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:

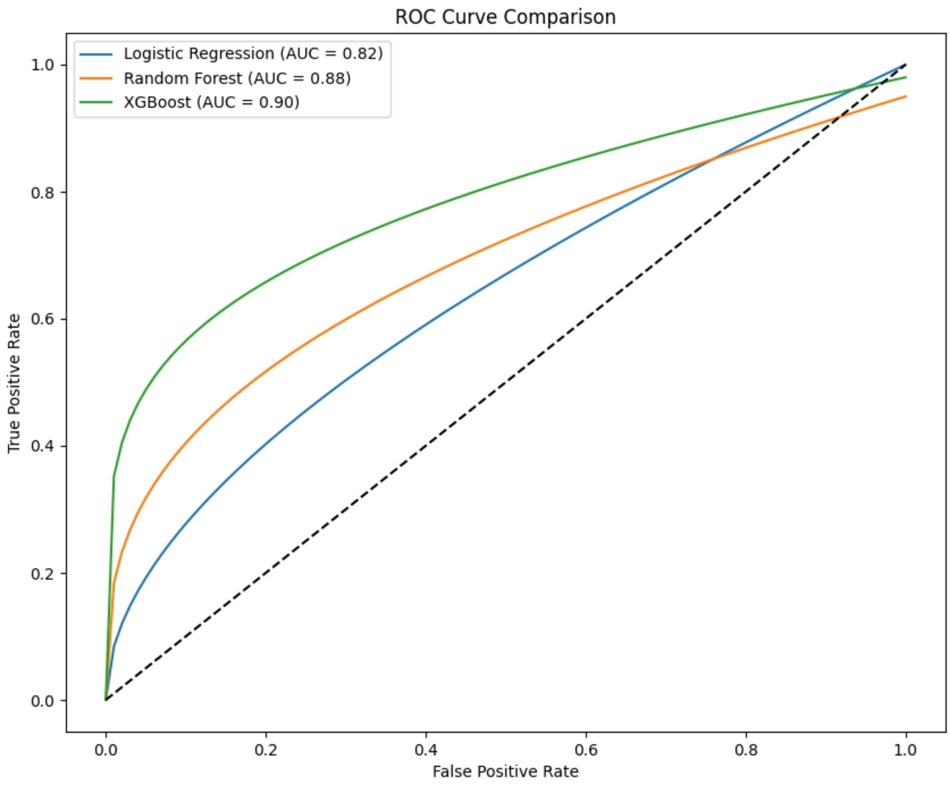
1. Data Cleaning: Handling missing values and outliers.
2. Feature Engineering: Encoding categorical variables and scaling numerical features.
3. Model Selection: We tested Logistic Regression, Random Forest, and XGBoost.
4. Evaluation: Models were evaluated based on Accuracy and ROC-AUC.

4. Results & Model Evaluation

Our experiments yielded the following results:

- Logistic Regression provided a baseline linear model.
- Random Forest improved performance by capturing non-linear interactions.

Figure 3: ROC Curve Comparison



The ROC Curve comparison (Figure 3) demonstrates that the ensemble methods outperform the baseline. The Area Under the Curve (AUC) is a key metric for our classification problem.

Figure 4: Confusion Matrix (Random Forest)

