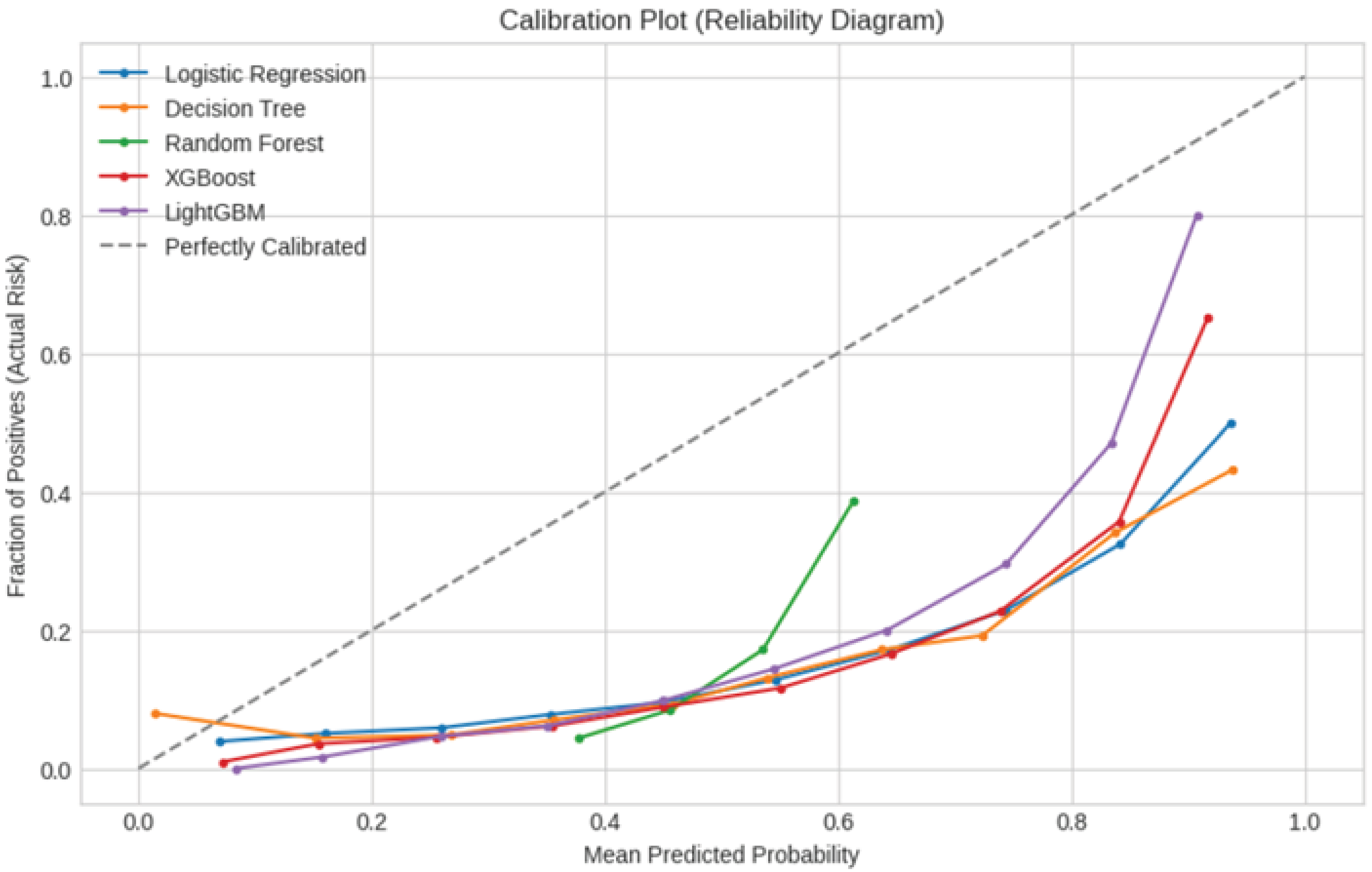
Key Results - Model Performance Comparison

Model Performance Summary :

Model	Accuracy	AUC-ROC	F1 Score
Logistic Regression (Baseline)	0.64	0.64	0.25
Random Forest	0.68	0.66	0.27
LightGBM (Top Performer)	0.66	0.69	0.28
XGBoost	0.57	0.67	0.26

Discussion Insights :

