

Data Science Intern at Data Glacier

Project: Healthcare - Persistency of a drug

Week 8: Deliverables

Name: Chenyu Wang

Email: ryan.wang0924@gmail.com

University: University of Ottawa

Country: Canada

Specialization: Data Science

Batch code: LISUM 13:30

Submission date: Oct 26, 2022

Submitted to: Data Glacier

1. Problem description:

One of the challenges for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

2. Problem understanding:

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

3. Project lifecycle:

Weeks	Due Date	Task
Week 7	19 Oct, 2022	<ul style="list-style-type: none">● Problem understanding● data intake report● Data Understanding
Week 8	26 Oct, 2022	<ul style="list-style-type: none">● Data Cleaning and Feature engineering
Week 9	02 Nov, 2022	<ul style="list-style-type: none">● Model Development
Week 10	9 Nov, 2022	<ul style="list-style-type: none">● Model Selection● Model Evaluation
Week 11	16 Nov, 2022	<ul style="list-style-type: none">● Report the accuracy, precision and recall of both the class of target variable● Report ROC-AUC as well
Week 12	23 Nov, 2022	<ul style="list-style-type: none">● Deploy the model
Week 13	30 Nov, 2022	<ul style="list-style-type: none">● Final Submission (Report + Code + Presentation)

4. GitHub Repo link

https://github.com/chenyu-wang55/Data_Scientist_Intern_Data_Glacier/tree/main/Healthcare_project

5. Data Report

This dataset about the persistency of drug which contains 69 features and 3424 observations. The target feature in this dataset is ‘Persistency_Flag’ which classify the dataset as persistent and non-persistent.

Tabular data details: Healthcare Data

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	xlsx
Size of the data	1.8 MB

6. Data Understanding & Cleaning

6.1 Used the head function in Pandas package to display the top 5 data records.

healthcare_df.head()

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

5 rows x 69 columns

6.2 Check the type of columns

```
# data info & dtype
healthcare_df.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
```

6.3 check the missing value.

```
# check missing value
healthcare_df.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

[ ] healthcare_df[healthcare_df.isnull().any(axis=1)]

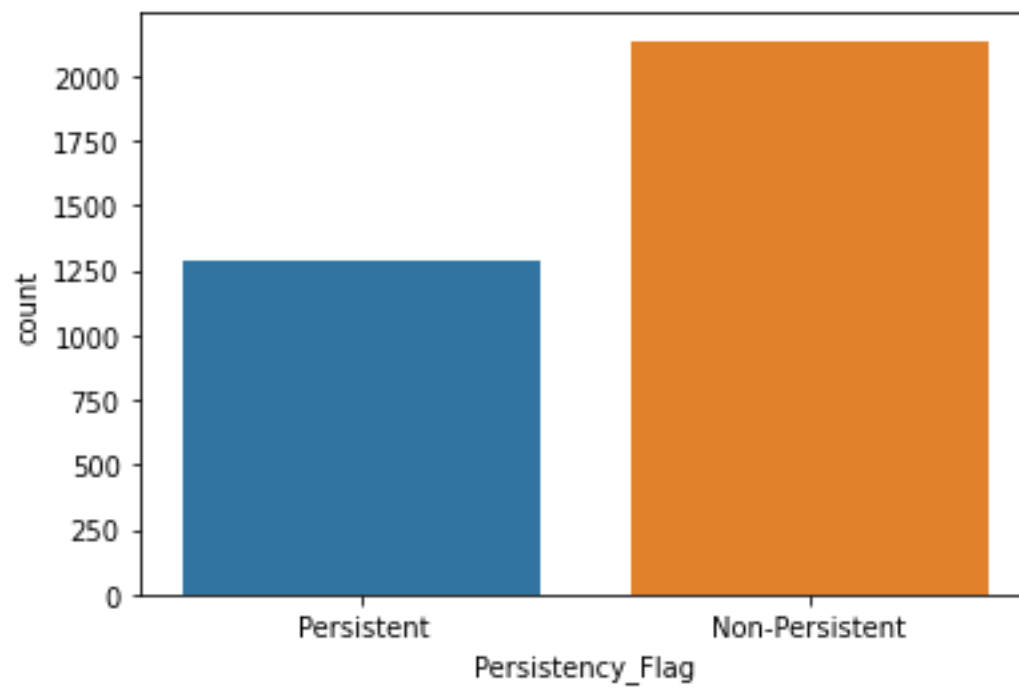
Ptid Persistency_Flag Gender Race Ethnicity Region Age_Bucket Ntm_Spe
0 rows x 69 columns
```

6.4 check the duplicate values.

