Group project: Week 8 and Week 9 report

Team: FeatureTransformers

Members: Wangu Ndungu, Nikola Andrejić

Batch code: LISUM10

Contents

1 TEAM AND TEAM MEMBERS DETAILS

Team name: **FeatureTransformers**

Team members:

name: **Wangu Ndungu**

email: **nwangu349@gmail.com**

country: **Kenya**

college/Company: **Kenyatta University**

specialization: **Data Science**

name: **Nikola Andrejić**
email: **nikola.ing.nl@gmail.com**
country: **Serbia**
college/Company: **University of Niš**
specialization: **Data Science**

2 PROJECT DETAILS

Project title: **Drug persistence and Medical Adherence**

2.1 INTRODUCTION

Let's start things off by defining the terms medical adherence and drug persistence.

The U.S. Food and Drug Administration (FDA) terms medical adherence as: *"The extent to which patients take medication as prescribed by their doctors. This involves factors such as getting prescriptions filled, remembering to take medication on time, and understanding the directions."*

Drug persistence can be defined as the extent to which a patient acts in compliance to the prescribed interval, and dose of a dosing regimen.

What is the difference between these two terms? Adherence refers to the proportion of pills taken within a specific time interval and persistence refers to the continuing use (in time) of the prescribed therapy.

2.2 PROBLEM STATEMENT

According to the World Health Organisation, only 50-70% of patients adhere properly to prescribed drugs during therapy. This is especially true among those with long term medication. This worrying statistic is caused by various factors, for example: patient's condition or disease, their socio-economic status, confusion by the schedule, forgetting, discontinuing because they feel better, just to name a few. Medical non-adherence can lead to devastating consequences on one's health, especially those with chronic illnesses.

The purpose of this project is to study trends among patients in a sample and build a model that'll classify a new patient as Persistent or Non-Persistent.

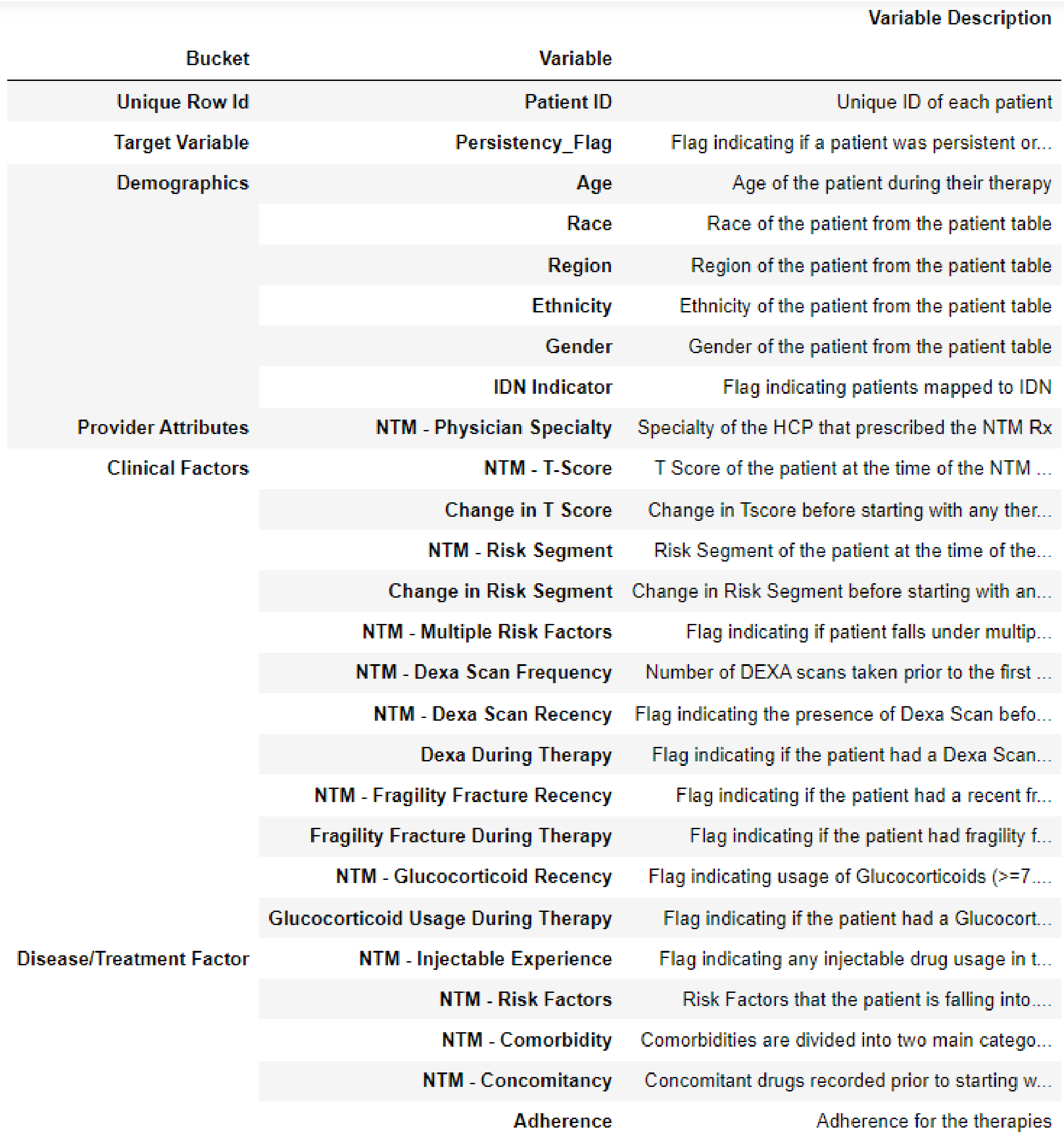
This project will give medical practitioners(especially pharmaceuticals) insight on which patients might require more rigorous follow-ups to ensure they will adhere to their prescriptions.

