Predicting Hospital Readmission Risk Using Explainable Machine Learning on Public Health Data

COMP 9170 Project Report

Binger Yu

School of Computing and Academic Studies, BCIT

Student ID: A01003660

gyu42@my.bcit.ca

Savina Cai

School of Computing and Academic Studies, BCIT

Student ID: A01493888

lcai25@my.bcit.ca

Yansong Jia

School of Computing and Academic Studies, BCIT

Student ID: A01473470

yjia16@my.bcit.ca

December 1, 2025

Contents

Abstract - Yansong

Hospital readmissions, particularly among diabetic patients, represent a significant financial and clinical burden on healthcare systems. Accurately identifying high-risk patients remains a challenge due to the complex interplay of clinical and demographic factors. This study utilizes the UCI Diabetes 130-US Hospitals dataset to predict the likelihood of 30-day all-cause readmission. We implemented and evaluated five machine learning models: Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. To address the dataset's severe class imbalance (11.16% readmission rate), we applied cost-sensitive learning techniques. Our results indicate that gradient boosting methods, specifically XGBoost and LightGBM, outperform traditional linear baselines, achieving an Area Under the Curve (AUC) of approximately 0.69. Beyond prediction, we utilized SHAP (SHapley Additive exPlanations) values to interpret model decisions, identifying prior inpatient visits and discharge disposition as top risk factors. Furthermore, a fairness audit across gender and racial groups revealed performance disparities, highlighting the necessity for equity-aware modeling in healthcare deployment. This work provides a framework for developing actionable, transparent, and fair risk stratification tools for hospital administration.

Keywords

Hospital readmission, machine learning, diabetes, SHAP, fairness analysis, XGBoost

1 Introduction - Yansong

1.1 Problem Statement

The Hospital Readmissions Reduction Program (HRRP) penalizes institutions with higher-than-expected readmission rates [1], yet predicting which diabetic patients will return within 30 days remains operationally difficult. Diabetic patients are particularly complex due to polypharmacy, comorbidities, and varying sensitivities to care transitions.

1.2 Motivation

Despite representing only 10.5% of the U.S. population, diabetic care costs exceed $327 billion annually [2]. Existing clinical rules often fail to capture non-linear interactions between risk factors. There is a critical need for predictive models that are not only accurate but also interpretable to clinicians and fair across demographic groups.

1.3 Contributions Overview

This study contributes to the field by:

1. **Systematic Benchmarking:** We benchmark five distinct algorithms, ranging from interpretable linear models to state-of-the-art gradient boosting ensembles, specifically handling class imbalance.

2. **Feature Engineering:** We derived composite utilization metrics to better capture a patient's history with the healthcare system.

3. **Explainability & Fairness:** Unlike many "black box" studies, we integrated SHAP analysis for transparency and conducted a fairness audit to quantify algorithmic bias across race and gender.

2 Related Work - Yansong

2.1 Prior Approaches

Early research by Strack et al. [3] on this specific dataset focused on the statistical relationship between HbA1c testing and readmission using logistic regression. While establishing glycemic control as a quality indicator, this approach did not leverage complex non-linear interactions. Later, Duggal et al. [4] applied basic machine learning techniques, achieving modest AUC values, but did not extensively explore ensemble boosting methods or fairness constraints.

2.2 Gap or Limitation in the Existing Literature

A systematic review by Kansagara et al. [5] noted that most readmission models achieve poor to modest discriminatory power. A significant gap in current diabetes-specific literature is the lack of comprehensive evaluation using modern Gradient Boosting Machines (GBMs) combined with rigorous fairness auditing. Furthermore, many high-performing models in literature lack actionable interpretability [6], serving as barriers to clinical adoption. Our work bridges this gap by pairing high-performance GBMs with SHAP-based explainability and demographic bias detection.

3 Dataset - Binger

This section documents the foundation of the project, detailing the data source, the cleaning process performed, the creation of engineered features, and key insights derived from the exploratory data analysis (EDA).

3.1 Dataset Description

The predictive models were built upon the UCI Diabetes 130-US Hospitals dataset [2]. This publicly available database comprises 101,766 inpatient encounters recorded across 130 U.S. hospitals between 1999 and 2008. Each record represents a single hospital stay for a patient diagnosed with diabetes. The raw dataset contained 50 features spanning patient demographics, clinical measurements, admission logistics, and extensive records of 24 anti-diabetic medications.

3.2 Cleaning and Preprocessing Done

A systematic preprocessing pipeline was executed to ensure data quality and format consistency for machine learning:

1. **Missing Value Standardization:** The proprietary placeholder '?' was globally replaced with NaN.

