Data Glacier

Your Deep Learning Partner

Final Project – Healthcare Data (Persistency of a Drug)

Virtual Internship

Name: Vincent Obiri

LISMU05

5th March 2022

Background – Healthcare Dataset

Business problem:

A pharmaceutical company approached us and asked whether we could automate the process of detecting persistency of drug use (these drugs are prescribed by doctors). We are provided with a dataset which contains many features, and a target variable ('Persistency_Flag'), which indicates whether a patient has been persistent.

The analysis involves:

- Data overview, Overview of outliers
- Investigating the features in the dataset and their relationships with Persistency_Flag
- Understand patient demographics
- Investigating the numerical variables (Dexa_Freq_During_Rx, Count_Of_Risks)
- Some comparison of data before and after transformation
- A correlation heatmap for the best variables

Plots on the following slides are created using data from either the original dataset or the modified dataset, as appropriate. Some variables' classes are lost after transformation, so it makes sense to use both datasets, since it may be useful to provide insights pertaining to these variables.

Overview of data and outliers

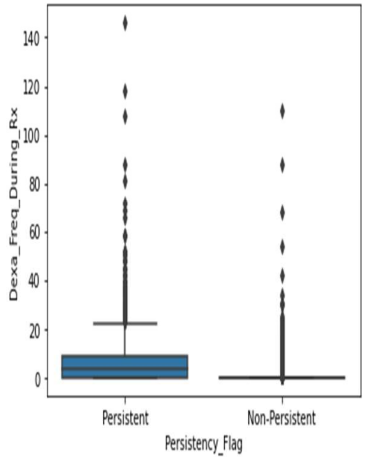
Overview:

- There are 69 features in the original dataset, and 3424 rows. As we converted variables to dummies, the number of variables increased. Eventually we chose the 6 best features. Outliers are present in the 2 numerical variables (Count_Of_Risks and Dexa_Freq_During_Rx). All rows with outliers were removed.

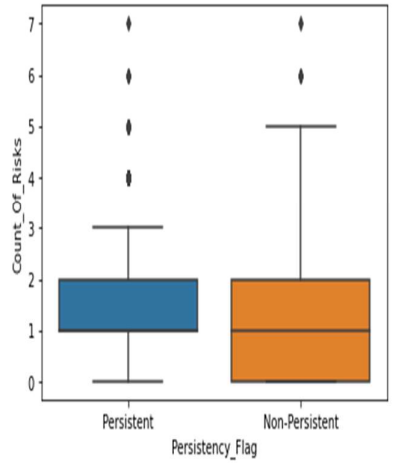
- Many classes were skewed. Those with heavy skew (e.g., Gender with 194 males and 3230 females in original dataset) won't be used during the modelling in upcoming weeks.

- The target class was only slightly skewed: 2135 non-persistent vs 1289 persistent in original dataset, 844 non-persistent vs 371 persistent in modified dataset. Unless otherwise stated, assume that the upcoming plots were produced with data from the original dataset.

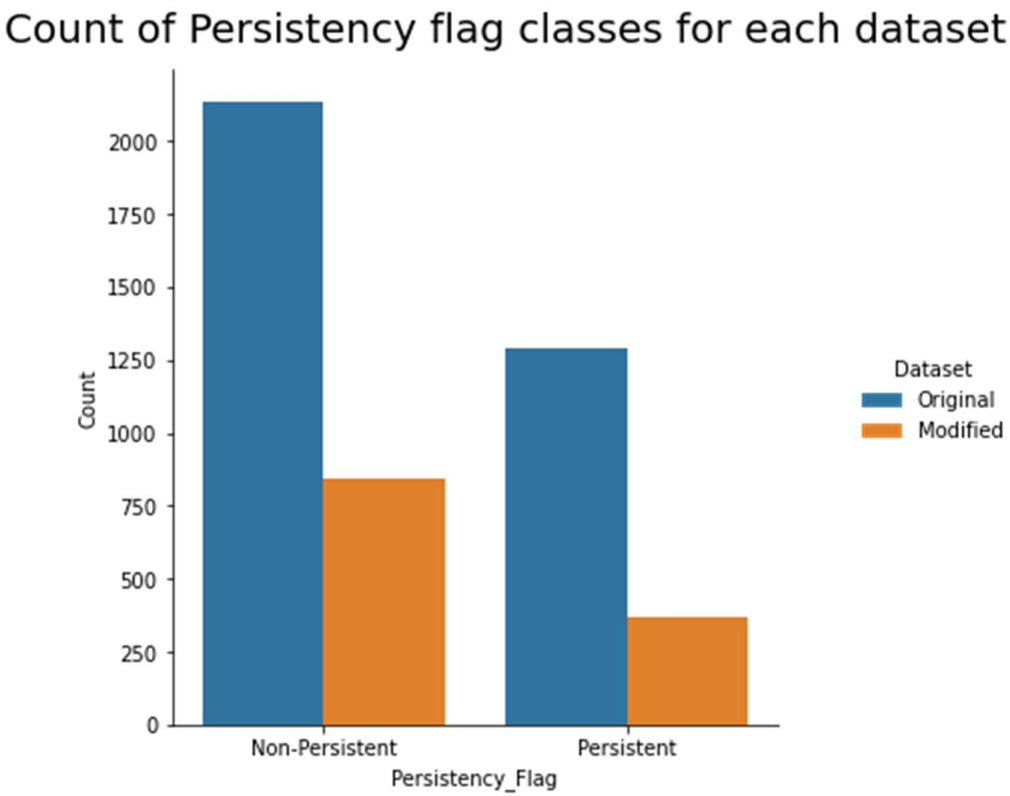
Boxplot of Dexa_Freq_During_Rx, grouped by Persistency Flag, using original dataset



