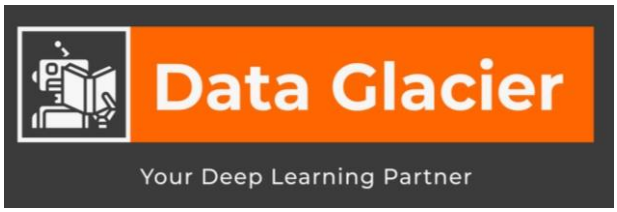
Data Science HealthCare Project Drug Persistance ABC Pharma



By : Sourour Cherif

Mail : Sourour.cherif@esprit .tn

Data Science Engineering Student Esprit

Tunisia

Problem description

ABC Pharma is looking for an automated way better than the traditional debilitating methods currently used to assess persistence of drugs as per the physician prescription, 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. We have been provided with a dataset which contains patients' details.

Business understanding

We will create a classification model as a solution that divides patients into categories depending on their information, to determine if a patient was persistent or not.

Our goal is to create a web application that might be used as an automated solution to this process of identification.

Data understanding

To fit any predictive model on a dataset, we need to understand the complexity of the dataset before deciding which predictive model to use to get optimal performance .

```
persist_pd.head()
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Risk_F
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	

5 rows × 69 columns

Type of data

```
persist_pd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 69 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Ptid                                                    3424 non-null   object
 1   Persistency_Flag                                        3424 non-null   object
 2   Gender                                                  3424 non-null   object
 3   Race                                                    3424 non-null   object
 4   Ethnicity                                               3424 non-null   object
 5   Region                                                  3424 non-null   object
 6   Age_Bucket                                              3424 non-null   object
 7   Ntm_Speciality                                          3424 non-null   object
 8   Ntm_Specialist_Flag                                     3424 non-null   object
 9   Ntm_Speciality_Bucket                                   3424 non-null   object
 10  Gluco_Record_Prior_Ntm                                  3424 non-null   object
 11  Gluco_Record_During_Rx                                  3424 non-null   object
 12  Dexa_Freq_During_Rx                                     3424 non-null   int64
 13  Dexa_During_Rx                                          3424 non-null   object
 14  Frag_Frac_Prior_Ntm                                     3424 non-null   object
 15  Frag_Frac_During_Rx                                     3424 non-null   object
 16  Risk_Segment_Prior_Ntm                                  3424 non-null   object
 17  Tscore_Bucket_Prior_Ntm                                 3424 non-null   object
 18  Risk_Segment_During_Rx                                  3424 non-null   object
 19  Tscore_Bucket_During_Rx                                 3424 non-null   object
 20  Change_T_Score                                          3424 non-null   object
 21  Change_Risk_Segment                                     3424 non-null   object
 22  Adherent_Flag                                           3424 non-null   object
 23  Idn_Indicator                                           3424 non-null   object
 24  Injectable_Experience_During_Rx                         3424 non-null   object
 25  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms  3424 non-null   object
 26  Comorb_Encounter_For_Immunization                       3424 non-null   object
 27  Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx  3424 non-null   object
 28  Comorb_Vitamin_D_Deficiency                             3424 non-null   object
 29  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified    3424 non-null   object
```

