Explainable AI for Predicting Hospital Readmissions in Diabetic Patients: An End-to-End Case Study

Adrián López-Lanchares Echezarreta

December 30, 2025

Abstract

Hospital readmissions among diabetic patients represent a significant cost burden and a quality-of-care challenge for healthcare institutions. This project develops a machine learning pipeline to predict readmission within 30 days using the UCI Diabetes Dataset. We implement and evaluate four distinct architectures: Multi-Layer Perceptron (MLP), Logistic Regression, Random Forest, and XGBoost. While Random Forest achieved the highest accuracy (89%), it failed to identify the minority class (Recall 0.00). XGBoost was selected as the best performing model (ROC-AUC 0.68) for its ability to balance sensitivity and specificity. We apply Explainable AI techniques, including Permutation Importance and SHAP, to uncover risk drivers. We rigorously evaluate these explanations using Data Randomization and ROAD faithfulness tests. Finally, we demonstrate actionable utility by retraining a lightweight model on the top 5% of features without performance loss, proving that interpretable, parsimonious models can effectively support clinical decision-making.

Data and Code Availability

The dataset used in this study is the **Diabetes 130-US hospitals for years 1999-2008**, publicly available on Kaggle: https://www.kaggle.com/datasets/brandao/diabetes. This dataset is based on the original research paper: https://pubmed.ncbi.nlm.nih.gov/24804245/

The complete source code, including the Jupyter Notebook with data preprocessing, model training, and XAI evaluations, is hosted on GitHub: https://github.com/ICAI-IMAT-XAI/final-project-adrianlopezlanchares.

1 Introduction

1.1 Problem Statement

Diabetes mellitus requires ongoing medical care and patient self-management to prevent acute complications. A major indicator of failure in this care continuum is early hospital readmission. The objective of this study is to predict whether a diabetic patient will be readmitted to the hospital within 30 days of discharge.

1.2 Motivation and Stakeholders

Reducing readmissions is a priority for two stakeholders:

1. **Hospital Administrators:** To reduce financial penalties associated with high readmission rates and optimize resource allocation.
2. **Clinical Staff:** To identify high-risk patients before discharge and target them with specific interventions (e.g., home health visits, medication reconciliation).

However, a "black box" model is insufficient for clinical adoption. Stakeholders must trust that the model relies on medically relevant factors rather than spurious correlations. Therefore, this project emphasizes Explainable AI (XAI) to validate model behavior and derive domain insights.

2 Data and Methods

2.1 Dataset and Preprocessing

We utilize the **UCI Diabetic Data Set**, containing clinical care, demographics, and laboratory data for diabetic patients. The target variable `readmitted` was binarized: patients readmitted in < 30 days (Class 1) versus those readmitted > 30 days or not at all (Class 0).

Preprocessing steps included:

- Removal of unique identifiers (e.g., `encounter_id`) and administrative artifacts.
- Handling duplicates to prevent data leakage.
- One-Hot Encoding for categorical variables.
- Splitting data into Train, Validation, and Test sets.
- Addressing severe class imbalance (approx. 9:1 ratio) using class weighting (`pos_weight`) in cost-sensitive learning algorithms.

2.2 Modeling Strategy

We implemented four models to establish baselines and explore non-linear capabilities:

1. **Logistic Regression (Linear):** A transparent baseline.
2. **Multi-Layer Perceptron (MLP):** A neural network approach implemented in PyTorch.
3. **Random Forest:** An ensemble bagging method.
4. **XGBoost:** A gradient boosting method tuned via GridSearchCV for optimal hyperparameters.

3 Results: Model Performance

We evaluated models using Accuracy, Precision, Recall (Sensitivity), F1-Score, and ROC-AUC. Given the medical context, Recall for the minority class (readmissions) is critical, as missing a high-risk patient is more costly than a false alarm.