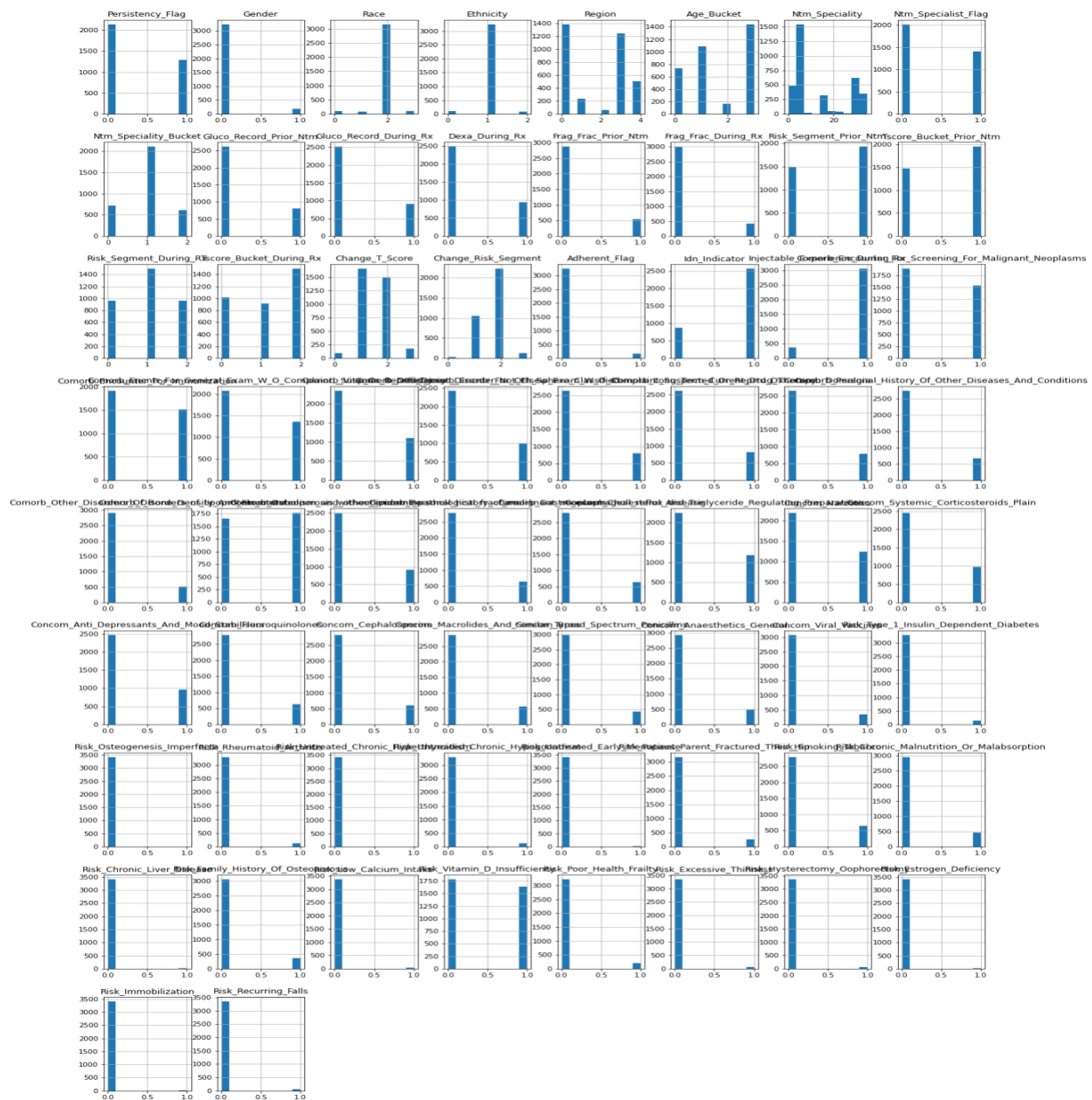
```
[ ] healthcare_df.duplicated().sum()
```

