Healthcare - Persistency of a drug

Team Member's Details

Group Name	DataPulse
Name	Keval Jaysukhbhai Savaliya
Email	skeval1601@gmail.com
Country	Germany
College/Company	Technische Universität Chemnitz
Specialization	Data Science

Problem Description

One of the key challenges for pharmaceutical companies is to understand the persistency of drug usage as per physician prescriptions. The objective of this project is to build a machine learning model to classify whether a patient is persistent or non-persistent with their prescribed therapy, and to identify factors that influence persistency.

Data Understanding

The dataset consists of 3424 rows and 69 columns. The target variable is *Persistency_Flag* (Persistent vs Non-Persistent). Features cover patient demographics, provider attributes, clinical factors, disease/treatment factors, and adherence. IDs such as Patient ID were excluded from training.

Type of Data

The dataset is a structured tabular dataset. It includes categorical features (gender, race, region, comorbidity flags, specialty etc.), numerical features (DEXA frequency, age bucket), and binary flags (Y/N or 0/1).

Problems in the Data

Key issues identified:

- NA values: Some categorical and numeric columns contain missing values.

- High cardinality: Certain features like Ntm_Speciality have many unique values.

- Class imbalance: 62% Non-Persistent vs 38% Persistent.

- Possible outliers in numeric features like DEXA frequency (range up to 58).

- Mixed encoding in categorical fields (e.g., Yes/No vs Y/N).

Approaches to Overcome Issues

To address the problems:

- NA values: Impute using median (numeric) and most frequent value (categorical).

- Outliers: Tree-based models are robust to outliers; for linear models, scaling and winsorization can be applied.

- Skew/imbalance: Use class weights in models and evaluate with ROC-AUC, precision/recall.

- Encoding: Apply one-hot encoding with handle_unknown='ignore' to safely encode categorical data.

- High cardinality: Group rare categories or use target encoding if needed.