Name: Data Science:: Healthcare - Persistency of a drug:: Group Project
Report date: 24th April 2024
Internship Batch: LISUM30:30 Jan24 - 30 Apr 24
Version: 1.0
Data intake by: David Paul Nalumenya
Group name: None
Specialization: Data science
Email: nalumenyad@gmail.com
Company: Makerere University
Data intake reviewer: None
Data storage location:

1.0 Problem Statement

One of the challenges for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

1.1 ML Problem

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

2.0 Data Understanding

2.1 Unveiling Patient and Treatment Factors for Persistency Prediction

This section delves into the characteristics and composition of the data used to build our patient drug persistency model. The data originates from two primary sources: patients and their physicians. By analyzing these combined datasets, we aim to gain a comprehensive understanding of factors influencing how well patients adhere to their prescribed medication regimens.

2.2 Patient Data

- **Demographics such as** Age, race, region, ethnicity, and gender

  These features help us explore potential demographic disparities in persistency. For instance, age might be relevant due to physiological changes or medication sensitivity in older populations.

- **Persistency Flag**

  This binary variable (adherent/non-adherent) serves as the target variable for our machine learning model. It allows us to identify the outcome we want to predict – a patient's completion of the prescribed treatment course.

2.3 Clinical Factors

- **NTM and change in T-Score**

  NTM (Dual-Energy X-ray Absorptiometry) scans measure bone mineral density (T-Score). Changes in T-Score can influence medication needs and treatment decisions by healthcare providers.

- **Risk Segment and Change**

  Classification of patients based on their risk for complications or treatment response (risk segment) can provide insights into patient vulnerability and potential reasons for persistency variations.

- **Scans and Risk Factors Information**

  Details regarding DEXA scan frequency, recency, and associated risk factors offer valuable information about bone health, potential drug interactions, and factors influencing scan scheduling.

- **NTM-related Fracture History**

  Data on recent fragility fractures (before and during therapy) helps assess a patient's susceptibility to bone breaks, potentially impacting treatment decisions and persistency.

- **Glucocorticoid Usage**

  Information on past and current use of glucocorticoids (steroids) is crucial as they can interact with the medication of interest and influence adherence due to potential side effects.

- **DEXA Scan Timing**

By analyzing DEXA scans performed before and after treatment, we can assess changes in bone health over time, which might influence treatment decisions and ultimately, persistency.

- **Risk Groups**
  Categorizing patients based on the specific therapy they received helps identify persistency patterns associated with different treatment options. Different medications might have varying side effects or dosing schedules, impacting adherence.
- **Risk Factors and Co-morbidities**
  Data on expected and reported risk factors for the prescribed therapy, along with any co-morbidities, allows us to explore potential challenges patients might face during treatment and how these factors might influence their adherence.
- **Concomitant Medications**
  Information on medications taken concurrently with the index therapy can reveal potential drug interactions or increased treatment complexity, factors that might impact adherence.
- **Adherence History and Injectable Usage**
  Knowing a patient's past medication adherence behavior and their experience with injectable medications provides insights into their comfort level with treatment regimens and potential adherence challenges.

2.4 Physician Data:

- **Specialty:** Understanding the physician's area of expertise allows us to account for potential variations in prescribing practices and treatment recommendations across specialties. Different specialties might have varying approaches to managing the condition and communicating with patients, impacting adherence.

By comprehensively analyzing these diverse data points, we can build a robust understanding of the patient- and treatment-related factors that influence medication persistency. This knowledge will be instrumental in developing an accurate machine learning model to predict persistency and ultimately improve patient health outcomes.

3.0 Data Source and Format

The data for this project was obtained from an Excel file (.xlsx) containing 3424 observations (patients) and 64 features potentially influencing patient drug persistency.

3.1 Feature Overview

The majority of the features (61 out of 64) are of the object data type. This suggests the data likely includes categorical information such as demographics, diagnoses, medication names, and other descriptive variables. The remaining 3 features are of integer data type, which could represent numerical values. Understanding the meaning and potential impact of each feature is crucial for building an effective machine learning model.

3.2 Data Cleaning and Feature Engineering

The data acquired for this project ideally should be inspected for missing values as even small amounts can impact model performance. While the initial data may appear complete, it's important to implement checks to identify missing entries (NA values). For categorical features (object data type) with a low percentage of missing values (<10%), imputing with the mode (most frequent value) can be a reasonable approach, as it reflects the central tendency of the category. However, for missing values exceeding 10% in categorical features, deletion might be less favorable, potentially introducing bias. In such cases, more sophisticated techniques like synthetic imputation or prediction based on existing data relationships are recommended. For numerical features (integer data type), if missing values fall below 10%, imputation with the median can be a statistically sound strategy. It's important to remember there is no one-size-fits-all solution for missing data, and the chosen approach should consider the specific characteristic of the feature and the amount of missing data present.

3.3 Handle outliers

While object datatypes typically represent categorical data and may not exhibit numerical outliers in the strictest sense, they can contain inconsistencies or duplicates that require cleaning. Techniques like label encoding or one-hot encoding can be used to transform these features into numerical representations suitable for machine learning models. On the other hand, numerical datatypes can contain outliers that skew analysis. IQR (Interquartile Range) is a robust method for identifying outliers, which can then be addressed based on domain knowledge. For instance, outliers might be indicative of data entry errors or genuine extreme cases requiring further investigation. Simply replacing outliers with median values might not always be optimal. Statistical techniques like winsorization (capping outliers to IQR thresholds) can be used to cap outliers to a specific percentile within the IQR, preserving valuable information while reducing their undue influence on the analysis.