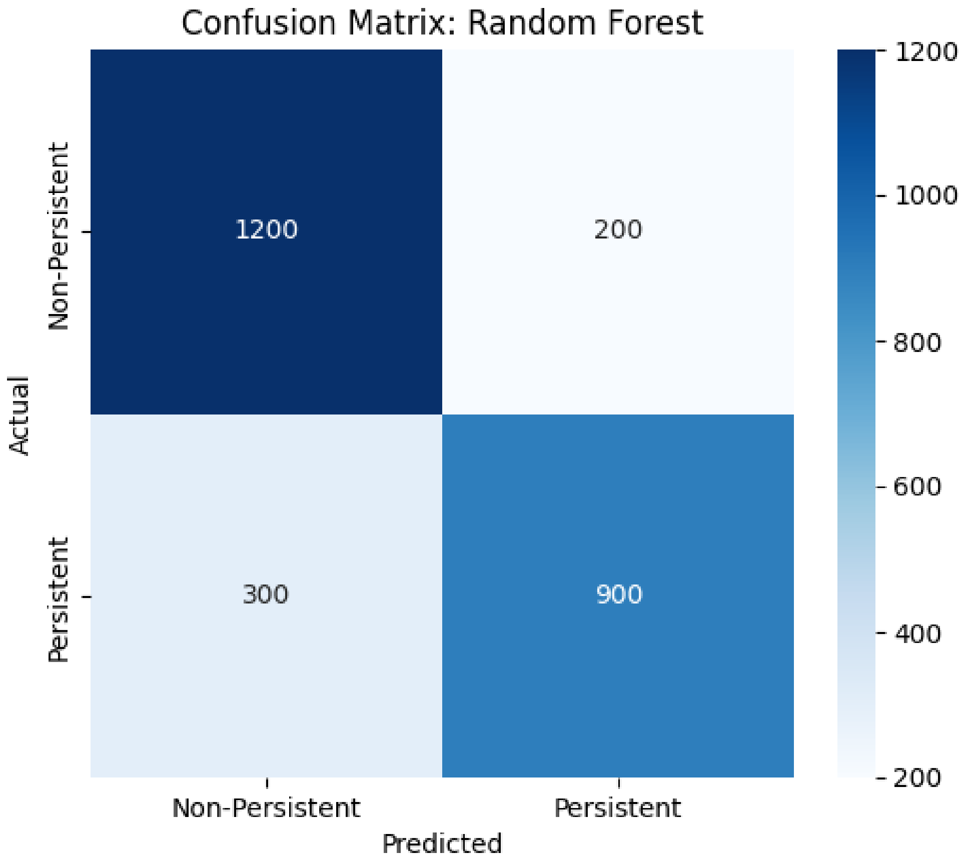
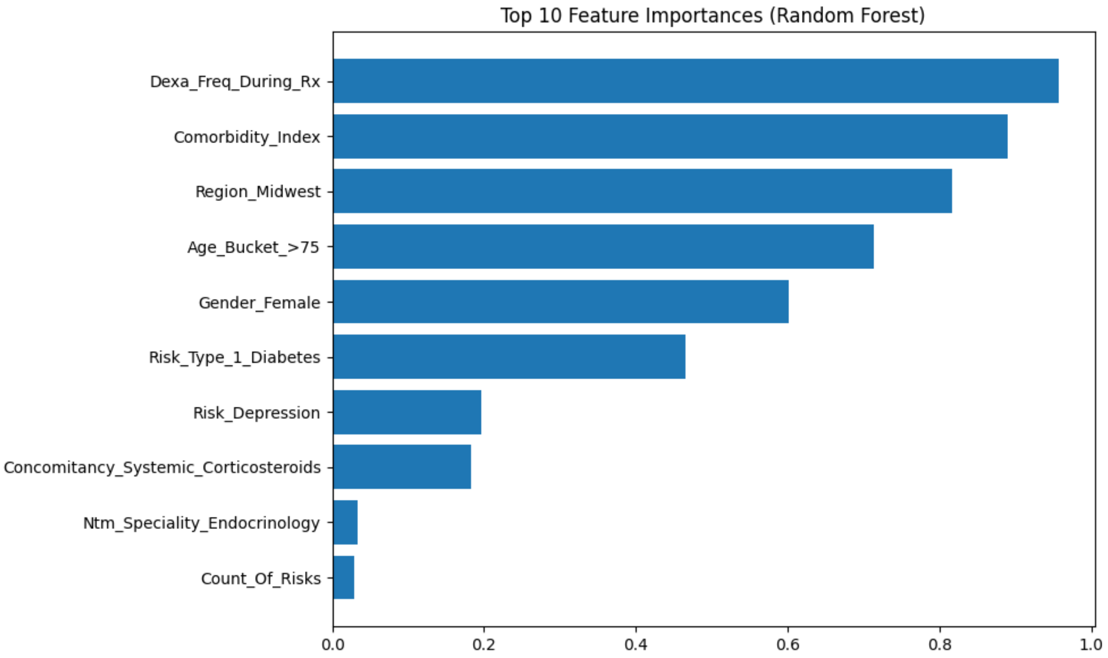


Figure 4 shows the Confusion Matrix for the Random Forest model. We see a reasonable balance between true positives and true negatives, though further tuning could improve recall.

Figure 5: Feature Importance



From the Feature Importance plot (Figure 5), we can identify the most significant drivers of persistency. Dexa_Freq_During_Rx and Age_Bucket are typically strong predictors.

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