Data Glacier

Your Deep Learning Partner

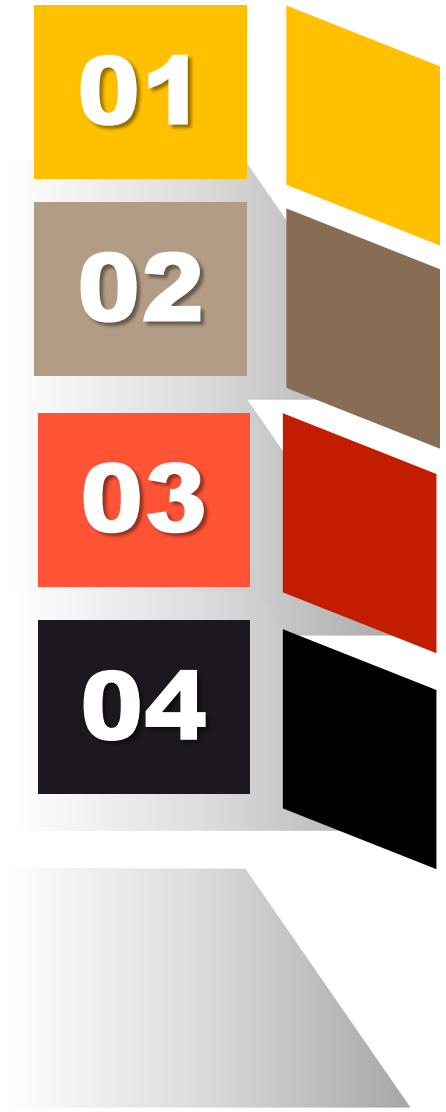
Healthcare - Persistency of a Drug

Groupe name: Ehiafarin
Country: Iran
Groupe member: Alireza Ehiaei
Email: arh.ehiaei@yahoo.com
Specialization: Data Science

Problem description & data understanding

Aug-2021

Table Of Contents



Problem Statement

Generally, drug efficiency assessment and so, clinical trials are daunting processes that may last several years. This can be due to different factors such as drug attributions or behavior of patients.

One common problem affecting drug efficieny is drug persistancy that is the amount of time that a patient remains on chronic drug therapy. Under this framework, patients are classified as either persistent or non persistent with medication therapy for some duration of time.

Problem Statement

Individuals who are persistent with therapy are continuous with their medication-taking behavior during a certain period. Persistent individuals refill their medications frequently and regularly. In contrast, no persistent individuals either have sporadic refilling practices or have discontinued refilling their medications completely.

It is important to be aware of medication (including drug administration) adherence (compliance) that is the extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen, and medication persistence (including drug persistence) that is the duration of time from initiation to discontinuation of therapy.

Problem Statement

Different factors can lead to drug persistancy such as:

- Diversity in the patient population different population segments react differently to drug prescribed interval.
- Race and Region of the patient
- Risk untreated chronic hypogonadism risk untreated chronic hyperthyroidism
- Risk untreated early menopause
- Risk Patient Parent Fractured Their Hip

Problem Statement

Therefore, It is hard to identify factors causing the drug persistence occurs. Because, the drug discovery process generally, and drug persistency specifically are multifaceted problems that in many cases they seem irrelevant.

However, there is little empirical evidence that the traditional methods have high accurecy in prediction of drug persistency of a patient and machine learning methods would be best approach to solve the problem. Before starting to use methods, we will have a look to the data we are going to use in this model.

