Data Science Project

Healthcare - Persistency of a drug

Group Name: BetterHealth Analytics	
Member Details	Name: Enias Vontas Country: Greece Email: vondas100@gmail.com Specialization: Data Science

Problem Description

One of the challenges for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification. We have been provided with an Excel file containing some of the company's recorded data. The particular affliction that patients in this dataset were treated for is Nontuberculous Mycobacterial (NTM), which originates from a family of common organisms found in water and soil. This type of infection is rare and can affect people with damaged lungs, or with a weakened immune system. If diagnosed, a patient might need up to two years of treatment and could get infected again in the future.

The dataset contains the target variable 'Persistency_Flag' which indicates whether a patient was persistent with their medication or not. We would like to better understand the factors affecting this variable (our dependent variable). In order to do this, we have been provided with 67 other variables (our independent variables) which can be grouped in four buckets.

- Demographics: with variables such as Age, Race, Gender, etc for each patient.
- Provider Attributes: some information about the provider that wrote the prescription to the patient, with variables such as the Specialty of the Physician, a T-Score which is the result of a scan done to patients of this disease, etc.
- Clinical Factors: certain physiological attributes which could be associated with the disease, with variables such as Usage of Glucocorticoids, Frequency of a Dexa Scan etc.
- Disease/Treatment: Comorbidity factor, divided into two categories – Acute and Chronic and Concomitancy factor, i.e. concomitant drugs recorded prior to starting with the therapy.

All of the above parameters will be considered in our Machine Learning approach in order to better understand the factors affecting a patient's Medication Persistence and to more accurately classify a patient to one of the two categories of our 'Persistency_Flag' variable.

Business Understanding

It can be clear to imagine why a patient not receiving the whole dosage regimen that was prescribed to them can have unwanted results toward the treatment of that patient's illness. Another important unwanted result from this scenario is all the prescribed medication that goes to waste, from manufacturing it, all the way to distributing it to local pharmacies. So it is very important for drug companies and healthcare systems to provide the required medication to patients, but also as important for patients to be consistent with that prescribed medication, otherwise all that drug availability and expenditure would have been for nothing.

Before we talk more about our project from a business understanding, we would like to offer two definitions for a better overall understanding [1].

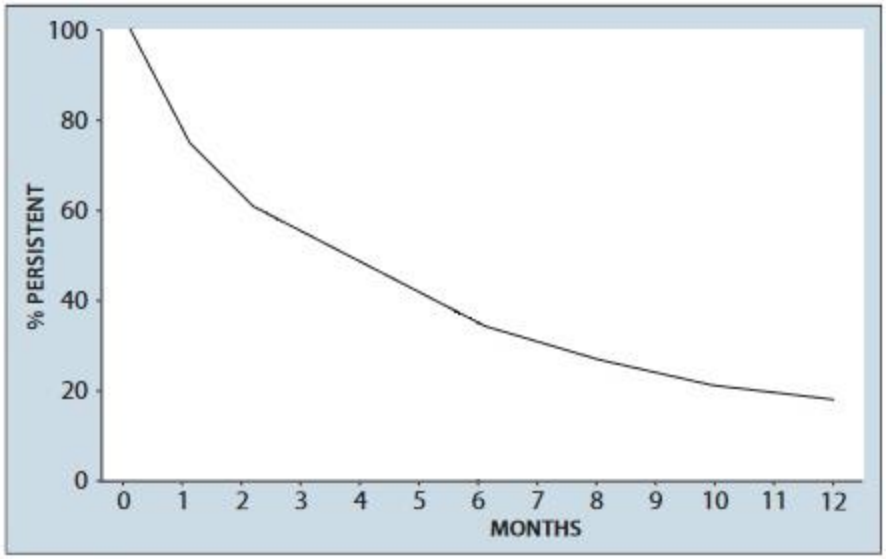
Medication Compliance (Adherence): refers to the degree of extent of conformity to the recommendation about day-to-day treatment by the provider with respect to timing, dosage, and frequency, or the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen.

Medication persistence: refers to the act of continuing the treatment for the prescribed duration, or the duration of time from initiation to discontinuation of therapy.

Inadequate medication compliance and persistence are age-old problems in the pharmaceutical business. When taken in varying degrees of deviation from the prescribed dosing regimen, medications have situation-specific alterations in benefit-risk ratios, either because of reduced benefits, increased risks, or both. Numerous studies have demonstrated that inadequate compliance and non persistence with prescribed medication regimens result in increased morbidity and mortality from a wide variety of illnesses, as well as increased healthcare costs. Factoring in actual compliance and persistence is central to an accurate assessment of effectiveness and cost-effectiveness of therapy.

This source of wasted US healthcare spending every year has the potential to reach even $300 billion, while also affecting pharmaceutical companies [2][3]. A lot of factors could contribute to a patient stopping, or altering their medication regimen, they could be a physician's time constraint, competing priorities for patients and shortcoming in follow-up initiatives. These factors need to be determined by healthcare providers, as well as pharmaceutical companies in order to address them and control them as much as possible.

It is known that the drug persistence curve has a downward trend and it tends to decrease at a decreasing rate as can be seen in the figure below, where we consider the drug persistence as a percentage, and observe in the duration of a year [4]:



We would like to determine the factors affecting a patient's persistence to their prescribed medication so that companies and doctors could then control for those factors when prescribing medication.

Project Lifecycle and Deadline

The project is due on 15th of August. It has been broken into various sections,which will be completed consecutively, as presented below:

- Problem Understanding
- Data Understanding
- Data Cleaning and Feature Engineering
- Model Development
- Model Selection
- Model Evaluation
- Report the accuracy, precision and recall of both the classes of target variable
- Report ROC-AUC
- Deploy the model
- Explain Challenges and Model Selection

As a first section, we will focus on the first two points, that are also underlined. As the project progresses, we will move forward with the other sections, as well as re-evaluate our findings if needed.

Data Intake Report

Name: Data Science Project – Persistency of a Drug
Report date: 25th of July 2021
Internship Batch: LISUM01
Version:<1.0>
Data intake by: Enias Vontas
Data intake reviewer:
Data storage location:
https://github.com/EniasVontas/Assignments/blob/main/Week7/Healthcare_dataset.xlsx

Tabular data details:

Total number of observations	3424
Total number of files	1, **Healthcare_dataset.xlsx**
Total number of features	69
Base format of the file	.xlsx
Size of the data	899 KB

No authorization was required to access the dataset and we assume that no mistake was done during the recording of our data points.

Proposed Approach:

- We will perform Data Preprocessing: check for missing data points, or null values, also check for outliers.
- We will perform Exploratory Data Analysis (EDA) and visualize some features and some relationships between variables, in order to better understand our dataset.
- We will create Dummy Variables, where necessary, in order to apply Machine Learning models.
- We will choose a model and evaluate it, report its ROC-AUC and deploy it.

References:

[1] https://www.sciencedirect.com/science/article/pii/S1098301510604950
[2] https://curanthealth.com/top-barriers-to-patient-persistence/
[3] https://www.pharmexec.com/view/top-barriers-patient-persistence
[4] https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/FS8.Lee-Fader-Hardie.pdf

Github Repo Link:
https://github.com/EniasVontas/Assignments/tree/main/Week7