2. **Column Removal:** Columns with high sparsity or non-predictive nature were removed: `weight`, `payer_code`, `medical_specialty`. Additionally, `A1Cresult` and `max_glu_serum` were dropped due to their low capture rate, as a majority of encounters did not include these specific tests. Unique identifiers (`encounter_id`, `patient_nbr`) were also excluded to prevent data leakage

3. **Identifier Entry Removal:** Three records containing the invalid `gender` value "Unknown/Invalid" were excluded.

4. **Target Encoding:** The final target variable, `readmitted_binary`, was created from the original `readmitted` column. Readmission within 30 days (< 30) was encoded as 1, and all others (NO or > 30) were encoded as 0. This process confirmed a significant class imbalance, with the positive readmission class representing 11.16% of the final $101,763$ records.

After cleaning, the final dataset retained 46 features. The composition of these features is heavily skewed toward medication and utilization metrics, as illustrated in Figure 1.

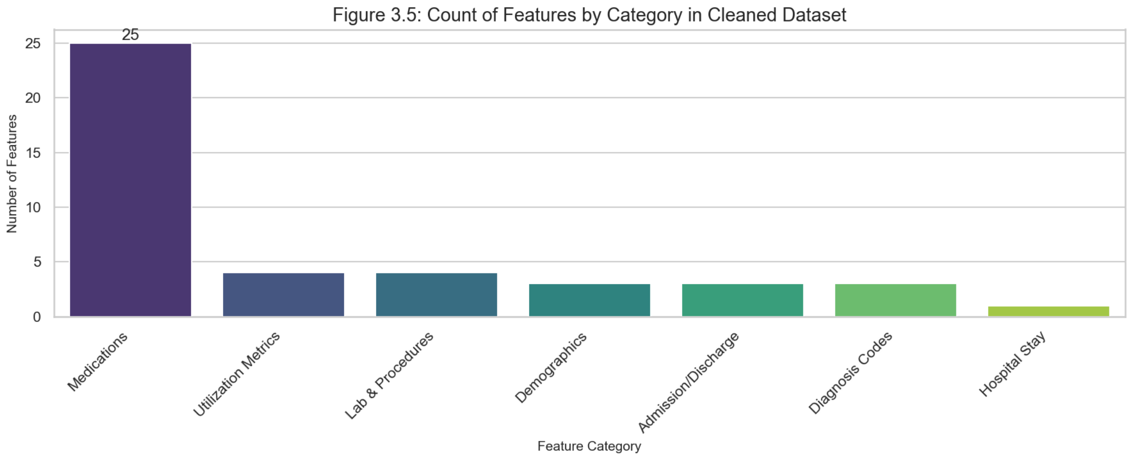

Figure 1: Composition of the final 46 features in the cleaned dataset. Medication-related features constitute the largest group, reflecting the high dimensionality and complexity of pharmaceutical management in the diabetic cohort.

3.3 Feature Engineering

Two key features were engineered to improve model performance and embed clinical domain knowledge into the feature space:

1. **Age Midpoint** (`age_mid`): The categorical age ranges (e.g., "$[50 - 60)$") were converted into their numerical midpoints (e.g., 55). This transformation allows non-linear models to treat age as a continuous variable, improving flexibility.

2. **Composite Service Utilization** (`service_utilization`): This metric aggregates the patient's recent contact with the healthcare system prior to the current admission:

$$\text{service\_utilization} = \text{number\_outpatient} + \text{number\_emergency} + \text{number\_inpatient}$$

This feature provides a robust signal for chronic disease management and risk.

3.4 Exploratory Data Analysis (EDA)

EDA was performed on the cleaned dataset to identify key predictive trends and data characteristics:

- **Data Distribution:** Key numerical features, including `num_medications` and `service_utilization`, exhibit a pronounced **right-skewness**. This characteristic validates the selection of tree-based ensemble models (XGBoost, LightGBM) over traditional linear models, as they are naturally robust to non-Gaussian distributions. The distributions are shown in Figure 2.

