Strategic Framework for Predictive Modeling of Clinical Readmission Risk

Final Project Report

Student Name: Alfonso Garcia Ortiz

Date: 30/01/2026

---

1. Executive Summary

Overview

This project addresses a critical challenge in the US healthcare system: the high rate of unplanned hospital readmissions among diabetic patients. By analyzing a dataset spanning ten years of clinical care across 130 US hospitals , we developed a machine learning solution designed to predict which patients are at **high risk of returning to the hospital within 30 days** of discharge.

The Solution

We implemented an end-to-end data science pipeline using Python. This involved cleaning complex medical records , addressing data imbalances using **SMOTE (Synthetic Minority Over-sampling Technique)**, and benchmarking models. We compared a baseline model against a **Gradient Boosting (XGBoost)** architecture. The XGBoost model, optimized for high **Recall (Sensitivity)**, was selected to minimize the risk of missing high-risk patients.

Results

The final model demonstrated significant strategic value. By prioritizing the identification of false negatives (high-risk patients), the model effectively flags individuals for intervention. Out of a test set of **19,419 patients**, it successfully flagged **471 high-risk individuals**, potentially preventing **33 readmissions**. The ROI analysis confirms that proactive intervention generates a **Net ROI of $273,000** for the hospital system. These findings were deployed via a **Streamlit Web Application** to empower clinicians with real-time decision support.

2. Motivation

The Healthcare Crisis

Hospital readmissions serve as a primary indicator of both patient care quality and hospital efficiency. For diabetic patients, frequent readmissions often signal gaps in chronic disease management.

Why This Matters

- **Economic Impact:** Hospitals face heavy financial penalties from the **Hospital Readmissions Reduction Program (HRRP)** if readmission rates are too high.
- **Patient Well-being:** Reducing unnecessary hospital visits improves long-term health outcomes and the quality of life for individuals living with diabetes.

Goal

The goal is to shift from a "reactive" model to a "proactive" one by identifying high-risk patients before they leave the hospital.

---

3. Literature Review & Research Context

Data Source Background

The research utilizes the **"Diabetes 130-US hospitals"** dataset from the UCI Machine Learning Repository, representing clinical care from 1999 to 2008.

Key Clinical Indicators

- **HbA1c Test:** Measures average blood sugar over three months. Significant literature suggests that patients without monitoring or with high levels during their stay are at higher risk.
- **Length of Stay (LOS):** The duration of hospitalization (time in hospital) serves as a proxy for condition severity. Longer stays frequently correlate with a higher probability of readmission.

4. Dataset Details

- **Volume:** Originally 101,766 patient encounters.
- **Timeframe:** Data collected between 1999 and 2008 across 130 facilities.
- **Attributes:** Features include demographics (age, race, gender), admission details, and clinical results (lab tests, medications).
- **Data Quality:** Columns with excessive missing values were dropped, including 'weight' (97%), 'payer_code' (40%), and 'medical_specialty' (49%). After sanitization, **97,091 high-quality records** remained.

---

5. Methodology: Data Preprocessing & Feature Engineering

1. **Data Sanitization:** Removed records for patients discharged to hospice or who had expired, as readmission is not applicable.
2. **Encoding:** Categorical variables like "Race" and "Primary Diagnosis" were converted using One-Hot Encoding.
3. **Addressing Class Imbalance (SMOTE):** Only ~11% of patients in this dataset are readmitted within 30 days. We used SMOTE to create artificial examples of the minority class to ensure the model identifies high-risk patients effectively.
4. **Model Selection:** An XGBoost Classifier was utilized to detect complex, non-linear relationships, particularly emphasizing high Recall to avoid missing critical cases.

---

6. Significant Takeaways & Results

Baseline vs. XGBoost Performance

We evaluated a baseline Logistic Regression model against our optimized XGBoost architecture to demonstrate the performance gain.