Boxplot of Count_Of_Risks, grouped by Persistency Flag, using original dataset



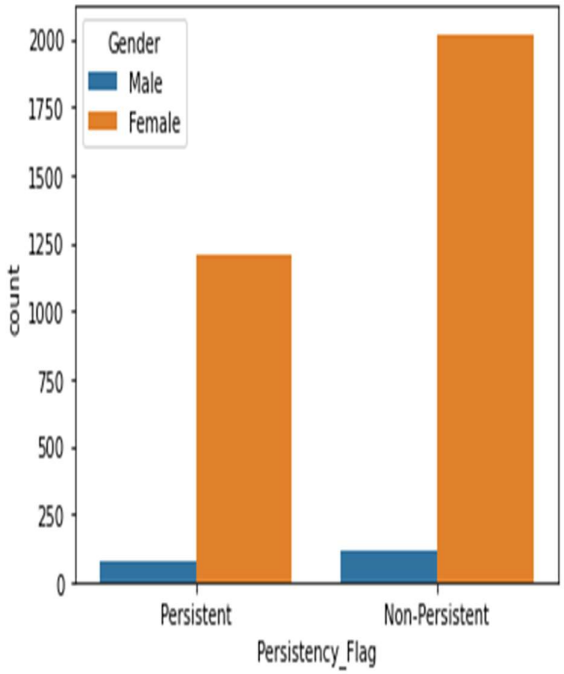
Frequency counts for Persistency flags for each dataset (original and modified)



When the dataset was modified, some rows were removed, but the general form of the distribution of classes is roughly the same.