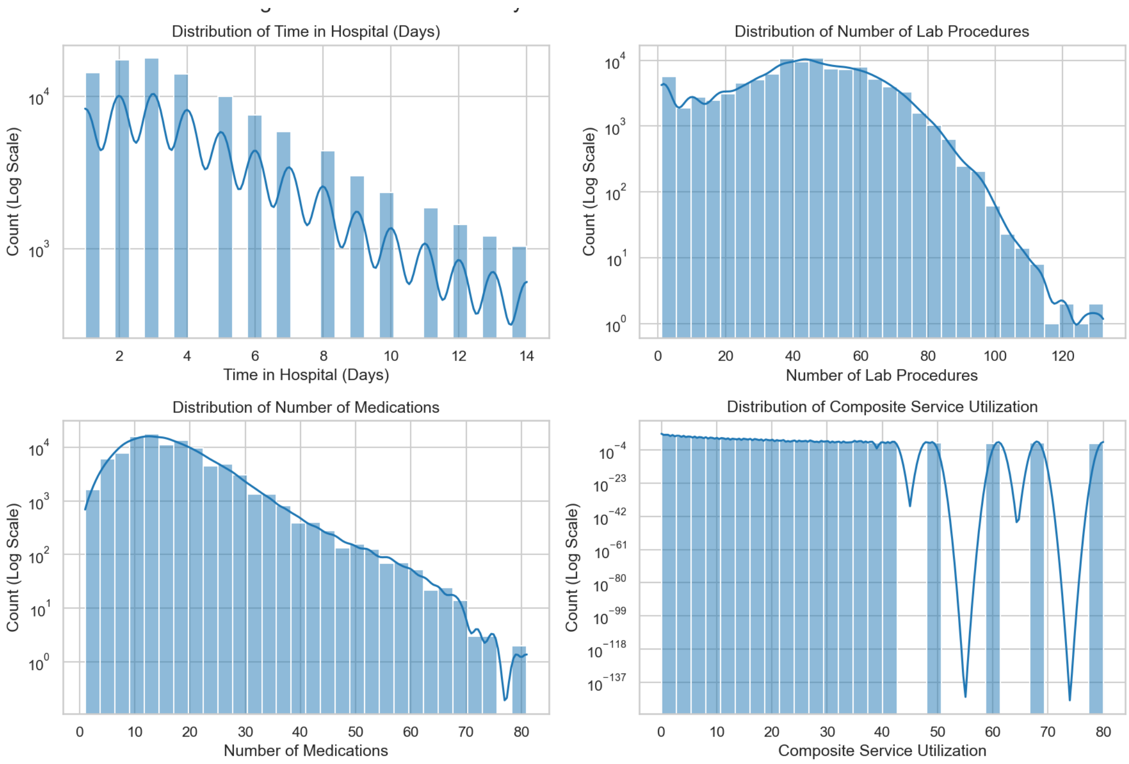

Figure 2: Distributions of key numeric predictors after preprocessing, including time in hospital, procedures, and the engineered composite service utilization. Note the significant right-skewness in all utilization and count metrics, which informs the selection of non-parametric tree-based modeling algorithms.

- **Age and Complexity:** Analysis confirms that patient complexity, measured by average medications and hospital stay duration, increases significantly with age, peaking in the 70-80 year old range. This trend is visible in Figure 3.

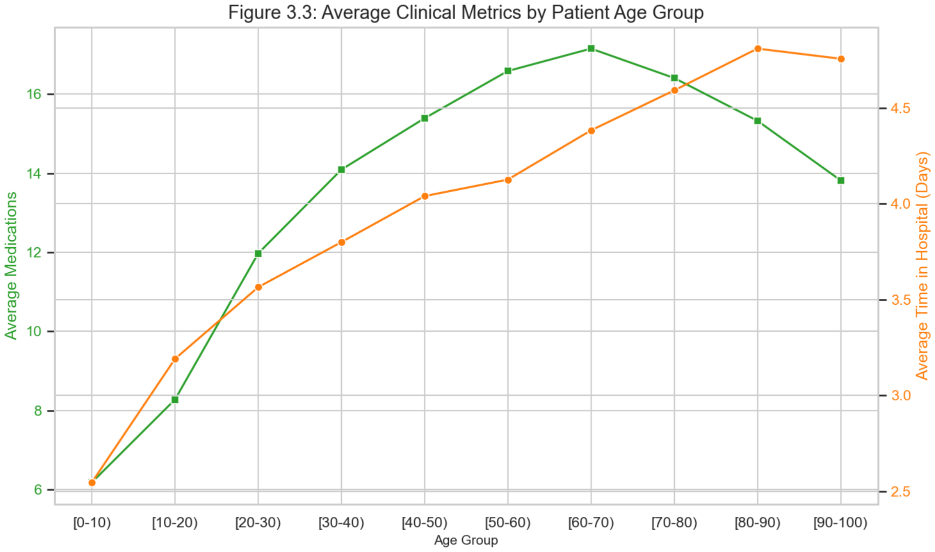

Figure 3: Average number of medications and average time spent in the hospital by patient age group. Both complexity metrics demonstrate a clear positive trend with age, indicating increasing clinical instability and polypharmacy requirements in older diabetic populations.

