Team Name:				
Name	Email	Country	College/Company	Specialization
Seyedeh Marzieh Hosseini	shosseini@uni-potsdam.de	Germany	University of Potsdam	Data Science
Bao Khanh Nguyen	Nguyenkhanhbao8695@gmail.com	USA	American Energy Project	Data science
Guillermo Leija	leija.guillermo@gmail.com			Data Science
Zain Ul Haq		Germany		Data Science

Project Life Cycle

Tasks	**08/11/2021 Week 0**	**15/11/2021 Week 1**	**22/11/2021 Week 2**	**29/11/2021 Week 3**	**6/12/2021 Week 4**
Week 7					
Week 8					
Week 9					
Week 10					
Week 11					
Week 12					

Problem Description

ABC is a pharmaceutical business that wants to know the persistency of a drug after a physician has prescribed it for a patient. This company has approached an analytics firm to automate the identifying procedure. This analytics firm has entrusted our team with the task of developing a solution to automate the persistence of a medicine for the client ABC.

Business Understanding

One of the long-lasting business issues in the world of pharmaceutical companies is the persistency of drugs which can significantly affect the outcome of medical treatments. One of the important factors that is related to persistency is the adherence of the patient to the prescribed regimens, meaning if the patient is committed to the prescribed regimens or not. There is a lot of information about Non-Tuberculous Mycobacterial (NTM) infections. In fact, related studies show that around 50%-60% of the patients with different illnesses in US miss doses, take the wrong doses, or drop off treatment in the first year. Additionally, the illness, either chronic or acute can be related to the adherence and persistency of drugs.

GitHub Repository:

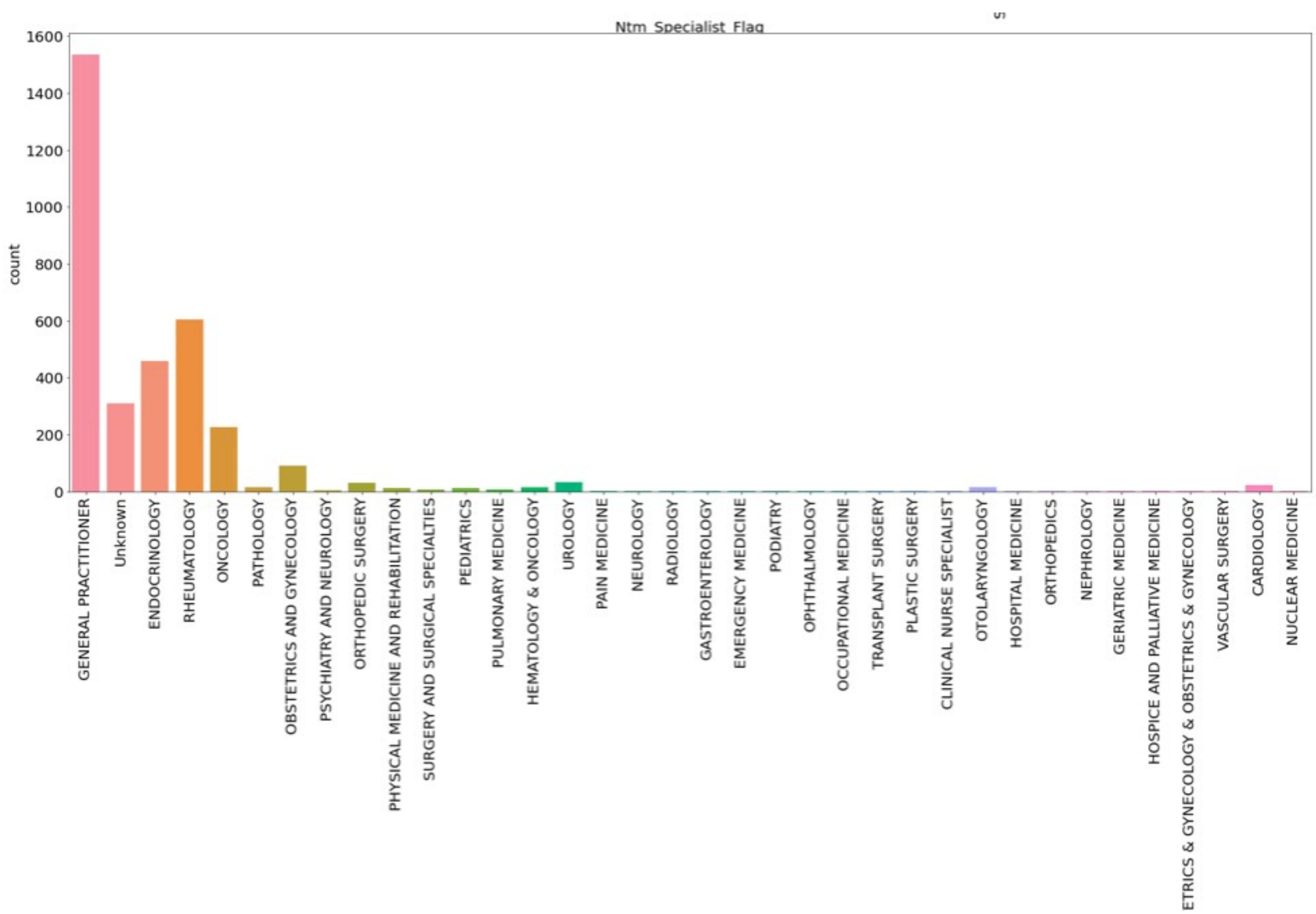
Project Link: https://github.com/Khanhbao8695/HealthCar_DS2021