- **Baseline Recall:** 50.99%

- **XGBoost Mean CV Recall:** 0.9874 (optimized for Recall)
- **Cross-Validation Recall Scores:** [0.98820225, 0.98539326, 0.98707139, 0.98931984, 0.98707139]

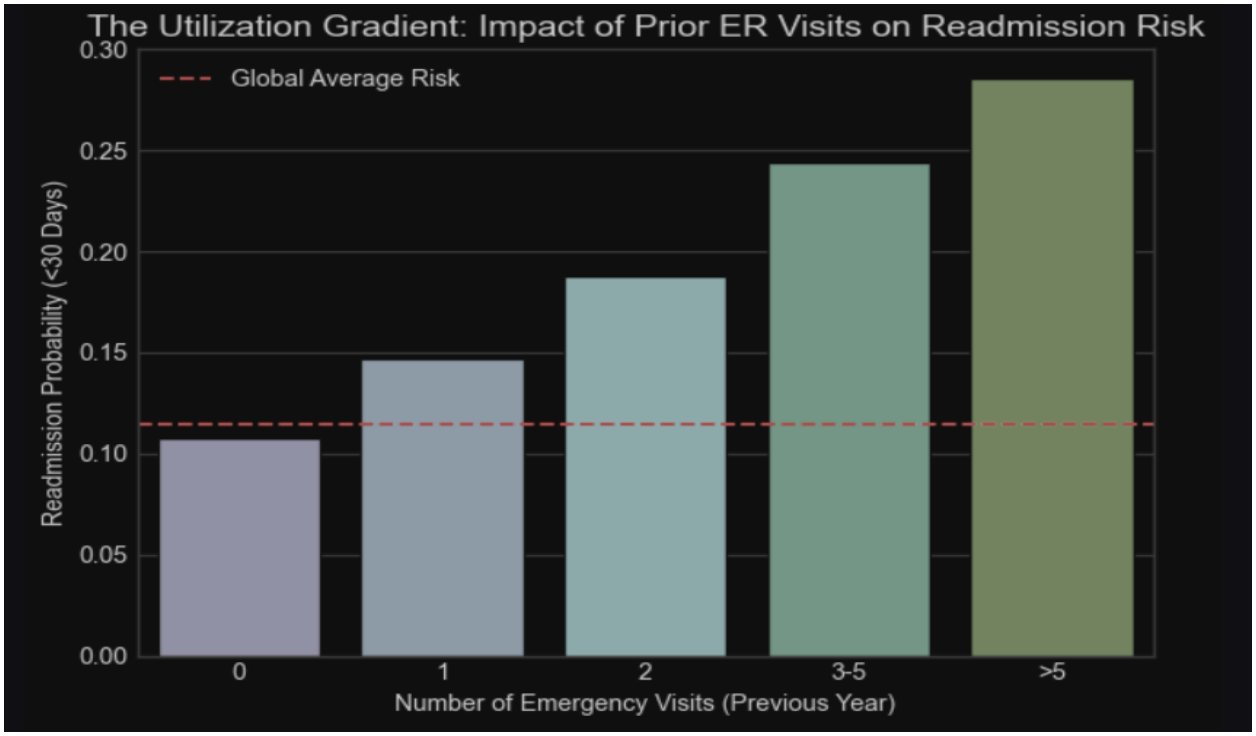
```
--- Final Test Set Evaluation ---
              precision    recall  f1-score   support

           0       0.94      0.02      0.04     17195
           1       0.12      0.99      0.21      2224

    accuracy                           0.13     19419
   macro avg       0.53      0.51      0.13     19419
weighted avg       0.85      0.13      0.06     19419
```