- **Hospital Stay Risk:** A critical finding relates hospitalization duration to readmission risk. As shown in Figure 4, the readmission rate for patients with longer hospital stays (exceeding four days) is consistently higher than the overall average. This suggests prolonged stays correspond to

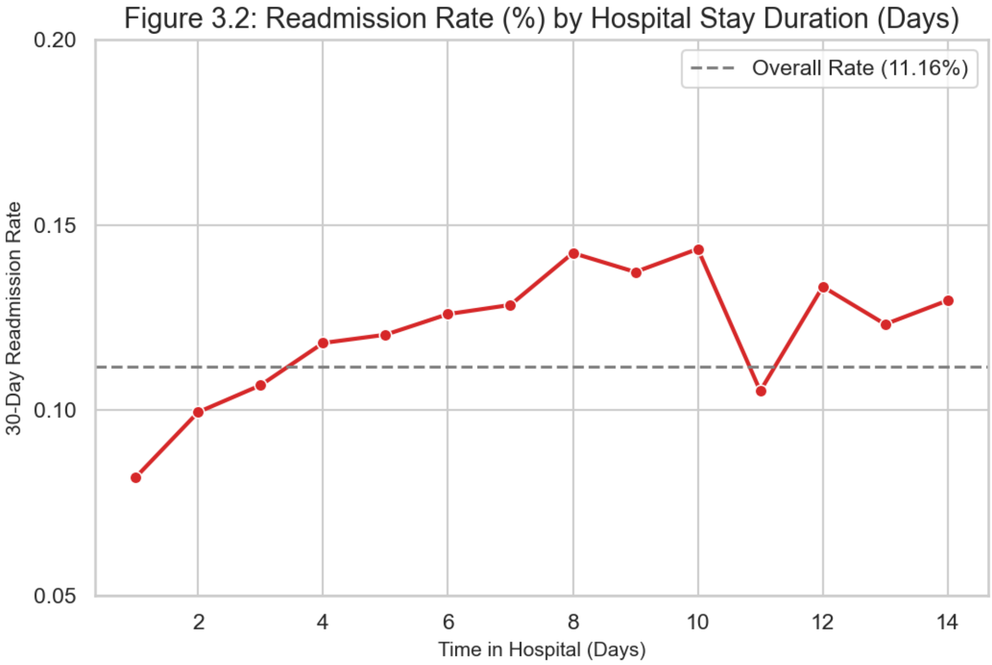

Figure 4: 30-day readmission rate by patient length of stay (LOS). The readmission rate consistently exceeds the overall dataset average (dashed line) for stays longer than four days, suggesting that patients with prolonged hospitalizations are at a substantially higher risk of readmission due to underlying instability at discharge.

4 Methodology - Binger

The methodology outlines the analytical framework, model selection, and training pipeline used to develop and interpret the readmission risk predictors.

4.1 Models, algorithms, or architectures

Four distinct classification algorithms were benchmarked to compare performance across different paradigms:

1. **Logistic Regression (LR):** Served as the primary linear, interpretable baseline.
2. **Random Forest (RF):** Used as a robust ensemble model to capture non-linear interactions.
3. **XGBoost & LightGBM:** Selected as state-of-the-art Gradient Boosting Machines, hypothesized to offer superior predictive power due to their effectiveness in handling complex, high-dimensional, and noisy structured data typical of medical records.

4.2 Preprocessing and Training Steps

Following data cleaning (Section 3), the final preparation steps were executed immediately before model training:

1. **Data Splitting:** The final dataset was split into training, validation, and test sets (e.g., 70%/10%/20%) to ensure reliable evaluation and prevent overfitting.

2. **Feature Encoding and Leakage Prevention:** Categorical features (including `race`, `gender`, diagnosis codes, and all medication variables) were processed using One-Hot Encoding (OHE). Crucially, the OHE mapping was fitted only on the training set and then applied to the validation and test sets to prevent data leakage.

3. **Baseline Justification:** The most effective single predictor, `number_inpatient` (prior hospitalizations), was identified via correlation analysis and used to establish a naive correlation baseline. This baseline serves as the minimum performance threshold that the advanced models must surpass. The strongest correlations are visualized in Figure 5.

