Healthcare Data Analysis

Final Project Report

Group Name: The Closer

Member 1 Name: Rongala Sreedhar

Member 1 Email: rongalasreedhar@gmail.com

1. Introduction

The objective of this project is to analyze the 'Healthcare Dataset' to understand the factors driving 'Drug Persistency' and to build a predictive model. Drug Persistency refers to the duration of time from initiation to discontinuation of therapy. Understanding this helps pharmaceutical companies vastly improve patient outcomes.

2. Exploratory Data Analysis (EDA)

We conducted a thorough analysis of the dataset, focusing on demographics, region, and risk factors.

Key Findings:
- Certain regions exhibit significantly higher persistency rates.
- Comorbidities are strong indicators of persistency behavior.

3. Methodology & Modeling

We followed a standard Data Science lifecycle:
1. Data Cleaning: Handling missing values and outliers.
2. Feature Engineering: Encoding categorical variables and scaling numerical features.
3. Model Selection: We tested Logistic Regression, Random Forest, and XGBoost.
4. Evaluation: Models were evaluated based on Accuracy and ROC-AUC.

4. Results

Our experiments yielded the following results:
- Logistic Regression provided a baseline with ~80% accuracy.
- Random Forest improved this with feature interactions (~86%).
- XGBoost achieved the best performance with ~88% accuracy and 0.90 AUC.

5. Conclusion & Recommendations

XGBoost is recommended for the production environment due to its superior predictive power. For

stakeholders requiring transparency, SHAP values can be used to explain individual predictions.

We have also delivered a Streamlit dashboard for real-time interaction with the model.

6. Code Repository

The complete code and resources can be found at: https://github.com/sreedharsiddhu/Data-Glacier