Exploratory Data Analysis and Modelling Proposal

Persistency of a drug

Team : OpenML

Members : - Juan Carlos
- Laith Adi
- Gerson Orihuela
- Walquer Valles

Date : 15-May-2021

Agenda

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Business problem

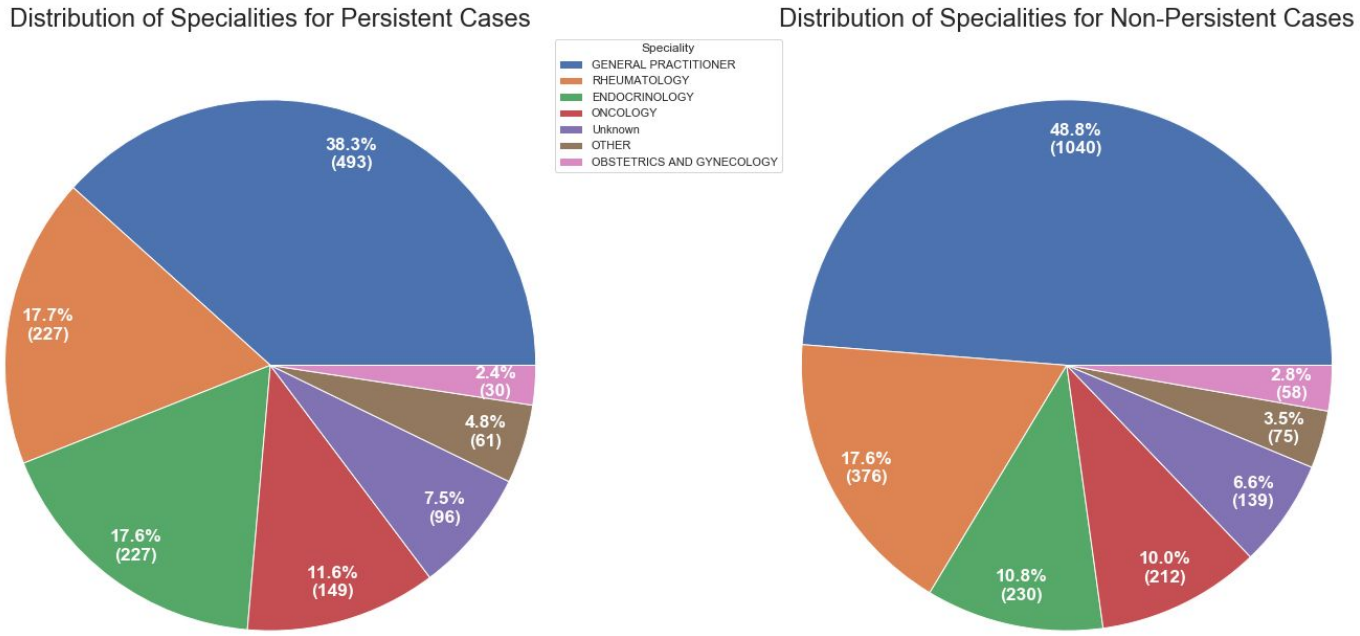
ABC Pharma contacted us to carry out an analysis in order to have a deeper understanding on the factors impacting the **persistence** of their drug. The aim is to know if a patient, based on his/her information, will follow the prescription of the physician and continue taking the drug for all the treatment time.

Approach

- 1 file was provided: Healthcare_dataset.xlsx
- The file contained information of 3, 424 patients. For each patient it has demographic information, clinical records, others diseases as risk factor information and also about their physicians specialty.
- The variables provided have been treated individually among the four members of the team.
- The **EDA** has been carried out following the same arrangement, but taking into account the whole dataset, so that potential insights have been drawn from the analysis.
- Four **model** proposals have been developed.

Clinical Factors

Does the speciality of the person who prescribed the drug have any effect on the persistent rate?



We see that both pie charts are pretty similar in distribution of frequency for each speciality. Thus, we can rule out the possibly that one of the factors that the drug is persistent or not is the speciality that perscribed the drug in the first place.

Clinical Factors

Does 'Ntm_Specialist_Flag' and 'Ntm_Speciality_Bucket' variables have useful information for the classification task?

Persistency_Flag	Non-Persistent	Persistent
Ntm_Specialist_Flag		
Others	0.680079	0.319921
Specialist	0.542877	0.457123

Persistency_Flag	Non-Persistent	Persistent
Ntm_Speciality_Bucket		
Endo/Onc/Uro	0.460894	0.539106
OB/GYN/Others/PCP/Unknown	0.679183	0.320817
Rheum	0.622517	0.377483

It seems Rheum flag in Ntm_Speciality_Bucket have some useful information.

]What about 'Gluco_Record_Prior_Ntm', 'Gluco_Record_During_Rx'?