```
29  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified                3424 non-null   object
30  Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx  3424 non-null   object
31  Comorb_Long_Term_Current_Drug_Therapy                               3424 non-null   object
32  Comorb_Dorsalgia                                                    3424 non-null   object
33  Comorb_Personal_History_Of_Other_Diseases_And_Conditions            3424 non-null   object
34  Comorb_Other_Disorders_Of_Bone_Density_And_Structure                3424 non-null   object
35  Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias     3424 non-null   object
36  Comorb_Osteoporosis_without_current_pathological_fracture           3424 non-null   object
37  Comorb_Personal_history_of_malignant_neoplasm                       3424 non-null   object
38  Comorb_Gastro_esophageal_reflux_disease                             3424 non-null   object
39  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations         3424 non-null   object
40  Concom_Narcotics                                                    3424 non-null   object
41  Concom_Systemic_Corticosteroids_Plain                               3424 non-null   object
42  Concom_Anti_Depressants_And_Mood_Stabilisers                        3424 non-null   object
43  Concom_Fluoroquinolones                                             3424 non-null   object
44  Concom_Cephalosporins                                               3424 non-null   object
45  Concom_Macrolides_And_Similar_Types                                 3424 non-null   object
46  Concom_Broad_Spectrum_Penicillins                                   3424 non-null   object
47  Concom_Anaesthetics_General                                         3424 non-null   object
48  Concom_Viral_Vaccines                                               3424 non-null   object
49  Risk_Type_1_Insulin_Dependent_Diabetes                              3424 non-null   object
50  Risk_Osteogenesis_Imperfecta                                        3424 non-null   object
51  Risk_Rheumatoid_Arthritis                                           3424 non-null   object
52  Risk_Untreated_Chronic_Hyperthyroidism                              3424 non-null   object
53  Risk_Untreated_Chronic_Hypogonadism                                 3424 non-null   object
54  Risk_Untreated_Early_Menopause                                      3424 non-null   object
55  Risk_Patient_Parent_Fractured_Their_Hip                             3424 non-null   object
56  Risk_Smoking_Tobacco                                                3424 non-null   object
57  Risk_Chronic_Malnutrition_Or_Malabsorption                          3424 non-null   object
58  Risk_Chronic_Liver_Disease                                          3424 non-null   object
59  Risk_Family_History_Of_Osteoporosis                                 3424 non-null   object
60  Risk_Low_Calcium_Intake                                             3424 non-null   object
61  Risk_Vitamin_D_Insufficiency                                        3424 non-null   object
62  Risk_Poor_Health_Frailty                                            3424 non-null   object
63  Risk_Excessive_Thinness                                             3424 non-null   object
64  Risk_Hysterectomy_Oophorectomy                                      3424 non-null   object
65  Risk_Estrogen_Deficiency                                            3424 non-null   object
66  Risk_Immobilization                                                 3424 non-null   object
67  Risk_Recurring_Falls                                                3424 non-null   object
68  Count Of Risks                                                      3424 non-null   int64
```

```
persist_pd.describe()
```

	Dexa_Freq_During_Rx	Count_Of_Risks
count	3424.000000	3424.000000
mean	3.016063	1.239486
std	8.136545	1.094914
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	1.000000
75%	3.000000	2.000000
max	146.000000	7.000000

Let's see how we can find out unique elements in a column of the dataset.

```
persist_pd["Race"].unique()
```

```
array(['Caucasian', 'Asian', 'Other/Unknown', 'African American'],
      dtype=object)
```