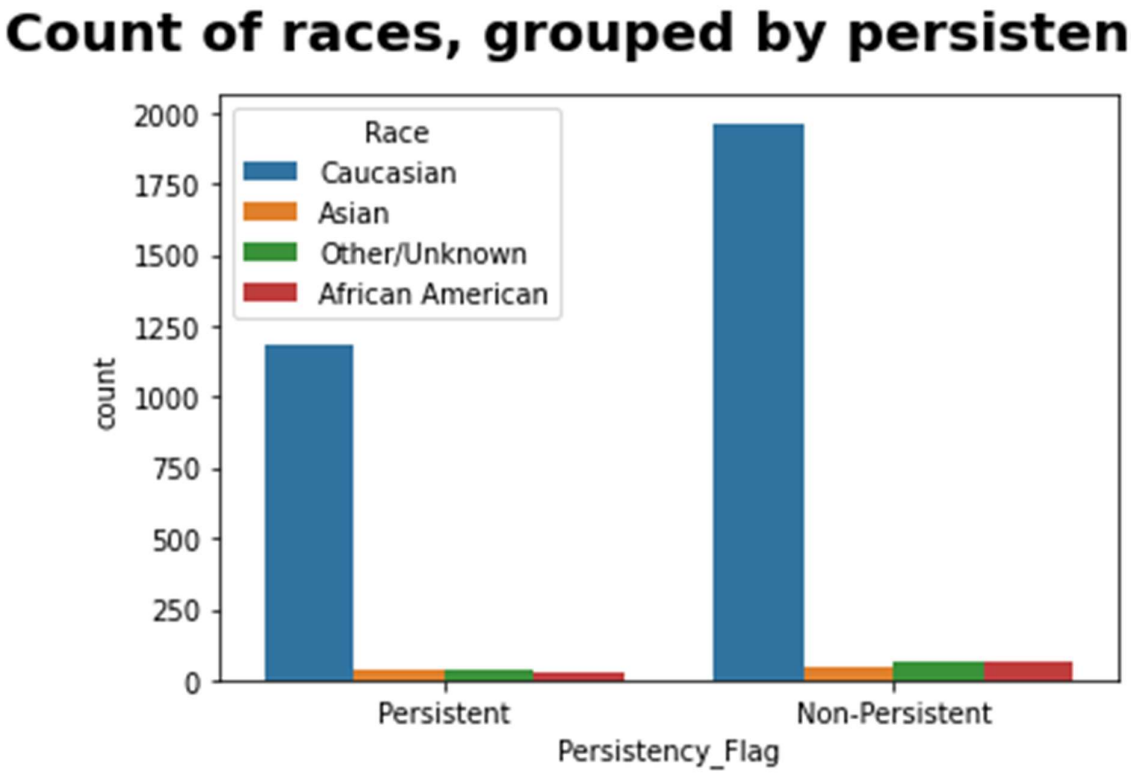
Gender & Persistency Flag

Count of male and female patients, grouped by persistency flag, using original dataset



We see from this plot that males are very underrepresented in the dataset so the Gender variable is not useful for analysis.

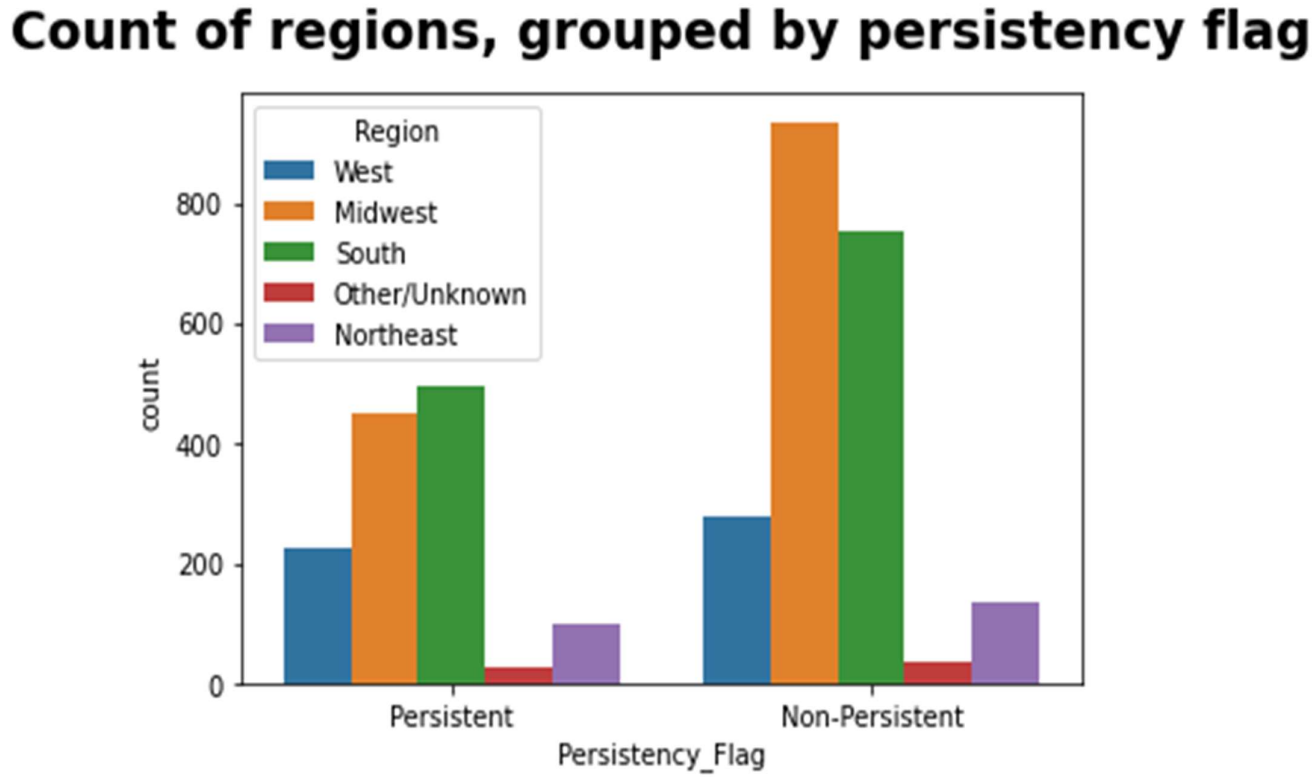
Race & Persistency Flag



Non-caucasian races are heavily underrepresented and this variable isn't useful for analysis.