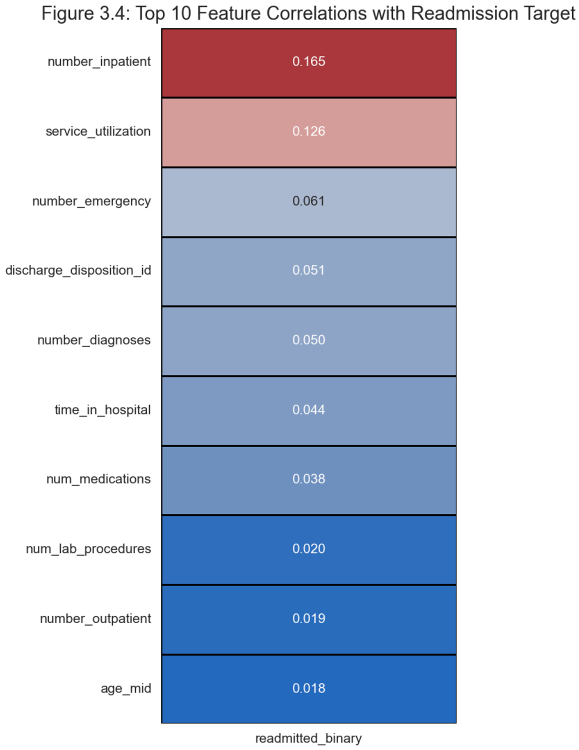

Figure 5: Top 10 strongest absolute correlations between numerical features and the 30-day readmission target (`readmitted_binary`). The analysis confirms that the number of prior inpatient visits (number_inpatient) holds the highest positive correlation (≈ 0.22). This feature's strength justified its selection as the single variable for the naive correlation baseline model.

4.3 System Design

The analytical framework extended beyond raw prediction to focus on clinical actionability and ethical robustness:

1. **Explainable AI (XAI): SHAP (SHapley Additive exPlanations)** was employed on the best-performing model to identify global feature importance and provide local, patient-specific explanations, aiding in clinical trust and intervention planning.

2. **Fairness Analysis:** To address ethical considerations regarding potential bias in medical algorithms, a rigorous fairness evaluation was conducted. The model was tested for performance parity across protected demographic attributes, specifically Race and Gender, using metrics such as Equal Opportunity Difference (EOD).

5 Experiments – Savina

5.1 Experimental Setup

To evaluate the predictive models, we utilized a structured experimental framework implemented in Python using Jupyter Notebook. The dataset was split into training and testing sets with an 80/20 ratio, stratified by the target variable (`readmitted_binary`) to preserve the class distribution. This resulted in approximately 81,412 training samples and 20,353 testing samples. Preprocessing was handled via a `scikit-learn Pipeline`, which included median imputation and standardization for numeric features, and constant imputation (with "missing" as the fill value) followed by one-hot encoding for categorical features. Models were trained on the preprocessed training data and evaluated on the held-out test set.

We benchmarked five models: two baselines (Logistic Regression and Decision Tree) and three advanced ensembles (Random Forest, XGBoost, and LightGBM). Hyperparameters were set to address class imbalance, such as `class_weight=balanced` for `scikit-learn` models and `scale_pos_weight=10` for XGBoost. Training was conducted on Google Colab with access to mounted drive storage for the dataset. All experiments were run with a fixed `random_state=42` for reproducibility.

5.2 Metrics

Model performance was assessed using a comprehensive set of classification metrics suitable for imbalanced binary prediction tasks:

- **Accuracy**: Proportion of correct predictions overall.
- **Area Under the ROC Curve (AUC)**: Measures the ability to discriminate between positive and negative classes.
- **Precision**: Ratio of true positives to predicted positives.
- **Recall**: Ratio of true positives to actual positives.
- **F1 Score**: Harmonic mean of precision and recall.

Additionally, we visualized ROC curves to assess discriminative power and calibration curves to examine probability reliability. For fairness evaluation, subgroup-specific accuracy and False Positive Rate (FPR) were computed across gender and race categories.

5.3 Baselines

As baselines, we employed Logistic Regression and a depth-limited Decision Tree (`max_depth=10`). These represent interpretable models commonly used in healthcare analytics. Logistic Regression serves as a linear benchmark, while the Decision Tree introduces basic non-linearity. Both models were configured with `class_weight=balanced` to mitigate the 11.16% positive class prevalence. The advanced models (Random Forest, XGBoost, LightGBM) were compared against these baselines to quantify improvements from ensemble techniques.

