Data Glacier Team: The Data Dynamos
Batch Code: `LISUM04`
Remote Internship Cohort: September to December 2021

1. Team member's details:

Shari Arangi
Shariarangi504@gmail.com
Ph +254 0742164535
We! Hub Victoria LTD, Kenya
Data Science

Ateeb Aqil Aziz
Atteeb.aqil@gmail.com
Ph +92 3044693262
Data Science

Jennifer Turley
Jennifer.turley@ucdconnect.ie
Ph +353 83 065 7252
University College Dublin, Ireland
Data Science

Queen Echerenwa
Quin_codes@outlooks.com
+234 806 352 8847
Data Science

2. Problem Description

Healthcare Project: Persistency of a Drug

This project utilizes data relating to treatments for improving bone density.

One of the challenges in medicine is to understand situational and patient-specific factors that contribute to treatment success. An important factor for promoting success is patient persistence with prescribed medical therapy. The 'persistency of a drug' for a patient in this data set refers to whether a given patient took the medication and continued taking it as prescribed. With 'patient persistency ' as the target variable and other variables of interest including age, race, ethnicity, religion, gender along with medical comorbidities considered as dependent variables, this project seeks to determine whether and to what extent these other variables correlate with the desired positive patient persistency for the prescribed bone density treatment.

3. Data Understanding

There are 3424 observations (each associated with a unique patient) and 69 features.

The unique **Row ID** is the patient identification number, and the target variable is the **Persistency_Flag**, which indicates whether the patient persisted with the medical treatment or not.

The data file then includes ***demographic information*** corresponding to the features Age (Age_Bucket), Race, Religion, Ethnicity, geographic Region, and Gender.

Another column is for ***Provider Attributes*** which in this case provides information on the speciality of the physician who prescribed the treatment: these are Ntm_Specialty, Ntm_Specialist_Flag, Ntm_Specialty_Bucket,.

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

Finally, there are the ***Clinical Factors*** : these include

The NTM T-Score, the prior to treatment T-score which indicates the bone density of the patient within two years prior to the commencement of treatment. The Change in T-Score since the start of therapy; it has 4 possible values: Worsened, Remained the Same, Improved, Unknown.

The NTM-Risk Segment gives the risk segment of the patient within two years before the commencement of treatment. The Change in Risk Segment gives the change after treatment, from the possible values : Worsened, Remained the Same, Improved, Unknown.

The NTM-Multiple Risk Factors flag indicates if the patient has more than one risk factor at the time of diagnosis or within a year prior to the beginning of treatment.

The NTM Dexa Scan frequency gives the number of Dexa scans taken within the year prior to commencement of treatment.

The NTM Dexa Scan Recency is a flag indicating the presence of the Dexa scan before the diagnosis or between the first and more recent treatment.

The Dexa During Therapy flag indicates if the patient had a dexa scan during the first continuous therapy. This column has outlier values.

The NTM Fragility Fracture Recency flag indicates if the patient has experienced a fragility fracture within a year prior to the commencement of treatment.

The Fragility Fracture During Therapy flag indicates the patient experienced a fragility fracture during their first continuous treatment.

The next three columns indicate whether the patient has had glucocorticoid treatment in the year prior to or during the first continuous therapy, or if the patient had any injectible drug treatment during the year prior to the commencement of the bone density treatment. These are NTM Glucocorticoid Recency, Glucocorticoid Usage During Therapy, NTM Injectible Experience.

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 prior to starting with the bone density therapy, within the 365 days prior to the bone density diagnosis date.

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

There are flags for Comorbid Screening for Malignancy, Comorbid encounter for immunization, Comorbid Encounter for other exam, indicating whether the patient was seen for any of these other reasons during the treatment. Also there are flags for specific and general comorbid conditions: Comorbid other long term drug therapy, Comorbid Dorsalgia, Comorbid history of other diseases or conditions (according to a list), Comorbid history of other disorders of bone density or structure, Comorbid disorder of lipoprotein metabolism or other lipidemias, Comorbid Osteoporosis without current pathological fracture, Comorbid personal history of malignant neoplasm (history of cancer), or Comorbid Gastro-esophageal reflux disease. These are individually flagged if present with 'Y' (otherwise 'N').

There are individual flags for concomitant therapies as follows: Concomitant Cholesterol and Trigliceride Regulating Preparations, or Concomitant Narcotics, or Concomitant Systemic Corticosteroids 'Plain', or Concomitant Anti-depressants or mood stabilizers, or Concomitant Fluoroquinolones, or Concomitant Cephalosporins, Concomitant Macrolides, Concomitant Broad Spectrum Penicillins, Concomitant Anaesthetics General, or Concomitant Viral Vaccines.

There are individual flags for risks for or from the following: Type 1 Insulin Dependent Diabetes, Osteogenesis Imperfecta, Rheumatoid Arthritis, Untreated Chronic Hyperthyroidism, Untreated Chronic Hypogonadism, Untreated Early Menopause, History of patient's parent had fractured hip, Risk from smoking tobacco, Risk from Chronic malnutrition or malabsorption, or Chronic Liver Disease, or Family history of Osteoporosis, or Low Calcium intake, or Vitamin D insufficiency, or 'Risk of Poor Health Frailty', Excessive thinness, Hysterectomy or Oophorectomy, Estrogen Deficiency, or Risk from Immobilization, Recurring Falls, and finally a Count of Risks which sums the 'Y' responses from the above.

4. Comment on types of data available for analysis

   Most of the data is represented as objects; there are two integer columns, the count of risks and a Dexa frequency column. Age of the patient is represented by objects which are buckets with four possible entries: <55, 55-65, 65-75, >75. The target variable has possible entries "Persistent" and "Non-Persistent."

   There are 69 columns, with much of this data in need of transformation into formats that will work for running the machine learning classification programs. Race will need to be converted to multiple columns using dummy variables. Age will need to have another column added where the possible ages are denoted in a manner so the ranking is incorporated.

Note another issue is that the target variable Drug Persistency somewhat unbalanced, with 1289 Persistent to 2135 Non-Persistent, which will be addressed when building a model for classification.

5. What are the problems in the data

   Problems in the data fall into two categories.

   General problems with the data, requiring attention before applying the machine learning programs for classification include:

   The target variable is somewhat unbalanced (not evenly divided between persistent and non-persistent). Also many columns appear dependent on each other (different ways of presenting similar or the same information) which is an issue for models requiring predictor variables to be independent. Another issue is that the data is primarily categorical – which will require some preliminary adjustments in the presentation of this data before the algorithms can be applied. Some ordered data (such as age groupings) are not coded in a way that accounts for the ordering.

   Specific problems included the following:

   One column had a problematic column name which included a comma. ''Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx' needed to have the comma removed for some other code to work.

   One observation belonging in speciality OBSTETRICS & GYNECOLOGY was miscoded as OBSTETRICS & GYNECOLOGY plus other garbage within the speciality label; this had to be corrected so that observation would be properly included when the ML classification is applied.

Github repo link for this PDF and other project files:

https://github.com/DataDynamosTeam/W8_Deliverables