Project: Data Science :: Persistency of a Drug

Week 9: 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
- **Specialization:** Data Science

Member 3:

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

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 Cleaning:

We began this process last week and continued to move forward with it.

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.

Acknowledgement of outliers:

In exploring our data, we discovered that because our data is mostly binary (such as above in the 'Yes versus No' section) that there were no significant outliers at this time.

'Age_Bucket' addressed:

In the original data, age was not presented in the same manner as in the description. Instead of producing quantifiable numbers for analysis, the values were grouped together as objects. To rectify this, each 'bucket' was replaced by a normal random output of values that fit the parameters of each 'bucket.'

Data Preprocessing:

Encoding Categorical Data:

The data had certain categorical values. These values were changed into integer format, which will later be used for the model. Target Encoding, Label Encoding, One-Hot Encoding and Smoothing were observed.

Key Errors fixed:

Small key errors fixed, allowing for clean, correct encoding.

Github Repo links:

General Repository:

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

Week 9 Jupyter Notebook:

Please note that this file includes all of our work, where we each worked on it one at a time.

https://github.com/LiviaNFoster/DataGlacierFinalProject/blob/main/Week9code.ipynb