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 Cleansing & Transformation

- It's verified that no missing values exist across all columns
- Dropped redundant columns like the detailed Ntm_Speciality and Ntm_Specialist_Flag, retaining the more meaningful bucketed version
- Performed encoding:
  - ➢ Ordinal encoding for most of the columns like 'Persistency Flag', 'Age Bucket', 'Glucocorticoid record during or prior Ntm', 'Tscore Bucket', 'Adherent Flag' etc.
  - ➢ N-1 Dummy encoding for other features like 'Gender', 'Race', 'Ethnicity', 'Region', 'Risk segment during or prior Rx', 'Changes in Tscore'
- Checked numerical column i.e., Dexa frequency during Rx for outliers using IQR method and distribution plots

---

Techniques Applied for Cleansing

The techniques used are:

- **Outlier Handling:**
  - ➢ Identified numeric outliers in Dexa frequency during Rx

    - Applied IQR based capping, using Q1 (25th percentile) and Q3 (75th percentile) to calculate the outlier range. Extreme values above the upper bound were capped instead of removing rows – preserving sample size due to small dataset
- **Skewness & Transformation:**
    - Reviewed numerical column distribution; planned log 1p transformation as there were only positive values and zero's and to normalize skewness for better model performance
- **Encoding Approaches:**
    - Performed n-1 dummy encoding for nominal columns and not label encoding as it would have changed the order of the data and n-1 had given the cleanest input matrix and best interpretability
    - Used ordinal mapping for ordinal columns like 'Persistency_Flag', 'Age Bucket' to preserve the inherent order
- **Class Imbalance:**
    - The target shows mild imbalance (~62% non-persistent, ~38% persistent). For fairness, will focus on precision and recall as well rather than accuracy alone while building the model
    - If needed, class weights or resampling strategies will be explored at the modelling stage

---

Simulated Peer Review

The project is completed solo, so no multiple team member coding/review applies. However, all code has clear inline comments, step-by-step explanations, version-controlled commits in GitHub to ensure transparency and reproducibility

---

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