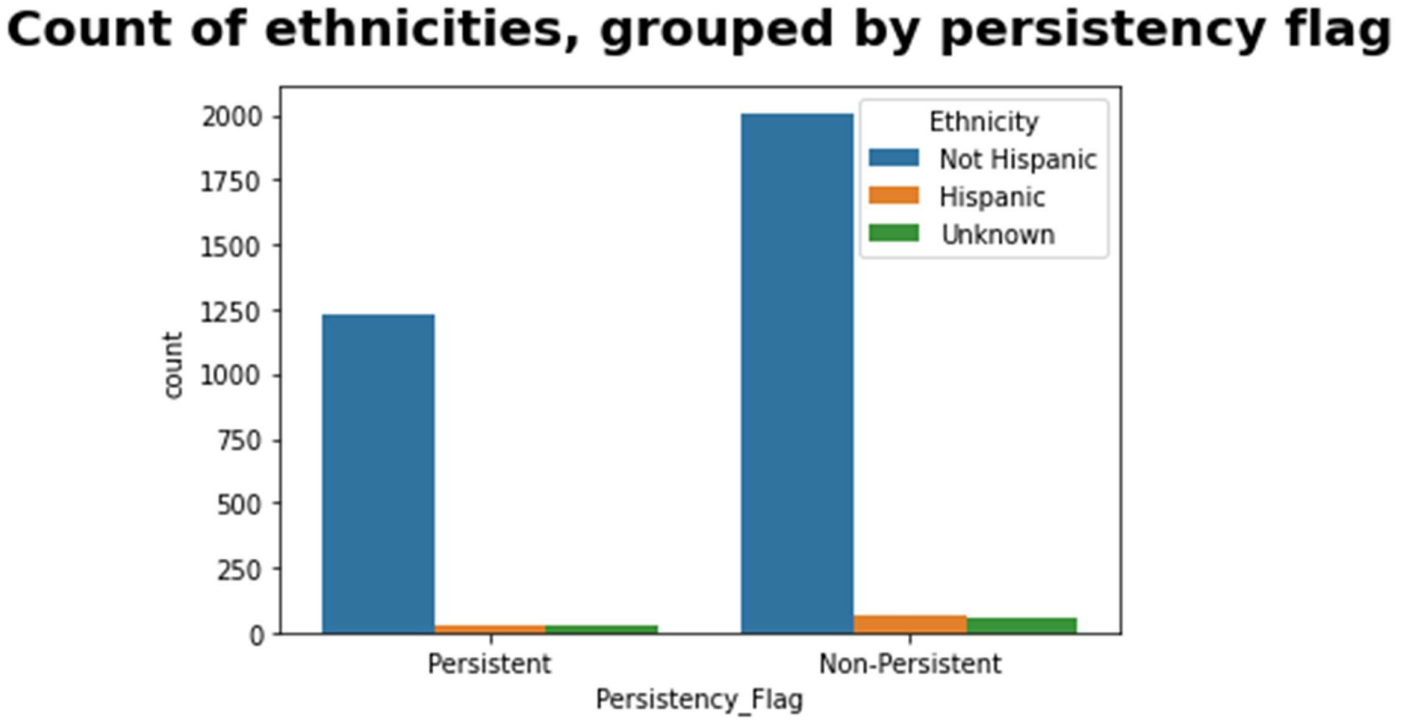
Region & Persistency Flag

The two sets of plots for persistent / non-persistent patients have similar forms, except that for the non-persistent group, the highest count within the set of plots is for the Midwest. This implies that people from the Midwest may be slightly more likely to not be persistent than people from other regions.



Ethnicity & Persistency Flag

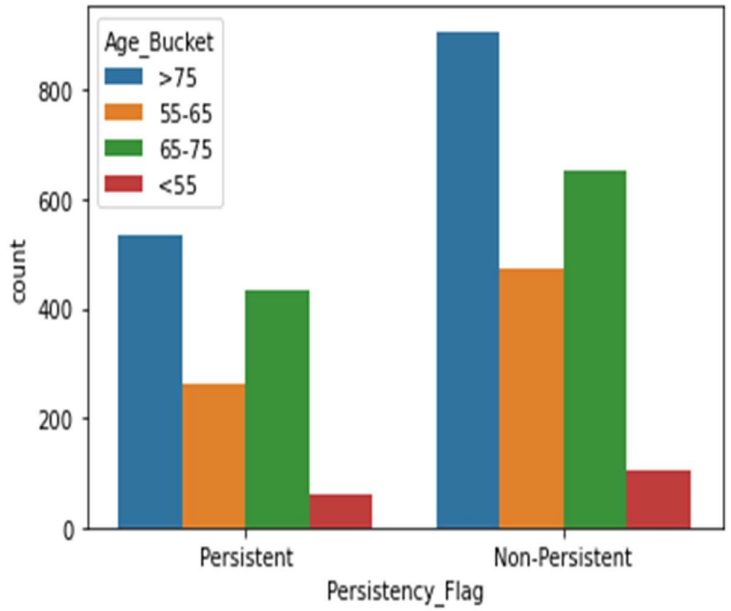
The two sets of plots for persistent / non-persistent patients have similar forms which implies that age isn't a useful indicator of whether or not a patient is persistent, and also some age ranges are heavily underrepresented (namely <55).



