Data Science Internship at Data Glacier

Healthcare: Persistency of a drug (Data Science)

Week 10: **Deliverables**

Name: Chooladeva Piyasiri

University: National Institute of Business Management (NIBM)

Email: *chooladevapiyasiri@gmail.com*

Country: Sri Lanka

Specialization: Data Science

Batch Code: LISUM18

Date: 09 April 2023

Submitted to: Data Glacier

Table of Contents

Problem Statement

ABC is a pharmaceutical company. Medication persistency following a patient's doctor's prescription is a problem that ABC, a pharmaceutical company, wants to comprehend. Persistence of drugs is the duration of time a patient takes medication, from initiation to discontinuation of therapy. To automate this identifying process, this organization has contacted an analytics firm.

The analytics firm has to create a classification for the given dataset with the aim of gathering insights on the factors that are affecting the persistency.

Project Plan

Weeks	Date	Plan
Weeks 07	19 March 2023	Problem Statement, Data Collection, Data Intake Report
Weeks 08	26 March 2023	Data Understanding, Data Problems
Weeks 09	02 April 2023	Feature Extraction
Weeks 10	09 April 2023	Data Cleansing and Transformation
Weeks 11	16 April 2023	EDA Presentation and proposed modeling technique
Weeks 12	23 April 2023	Model Selection and Model Building/Dashboard
Weeks 13	30 April 2023	Final Project Report and Code

Data Understanding and Data Problems

The dataset contains 3423 observations with 26 columns and 68 features. The feature names and their data types are shown below.

```
Ptid                                                                 object
Persistency_Flag                                                     object
Gender                                                               object
Race                                                                 object
Ethnicity                                                            object
Region                                                               object
Age_Bucket                                                           object
Ntm_Speciality                                                       object
Ntm_Specialist_Flag                                                  object
Ntm_Speciality_Bucket                                                object
Gluco_Record_Prior_Ntm                                               object
Gluco_Record_During_Rx                                               object
Dexa_Freq_During_Rx                                                   int64
Dexa_During_Rx                                                       object
Frag_Frac_Prior_Ntm                                                  object
Frag_Frac_During_Rx                                                  object
Risk_Segment_Prior_Ntm                                               object
Tscore_Bucket_Prior_Ntm                                              object
Risk_Segment_During_Rx                                               object
Tscore_Bucket_During_Rx                                              object
Change_T_Score                                                       object
Change_Risk_Segment                                                  object
Adherent_Flag                                                        object
Idn_Indicator                                                        object
Injectable_Experience_During_Rx                                      object
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms               object
Comorb_Encounter_For_Immunization                                    object
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx      object
Comorb_Vitamin_D_Deficiency                                          object
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified                 object
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx   object
Comorb_Long_Term_Current_Drug_Therapy                                object
Comorb_Dorsalgia                                                     object
Comorb_Personal_History_Of_Other_Diseases_And_Conditions             object
Comorb_Other_Disorders_Of_Bone_Density_And_Structure                 object
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias      object
Comorb_Osteoporosis_without_current_pathological_fracture            object
Comorb_Personal_history_of_malignant_neoplasm                        object
Comorb_Gastro_esophageal_reflux_disease                              object
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations          object
Concom_Narcotics                                                     object
Concom_Systemic_Corticosteroids_Plain                                object
Concom_Anti_Depressants_And_Mood_Stabilisers                         object
Concom_Fluoroquinolones                                              object
Concom_Cephalosporins                                                object
Concom_Macrolides_And_Similar_Types                                  object
Concom_Broad_Spectrum_Penicillins                                    object
Concom_Anaesthetics_General                                          object
Concom_Viral_Vaccines                                                object
Risk_Type_1_Insulin_Dependent_Diabetes                               object
Risk_Osteogenesis_Imperfecta                                         object
Risk_Rheumatoid_Arthritis                                            object
Risk_Untreated_Chronic_Hyperthyroidism                               object
Risk_Untreated_Chronic_Hypogonadism                                  object
Risk_Untreated_Early_Menopause                                       object
Risk_Patient_Parent_Fractured_Their_Hip                              object
Risk_Smoking_Tobacco                                                 object
Risk_Chronic_Malnutrition_Or_Malabsorption                           object
Risk_Chronic_Liver_Disease                                           object
Risk_Family_History_Of_Osteoporosis                                  object
Risk_Low_Calcium_Intake                                              object
Risk_Vitamin_D_Insufficiency                                         object
Risk_Poor_Health_Frailty                                             object
Risk_Excessive_Thinness                                              object
Risk_Hysterectomy_Oophorectomy                                       object
Risk_Estrogen_Deficiency                                             object
Risk_Immobilization                                                  object
Risk_Recurring_Falls                                                 object
Count_Of_Risks                                                        int64
dtype: object
```

There are two numerical columns in the dataset, per the data type description. Which are:

- **Dexa_Freq_During_Rx**
- **Count_Of_Risks**

Analyzing data problems

1. NA values

The dataset does not contain any NA values.