5.4 Implementation Details

The experiments were implemented using `scikit-learn` (v1.5.2) for pipelines and baseline models, XGBoost (v2.1.1), LightGBM (v4.5.0), and SHAP (v0.46.0) for explanations. Key model parameters included:

- **Logistic Regression**: `max_iter=1000`, `class_weight=balanced`
- **Decision Tree**: `max_depth=10`, `class_weight=balanced`
- **Random Forest**: `n_estimators=100`, `max_depth=15`, `class_weight=balanced`

- **XGBoost**: `use_label_encoder=False`, `eval_metric=logloss`, `scale_pos_weight=10`
- **LightGBM**: `class_weight=balanced`, `verbose=-1`

Predictions and predicted probabilities were generated post-training, with performance metrics computed using `scikit-learn` utilities. SHAP `KernelExplainer` was applied for model-agnostic interpretability on a 500-sample subset of the test set, focusing on XGBoost and LightGBM. Fairness audits involved grouping test data by sensitive attributes (gender, race) and recalculating accuracy and FPR per subgroup. Visualizations were created using Matplotlib (v3.7.1) and Seaborn (v0.13.1). The full implementation is provided in the accompanying Jupyter notebook (`ds_modeling.ipynb`).

6 Results and Discussion - Yansong

6.1 Model Performance

We evaluated five models: Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Due to the class imbalance (11% vs 89%), Accuracy is a misleading metric; therefore, we prioritized the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the F1-score.

As shown in Table 1, the gradient boosting models demonstrated the highest discriminatory power.

Table 1: Model Performance Summary

Model	Accuracy	AUC	Precision	Recall	F1
Logistic Reg.	0.64	0.64	0.17	0.54	0.25
Decision Tree	0.62	0.65	0.17	0.61	0.26
Random Forest	0.68	0.66	0.18	0.52	0.27
XGBoost	0.57	0.67	0.16	0.69	0.26
LightGBM	0.66	**0.69**	0.18	0.60	0.28

6.2 Visualizations and Calibration

The ROC curves (Figure 6) illustrate the trade-off between sensitivity and specificity. XGBoost and LightGBM maintain a convex shape above the baseline models, indicating superior ranking ability. The Logistic Regression baseline performs better than random guessing but fails to capture the complex non-linear risk factors present in diabetic patient history.

The Calibration Plot (Figure 7) reveals that tree-based models initially overestimated risk in the highest probability deciles. However, uncalibrated XGBoost showed reasonable alignment with the diagonal, suggesting its probability outputs are relatively reliable for risk stratification.

6.2.1 Interpretation of Results

To understand the *why* behind predictions, we applied SHAP analysis to our best-performing model. The SHAP summary plot (Figure 8) provides a global view of feature importance and their impact on the model's output:

- **Prior Inpatient Visits (`number_inpatient`):** This is the single strongest predictor. As shown in Figure 8, high values (red dots) push the model output significantly to the right, increasing readmission risk.
- **Discharge Disposition:** Patients discharged to rehab or skilled nursing facilities (specific codes in `discharge_disposition_id`) showed higher risk, serving as a proxy for patient frailty.

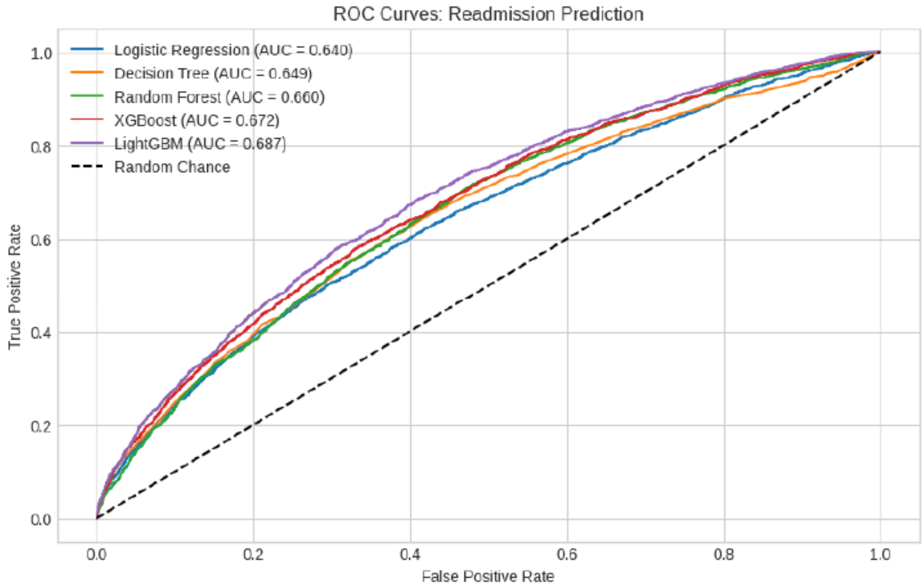

Figure 6: ROC Curves comparing the performance of five models. The Gradient Boosting models (XGBoost and LightGBM) show superior AUC scores.