2.3 DATA UNDERSTANDING

Our data contains different features that describe the patient and a binary target variable that flags whether we have persistence or not.

3 THE DATASET

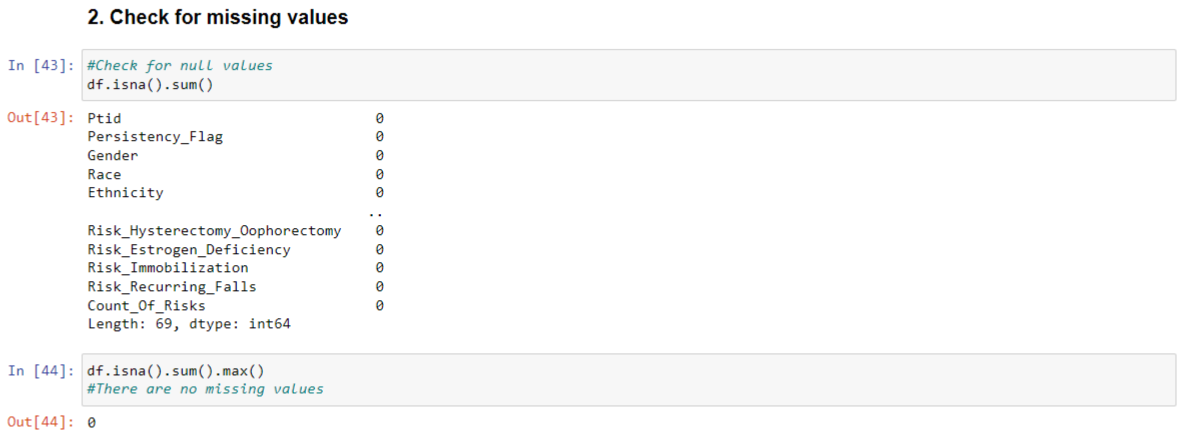
Without counting the unique patient identifier column, the original dataset has 68 columns, of which 67 are features and one called Persistency_Flag is the target. There are 3424 data points (rows) in our dataset. Upon inspection we see that we have only two numerical features, namely Count_Of_Risks and Dexa_Freq_During_Rx, while the other 65 are categorical. In the table below we present the column descriptions.