Table 1: Model Performance Metrics on Test Set

Model	**Accuracy**	**Prec (1)**	**Recall (1)**	**F1 (1)**	**ROC-AUC**
MLP	0.58	0.16	0.64	0.26	0.62
Logistic Regression	0.71	0.19	0.46	0.27	0.67
Random Forest	**0.89**	1.00	0.00	0.00	0.65
XGBoost (Tuned)	0.64	0.18	**0.59**	**0.27**	**0.67**

As shown in Table 1, the Random Forest model achieved high accuracy (89%) but failed completely on the objective, yielding a Recall of 0.00 for readmissions. This highlights the danger of using Accuracy on imbalanced datasets.

The XGBoost model provided the best balance, achieving a Recall of 0.59 while maintaining a comparable ROC-AUC to the linear model. Consequently, XGBoost was selected as the main model for explainability analysis.

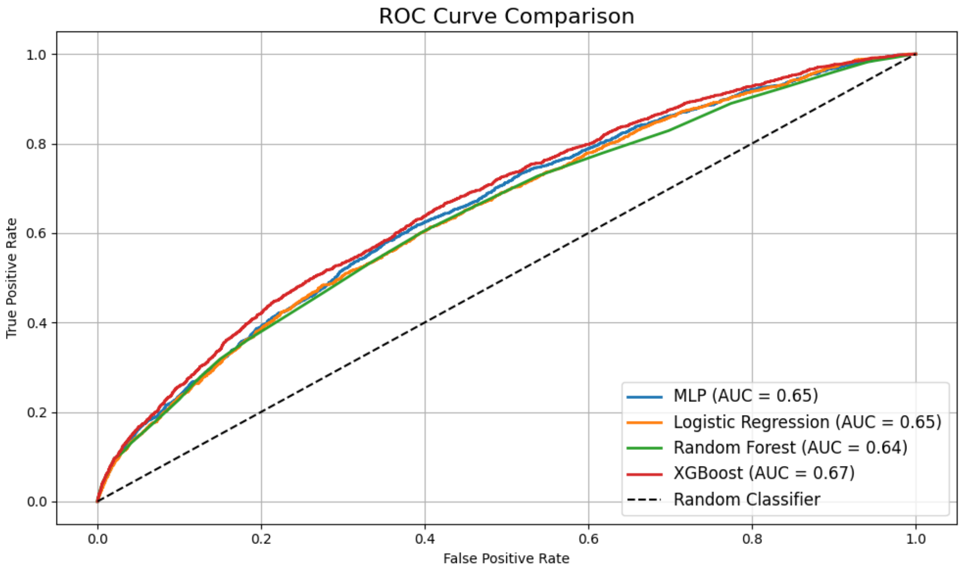
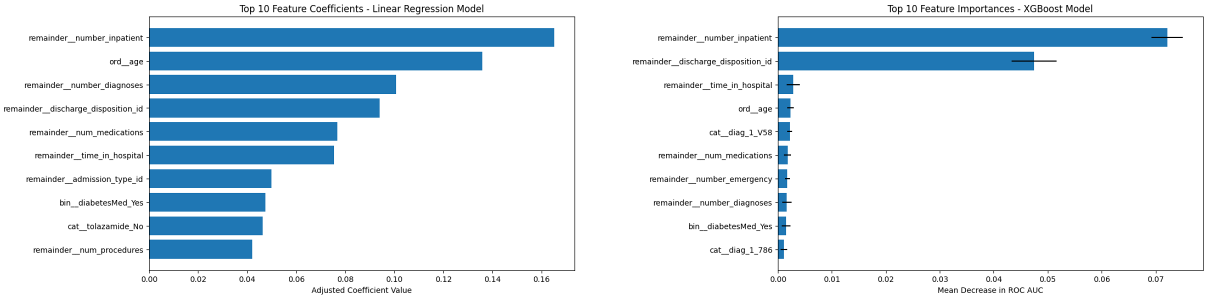


Figure 1: ROC Curve comparison. XGBoost and Logistic Regression show superior separability compared to MLP and Random Forest.

