Data Intake Report - Data Science Healthcare - Persistency of a Drug – Classification – Week 10

Group Name: <Igor Azevedo de Queiroz>
Email: <igor_queiroz17@yahoo.com.br>
Country: <Ireland>
College: <Dublin Business School>
Specialization: <Data Science - Classification>
Report date: <20/09/2021>
Internship Batch:<LISUM02>
Version:<1.0>
Data storage location: https://github.com/IgorQueiroz32/Data-Science-Healthcare---Persistency-of-a-Drug-Classification/tree/main/week%2010

Problem Description: <Data Science Healthcare - Persistency of a Drug - Classification>

Tabular data details:

Total number of observations	< 3424rows >
Total number of files	<1>
Total number of features	< 69 columns >
Base format of the file	<.ipynb, .csv, .txt, .png>
Size of the data	<891 in KB>

Healthcare - Persistency of a Drug - Classification

1. Business Description/ Problem.

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.

With an objective to gather insights on the factors that are impacting the persistency, it is necessary to build a classification for the given dataset, using the variable 'Persistency_Flag' as target variable and other attributes as prediction variables.

2. Business Understanding.

ABC it is a private pharma company. Due to the problem to the persistency of drug as per the physician prescription, a data science project is applied to predict the classification of 'Persistency_Flag' variable. In other words, based on the previously patients characteristics it is possible predict if futures patients will use the drugs during the role treatment or if they won't.

The object of this project is providing answer of the main questions made by the company's CEO, which are:

- What is the 'Persistency_Flag' classification for future patients?

The answer for those questions is presented in two different methods:

- A webapp with all necessary prediction attributes in order to predict the classification of the 'Persistency_Flag' for future patients.
- A dashboard with several hypotheses and insights to help the company CEO with future decisions.

The tools used for this project are: Python 3.8, Pycharm, Jupyter Notebook, Streamlit and Heroku.

3. Data Understanding.

There is 1 dataset provided:

https://www.kaggle.com/harbhajansingh21/persistent-vs-nonpersistent

healthcare_dataset.csv – this file includes characteristics of several patients.

Variables Description:

Here I'm describing the columns in detail:

Patient Details:

- **Patient ID:** Unique ID of each patient;
- **Persistency_Flag:** Flag indicating if a patient was persistent or not;
- **Age:** Age of the patient during their therapy;
- **Race:** Race of the patient from the patient table;
- **Region:** Region of the patient from the patient table;
- **Ethnicity:** Ethnicity of the patient from the patient table;
- **Gender:** Gender of the patient from the patient table;
- **IDN Indicator:** Flag indicating patients mapped to IDN;

Provider Attributes:

- **NTM - Physician Specialty:** Specialty of the HCP that prescribed the NTM Rx;

Clinical Factors:

- **NTM - T-Score:** T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate);
- **Change in T Score:** Change in Tscore before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
- **NTM - Risk Segment:** Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate);

- **Change in Risk Segment:** Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
- **NTM - Multiple Risk Factors:** Flag indicating if patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate);
- **NTM - Dexa Scan Frequency:** Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate);
- **NTM - Dexa Scan Recency:** Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable);
- **Dexa During Therapy:** Flag indicating if the patient had a Dexa Scan during their first continuous therapy;
- **NTM - Fragility Fracture Recency:** Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate);
- **Fragility Fracture During Therapy**: Flag indicating if the patient had fragility fracture during their first continuous therapy;
- **NTM - Glucocorticoid Recency:** Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx;
- **Glucocorticoid During Therapy:** Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy;

Disease/Treatment Factors:

- **NTM - Injectable Experience:** Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx;
- **NTM - Risk Factors:** Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx;
- **NTM - Comorbidity:** Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied;
- **NTM - Concomitancy:** Concomitant drugs recorded prior to starting with a therapy (within 365 days prior from first rxdate)
  Adherence: Adherence for the therapies.

4. Top 14 Data Insights.

Hypothesis 01: Female patients are more persistent of a drug than male.

TRUE: There are 1212 female persistent patients and 77 male persistent patients, which means, there are a difference between the genders of 1135 persistent patients.

Hypothesis 02: Patients from Northeast are more persistent of a drug than patients from South.

False: There are 98 persistent patients from Northeast and 494 persistent patients from South, which means, there are a difference between the regions of 396 persistent patients.

Hypothesis 03: Patients over 65 years of age are more persistent of a drug than patients 65 years of age or younger.

TRUE: There are 965 persistent patients over 65 of age and 324 persistent patients' young than 65 years of age, which means, there are a difference between the ages of 641 persistent patients.