		Variable Description
Bucket	Variable	
Unique Row Id	Patient ID	Unique ID of each patient
Target Variable	Persistency_Flag	Flag indicating if a patient was persistent or...
Demographics	Age	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	Ethnicity	Ethnicity of the patient from the patient table
	Gender	Gender of the patient from the patient table
	IDN Indicator	Flag indicating patients mapped to IDN
Provider Attributes	NTM - Physician Specialty	Specialty of the HCP that prescribed the NTM Rx
Clinical Factors	NTM - T-Score	T Score of the patient at the time of the NTM ...
	Change in T Score	Change in Tscore before starting with any ther...
	NTM - Risk Segment	Risk Segment of the patient at the time of the...
	Change in Risk Segment	Change in Risk Segment before starting with an...
	NTM - Multiple Risk Factors	Flag indicating if patient falls under multip...
	NTM - Dexa Scan Frequency	Number of DEXA scans taken prior to the first ...
	NTM - Dexa Scan Recency	Flag indicating the presence of Dexa Scan befo...
	Dexa During Therapy	Flag indicating if the patient had a Dexa Scan...
	NTM - Fragility Fracture Recency	Flag indicating if the patient had a recent fr...
	Fragility Fracture During Therapy	Flag indicating if the patient had fragility f...
	NTM - Glucocorticoid Recency	Flag indicating usage of Glucocorticoids (>=7....
	Glucocorticoid Usage During Therapy	Flag indicating if the patient had a Glucocort...
Disease/Treatment Factor	NTM - Injectable Experience	Flag indicating any injectable drug usage in t...
	NTM - Risk Factors	Risk Factors that the patient is falling into....
	NTM - Comorbidity	Comorbidities are divided into two main catego...
	NTM - Concomitancy	Concomitant drugs recorded prior to starting w...
	Adherence	Adherence for the therapies

4 MISSING VALUES

Upon inspection, we see that there are no missing values in this dataset.



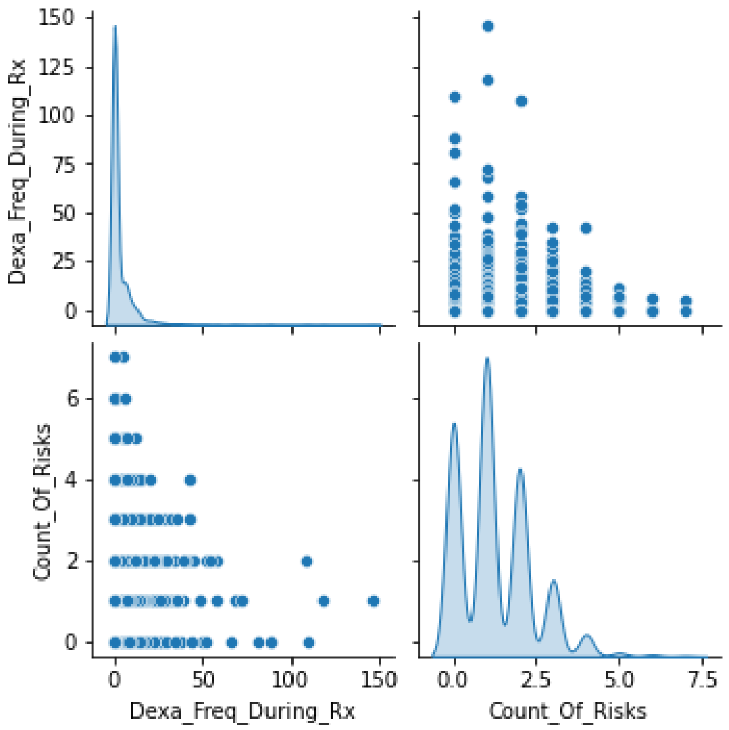
2. Check for missing values

```
In [43]: #Check for null values
         df.isna().sum()

Out[43]: 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

In [44]: df.isna().sum().max()
         #There are no missing values

Out[44]: 0
```