Data Types

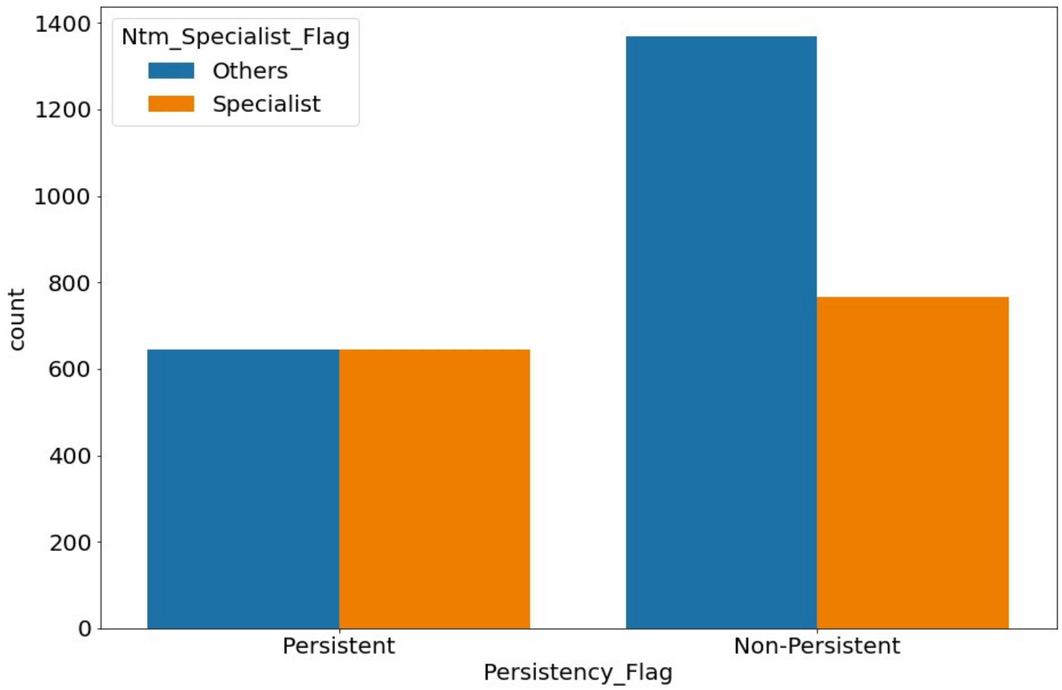
There are 69 features in this dataset and about 3424 rows. The majority of the data types in this dataset are "object" types with about 69 features and only 2 features are “inte64” data type.

Data Problems

The first problem with the dataset is the high number of categorical columns. Therefore it is important to drop few columns that does not seem to impact the persistency factor to high extent. One example would be the NTM_Speciality features which are three similar columns, Ntm_Speciality, Ntm_Specialist_Flag and Ntm_Speciality_Bucket. These columns are about the speciality of the person who prescribes the drug. Further investigation of feature Ntm_speciality shows the number of general practitioner is very high compared to other specialists and other specialits does not play that much of role.