Hypothesis 04: Caucasian patients, not Hispanic are more persistent of a drug than patients with different race and ethnicity.

TRUE: There are 1141 Caucasian Not Hispanic persistent patients and 148 persistent patients from different race and ethnicity, which means, there are a difference between race and ethnicity of 993 persistent patients.

Hypothesis 05: Patients mapped to IDN are more persistent of a drug than patient not mapped.

TRUE: There are 1043 persistent patients mapped to IDN and 246 persistent patients not mapped to IDN, which means, there are a difference between persistent patients mapped to IDN of 797 patients.

Hypothesis 06: Patients that received the drug prescription from General Practitioner Specialty are less persistent of a drug than patients that received the drug prescription from others Specialty.

FALSE: There are 494 persistent patients that received the drug prescription from General Practitioner Specialty and 795 persistent patients that received the drug prescription from others Specialty, there are a difference of 301 persistent patients.

Hypothesis 07: Patients that used Glucocorticoid and had a Fragility Fracture, before and during the therapy, are more persistent of a drug than patients that not used Glucocorticoid either had a Fragility Fracture, in any situation.

FALSE: There are 24 persistent patients that used Glucocorticoid and had a Fragility Fracture, before and during the therapy and 555 persistent patients that did not use Glucocorticoid neither had a Fragility Fracture, in any situation, which means, there are a difference between both status of 531 persistent patients.

Hypothesis 08: Patients that had Dexa Scan more than 0 times are more persistent of a drug than patients that had Dexa Scan 0 times.

FALSE : There are 716 persistent patients that had Dexa Scan more than 0 times and 573 persistent patients that had Dexa Scan 0 times, which means, there are a difference between both status of 185 persistent patients.

Hypothesis 09: Patients that presents VLR_LR as Risk Segment during the therapy and presents Unknown value in Risk Segment after the therapy are more persistent of a drug than patients that present other types of Risk Segment and Change in Risk Segment.

FALSE: There are 324 persistent patients that presents VLR_LR as Risk Segment during the therapy and presents Unknown value in Risk Segment after the therapy and 965 persistent patients that presents other types of Risk Segment and Change on Risk Segment, which means, there are a difference between both status of 641 persistent patients.

Hypothesis 10: Patients that presents T Socre of >-2.5 before the therapy and presents No Change T Score status after the therapy are less persistent of a drug than patients that presents other types of T Score and T Score status.

TRUE: There are 413 persistent patients that presents T Score of >-2.5 before the therapy and presents No Change T Score status after the therapy and 876 persistent patients that presents other types of T Score and Change on T Score, which means, there are a difference between both status of 463 persistent patients.

Hypothesis 11: Patients that presents the amount of risk factor higher than 1 are more persistent of a drug than patients that presents the amount of risk factor lower than 1.

FALSE: There are 783 persistent patients that do not presents the amount of risk factor higher than 1 and 506 persistent patients that presents the amount of risk factor higher than 1, which means, there are a difference between both status of 277 persistent patients.

Hypothesis 12: Patients adherent for therapies and that used drugs before the therapy are more persistent of a drug than patients in different situations related to both attributes.

TRUE There are 1100 persistent patients adherent for therapies and that used drugs before the therapy and 506 persistent patients, which means, there are differences between both status of 189 persistent patients.

Hypothesis 13: Patients that used more than 2 concomitancy drugs before the therapy are less persistent of a drug than patients that used more than 2 concomitancy drugs before the therapy.

FALSE: There are 690 persistent patients that used more than 2 concomitancy drugs before the therapy and 599 persistent patients that used less than 3 concomitancy drugs before the therapy, which means, there are a difference between both status of 91 persistent patients.

Hypothesis 14: Patients that presents more than 3 different types of comorbidity are more persistent of a drug than patients that present less than 3 different types of comorbidity.

True: There are 1055 persistent patients that presents more than 3 different types of comorbidity and 234 persistent patients that do not presents more than 3 different types of comorbidity, which means, there are a difference between both status of 821 persistent patients.

5. Final Recommendation.

Based on all the Insights created from this dataset, patients who have a higher probability of persisting of a drug for the entire treatment, have the following characteristics:

Female;

Patients older than 65 years;

Caucasian non-Hispanic Patients;

Patients mapped to IDN;

Patients that received the drug prescription from General Practitioner Specialty;

Patients who did not use Glucocorticoid and had Fragility Fracture;

Patients who had no Dexa Scan;

Patient's adherent for therapies and that used drugs before the therapy.

6. Project lifecycle along with deadline.

- Problem understanding
- Data Understanding
- Data Cleaning and Feature engineering
- Model Development
- Model Selection