Age & Persistency Flag

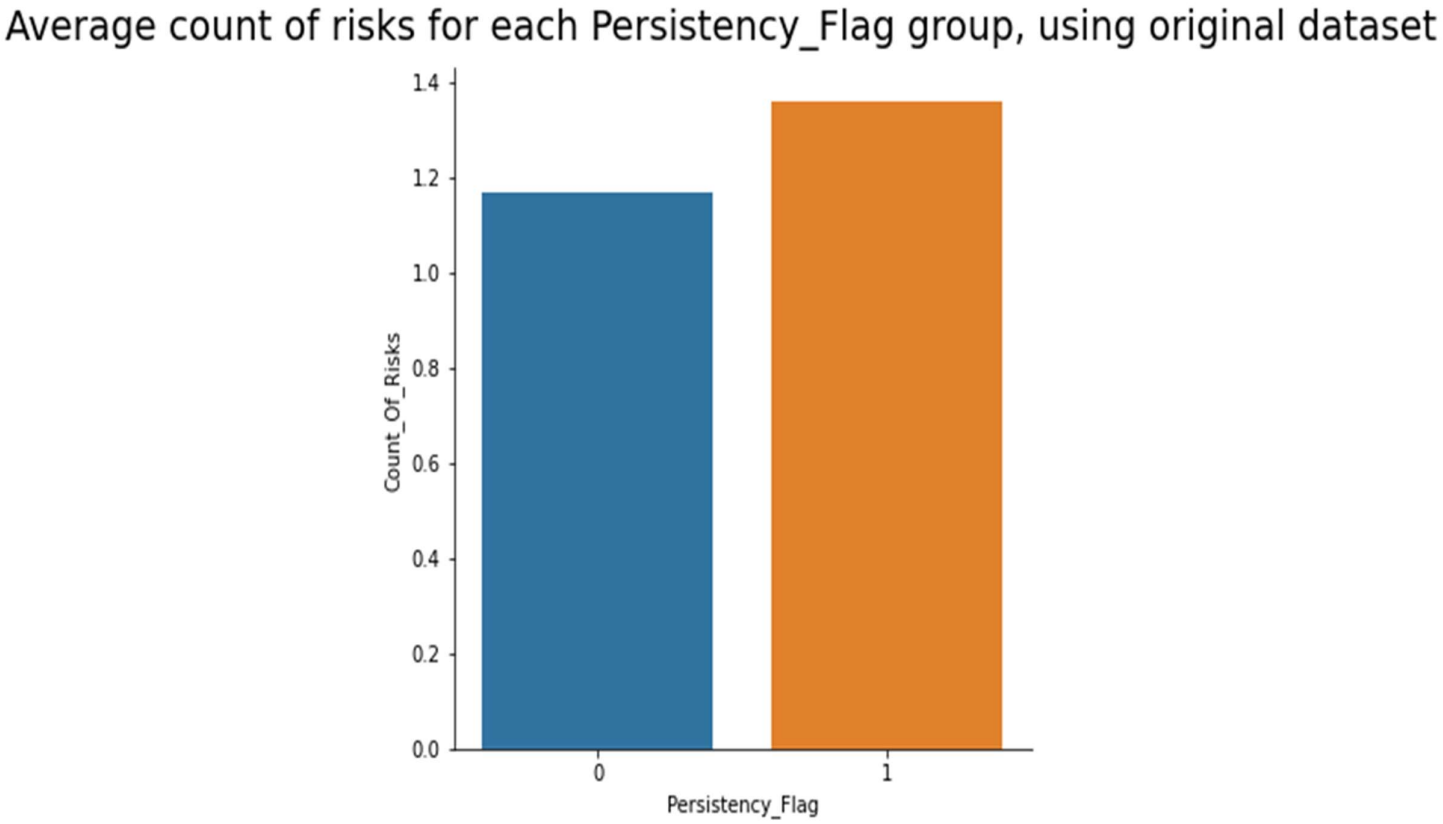
The two sets of plots for persistent / non-persistent patients have similar forms which implies that age isn't a useful indicator of whether or not a patient is persistent, and also some age ranges are heavily underrepresented (namely <55)