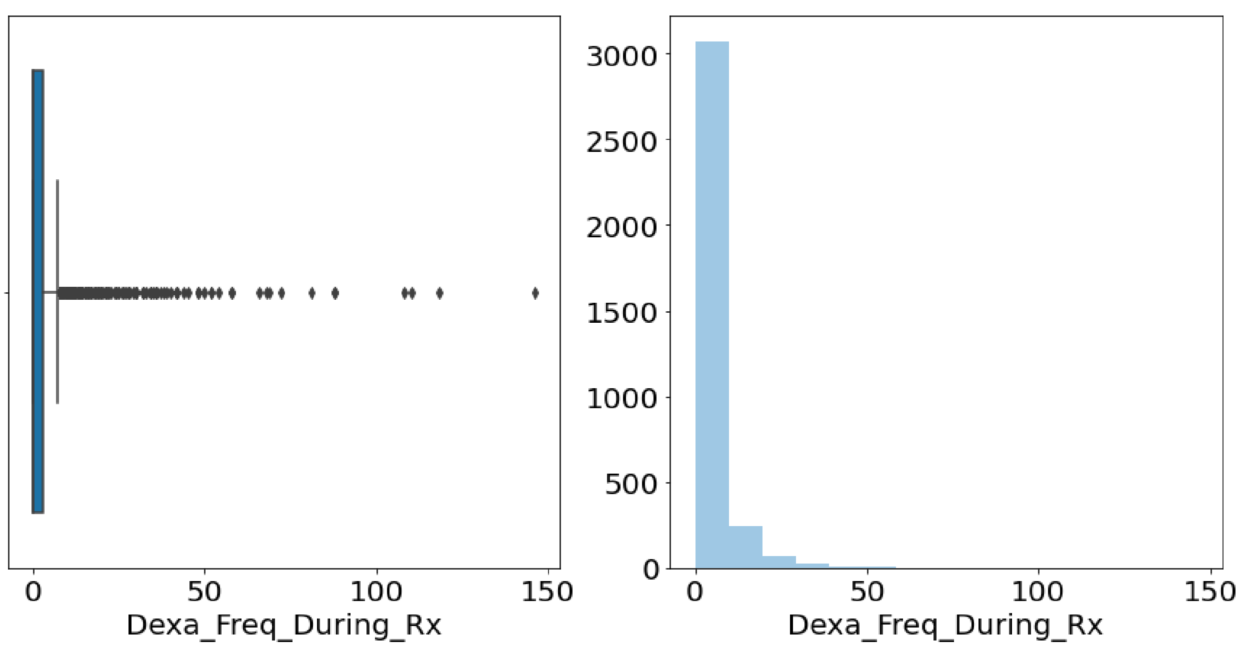


However, since there are three columns with similar information, its better to keep one. Additionally, its not clear if the speciality of the person who prescribed the medicine, is related to the persistency of drug. To investigate this, we plot the persistency and non_persistency of NTM_speciality flag. As can be seen in the second plot, the persistency is the same for the others and specialist flags. However, non-persistency is higher for other practioners than specialist. So to not lose additional information, we keep the NTM_speciality_flag .