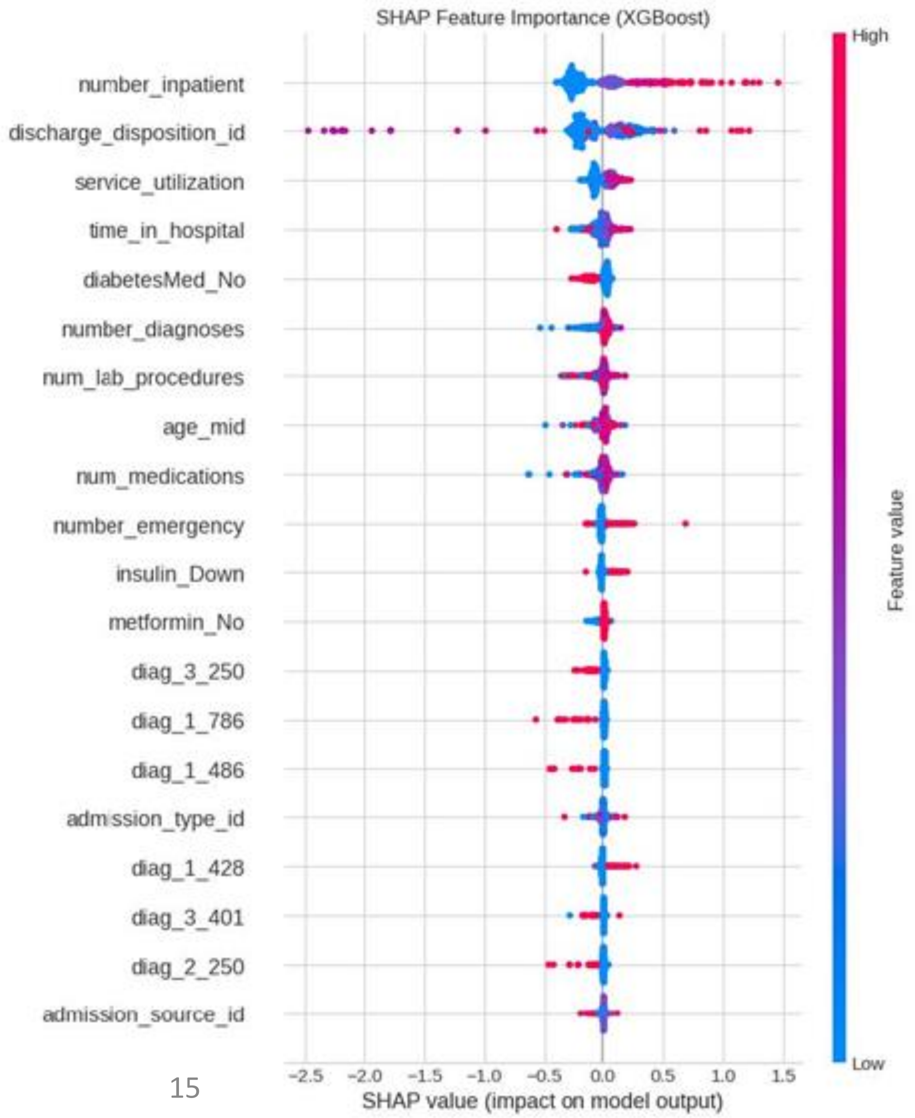
- Gradient boosting models outperform linear baselines
- LightGBM achieves the best AUC (0.69).
- Baselines fail to capture non-linear patterns.



Interpretability - SHAP Feature Importance

Discussion Insights :

- Top Predictor: prior Inpatient visits -> strongest signal for readmission.
- High-risk patterns: long hospital stays, complex medication profiles.
- SHAP identifies actionable clinical features that clinicians can monitor.



Discussion: Key Takeaways & Fairness

What We Accomplished

- Built an explainable machine learning pipeline for predicting 30-day readmissions.
- Identified clinically meaningful risk factors using SHAP.
- Evaluated fairness across demographic groups to uncover potential bias.

Key Takeaways

- LightGBM performed best, capturing complex non-linear relationships.
- Model achieved strong AUC and consistent recall for most groups.
- Fairness gaps exist - certain racial groups showed lower recall.
- Results support targeted interventions and more equitable deployment in real healthcare settings.

Our project delivers both accurate predictions and transparent insights to support better hospital readmission management.

Future Work

Multi-visit Sequential Modeling

Use longitudinal data across multiple visits to predict readmission risks.

External Validation & Generalization

Validate model with diverse, external data for broad applicability.

Intervention Cost Simulation

Simulate financial impact and cost savings of targeted interventions.

References

[1] K. E. Joynt and A. K. Jha, “Thirty-day readmissions—truth and consequences,” New England Journal of medicine, vol. 366, no. 15, pp. 1366–1369, 2012.

[2] American Diabetes Association, “Economic costs of diabetes in the u.s. in 2017,” Diabetes Care,vol. 41, no. 5, pp. 917–928, 2018.

[3] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, “Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records,” BioMed Research International, vol. 2014, p. 81670, 2014.

[4] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, “Predictive risk modelling for early hospital readmission of patients with diabetes in india,” International Journal of Diabetes in Developing Countries, vol. 36, pp. 519–528, Dec 2016.

[5] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, “Risk prediction models for hospital readmission: A systematic review,” JAMA, vol. 306, pp. 1688–1698, Oct 2011.

[6] A. Artetxe, A. Beristain, and M. Gra˜na, “Predictive models for hospital readmission risk: A systematic review of methods,” Computer Methods and Programs in Biomedicine, vol. 164, pp. 49–64, Oct 2018

Q&A + Project Resources

We welcome any questions.

🔗 GitHub Repository:

- https://github.com/bing-er/hospital-readmission-prediction

🔗 Overleaf:

- https://www.overleaf.com/read/xzxhkbxrmydt#310bf6

🔗 Google Drive (Dataset / Report / PPT)

- https://drive.google.com/drive/folders/1ANFkS1HQPx4kzd-wwBV_tNoGyoGo-GfL?usp=drive_link