5 OUTLIERS

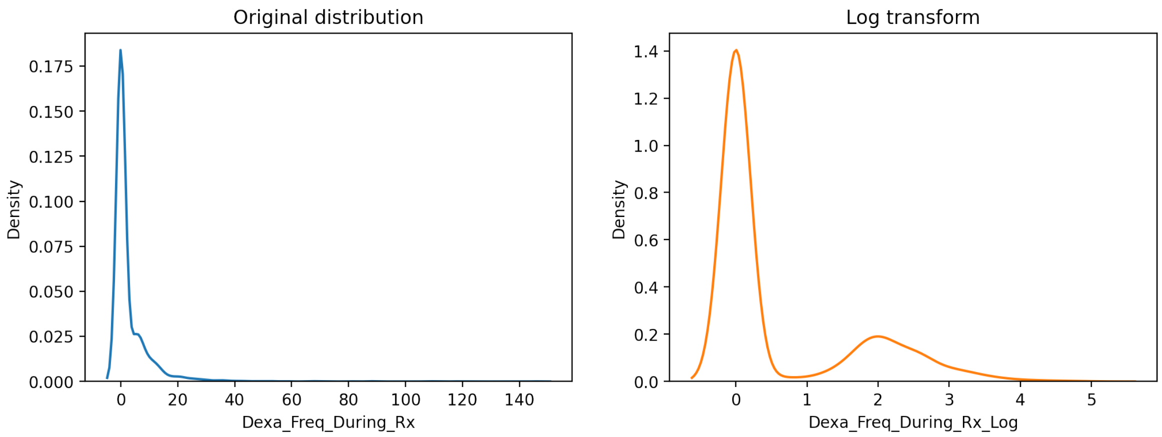
5.1 NUMERICAL COLUMNS

To detect outliers for the numerical features we first visually inspect the pairplot.

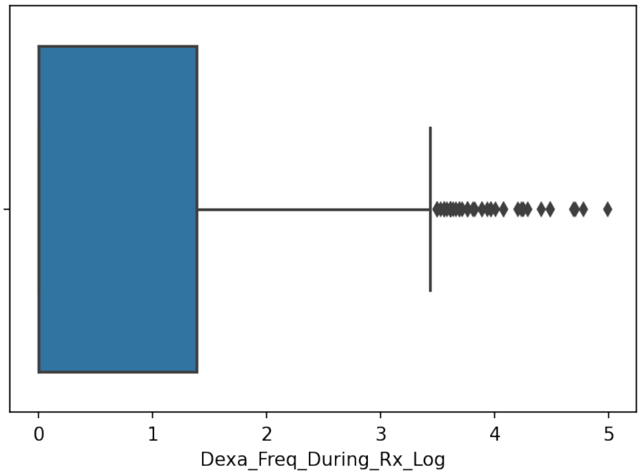


We notice that both of the distributions over numerical columns have positive skewness. In fact, we calculate that the skewness for Dexa_Freq_During_Rx is 6.8 and for Count_Of_Risks is 0.73.

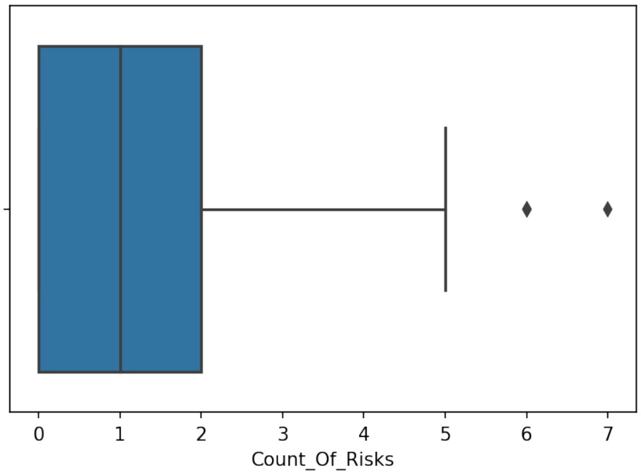
The second thing that we notice is that the distribution over Dexa_Freq_During_Rx is fat tailed and exponential-like. Because of this we choose to apply the log transform to it. In fact, we will apply the function $1 + \log x$ to it to avoid logarithmic divergence at $x = 0$. The original and the transformed distributions are shown in the figure below.



Next we make a box plot for the newly created Dexa_Freq_During_Rx_Log column.