```
persist_pd["Age_Bucket"].unique()
```

```
array(['>75', '55-65', '65-75', '<55'], dtype=object)
```

```
persist_pd["Ntm_Speciality"].unique()
```

```
array(['GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY',
       'ONCOLOGY', 'PATHOLOGY', 'OBSTETRICS AND GYNECOLOGY',
       'PSYCHIATRY AND NEUROLOGY', 'ORTHOPEDIC SURGERY',
       'PHYSICAL MEDICINE AND REHABILITATION',
       'SURGERY AND SURGICAL SPECIALTIES', 'PEDIATRICS',
       'PULMONARY MEDICINE', 'HEMATOLOGY & ONCOLOGY', 'UROLOGY',
       'PAIN MEDICINE', 'NEUROLOGY', 'RADIOLOGY', 'GASTROENTEROLOGY',
       'EMERGENCY MEDICINE', 'PODIATRY', 'OPHTHALMOLOGY',
       'OCCUPATIONAL MEDICINE', 'TRANSPLANT SURGERY', 'PLASTIC SURGERY',
       'CLINICAL NURSE SPECIALIST', 'OTOLARYNGOLOGY', 'HOSPITAL MEDICINE',
       'ORTHOPEDICS', 'NEPHROLOGY', 'GERIATRIC MEDICINE',
       'HOSPICE AND PALLIATIVE MEDICINE',
       'OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY',
       'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE'], dtype=object)
```

Data Problems

Data problems such as irrelevant columns , Null values , duplicates , skewed data ,outliers and many others may cause bad predictions …

So we need to check if we have one of them to know then how to overcome it .

Bucket	Columns	Information
Demographics	Gender	Type: Object No missing values # of unique values: 2 Values: "Male", "Female" Mode: Female (3230/3424 or 94.33%)
	Age_Bucket	Type: Object No missing values # of unique values: 4 Values: >75, 55-65, 65-75, <55 Mode: >75 (1439/3424 or 42.03%)
	Race	Type: Object Missing values: "Other/Unknown" (97/3424 or

		2.83%) # of unique values: 4 Values: [Caucasian, Asian, Other/Unknown, African American] Mode: “Caucasian” (3148/3424 or 91.94%)
	Region	Type: Object Missing values: “Other/Unknown” (60/3424 or 1.75%) # of unique values: 5 Values: West, Midwest, South, Other/Unknown, Northeast Mode: “Midwest” (1383/3424 or 40.39%)
	Ethnicity	Type: Object Missing values: “Unknown” (91/3424 or 2.66%) # of unique values: 3 Values: “Not Hispanic”, “Hispanic”, “Unknown” Mode: “Not Hispanic” (3235/3424 or 94.48%)
	Idn_Indicator	Type: Object No missing values # of unique values: 2 Values: “Y”, “N” Mode: “Y” (2557/3424 or 74.68%)

Provider Attributes	Ntm_Specialty	Type: Object Missing values: “Unknown” (310/3424 or 9.05%) # of unique values: 36 Values: 'GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY', 'ONCOLOGY', 'PATHOLOGY', 'OBSTETRICS AND GYNECOLOGY', 'PSYCHIATRY AND NEUROLOGY', 'ORTHOPEDIC SURGERY', 'PHYSICAL MEDICINE AND REHABILITATION', 'SURGERY AND SURGICAL SPECIALTIES', 'PEDIATRICS', 'PULMONARY MEDICINE', 'HEMATOLOGY & ONCOLOGY', 'UROLOGY', 'PAIN MEDICINE', 'NEUROLOGY', 'RADIOLOGY', 'GASTROENTEROLOGY', 'EMERGENCY MEDICINE', 'PODIATRY', 'OPHTHALMOLOGY', 'OCCUPATIONAL MEDICINE', 'TRANSPLANT SURGERY', 'PLASTIC SURGERY', 'CLINICAL NURSE SPECIALIST', 'OTOLARYNGOLOGY', 'HOSPITAL MEDICINE', 'ORTHOPEDICS', 'NEPHROLOGY', 'GERIATRIC MEDICINE', 'HOSPICE AND PALLIATIVE MEDICINE', 'OBSTETRICS & OBSTETRICS &
		GYNECOLOGY & OBSTETRICS & GYNECOLOGY', 'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE' Mode: “General Practitioner” (1535/3424 or 44.83%)
Clinical factors	Ntm_Speciality	Type: Object % missing values: 9.05% as Unknown # of unique values: 36 Values: 'GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY', […],'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE' Mode: GENERAL PRACTITIONER (1535/3424 or 44.83%)

	Ntm_Specialist_Fla g	Type: Object % missing values: 0% # of unique values: 2 Values: 'Others', 'Specialist' Mode: Others (2013/3424 or 58.79%)
	Ntm_Speciality_Buc ket	Type: Object % missing values: 0% # of unique values: 3 Values: 'OB/GYN/Others/PCP/Unknown', 'Endo/Onc/Uro', 'Rheum' Mode: OB/GYN/Others/PCP/Unknown (2104/3424 or 61.45%)
	Gluco_Record_Prior _Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2619/3424 or 76.49%)
	Gluco_Record_Duri ng_Rx	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2522/3424 or 73.66%)
	Dexa_Freq_During_ Rx	Type: Int64 % missing values: 0% # of unique values: 58 Values info: Mean 3.01, std 8.13 min 0.00 25% 0.00 50% 0.00 75% 3.00 max 146.0 Mode: 0 (2488/3424 or 72.66%)

	Dexa_During_Rx	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2488/3424 or 72.66%)

	Frag_Frac_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2872/3424 or 83.88%)
	Frag_Frac_During_Rx	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y'
	Risk_Segment_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: 'VLR_LR', 'HR_VHR'
	Tscore_Bucket_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: '>-2.5', '<=-2.5'
	Risk_Segment_During_Rx	Type: Object % missing values: 43% as Unknown # of unique values: 3 Values: 'VLR_LR', 'Unknown', 'HR_VHR'
	Tscore_Bucket_During_Rx	Type: Object % missing values: 43% as Unknown # of unique values: 3 Values: <=-2.5', 'Unknown', '>-2.5'
	Change_T_Score	Type: Object % missing values: 43% as Unknown # of unique values: 4 Values:'No change', 'Unknown', 'Worsened', 'Improved'
	Change_Risk_Segment	Type: Object % missing values: 65% as Unknown # of unique values: 4 Values: 'Unknown', 'No change', 'Worsened', 'Improved'
Disease/treatment factors	NTM - Injectable Experience	Type: Object No null values # of unique values: 2 Values: “Y”, “N”

	NTM - Risk Factors	Type: Object
		No null values # of unique values: 2 Values: “Y”, “N”
	NTM - Comorbidity	Type: Object No null values # of unique values: 2 Values: “Y”, “N”
	NTM - Concomitancy	ype: Object No null values # of unique values: 2 Values: “Y”, “N”
	Adherence	Type: Integer No null values # of unique values: 8 Values: 0, 1, 2, 3, 4, 5, 6, 7

Solutions :

Having duplicates in the dataset is not advisable and it often leads to overfitting.