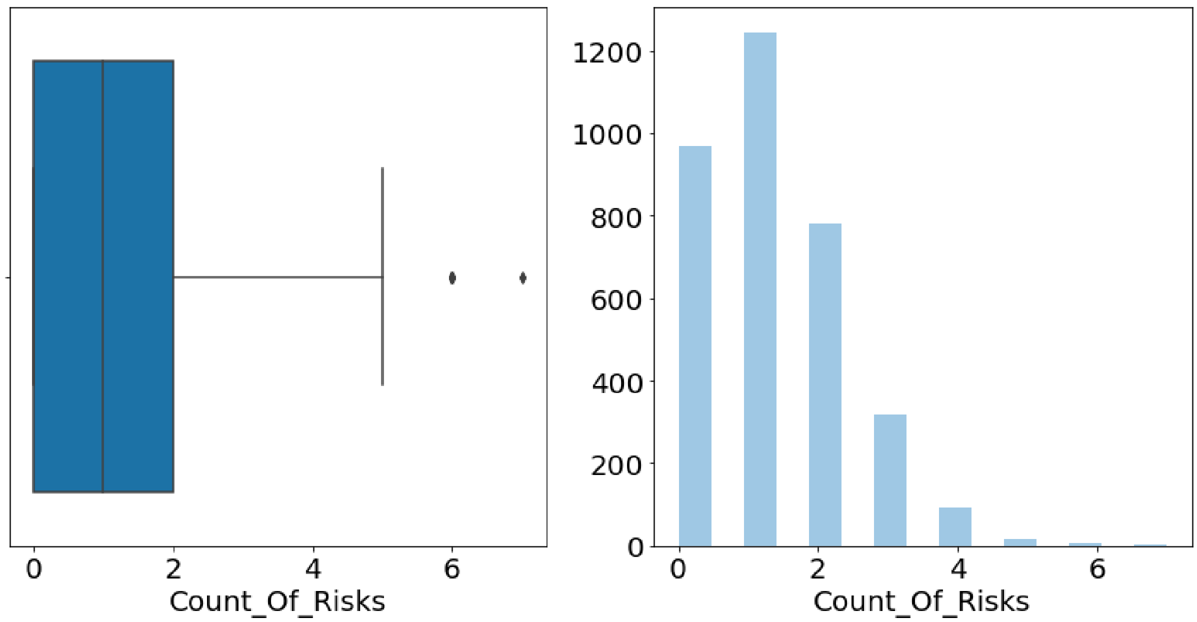


numerical variables, the two columns, Dexa_Freq_During_Rx and the column Count_Of_Risks, contains outliers and skewness as shown in the figures that need to be taken into account.