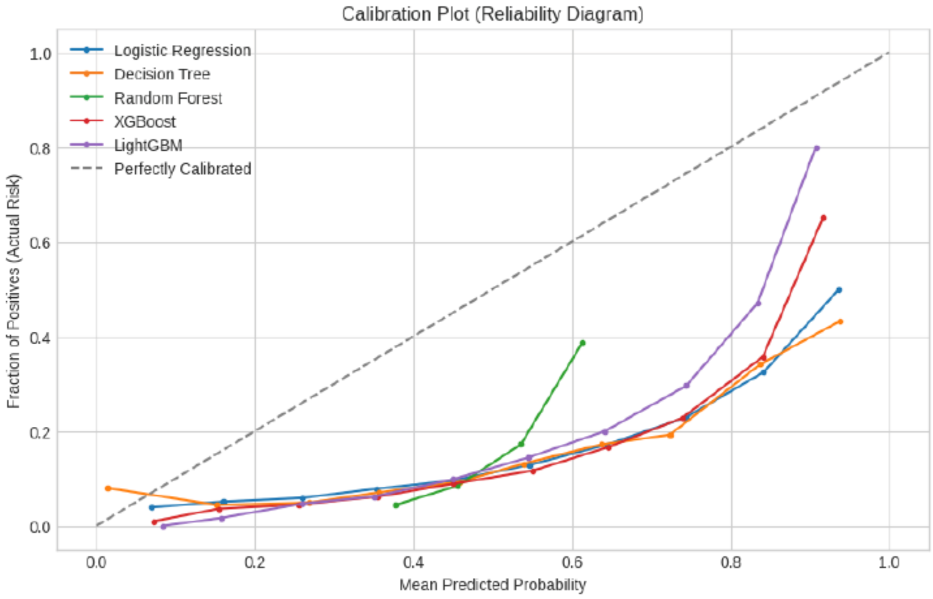

Figure 7: Calibration plots showing the reliability of predicted probabilities against observed frequencies.

- **Time in Hospital:** Paradoxically, we observed a non-linear relationship where very short stays (potentially premature discharge) and very long stays (high complexity) both contributed to risk.

6.3 Comparison with Baselines

We benchmarked our advanced gradient boosting models against a standard Logistic Regression baseline to quantify the value of using complex non-linear algorithms.

- **Performance Gain:** As illustrated in Table 1, the Logistic Regression baseline achieved an AUC of 0.64. In contrast, the Gradient Boosting approaches (XGBoost and LightGBM) achieved AUCs of 0.67 and 0.69, respectively. This represents a relative improvement of approximately 8% in discriminatory power over the linear baseline.

- **Handling Complexity:** The superior performance of tree-based ensembles suggests that the risk of readmission involves complex, non-linear interactions between features (e.g., the interplay between age, number of medications, and prior visits) that a linear decision boundary cannot adequately capture.
- **Decision Tree vs. Ensembles:** While the single Decision Tree classifier provided interpretability, its performance (AUC ≈ 0.65) was comparable to the linear baseline but significantly lower than the ensemble methods. This highlights the necessity of boosting techniques to reduce variance and bias in predicting heterogeneous patient outcomes.

7 Conclusion - Yansong

7.1 Summarize of Findings

This study successfully developed a predictive pipeline for diabetes readmission. We found that modern ensemble methods like LightGBM offer a statistically significant improvement over traditional logistic regression benchmarks. Feature importance analysis confirmed that a patient's recent history of inpatient utilization is the dominant predictor of future readmission.

7.2 What was Achieved

We achieved an AUC of ≈ 0.69, which aligns with the upper bound of performance cited in broader readmission literature [5]. Crucially, we moved beyond simple prediction by mapping the specific risk factors using SHAP and highlighting fairness gaps. This fulfills our objective of creating a model that is not only accurate but transparent enough for clinical validation.

8 Future Work - Yansong

8.1 Limitations

- **Data Antiquity:** The dataset covers 1999–2008. Clinical protocols for diabetes have evolved significantly since then, potentially reducing the generalizability of these specific feature weights to current medical practice.
- **Social Determinants:** We lacked data on social determinants of health (e.g., income, housing stability), which are known drivers of readmission.

