Week 10: EDA & Final Recommendations

Team Member's Details

Group Name: The Closer

Member 1 Name: Rongala Sreedhar

Member 1 Email: rongalasreedhar@gmail.com

Member 1 Country: Italy

Member 1 College/Company: University of Naples Federico II

Member 1 Specialization: Data Science

Problem Description

The objective is to perform comprehensive Exploratory Data Analysis (EDA) on the Healthcare dataset to identify key factors affecting drug persistency. This involves understanding data distribution, correlation between variables, and deriving actionable insights.

Github Repo Link

https://github.com/sreedharsiddhu/Data-Glacier

EDA Performed

We executed the following analysis (detailed in `eda.ipynb`):

1. Univariate Analysis: Examined distributions of `Persistency_Flag` (Target), `Count_Of_Risks`, and `Gender`.
2. Bivariate Analysis: Investigated the relationship between `Gender` and `Persistency_Flag`, finding minimal deviation. Analyzed `Count_Of_Risks` against `Persistency_Flag`, showing that higher risk counts correlate slightly with persistency issues.
3. Correlation Analysis: Generated heatmaps to detect multicollinearity among numerical features.

Final Recommendation

Based on our analysis, we recommend the following procedure:

1. Data Preprocessing: Impute missing values in `Dexa_Freq_During_Rx` and scale numerical features (`Count_Of_Risks`).
2. Feature Engineering: Create an interaction term between `Risk` factors and `Age` buckets if possible.

3. Modeling Strategy: Proceed with Logistic Regression for interpretability, then attempt Gradient Boosting (XGBoost) to capture non-linear patterns.
4. Deployment: Monitor the `Persistency_Flag` distribution in production to detect drift.