0

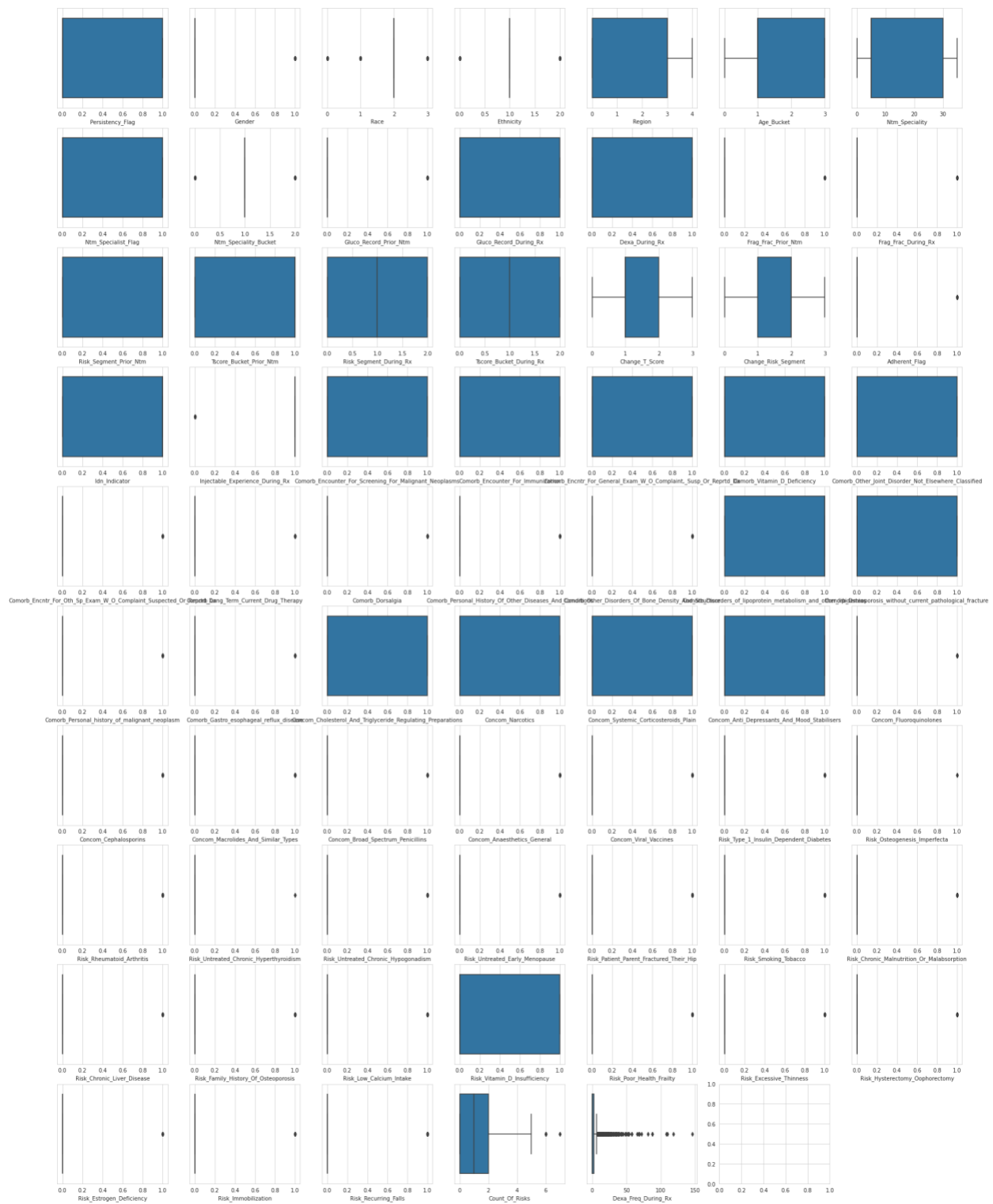
6.5 check whether the dataset is balanced or not.



6.6 Check the distribution of each feature.



6.7 check the outliers in each feature.



6.8 Conclusion:

1. There are not missing value and duplicate value.
2. 67 features are categorical values. I used label encoder covert them to numerical values.
3. The dataset is imbalanced. I used SMOTE technology to deal with this problem.
4. There are outliers in feature 'count_of_risks' and 'Dexa_freq_during_Rx'. I used Quantile Transformer transforms the features to follow a uniform or a normal distribution.