```
persist_pd.duplicated().sum()
```

```
0
```

⇨ **There is no duplicates**

Missing Values

```
persist_pd.isnull().sum()
```

```
Ptid                                0
Persistency_Flag                    0
Gender                              0
Race                                0
Ethnicity                           0
                                   ..
Risk_Hysterectomy_Oophorectomy      0
Risk_Estrogen_Deficiency            0
Risk_Immobilization                 0
Risk_Recurring_Falls                0
Count_Of_Risks                      0
Length: 69, dtype: int64
```

⇨ **No missing Values**

Skewed data check

```
Q1 = persist_pd.quantile(0.25)
Q3 = persist_pd.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
Dexa_Freq_During_Rx    3.0
Count_Of_Risks         2.0
dtype: float64
```

```
import warnings
warnings.filterwarnings("ignore")
```

```
print(persist_pd < (Q1 - 1.5 * IQR)) |(persist_pd> (Q3 + 1.5 * IQR))
```

```
      Adherent_Flag  Age_Bucket  Change_Risk_Segment  Change_T_Score  \
0             False       False                False           False
1             False       False                False           False
2             False       False                False           False
3             False       False                False           False
4             False       False                False           False
...             ...         ...                  ...             ...
3419          False       False                False           False
3420          False       False                False           False
3421          False       False                False           False
3422          False       False                False           False
3423          False       False                False           False

      Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias  \
0                                                     False
1                                                     False
2                                                     False
3                                                     False
4                                                     False
```

Trying to :

- Remove no significant columns
- Handling with Skewed data
- Detect Outliers using different vizualisation tools and mathematical functions
  By calculated Z score