2. Skewness & Kurtosis

Count_Of_Risks

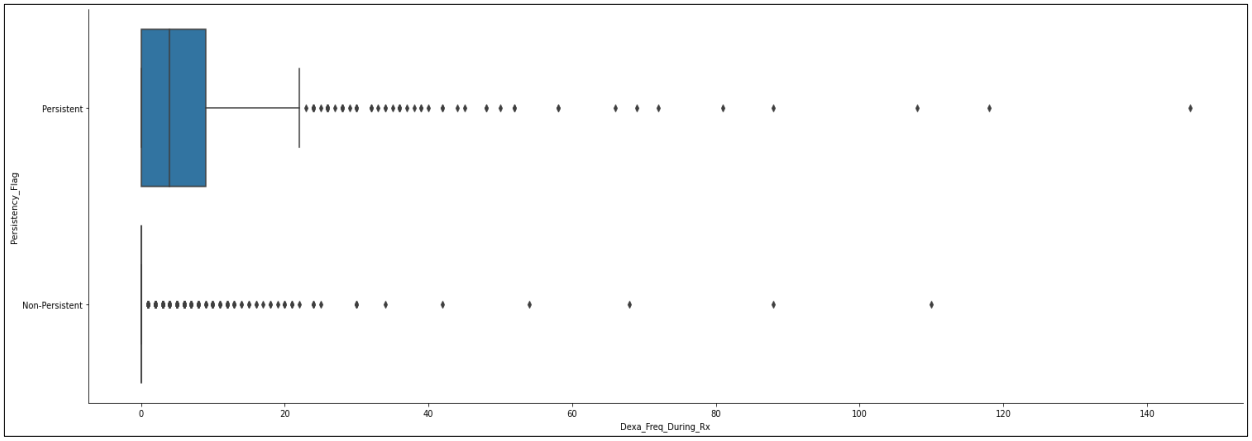
- ✓ The Count_Of_Risks distribution is moderately skewed. (0.879)
- ✓ The Count_Of_Risks distribution is Platykurtic (kurtosis <3). Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader. (0.900)

Dexa_Freq_During_Rx

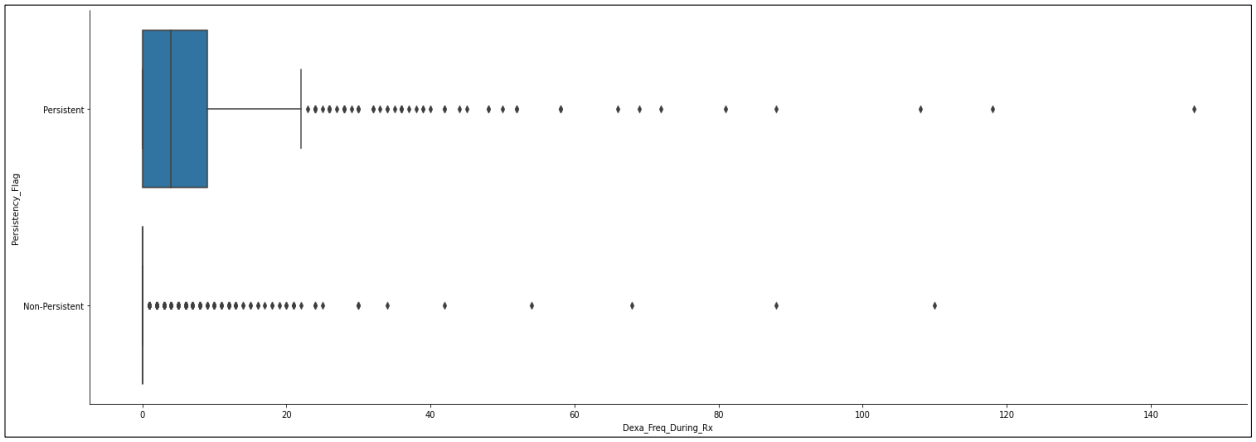
- ✓ The Dexa_Freq_During_Rx distribution is highly skewed. (6.808)
- ✓ The Dexa_Freq_During_Rx distribution is Leptokurtic (kurtosis >3). Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper. (74.758)

3. Outliers

Dexa_Freq_During_Rx



Count_Of_Risks



Both numerical features contain some outliers.

