DATA SCIENCE HEALTHCARE PROJECT

INDIVIDUAL PROJECT NAME: Health and Care

Name : Abina Azees

MAIL ID: abinaazees@gmail.com

Country : UAE

College: Amal Jyothi College of Engineering and Technology

Specialisation: Data Science

Problem Description

Pharmaceutical companies often struggle to assess how consistently patients adhere to their prescribed treatments, a concept known as **drug persistency**. This persistency plays a vital role in determining not only the **effectiveness of medical therapies** but also their **impact on patient outcomes** and the **commercial success** of the drugs.

To address this challenge, **ABC Pharma** has collaborated with an analytics firm to build a **predictive model** capable of identifying patients who are likely to be **persistent or non-persistent** with their therapies. The goal is to leverage a combination of **demographic, clinical, treatment-related, and adherence data** to uncover key factors influencing persistency. These insights will help enhance **patient engagement strategies**, optimize **treatment planning**, and ultimately improve **healthcare outcomes and business performance**.

Business Understanding

Drug persistency is a key indicator of how effectively patients adhere to their prescribed therapies. Gaining insights into persistency patterns provides valuable opportunities to enhance both **clinical outcomes** and **business performance**.

Why Drug Persistency Matters:

- **Improved Patient Targeting:** Enables more tailored and proactive interventions for patients at risk of non-adherence.
- **Optimized Sales and Marketing:** Facilitates data-driven strategies to promote therapies to the right patient segments.
- **Better Treatment Outcomes:** Helps reduce relapses and health complications by ensuring continuous medication intake.
- **Operational Efficiency:** Supports cost-effective resource allocation by focusing on patients who are less likely to persist.

Stakeholder Objectives:

- **ABC Pharma:**
  - Increase overall drug persistency rates
  - Enhance therapeutic effectiveness and patient satisfaction
  - Maximize market share and revenue growth
- **Analytics Partner:**
  - Build a robust, interpretable, and high-performing predictive model
  - Provide actionable insights into the drivers of persistency
  - Deliver a solution that can be seamlessly integrated into ABC Pharma's decision-making processes

DATASET Understanding

The dataset, as summarized in Table, contains four primary categories of predictor variables: **Demographics**, **Provided Attributes**, **Clinical Factors**, and **Disease/Treatment Factors**. The **target variable** in this study is the **Persistency Flag**. Each category of predictors will be analyzed in terms of their interclass and intraclass means and variances to understand their influence on drug persistency.

Bucket	Variable	Variable Description
Unique Row Id	Patient ID	Unique ID of each patient
Target Variable	Persistency Flag	Flag indicating if a patient was persistent or not
Demographics	Age	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	Ethnicity	Ethnicity of the patient from the patient table
	Gender	Gender of the patient from the patient table
	IDN Indicator	Flag indicating patients mapped to IDN
Provider Attributes	NTM - Physician Specialty	The specialty of the HCP that prescribed the NTM Rx
Clinical Factors	NTM - T-Score	T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate)
	Change in T Score	Change in T-score before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Risk Segment	Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate)
	Change in Risk Segment	Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Multiple Risk Factors	Flag indicating if a patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate)
	NTM - Dexa Scan Frequency	Number of DEXA scans taken before the first NTM Rx date (within 365 days prior from rxdate)
	NTM - Dexa Scan Recency	Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable)
	Dexa During Therapy	Flag indicating if the patient had a Dexa Scan during their first continuous therapy
	NTM - Fragility Fracture Recency	Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate)
	Fragility Fracture During Therapy	Flag indicating if the patient had fragility fracture during their first continuous therapy
	NTM - Glucocorticoid Recency	Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the oneyear look-back from the first NTM Rx
	Glucocorticoid Usage During Therapy	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy
Disease/Treatment Factors	NTM - Injectable Experience	Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx
	NTM - Risk Factors	Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, oneyear lookback from the date of the first OP Rx
	NTM - Comorbidity	Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease, we are taking a complete look back from the first Rx date of NTM therapy and for acute diseases, a period before the NTM OP Rx with one-year lookback has been applied
	NTM - Concomitancy	Concomitant drugs recorded before starting with a therapy(within 365 days before the first rxdate)

	Adherence	Adherence to the therapies

Subject data related to **Gender, Race, Age, Region, Ethnicity**, and **IDN Indicator** were analyzed using **Pivot Tables and Pivot Charts** in MS Excel, as illustrated in Figure 2. The descriptive analysis indicates noticeable demographic imbalances within the dataset. For instance, approximately **94% of the patients are female**, while only **6% are male**, revealing a significant gender skew. This raises questions about whether the dataset is centered around a disease that predominantly affects females or if women, particularly within certain age groups, are more likely to seek treatment. Regardless of the underlying cause, this gender imbalance suggests that the findings derived from the dataset will be more representative of **female patients**.

Additionally, the dataset is heavily skewed toward **non-Hispanic, Caucasian patients**, indicating limited ethnic diversity. While the **Region** and **Age** distributions appear relatively more balanced, a closer look reveals that the majority of patients are in the **older age groups**, with only a small proportion younger than 55. Most patients are from the **Midwest region**, highlighting a geographical concentration in the sample. A clinically significant observation is the high percentage of subjects associated with an **Integrated Delivery Network (IDN)**—approximately **75%** of the patients. This suggests effective clustering of healthcare service providers, potentially leading to improved coordination and **enhanced patient experiences**.

resence of Null Values and Outliers

Using Python's **Pandas** library, null values were assessed across all columns. The dataset **contains no missing values**, indicating it is clean in this regard. Detecting and addressing null values is essential, as their presence can cause errors during machine learning or lead to misleading analysis outcomes.

Furthermore, since most variables in the dataset are **categorical**, traditional **outlier detection** was not applicable. No significant anomalies or extreme values were found, confirming that the dataset is free from outliers.

However, some **class imbalance and skewness** were observed in specific predictors.

- **Demographic variables** exhibited low skewness, indicating they are closer to a **normal distribution**, despite some imbalances (e.g., gender).
- **Risk-related variables** demonstrated the **highest skewness**, indicating a concentration of values in certain categories.
- Similar patterns were observed in **kurtosis**, which is closely related to skewness and measures the "tailedness" of the distribution.

Overall, while the dataset is clean in terms of null values and outliers, **skewed distributions** in several predictors—particularly risk factors—may influence model performance and require preprocessing techniques such as resampling or transformation.

Approaches

Among the 68 parameters in the dataset, only **2 are numerical**, while the remaining are **categorical**. To prepare the data for machine learning algorithms—which typically require numerical input—**encoding techniques** must be applied to convert categorical variables into numerical form. One common approach used is the creation of **dummy variables** through **one-hot encoding**.