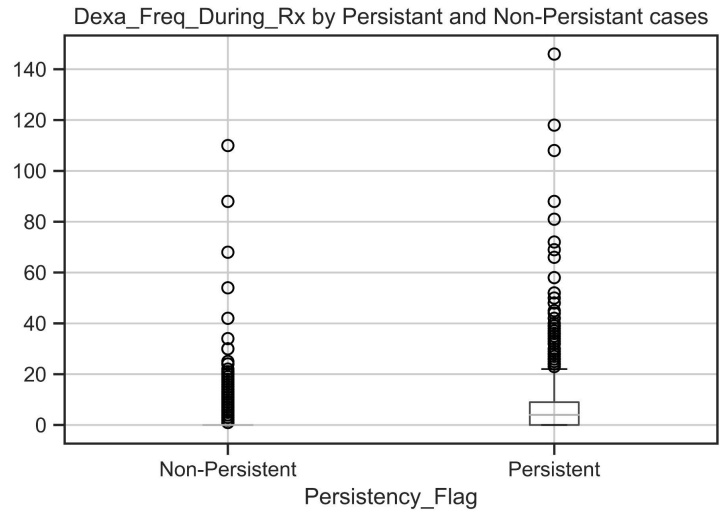
Persistency_Flag	Non-Persistent	Persistent
Gluco_Record_Prior_Ntm		
0	0.621993	0.378007
1	0.628571	0.371429

Persistency_Flag	Non-Persistent	Persistent
Gluco_Record_During_Rx		
0	0.68517	0.31483
1	0.45122	0.54878

Gluco_Record_During_Rx seems to be more useful than Gluco_Record_Prior_Ntm to predict the target

Clinical Factors

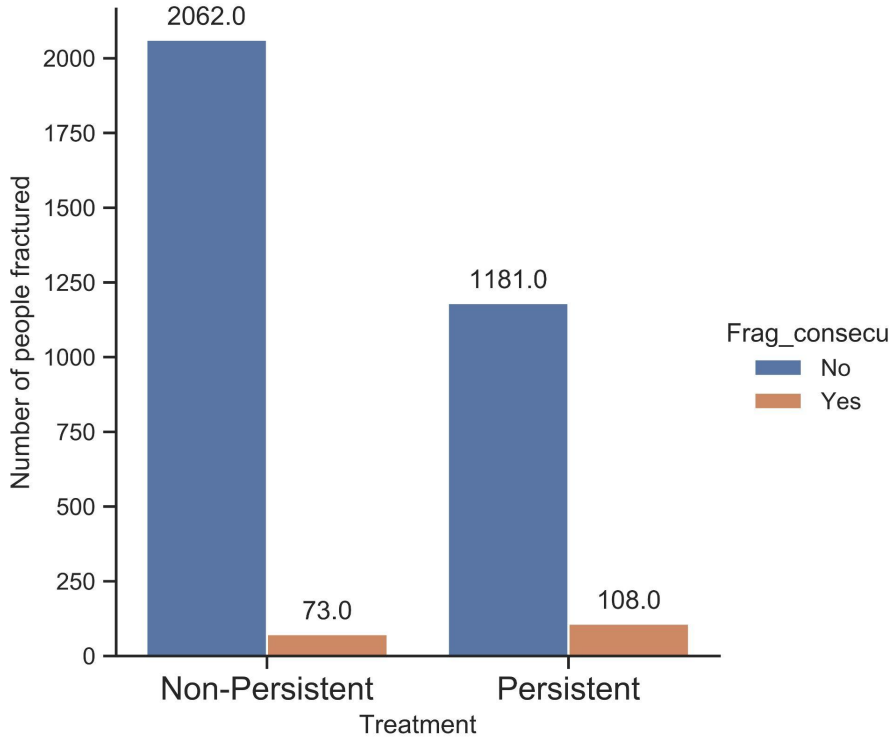
The distribution of Dexa_Freq_During_Rx numbers seems to be higher in the Persistent patients



Variables that are recorded during the treatment have more useful information for the classfication than others. It can be checked with the percentajes shown by Dexa_During_Rx variable.

Persistency_Flag	Non-Persistent	Persistent
Dexa_During_Rx		
0	0.769695	0.230305
1	0.235043	0.764957