Data Cleaning and Transformation

As there are no NA values in the dataset, I started to detect outliers using different approaches, such as:

- **Boxplot Distribution**: A box plot is the visualization design that is typically depicted by quartiles and inter-quartiles and helps in defining the upper and lower limits beyond which any data lying will be considered outliers. The very purpose of this diagram is to identify outliers and discard them from the data series before making any further observations so that the conclusion drawn from the study gives more accurate results that are not influenced by any extremes or abnormal values.

- **Z-Score:** Z-scores are the number of standard deviations above and below the mean that each value falls. This score helps to understand if a data value is greater or smaller than the mean and how far away it is from the mean. If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

- **IQR Range:** The IQR is a statistical concept describing the spread of all data points within one quartile of the average, or the middle 50 percent range. The IQR is commonly used when people want to examine what the middle group of a population is doing. The IQR method is used to identify the outliers and set up a "fence" outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers.

Finally, to remove the detected outliers, I have approached two techniques: Capping and Trimming.

- **Capping**: Because there are many outliers in the Dexa_Freq_During_Rx column and because removing a large amount of data from the dataset is not a good idea, I used the

Capping method to handle the outliers. It won't remove them. Instead, it brings back those data points within the range that was specified according to the Z-Score value.

- **Trimming/Removing the outliers:** I pluck out all the outliers in the 'Count_Of_Risks' column using the filter condition in this technique.

After removing the outliers by calculating the IQR and removing data smaller or greater than two whiskers in the Count_Of_Risks column, the number of rows in the dataset was reduced from 3424 to 3401.

```
df_Healthcare.shape
```

```
(3401, 69)
```

I have dropped 'Ptid': Patient unique ID column as it's not necessary for modeling.

EDA

The transformed dataset's dimensions are (3401,68).

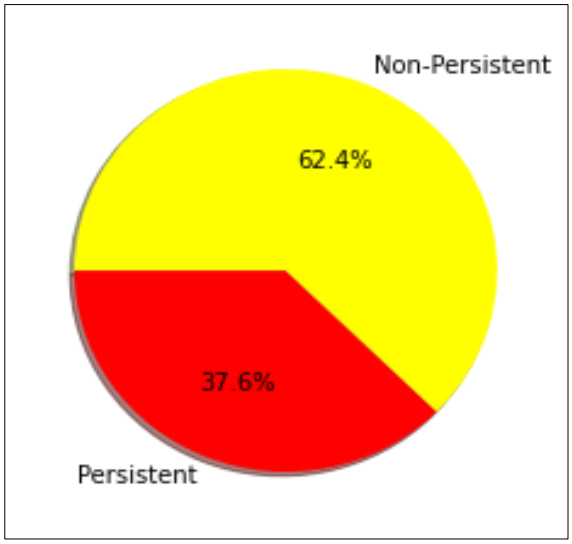
The Descriptive Statistics of the numerical features of the dataset are shown below.

	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

Here, I classified the EDA into two groups:

1. **Univariate Analysis**: Exploration and analysis of the target variable

The target variable of the dataset is "Persistency_Flag," - a flag indicating if a patient was persistent or not. There are 2135 non-persistent patients and 1289 persistent patients in this sample.



2. **Bivariate Analysis**: Exploration and analysis of both numerical and categorical variables.

In this section, I analyze:

- Demographical features according to the Persistency_Flag variable.
- Risk, Comorbidity, and Concomitant feature Analysis: whether the patients already have risk, comorbidity, or concomitant factors.
- Analysis of average Count_Of_Risks, Dexa_Freq_During_Rx according to the Gender, Age_Bucket and Persistency_Flag.
- Some other clinical and provider attributes according to the Persistency_Flag.
- Patients' Health Improvement Analysis according to Persistency_Flag.

Recommendation

For prediction, I recommend classification models that are also interpretable and simple.
Because Using pre-categorized training datasets, we can use a variety of algorithms to classify future datasets into categories.