4 Explainability Analysis

4.1 Global Explanations: Feature Importance

To understand the model's global behavior, we compared the top features identified by the linear baseline (coefficients) versus the non-linear XGBoost (Permutation Importance).



(a) Logistic Regression Coefficients

(b) XGB Permutation Importance

Figure 2: Comparison of Global Feature Importance. XGBoost relies heavily on `number_inpatient` and `discharge_disposition`, whereas the Linear model relies on a wider spread of features.

Both models agree that prior inpatient visits (`number_inpatient`) is a dominant predictor of readmission risk. This aligns with medical intuition: patients who frequent the hospital are often sicker or have chronic management issues.

4.2 Local Explanations: SHAP

We utilized SHAP (SHapley Additive exPlanations) to explain individual predictions. The waterfall plot below illustrates a specific high-risk instance.

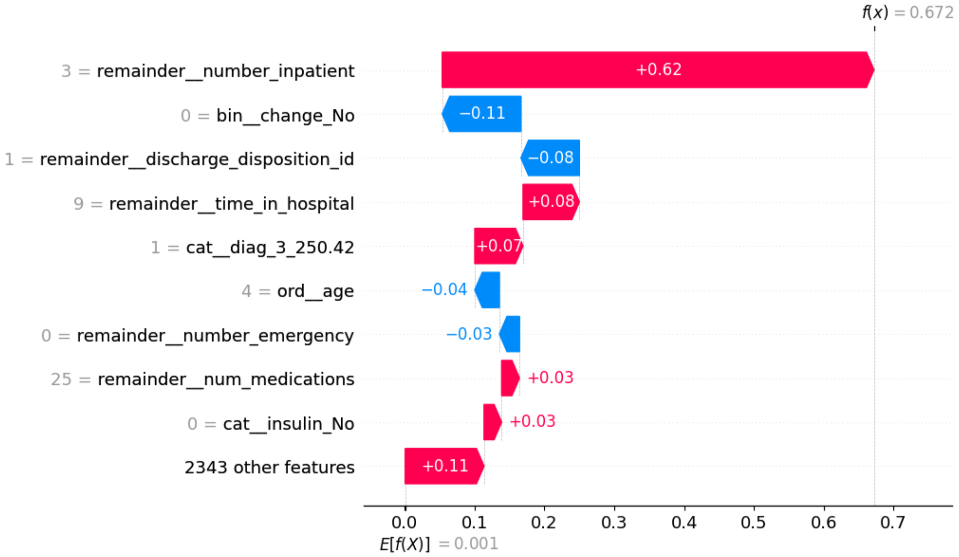


Figure 3: Local explanation for a specific patient predicted as high risk. The high number of prior inpatient visits pushes the risk score significantly higher.

5 Evaluation and Critique of XAI

To ensure our explanations are trustworthy, we performed two rigorous sanity checks.

5.1 Faithfulness: ROAD Test

The ROAD (Remove And Debias) test evaluates if the features identified as "important" are actually used by the model for prediction. We iteratively replaced the most important features with their mean values and observed the degradation in model performance.