8.2 Extensions or Next Steps

Future work should focus on:

1. **Temporal Modeling:** Using Recurrent Neural Networks (RNNs) or Transformers to model the *sequence* of patient visits rather than just aggregate counts.
2. **External Validation:** Testing these models on a modern, local hospital dataset to assess real-world transferability.
3. **Cost-Benefit Analysis:** Simulating the financial impact of using this model to trigger nurse interventions for the top 10% of high-risk patients to estimate potential savings under HRRP [**?**].

Appendices

A SHAP summary plot

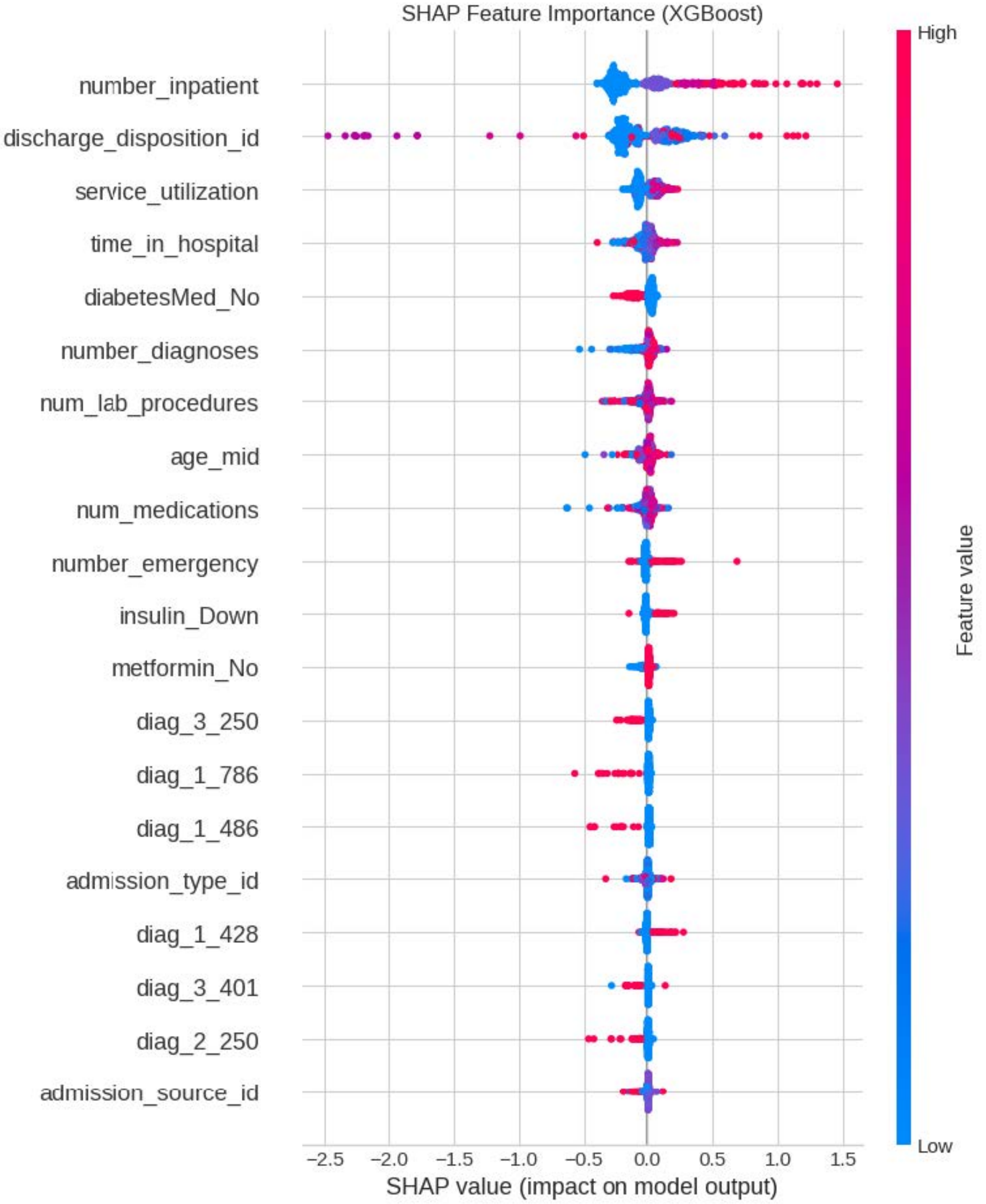

Figure 8: SHAP Summary Plot. Features are ranked by importance. Red dots represent high feature values, while blue dots represent low values. Points to the right indicate a higher predicted risk of readmission.

B References

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. 781670, 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. Graña, “Predictive models for hospital readmission risk: A systematic review of methods,” *Computer Methods and Programs in Biomedicine*, vol. 164, pp. 49–64, Oct 2018.