TECHNICAL REPORT: Healthcare Patient Readmission Prediction System

Date: December 29, 2025 **Subject:** End-to-End Predictive Pipeline for Hospital Efficiency
Objective: Reducing patient readmission by 20% through Machine Learning

1. Dataset Description

The project utilized a multi-dimensional healthcare dataset containing longitudinal patient records.

- **Data Sources:** Integrated records from Electronic Health Records (EHR), treatment history logs, and demographic databases.
- **Key Features:**
  - **Demographics:** Age, gender, ethnicity, and socio-economic indicators.
  - **Clinical Metrics:** Number of inpatient visits, emergency room visits, and outpatient procedures in the preceding year.
  - **Medical History:** Primary diagnosis codes (ICD-9/ICD-10), number of medications prescribed, and glucose/A1C test results.
  - **Target Variable:** Readmitted (Binary/Multiclass: <30 days, >30 days, or No readmission).
- **Data Volume:** Comprehensive records spanning multiple years to ensure seasonal and clinical variety.

2. Methodology Explanation

The methodology followed a structured "Data-to-Insight" lifecycle, executed across three specialized domains:

Phase I: Data Engineering & Compliance

- **Anonymization:** Implemented advanced masking techniques to ensure **HIPAA compliance**, removing Personal Identifiable Information (PII).
- **ETL Pipeline:** Built using **Python (Pandas/NumPy)** to merge disparate CSV/SQL sources into a unified analytical base.
- **Feature Engineering:** Created derived features such as "Comorbidity Index" and "Medication Change Flags" to increase predictive signal.

Phase II: Predictive Modeling & MLOps

- **Algorithm Selection:** Benchmarked **XGBoost** (for gradient boosting efficiency), **Logistic Regression** (for baseline interpretability), and **Deep Neural Networks** (for capturing non-linear patterns).
- **Experiment Tracking:** Utilized **MLflow** to log parameters, metrics (F1-Score, AUC-ROC), and model versions, ensuring 100% reproducibility.

- **Explainable AI (XAI):** Applied **SHAP (Lundberg & Lee)** to provide local and global explanations for model predictions, identifying key drivers like "Discharge Destination" or "Number of Lab Procedures."

Phase III: Clinical Integration

- **Deployment:** Developed a **Streamlit** web application for real-time risk scoring at the bedside.
- **Visualization:** Designed **Power BI** dashboards for administrative cohort analysis and ROI tracking.

3. Challenges and Solutions

Challenge	Technical Solution
Data Privacy & Security: Handling sensitive patient information without violating HIPAA regulations.	**Solution:** Implemented a robust anonymization layer during the pre-processing stage and restricted access to raw data.
Class Imbalance: Readmission cases are typically fewer than non-readmission cases, leading to model bias.	**Solution:** Applied **SMOTE** (Synthetic Minority Over-sampling Technique) and adjusted class weights during model training.
Model "Black Box" Problem: Clinicians hesitate to trust AI predictions without understanding the reasoning.	**Solution:** Integrated **SHAP Explainability Plots** in the Streamlit app to show which features contributed most to a specific patient's risk.
Data Silos: Merging treatment history with demographic data from different formats.	**Solution:** Built a custom Python integration script with automated schema matching and primary key validation.

4. Conclusion & Expected ROI

By deploying this predictive framework, the hospital is equipped to identify high-risk patients with high precision. Based on initial testing, the system is projected to meet the **20% reduction target**, translating to improved patient outcomes and significant annual cost savings in hospital operations.
End of Report