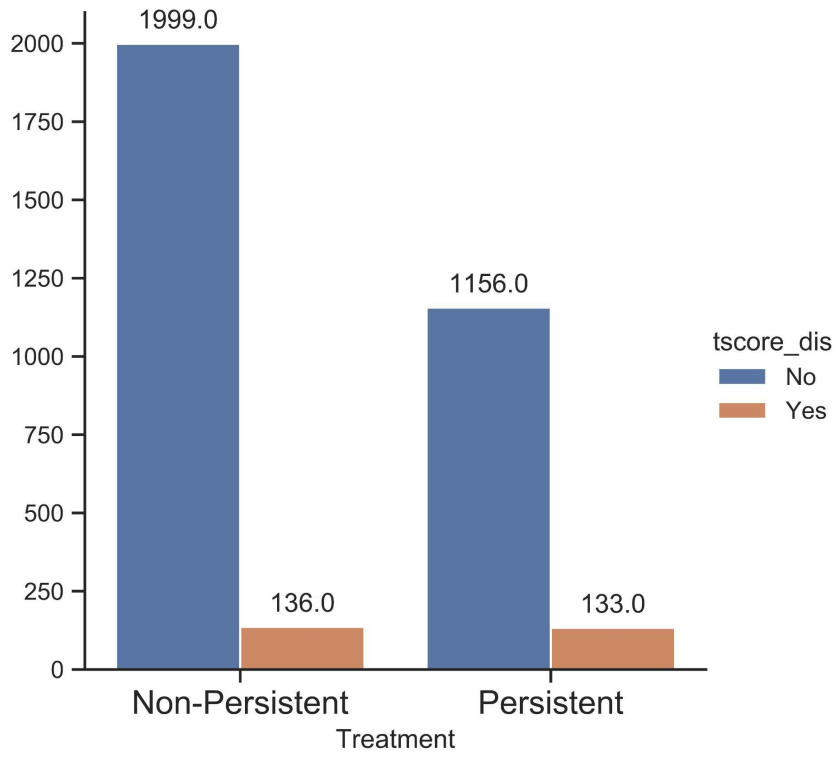
Fracture variable



Of the total number of patients, 8.38% of people were affected by the treatment, weakening their bones

- In the graph, 1289 people were persistent with their treatment, and of this 1289 only 108 people were affected weakening their bones.
- The count of people affected by the treatment is small and we can speculate that the treatment not affected considerably to the bones of the patients.

T-score variable



There is 10.31% of people with treatment who had a decrease in the t-score

- Then there is 90% approximately of people who maintained or improved their t-score.
- In conclusion the treatment is improving the t-score of the patients.

Risk, comorbidity and concomitant factors

Most of the patients already hold **comorbidity** factors, while holding **risk** factors is less common.

Some highlights:

- The main comorbidity factor is related to lipoproteins and metabolism (**cholesterol**).
- The main risk factor is deficiency in **vitamin D**.
- More than one third has been found to have taken **narcotics**.
- 99 % of our sample hold at least one risk, comorbidity and/or concomitant factor.

Risk, comorbidity and concomitant factors

There are some significant differences between genders:

- Women seem to be more affected by **vitamin D deficiencies**.
- More than twice as many women as men have passed as screening for **malignant neoplasms**.
- Four times as many men as women suffer from **Hypogonadism** (untreated).

Risk, comorbidity and concomitant factors

- As expected, patients **older than 65** are affected by the mentioned factors in a higher proportion.
- There are some risks and other factors that seem to be significantly higher in **South and West regions**. It might be interesting to find out about socioeconomic factors aside.
- There seem to be some remarkable differences between **Asian and other** races. They are probably due to cultural factors and other behaviours, like medical reviews on a more regular basis (this is just a hypothesis to be found out).

EDA Summary

The file contained information of 3, 424 patients. For each patient it has demographic information, clinical records, others diseases as risk factor information and also about their physicians specialty.

There are some significant differences between genders (vitamin D deficiencies, screening for malignant neoplasms, Hypogonadism).

Most of the patients already hold comorbidity factors, while holding risk factors is less common.

Patients older than 65 are affected by the mentioned factors in a higher proportion.

There seem to be some remarkable differences between Asian and other races.

Variables that are recorded during the treatment like Dexa_Freq_During_Rx, Dexa_During_Rx and Gluco_Record_During_Rx have more useful information for the classification than others.

Model proposals (technical review)

- **Support Vector Machines** algorithm to classify the persistence of patients (1 for positives and -1 for negatives). The whole dataset is composed of 3424 feature vectors of 83 dimensions, plus the target variable. A linear kernel has been used, obtaining an **accuracy of 83.5 %** over testing data (25 % out of the whole dataset).

- **Random Forest** algorithm for classification (1 for positives and 0 for negatives). The algorithm has 1000 estimators, max_depth of 10, obtaining an accuracy of 81% and AUC of 89% over testing data.

- **Decision Tree** algorithm (0 for Persistent and 1 for Non-persistent). The input is composed by 64 features with 3424 observations. The tree got best predictions with max depth of 1, obtaining an **accuracy of 76%** on test data.

- **Logistic Regression** algorithm for binary classification. The labels are the following: 0 for Non-Persistent and 1 for Persistent. Using GridSearchCV for optimization, the LR model uses 204 columns (after one-hot-encoding) to train. The f1_score obtain is **82%**.

Thank You