

**Team member's details:****Group Name: Data Dominators**

<a href="mailto:chau.devin031602@gmail.com">chau.devin031602@gmail.com</a>	Devin Chau	United States	San Jose State University	Data Science
--	------------	---------------	---------------------------	--------------

<a href="mailto:rohankhatri0507@gmail.com">rohankhatri0507@gmail.com</a>	Rohan Khatri	United States	San Jose State University	Data Science
--	--------------	---------------	---------------------------	--------------

<a href="mailto:ethan05dy@gmail.com">ethan05dy@gmail.com</a>	Ethan Dy	United States	San Jose State University	Data Science
--	----------	---------------	---------------------------	--------------

**Project: Data Science:: Healthcare - Persistency of a drug****Problem description**

We are building a predictive model that classifies patients into “persistent” or “non-persistent” categories based on factors like their demographics, medical history, physician characteristics, and treatment details. Factors like the patient level such as their age, risk factors, previous test results, or provider type allows for insights into why some patients continue therapy while others drop off. Thus understanding “persistence” levels. By analyzing these data points and finding patterns, the predictive model helps explain patient behavior and supports the creation of targeted interventions to improve adherence.

**Data understanding**

Here is our understanding of the data:

- Feature Descriptions (High-level)
  - Unique Row ID: Patient ID (unique identifier)
- Target Variable:
  - Persistency\_Flag: “Yes/No” or “Persistent/Non-Persistent”
- Demographics:
  - Age, Race, Region, Ethnicity, Gender
  - IDN Indicator (whether a patient is mapped to an IDN)
- Provider Attributes:

- NTM – Physician Specialty (e.g., General Practitioner, Endocrinologist, etc)
- Clinical Factors:
  - NTM – T-Score, Change in T Score
  - NTM – Risk Segment, Change in Risk Segment
  - NTM – Multiple Risk Factors
  - NTM – DEXA Scan Frequency, NTM – DEXA Scan Recency
  - DEXA During Therapy
  - NTM – Fragility Fracture Recency, Fragility Fracture During Therapy
  - NTM – Glucocorticoid Recency, Glucocorticoid Usage During Therapy
- Disease/Treatment Factors:
  - NTM – Injectable Experience
  - NTM – Risk Factors (chronic vs. acute)
  - NTM – Comorbidity
  - NTM – Concomitancy (drugs used within a certain timeframe)
  - Adherence

### **What type of data you have got for analysis**

Here is the type of data we got from our analysis:

- Dataset Size
  - Rows: **3,424**
  - Columns: **69**
- Types of Variables
  - **Numeric (2 columns):**
    - DEXA\_Freq\_During\_Rx
    - Count\_Of\_Risks
  - **Categorical (67 columns):**
    - Examples: Persistency\_Flag, Gender, Ntm\_Speciality, etc
    - Many are binary flags (Y/N, etc.), while some have multiple categories (e.g., Ntm\_Speciality has 36)
- Hidden “Unknown” Data
  - Some columns use the string "Unknown" instead of NaN. Like, Risk\_Segment\_During\_Rx, Change\_T\_Score, and others have a lot of "Unknown" entries

### **What are the problems in the data ( number of NA values, outliers , skewed etc)**

1. Missing or “Unknown” Values:
  - a. Although `df.isnull().sum().sum()` equals **0** that shows no null entries
    - i. Multiple columns (such as `Risk_Segment_During_Rx`, `Change_T_Score`, etc.) contain a large number of “Unknown” entries. This shows hidden missing data that could influence model training and interpretation
2. Outliers:
  - a. Two numeric variables, `Dexa_Freq_During_Rx` and `Count_Of_Risks`, have outliers:
    - i. `Dexa_Freq_During_Rx` shows **460** outliers (based on the Interquartile Range method)
    - ii. `Count_Of_Risks` shows **8** outliers (also IQR-based)
3. Skewed Distributions
  - a. `Dexa_Freq_During_Rx` exhibits high skewness (approx. **6.81** before applying a log transform).
  - b. `Count_Of_Risks` is moderately skewed (approx. **0.88**)
4. High Size in Categorical Features
  - a. `Ntm_Speciality` alone has **36** unique categories. If modeled incorrectly, this can lead to issues like overfitting

### **What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?**

1. Handling “Unknown” Entries
  - Maintain “Unknown” as a Separate Category: This approach keeps potential patterns related to missing data
  - Imputation: Where domain knowledge allows, “Unknown” can be combined with existing categories or replaced using an imputation strategy
2. Removing Outliers
  - Log Transformation: Already applied to `Dexa_Freq_During_Rx` to reduce the impact of extreme values and normalize its distribution
  - Capping Extreme Values (Winsorizing): Could be considered if certain outliers are not really possible or are likely data-entry errors

### 3. Addressing Skewness

- DEXA\_Freq\_During\_Rx: The log transform helps reduce very high skew, stabilizing variance for modeling
- Count\_Of\_Risks: With a lower skew, a transform may or may not be necessary. Comparisons during modeling would confirm its impact on performance

### 4. Encoding Categorical Variables

- With 67 categorical columns, methods such as one-hot encoding or label encoding are needed
- Binary columns (e.g., "Y"/"N") can be straightforwardly mapped to 0 and 1

### **Github Repo link**

<https://github.com/ethan05d/DataGlacier-Internship/tree/main/Week%208>