Count of people by age group, grouped by persistency flag, using original dataset



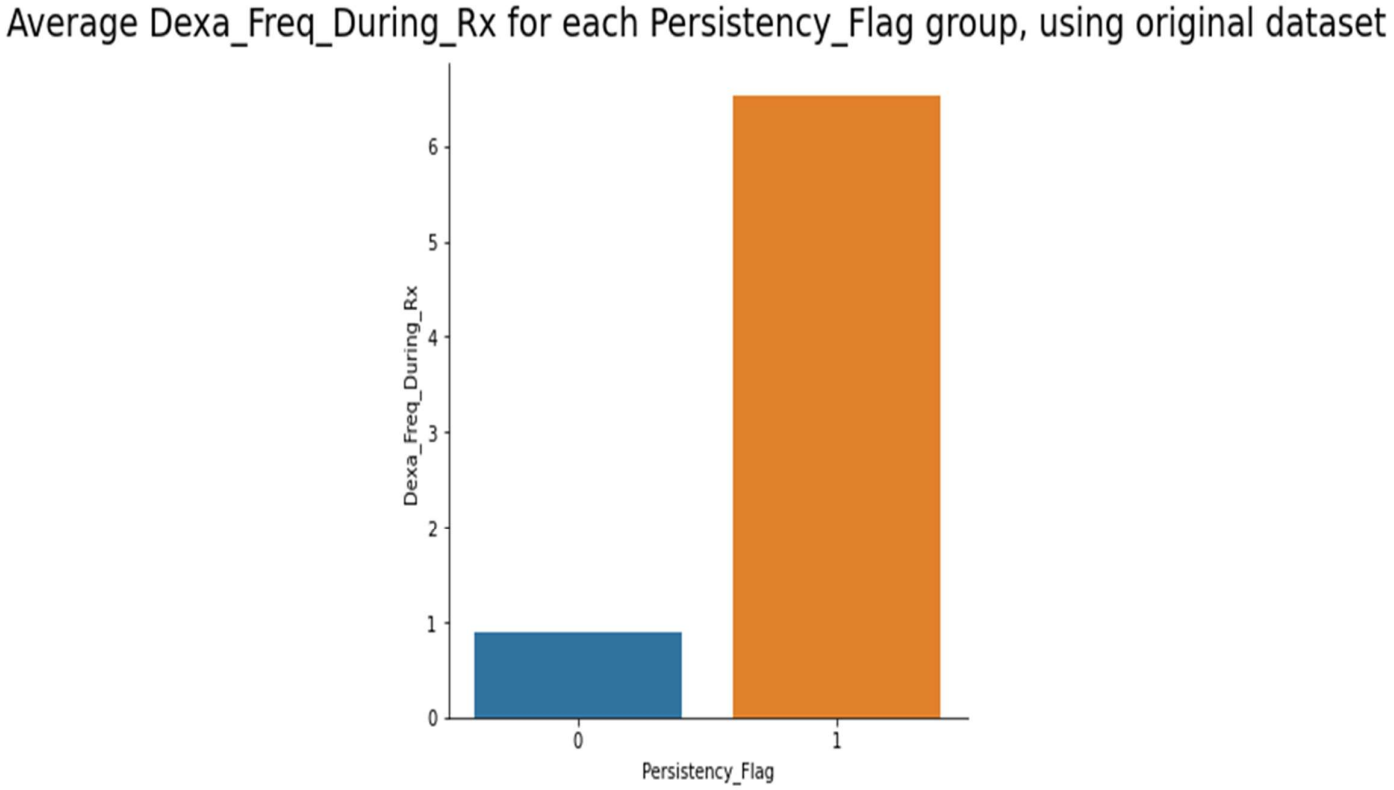
Average Count_Of_Risks per target class

People who were persistent, had, on average, a slightly higher Count_Of_Risks (1.37) than people who weren't persistent (1.16)