6.8087302112992285

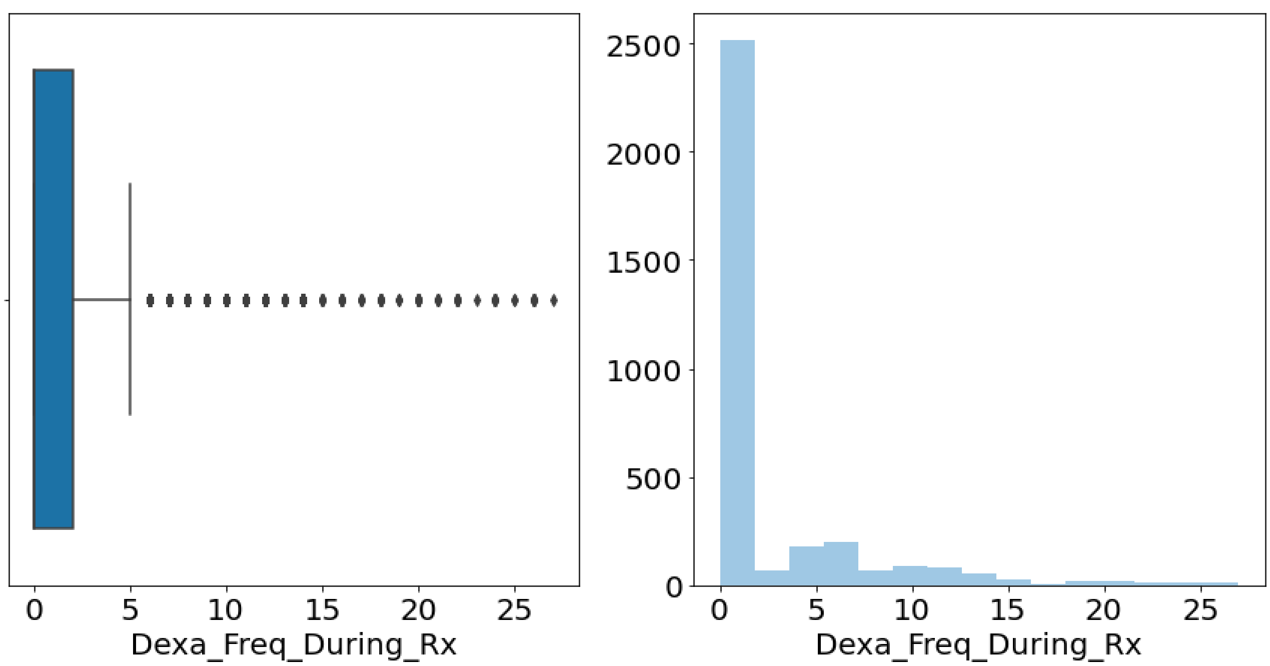


0.8797905232898707

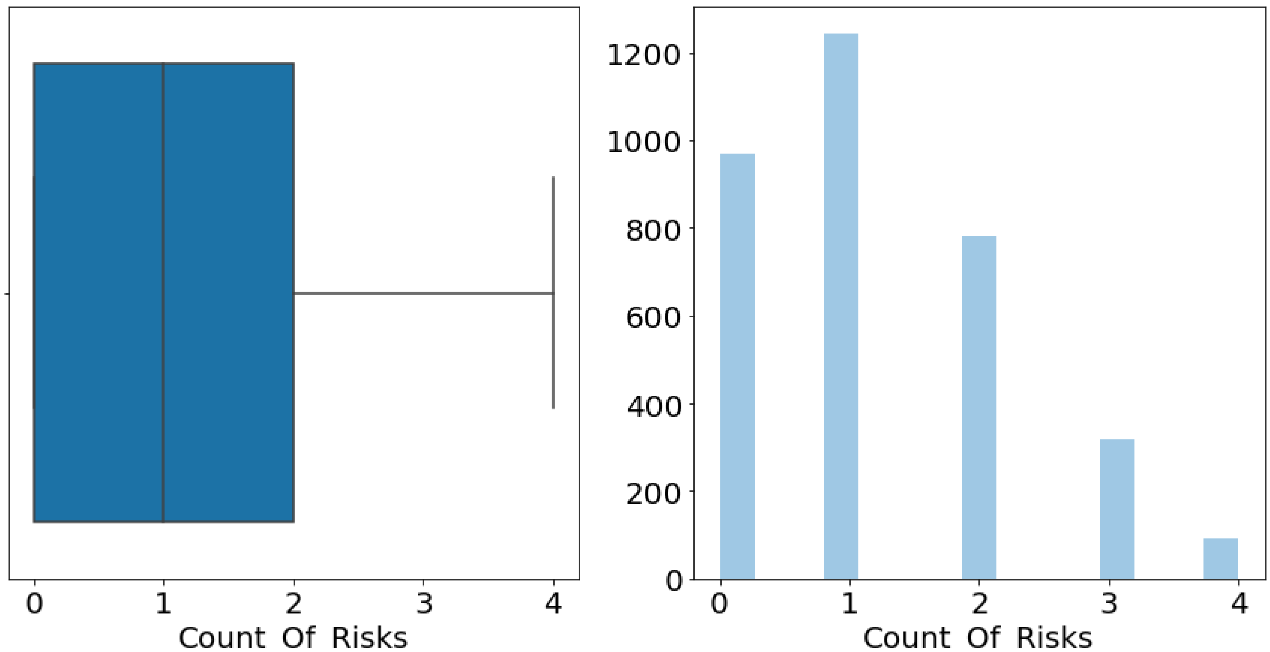


Remove Outliers

My first approach is to use Z scores to remove outliers from count of risks and Dexa_Freq_During_Rx columns. Z score finds the relationship of each data point with standard deviation and mean of the group of data points. So Z score rescales data and look for data points which are too far from zero. However, this method did not get ride of the all of outliers in the Dexa_Freq_During_Rx column. As shown in the figure, there are still few outliers left.



However, it worked good for the count_of_risks column as shown here.



The second approach I used to remove the outliers, is the IQR method. The interquartile range is calculated in much the same way as the range. All one find is subtract the first quartile from the third quartile: IQR = Q3 – Q1. The interquartile range shows how the data is spread about the median. Anything outside of the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR) is considered an outlier and should be eliminated.

I applied this method on both columns, and the filtered dataset reduced to the 2964 rows at the end.

One hat encoding for categorical variables

Among the 67 columns left, there are about 65 columns that are categorical variables. However, non of the columns contains ordinal values. Thus, we can use One-hat encoding to transform the categorical variables to numerical ones. I used the 'get.dummies' method.

t[39]:

	Ptid	Dexa_Freq_During_Rx	Count_Of_Risks	Persistency_Flag_Non-Persistent	Persistency_Flag_Persistent	Gender_Female	Gender_Male	Race_African American	Race_As
0	P1	0	0	0	1	0	1	0	
1	P2	0	0	1	0	0	1	0	
2	P3	0	2	1	0	1	0	0	
3	P4	0	1	1	0	1	0	0	
4	P5	0	1	1	0	1	0	0	
...	...	...	...	...	...	...	...	...	
3419	P3420	0	1	0	1	1	0	0	
3420	P3421	0	0	0	1	1	0	0	
3421	P3422	7	1	0	1	1	0	0	
3422	P3423	0	0	1	0	1	0	0	
3423	P3424	0	1	1	0	1	0	0	

2964 rows × 145 columns

Next Step

The next step would be to apply classification algorithms on the dataset.