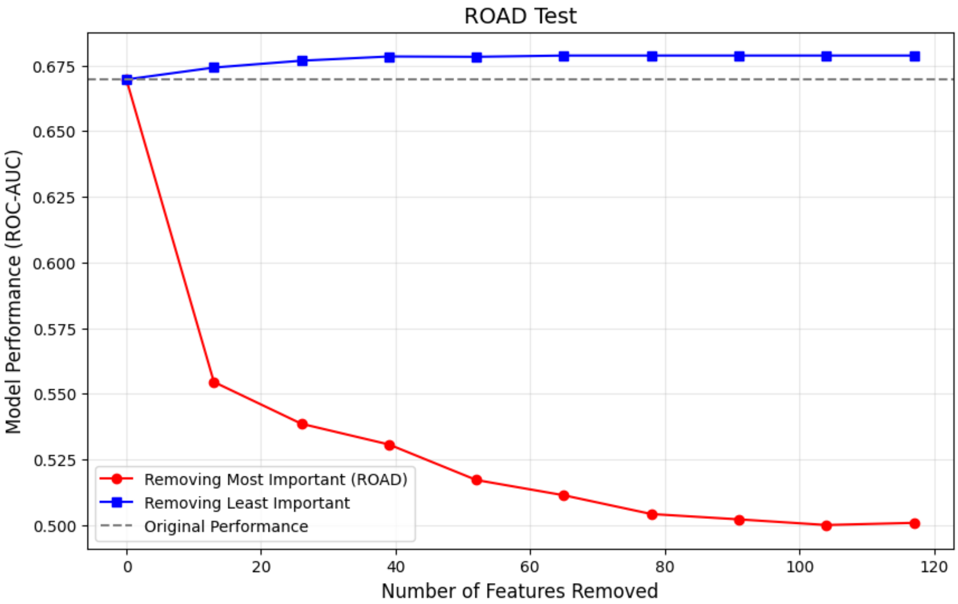


Figure 4: Faithfulness evaluation. Removing only top 5% features causes a rapid drop in AUC, confirming the explanation identifies the true drivers of the model. Number of features is high, as many were one-hot encoded.

Result: The initial AUC of 0.6697 dropped to 0.5009 (random guessing) after masking the top features. This confirms the explanations are highly faithful to the model's logic.

5.2 Sanity Check: Data Randomization

We compared feature importance on the real dataset versus a dataset with randomized labels.

- **Spearman Correlation:** 0.0140
- **Result:** PASS.

The near-zero correlation indicates that the model is learning relationships specific to the target variable, not just picking up on data artifacts or noise.

6 Actions and Insights

6.1 Model Improvement: Feature Selection

Using the insights from Permutation Importance, we identified that the vast majority of the one-hot encoded features contributed minimal information. We hypothesized that a simpler model could achieve similar performance.

We retrained the XGBoost model using only the **top 5% (approx. 117)** features.

Table 2: Performance Comparison: Full vs. Simplified Model

Metric	**Full XGBoost (All Features)**	**Simplified XGBoost (Top 5%)**
Test ROC-AUC	0.67	**0.68**
Recall (Class 1)	0.59	0.59

Action: The simplified model maintained (and slightly improved) performance. We recommend deploying the simplified model to reduce computational cost and data monitoring requirements.

6.2 Domain Insights

The XAI analysis highlighted specific risk factors for clinicians to monitor:

1. **Prior Utilization:** Patients with high `number_inpatient` or `number_emergency` are critical risks.

2. **Discharge Disposition:** Where the patient goes after discharge (e.g., home vs. skilled nursing facility) heavily influences readmission.

3. **Age:** Older age groups showed consistently higher SHAP values for risk.

7 Discussion and Conclusion

7.1 Limitations and Confounding Factors

While XAI provides transparency, the model's predictive ceiling (ROC-AUC ≈ 0.68) suggests the presence of unobserved confounders.

First, the dataset lacks explicit socioeconomic determinants of health. A patient's financial status significantly impacts their ability to afford medication and adhere to post-discharge care instructions, yet this variable is not directly captured.

Second, systemic operational factors, such as hospital bed capacity and staffing levels, often influence discharge decisions independent of patient physiology. A premature discharge due to bed shortages creates a risk that the model cannot "see."

Finally, the problem of fragmented care implies label noise: a patient readmitted to a *different* hospital system would be recorded as a "success" (Class 0) in this dataset, confusing the model during training. These unmeasured external factors likely explain the difficulty in achieving higher precision.

7.2 Conclusion

This study successfully developed a predictive model for diabetic readmissions and utilized XAI to validate its safety and relevance. By confirming the model focuses on medical history (inpatient visits) rather than artifacts, we provide a trustworthy tool for hospital administrators. The ability to prune 95% of features without performance loss demonstrates the power of XAI not just for explaining, but for optimizing Machine Learning pipelines.