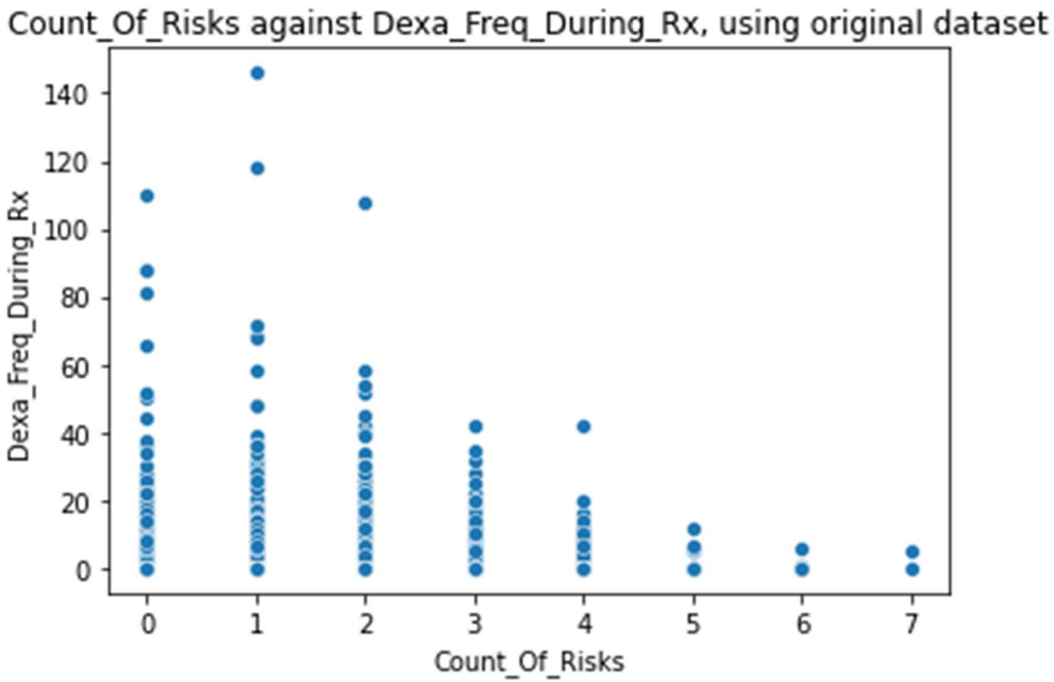
Average Dexa_Freq_During_Rx per target class

People who were persistent, had, on average, a much higher Dexa_Freq_During_Rx (~6) than those who weren't persistent (~1).