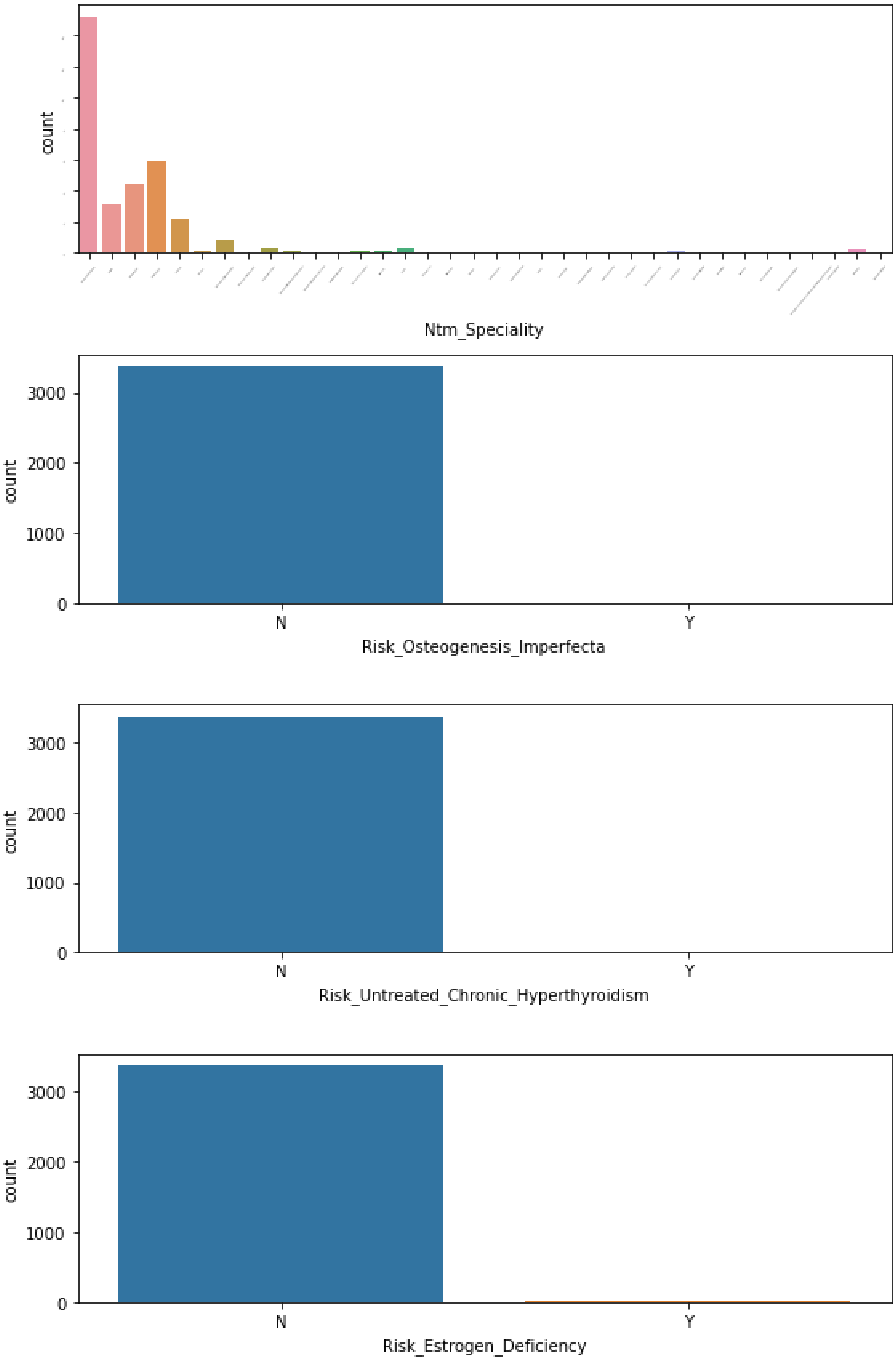
Only 42 points or about 1.22% of the dataset are classified as outliers in the figure above so we choose to simply drop those points from the dataset.



According to the box plot for the Count_Of_Risks column, we identify only two outliers and drop them from the dataset.

5.2 CATEGORICAL COLUMNS

In four categorical columns we detect categories that are populated with 10 (which is about 0.3% of the rows) or less data points. In other words, these categories are rare and can be treated as outliers, as they cause the categorical feature to be extremely imbalanced (see the figure below).



In total, there are 60 datapoints belonging to such rare categories. Since this is still a small percentage of the number of records, we can drop all of these datapoints. Since the last three features on the figure above are binary, after removing the outliers we will remain only with a constant feature that is uninformative, so we drop those three columns from the dataset as well.

In the end, the total of 3314 records remain in the dataset i.e. we have removed 110 points or about 3% of the data.