Project: Data Science :: Persistency of a Drug

Week 8: Deliverables

Presented by: Sammy Suliman, Olivia Foster, and Tahsin Azad

Table of Contents:

Team Member Details:

Group Name: Call it Version 1.0

Member 1:

- **Name:** Sammy Suliman
- **Email:** sammysuliman@gmail.com
- **Country:** USA
- **College/Company:** UC Santa Barbara
- **Specialization:** Data Science

Member 2:

- **Name:** Olivia Foster
- **Email:** livia.n.foster@gmail.com
- **Country:** USA
- **College/Company:** NT Logistics[1]
- **Specialization:** Data Science

Member 3:

- **Name:** Tahsin Azad
- **Email:** tahsinazad31@gmail.com
- **Country:** USA
- **College/Company:** University of California, Santa Barbara
- **Specialization:** Data Science

[1] Changed from Frisco Independent School District upon acceptance of job offer at NT Logistics.

Problem Statement:

One of the challenges faced by Pharmaceutical companies is the persistence of a drug (that is, the extent to which a patient will act in accordance with the prescribed time interval, and dose of a medication) as the physician prescribed it. In this problem, we will automate the process of classifying factors that determine the persistence of a drug through Machine Learning and Python.

Drug persistence is a task of classifying different disorders and a patient's medical history to determine the dose and length of dose. In order to train our model, we will need to classify risk factors, medical histories, and disorders. To do this, we will be using a dataset based on over 3000 patients' records.

Data Intake Report:

Name: Final Project -- Data Science:: Healthcare - Persistency of a drug:: Group Project
Report date: 1/17/2023
Internship Batch: LISUM16
Version:<1.0>
Data intake by: Olivia Foster
Data intake reviewer: Tahsin Azad
Data storage location:
https://github.com/LiviaNFoster/DataGlacierFinalProject/blob/main/Healthcare_dataset.xlsx

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	xlsx
Size of the data	922 KB

Proposed Approach:

- Unique Row identified using with the "Ptid" column, a unique ID for each patient
- No duplicates of Unique ID confirmed
- Assumptions
    - No false positives or errors during any of the testing
    - No dishonest and unbiased data recorded

Data Understanding:

The data covers a study that tracked health information about patients and their persistence to medication. While some of the data is categorical (such as region, gender, physician's speciality, etc.) most of the data follows a binary pattern: answering yes or no to condition-based questions.

In terms of raw data, the data would be considered to be perfect: very little cleaning required. There were no null values and everything followed some sort of convention.

The main concern with our preprocessing was to prepare it for numeric analysis.

There does not seem to be any outliers in the numeric data since most of it is binary. In terms of skewness, the data is skewed towards the following:

- Older sample population (greater than 65 years of age)
- Higher number of non-hispanic caucasian patients (reflective of the population of the United States)
- Patients are more likely to be from the Midwest or South
- Most patients are female
- Physicians tend to be general practitioners
- DEXA Scan Frequency is heavily skewed towards 0

As for the binary data, upon an initial first look, most of it seems to favor a 'No,' or 0 with a few executions.

Cleaning:

Null Values:

There were no null values.

Yes versus No:

We had over half of the data be defined as either Y or N, "Yes" or "No" respectively. We replaced Y and N with 1 and 0 respectively.

Renaming Columns:

Because many of the columns had long and difficult names or were not great descriptors, we went ahead and renamed many of them to make referencing easier and to add clarity.

Case sensitive:

Made all words lowercase to make referencing easier without having to worry about case sensitivity.

Github Repo link:

https://github.com/LiviaNFoster/DataGlacierFinalProject.git

-