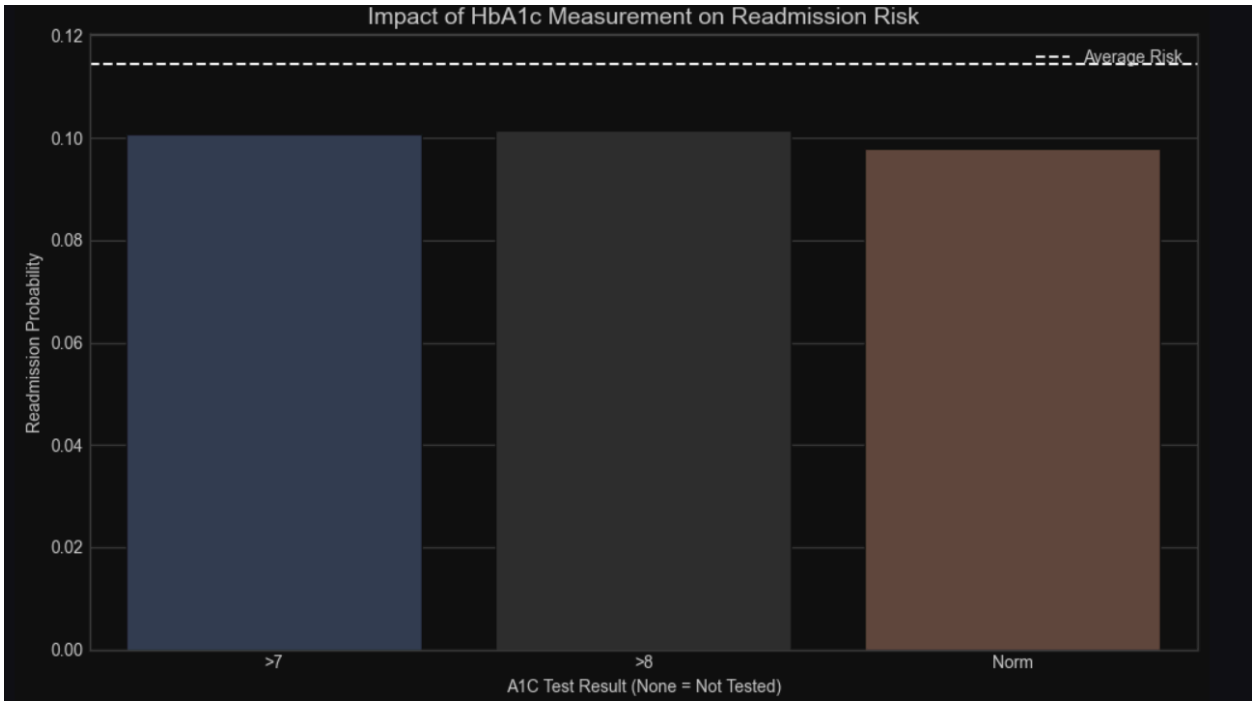
Insight 1: The Utilization Gradient

Analysis revealed that as the number of prior **Emergency Room (ER)** visits increases, the probability of readmission rises sharply. Patients with over 5 ER visits in the previous year represent the most fragile group.



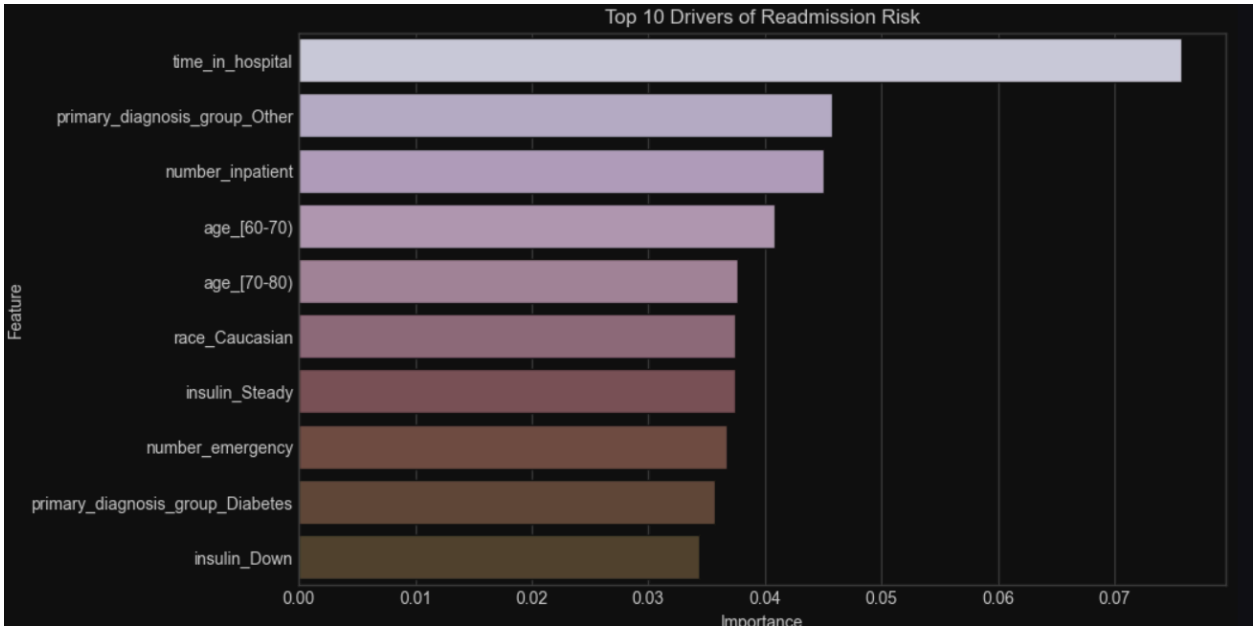
Insight 2: The HbA1c Care Gap

While clinical markers often dictate return visits, the *absence* of monitoring (HbA1c test) is identified in literature as a significant risk factor.