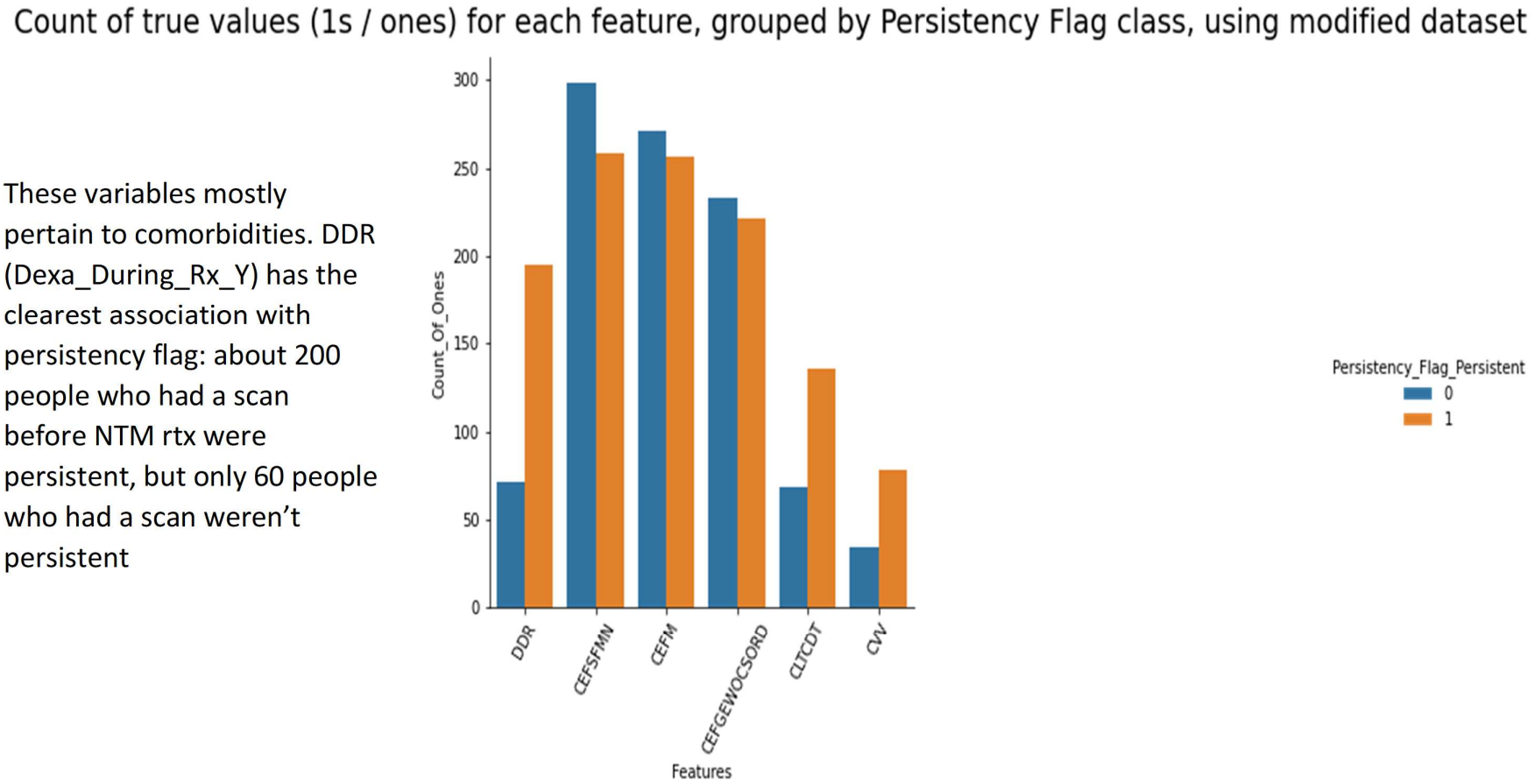
Count of Risks & Dexa_Freq_During_Rx

We see that as the Count_Of_Risks increases, Dexa_Freq_During_Rx tends to decreases.



6 Best features (part 1)

The 6 best features we found during weeks 8-9 were: "Dexa_During_Rx_Y" , "Comorb_Encounter_For_Screening_For_Malignant_Neoplasms_Y", "Comorb_Encounter_For_Immunization_Y" ,"Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx_Y" , "Comorb_Long_Term_Current_Drug_Therapy_Y" , "Concom_Viral_Vaccines_Y", "Persistency_Flag_Persistent". These names are abbreviated in the following plot.

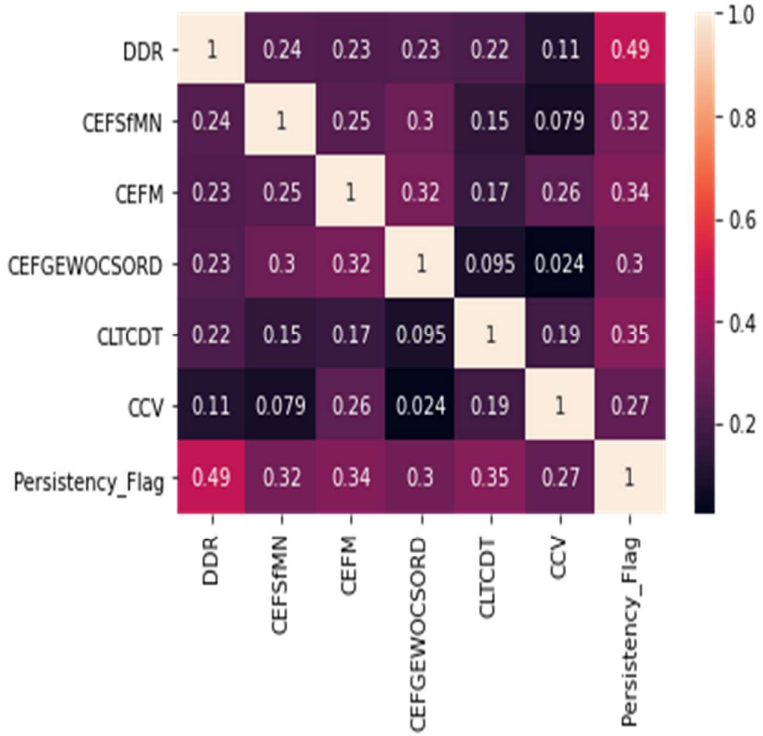


These variables mostly pertain to comorbidities. DDR (Dexa_During_Rx_Y) has the clearest association with persistency flag: about 200 people who had a scan before NTM rtx were persistent, but only 60 people who had a scan weren't persistent

6 Best features (part 2)

We created a correlation heatmap to visualise the correlations between each of the best features and the target. Note that Dexa_During_Rx_Y and Persistency_Flag_Persistent has the highest correlation (0.49), and there isn't multicollinearity.

Correlation heatmap for dataframe of best features and target, using modified data



Recommended Model

There is no multicollinearity between the 6 best features, and we have a good number of samples to work with (1215 in the modified dataset), but not the very high amount that some algorithms need. The observations are also independent of each other. Additionally, we want to predict a binary variable (Persistency_Flag), so **binary logistic regression** is a logical choice. We can also consider creating an ensemble of different algorithms, including binary logistic regression.

Thank you