Project lifecycle

- Understanding the data
- Modifying and Cleaning data
- Adding new variables using existing data to have some better insights to data
- Proposing machine learning models and testing hypothesizes
- Evaluating hypothesizes in order to predict persistency of drugs
- Report the accuracy of the model

Data understanding

There are 3424 rows and 69 coloumns in which except columns ‘Dexa Freq During Rx’ and ‘Count Of Risks’ which are integer, other coloumns have string type.

```
1 drug_Data.head()
```

	Ptid	Persistency_Flag	Gender	Race
0	P1	Persistent	Male	Caucasian
1	P2	Non-Persistent	Male	Asian
2	P3	Non-Persistent	Female	Other/Unknown
3	P4	Non-Persistent	Female	Caucasian
4	P5	Non-Persistent	Female	Caucasian

```
1 drug_Data.shape
```

```
(3424, 69)
```

```
1 type_dct = {str(k): list(v) for k, v in df.groupby(df.dtypes, axis=1)}
```

```
1 type_dct
```

```
{'int64': ['Dexa_Freq_During_Rx', 'Count_Of_Risks'],
 'object': ['Ptid',
  'Persistency_Flag',
  'Gender',
  'Race',
  'Ethnicity',
  'Region',
  'Age_Bucket',
  'Ntm_Speciality',
  'Ntm_Specialist_Flag',
  'Ntm_Speciality_Bucket',
  'Gluco_Record_Prior_Ntm',
  'Gluco_Record_During_Rx',
  'Dexa_During_Rx',
```

Data understanding

From 3424 cases, there are 1289 drug persistent patients and 2135 non persistent patients.

```
1 df_Persistent = df[df['Persistency_Flag'] == 'Persistent']
2 df_Non_Persistent = df[df['Persistency_Flag'] == 'Non-Persistent']
```

```
1 len(df_Persistent.iloc[:,1])
```

```
1289
```

```
1 len(df_Non_Persistent.iloc[:,1])
```

```
2135
```

Data understanding

From 3424 cases, there are just 194 man and 3230 women showing a significant difference in gender factor on drug persistency.

```
1 df_Male = df[df['Gender'] == 'Male']
2 df_Female = df[df['Gender'] == 'Female']
```

```
1 len(df_Male)
```

```
194
```

```
1 len(df_Female)
```

```
3230
```

Data understanding

The risk factor of region, has low effect on drug persistency.

```
risk_region = df.groupby(['Region']).agg({'Count_Of_Risks':'mean'}).reset_index()
risk_region
```

	Region	Count_Of_Risks
0	Midwest	1.10
1	Northeast	1.32
2	Other/Unknown	1.17
3	South	1.41
4	West	1.18

Data understanding

As it is shown in table below, there is no non value in data:

```
:  1 def missing_values_table(df):
   2     mis_val = df.isnull().sum()
   3     mis_val_percent = 100 * df.isnull().sum() / len(df)
   4     mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)
   5     mis_val_table_ren_columns = mis_val_table.rename(
   6     columns = {0 : 'Missing Values', 1 : '% of Total Values'})
   7     mis_val_table_ren_columns = mis_val_table_ren_columns[
   8         mis_val_table_ren_columns.iloc[:,1] != 0].sort_values(
   9     '% of Total Values', ascending=False).round(1)
  10     print ("Your selected dataframe has " + str(df.shape[1]) + " columns.\n"
  11         "There are " + str(mis_val_table_ren_columns.shape[0]) +
  12             " columns that have missing values.")
  13     return mis_val_table_ren_columns
```

```
:  1 missing_values_table(df)
```

```
Your selected dataframe has 69 columns.
There are 0 columns that have missing values.
```

:

Missing Values	% of Total Values

Data problem

Supposedly, there is no nan data, but there are some unknown values? For example, ‘Risk segment during rx’ column has 1497 unknown values.

```
new_df = df.Risk_Segment_During_Rx.str.split(expand=True).stack().value_counts().reset_index()

new_df.columns = ['Word', 'Frequency']

new_df
```

	Word	Frequency
0	Unknown	1497
1	HR_VHR	965
2	VLR_LR	962

Data problem

Threating with unknown data

In order to treat with unknown values we can use techniques such as:

- Exclude all unknown values. if the remaining data set is not imbalanced.
- Replace / group the unknown values with an appropriate value - e.g. replace missing values with the most populous label in the column.
- Targeting the column including unknown values and forecasting the label that each row (patient) likely should has instead of unknown value.

Thank You