Data Glacier Team: DS HealthCare Group

Batch Code: LISUM14

Internship Cohort: September to December 2022

**Team members:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Email** | **Country** | **College** | **Specialization** |
| **Yuchi Chen** | **ychen306@ur.rochester.edu** | **United States** | **University of Rochester** | **Data Science** |
| **Bolutife Akinlawon** | **bolu.akinlawon@gmail.com** | **United Kingdom** | **UWE, Bristol** | **Data Science** |
| **Alexis Collier** | **colliera75@gmail.com** | **United States** | **Emory University Bootcamp** | **Data Science** |
| **Han-Fu Lin** | **hanfu.lin@mail.utoronto.ca** | **Canada** | **University of Toronto** | **Data Science** |

**GitHub Repo link:** [**Healthcare Persistency of a Drug**](https://github.com/colla00/Data-Science-Healthcare---Persistency-of-a-Drug-Group-Project.git)

1. Problem Description

A pharmaceutical company conducts a large number of clinical trials in order to study the durability of a new drug. These trials record a large number of different attributes of experimental subjects and the results of the experiment by means of control variables. The company wants to use the data to understand what properties affect the drug's durability.

1. Data Understanding

* Type of data: .xlsx
* Data size: 3424 by 69
* Variable types: Categorical variables: 67 and Numerical variables: 2

The unique Row ID is the patient identification number, and the target variable is the Persistency\_Flag, indicating whether the patient persisted with the medical treatment.

The data file then includes *demographic information* corresponding to the features of Age, Race, Religion, Ethnicity, Geographic Region, and Gender.

Another column is for *Provider Attributes, which* provides information on the specialty of the physician who prescribed the treatment: Ntm\_Specialty, Ntm\_Specialist\_Flag, Ntm\_Specialty\_Bucket.

The IDN Indicator is a flag for patients mapped to an integrated medication delivery network. These networks have negotiated with the pharmaceutical company for special pricing on their medications, such as group pricing.

Finally, there are the *Clinical Factors,* including the NTM T-Score, the Change in T-Score, the NTM-Risk Segment, the NTM-Multiple Risk Factors, the NTM Dexa Scan, the NTM Dexa Scan Recency, the Dexa During Therapy, the NTM Fragility Fracture Recency, and the Fragility Fracture During Therapy.

The following three columns: NTM Glucocorticoid Recency, Glucocorticoid Usage During Therapy, and NTM Injectable Experience, indicate whether the patient has had glucocorticoid treatment in the year before or during the first continuous therapy or if the patient had any injectable drug treatment during the year before the commencement of the bone density treatment.

The NTM Comorbidity Column gives a sum of comorbidities where these are tallied according to two categories: Acute and Chronic. For Chronic, a complete look back is used, while for acute, a one-year look back is used.

The NTM Concomitancy records any concomitant drugs recorded before starting with the bone density therapy within 365 days before the bone density diagnosis date.

Adherence is an indicator of whether the patient is adherent to the therapy.

1. What type of data is being analyzed?

Most data is represented as objects; there are two integer columns, the count of risks and a Dexa frequency column. There are 69 columns, with much of this data needing transformation into formats that will work for running the machine learning classification programs. For example, the race will need to be converted to multiple columns using dummy variables. In addition, age will need to be added to another column where the possible ages are denoted, so the ranking is incorporated.

1. What are the problems with the data? What approaches are you trying to apply to the dataset to overcome problems and why?

Problems with the data requiring attention before applying the machine learning programs for classification include: the target variable being unbalanced with 1289 Persistent to 2135 Non-Persistent, which will be addressed when building a classification model. Also, many columns appear dependent on each other, which is an issue for models requiring predictor variables to be independent. Another issue is that the data is primarily categorical and will require some preliminary adjustments in the presentation of this data before the algorithms can be applied. Finally, some ordered data, such as age groupings, are not coded in a way that accounts for the ordering. Also, one column had a problematic name that included a comma; ‘'Comorb\_Encntr\_For\_General\_Exam\_W\_O\_Complaint, \_Susp\_Or\_Reprtd\_Dx' needed the comma removed for some other code to work. Lastly, one observation belonging to the specialty Obstetrics & Gynecology was miscoded, plus other garbage within the specialty label; this had to be corrected so that the observation would be included when the ML classification is applied.