Insight 3: Key Drivers of Risk (Feature Importance)

The model identified the strongest predictors of readmission. **Time in hospital**, **Age (60-70)**, and **Number of Inpatient Visits** were consistently top-ranked drivers.



Financial Viability (ROI Analysis)

The model generates tangible value by preventing costly readmissions.

Metric	Value
Total Test Patients	19,419
Patients Flagged	471
Projected Readmissions Prevented	33
Gross Savings	**$508,500.00**
Intervention Costs	**-$235,500.00**
Net ROI (Savings)	

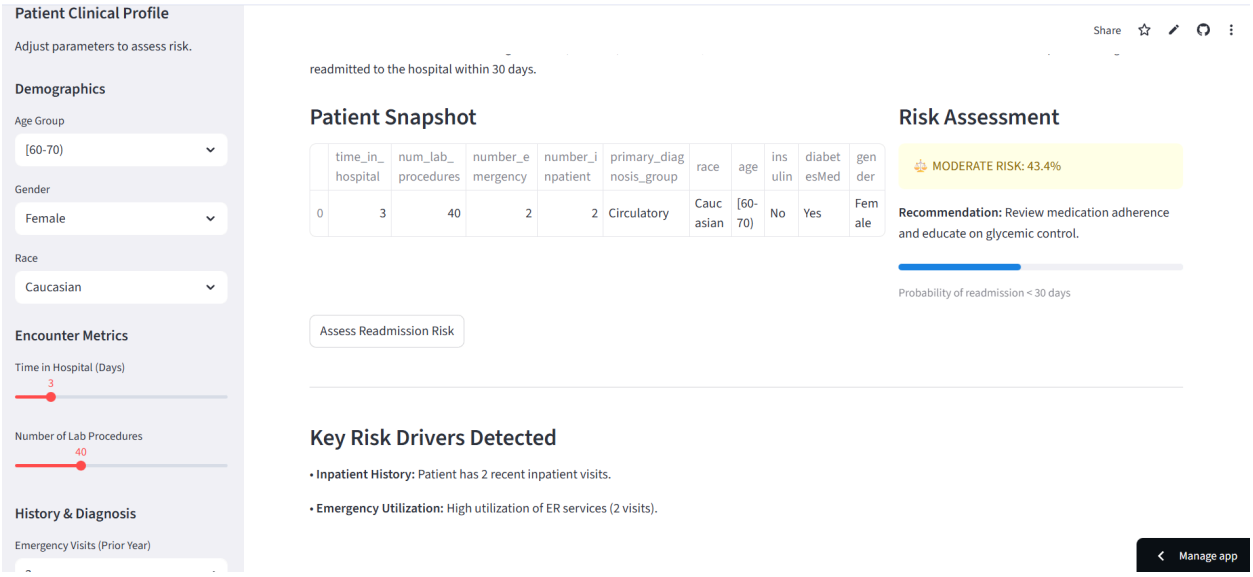
	$273,000.00

---

7. Streamlit Web Application Strategy

The project culminated in the deployment of the **"Diabetic Readmission Predictor"** app.

- **Accessibility:** Allows non-technical users (doctors/administrators) to interact with the model.
- **Real-time Prediction:** Provides instant risk assessment based on inputs like age, emergency visits, and lab procedures.
- **Risk Drivers:** The app highlights *why* a patient is considered high-risk (e.g., "Inpatient History: 4 recent visits") to build clinical trust.



---

8. Limitations & Future Work

- **Data Age:** The dataset covers 1999–2008. Clinical workflows and EHR systems have evolved significantly since the Affordable Care Act (ACA).
- **Potential Bias:** Demographic representation must be monitored to ensure fairness in modern deployment.

- **Future Exploration:** Testing deep learning models or incorporating "social determinants of health" (income, neighborhood data) could improve accuracy.

---

9. Conclusion

This project successfully demonstrated that machine learning can predict clinical readmission risk with high accuracy and clear financial benefits. By bridging the gap between complex data science and practical healthcare decision-making via a Streamlit application, we created a tool that saves resources and provides a framework for better patient care.