Team Member Details

Group Name: *SoloVision Analytics*

Name: Ankita Roy
Email: roy.anki39@gmail.com
Country: India
College/Company: Institute: Great Learning, Gurgaon/ Internship: Data Glacier
Specialization: Data Science

Problem Description

Drug persistency is a key metric in determining the long-term success of a treatment regimen. It reflects whether patients continue using a prescribed therapy as advised. Pharmaceutical companies like ABC Pharma often struggle to identify the factors that cause patients to discontinue treatment.

To address this, ABC Pharma has sought the help of data science to automate the detection of persistency patterns among patients using their drug. The aim is to develop a **classification model** that predicts whether a patient is likely to persist with the treatment based on demographic, clinical, and behavioural data.

The binary target variable for this task is **Persistency_Flag**, where the goal is to classify patients as either **persistent** or **non-persistent**.

Data Understanding

The dataset contains 3424 number of rows and 69 columns and covers:

- **Demographics:** Age bucket, race, gender, ethnicity, region
- **Physician Details:** Specialty, Specialty flag
- **Clinical Factors:** T-scores, Risk segments, Changes in risk/T-score, Scan recency/Frequency, Fractures
- **Treatment Usage:** Glucocorticoid usage, Injectable experience, Concomitant drugs
- **Adherence:** Therapy adherence, IDN mapping
- **Target Variable: Persistency_Flag (Persistent/Non-Persistent)**

Basic exploration shows that:

- The data includes both categorical (e.g., Gender, Race) and numerical variables (e.g., Dexa Frequency during Rx)
- The ‘Age’ feature is already bucketed as ‘Age_Bucket’ which is ordinal categorical
- Patient ID is a unique identifier with no modelling value but useful for validation

---

Types of Data

The dataset contains:

- **Categorical Variables** (nominal and ordinal): ~80% of columns (e.g., Race, Region, Gender, Risk Segment, T-score bucket, Age Bucket)
- **Numerical Variables:** e.g., Dexa frequency during Rx, risk counts
- **Binary Flags:** Yes/No feature indicating clinical history or drug usage

There are no null values in any column in the dataset

Problems Identified in Data

- **No null values,** so no missing values to impute
- **Outliers:** Numerical column i.e., Dexa frequency during Rx shows unusual high counts that need capping
- **Skewness:** Distribution for numerical column is right-skewed; transformation may be needed
- **Redundancy:** Some columns are derived from others (e.g., Ntm_Speciality_Flag is redundant with Ntm_Speciality)
- **Imbalance:** Class imbalance in Persistent_Flag (more non-persistent than persistent)

Proposed Approaches

- **Redundant columns:** Drop columns like Ntm_Speciality_Flag because they do not add new information
- **Outliers:** Cap outliers in numerical column to reduce their effect on model training and as the dataset is small so tried capping specifically and not removing
- **Skewness:** Apply log or yeo-johnson transformation on skewed numeric variables, if needed, for better model performance
- **Encoding:** Nominal categories can be encoded using Label encoding, ordinal categories can be encoded using Ordinal encoding and rest other columns can be encoded using dummy encoding
- **Class Imbalance:** The target variable shows a mild class imbalance, with ~62% non-persistent and ~38% persistent cases. This imbalance may slightly affect model bias towards the majority class, so will properly focus on recall/precision and not just accuracy and still if the scores doesn't improve then appropriate techniques like class weighting or resampling can be considered

---

GitHub Repository Link
https://github.com/Ankita-ar/PERSISTENCY-OF-A-DRUG