Data Intake Report

Team Members: Saif Amer, saifamer67@gmail.com, United States of America, Data Science track

Problem Description

ABC Pharma seeks to improve patient therapy outcomes by understanding **persistence patterns** in drug usage prescribed by physicians. Persistency refers to whether patients continue their prescribed therapies over time. The business objective is to **automate the identification of persistence behavior** using a machine learning classification model. This will enable targeted interventions to improve long-term patient outcomes and therapy adherence.

Task: Build a **classification model** to predict whether a patient will be persistent (Persistency_Flag = 1) or not (Persistency_Flag = 0).

Data Understanding

The dataset consists of patient-level health and treatment data, including demographics, provider and clinical information, comorbidities, drug usage, and adherence history. Each row represents a unique patient.

Key Points:

Target Variable: Persistency_Flag

Granularity: Patient level

Time Sensitivity: Some variables are time-bound (e.g., events in the last 365 days)

Type of Data Available for Analysis

Category	Examples
Demographics	Age, Gender, Race, Region, Ethnicity
Provider Attributes	NTM - Physician Specialty
Clinical Factors	T-Score, Risk Segment, Change Indicators, DEXA scans
Therapy Usage	Glucocorticoid and Injectable Usage
Comorbidities	Chronic and Acute Conditions
Adherence	Therapy adherence metrics
Outcome	Persistency_Flag (0 or 1)

Data Problems

Problem Type	Details
Missing Values (NA)	Present in several features like T-Score, Change in Risk Segment, DEXA-related features, and Comorbidities. NA values can represent either missing data or meaningful absence of diagnosis/event.
Categorical Imbalance	The Persistency_Flag is likely to be imbalanced (most patients are either persistent or non-persistent).
Outliers	Possible outliers in Age, NTM - Dexa Scan Frequency, and numerical adherence values.
Skewed Variables	Features like scan frequency, comorbidities count, and adherence are likely right-skewed.
High Cardinality	Fields like NTM - Risk Factors and NTM - Comorbidity may contain high-cardinality categorical data or text strings.

Overcoming Problems

A. Handling Missing Values

Approach	Why
Categorical NA → 'Unknown'	For features like Change in T Score, Change in Risk Segment, NA can be a meaningful category.
Numerical NA → Median/Mode Imputation	For variables like T Score, where NA may be due to unrecorded data, median imputation is robust.
Drop Variables/Rows	Only if missingness is extreme (>50%) and the variable adds little value.

B. Handling Outliers

Step	Why
IQR-based capping or transformation	Mitigate extreme values in Age, Dexa Scan Frequency, Adherence.
Domain-specific thresholds	Apply medically informed caps if available (e.g., age < 120).

C. Addressing Class Imbalance

Technique	Purpose
SMOTE / Oversampling	Balance the minority class in training data.
Stratified K-Fold Cross-validation	Maintain class ratios during model evaluation.
Class-weight Adjustment	Penalize the misclassification of the minority class more heavily.

D. Encoding Categorical Variables

Type	Technique
Nominal (e.g., Gender, Ethnicity)	One-Hot Encoding
Ordinal (e.g., Risk Segment: Worsened < Remained Same < Improved)	Label Encoding
High Cardinality (e.g., Risk Factors, Comorbidity)	Feature hashing or dimensionality reduction (e.g., PCA, clustering based encoding)

E. Feature Engineering Ideas

Feature	Transformation
Age	Binning into age groups

Adherence	Categorize into Low, Medium, High adherence
Interaction Terms	Risk × Specialty or Gender × Comorbidity

F. Scaling and Transformation

Method	Purpose
StandardScaler / MinMaxScaler	Normalize numerical features for distance-based models
Log Transformation	For skewed count-based features like scan frequency