GROUP NAME: OpenML

MEMBER'S DETAILS:

Name	Email	Country	College/Company	Specialization
Juan Carlos	juanca.gutierrez@outlook.com	Spain	Everis	Data Science
Laith Adi	Laith_adi@hotmail.com	Canada	Laurier University	Data Science
Gerson Orihuela	yovanni.orihuela@gmail.com	Peru	Inspira IT	Data Science
Walquer Valles	wx.vr@outlook.com	Peru	KeepCoding	Data Science

PROBLEM DESCRIPTION: ABC Pharma contacted OpenML to carry out an analysis in order to have an understanding on the persistence of taking of a drug they released to market. 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. We have been provided with a dataset with patients' details.

GITHUB REPO LINK: https://github.com/jaycee-ds/Drug_Persistency_ABC_Pharma

EDA PERFORMED ON DATA

Questions¶

- What are the most common risk factors?

  We can easily see that most of the patients already hold comorbility factors, while holding risk factors is less common.

  Some highlights:

  - The main comorbility 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.

- What is the percentage of patients holding at least one factor?

  99 % of our sample hold at least one risk, comorbidity and/or concomitant factor.

- How do risk factors relate to demographics?

  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).

    - 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 behaviors, like medical reviews on a more regular basis (this is just a hypothesis to be found out).

- What is the proportion of patients who were affected by the treatment and had a fracture?

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

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

  The distributions of frequency for the target variable by specialty are pretty similar. Thus, we may rule out the possibly that one of the factors that the drug is persistent or not is the specialty that prescribed the drug in the first place

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

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

- What is the proportion of patients who were affected by the treatment, decreasing their t-score?

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

- Does the gender play a role in the chances of a drug being persistent or not?¶

  60.31% of males are flagged as non-persistent.
  62.48% of females are flagged as non-persistent.

  Seems about the same, which tells us both genders are experiencing the results from the drug when it comes to the persistency.

FINAL RECOMMENDATION

We recommend testing models for the prediction, taking into account also interpretable and simple models, such as decision tree. We will take the explainability of simple models to support the predictions of more complex models.

It seems that RX variables and comorbidity factors have high prediction power. Test for simple model just with these variables.