



Data Glacier

Your Deep Learning Partner

Week 10 Deliverables

Group Name: The Data Doctors

09-Dec-2023

Team Details

<p>Name: Noah Gallego Email: noahgallego394@gmail.com Country: United States College/Company: California State University Bakersfield Specialization: Data Science</p>	<p>Name: Tomisin Abimbola Adeniyi Email: tomsin_adeniyi11@yahoo.com Country: Nigeria College/Company: N/A Specialization: Data Science</p>
<p>Name: Mohammad Shehzar Khan Email: mshehzarkhan@gmail.com Country: Turkey College/Company: Koç University Specialization: Data Science</p>	<p>Name: Ashish Sasanapuri Email: sashrao21@gmail.com Country: India College/Company: N/A Specialization: Data Science</p>

Problem Description

One challenge for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription. To solve this problem ABC Pharma company approached an analytics company to automate this process of identification.

Exploratory Data Analysis

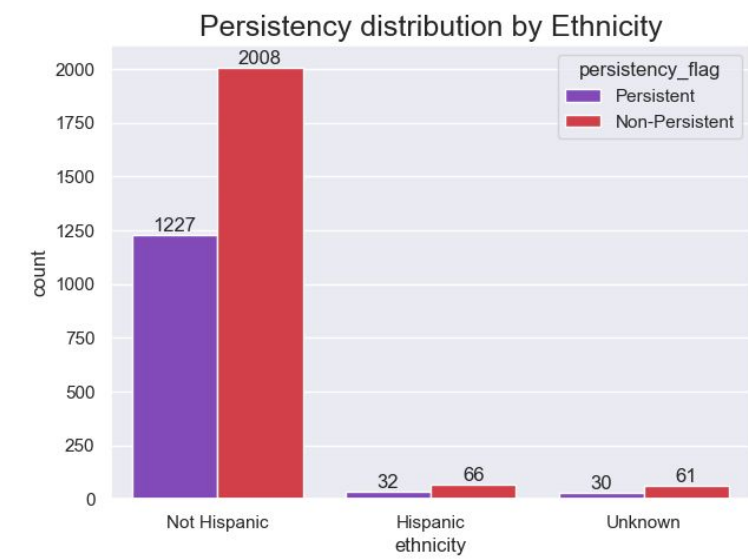
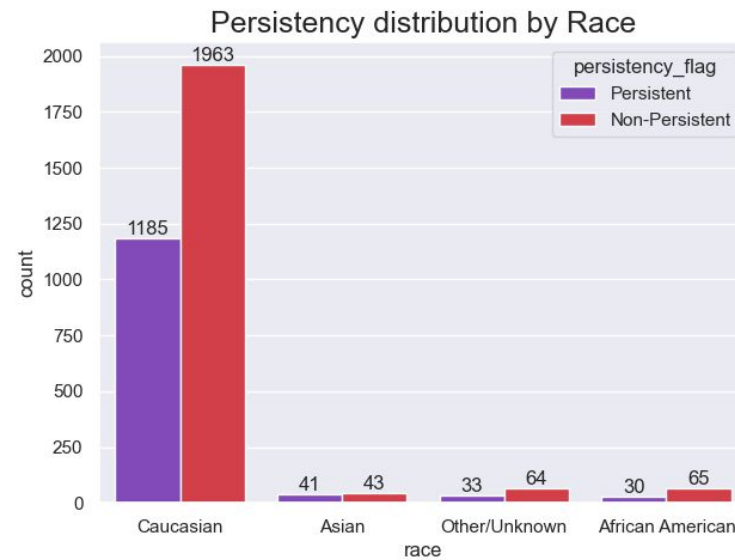
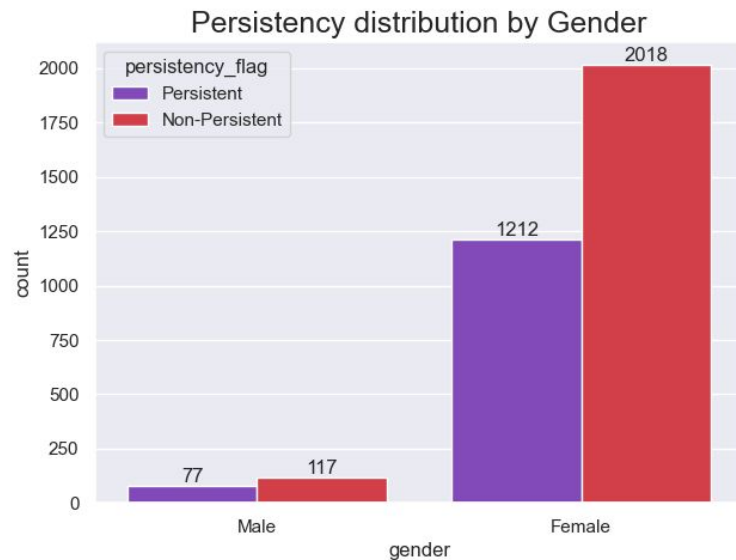


Data Glacier

Your Deep Learning Partner

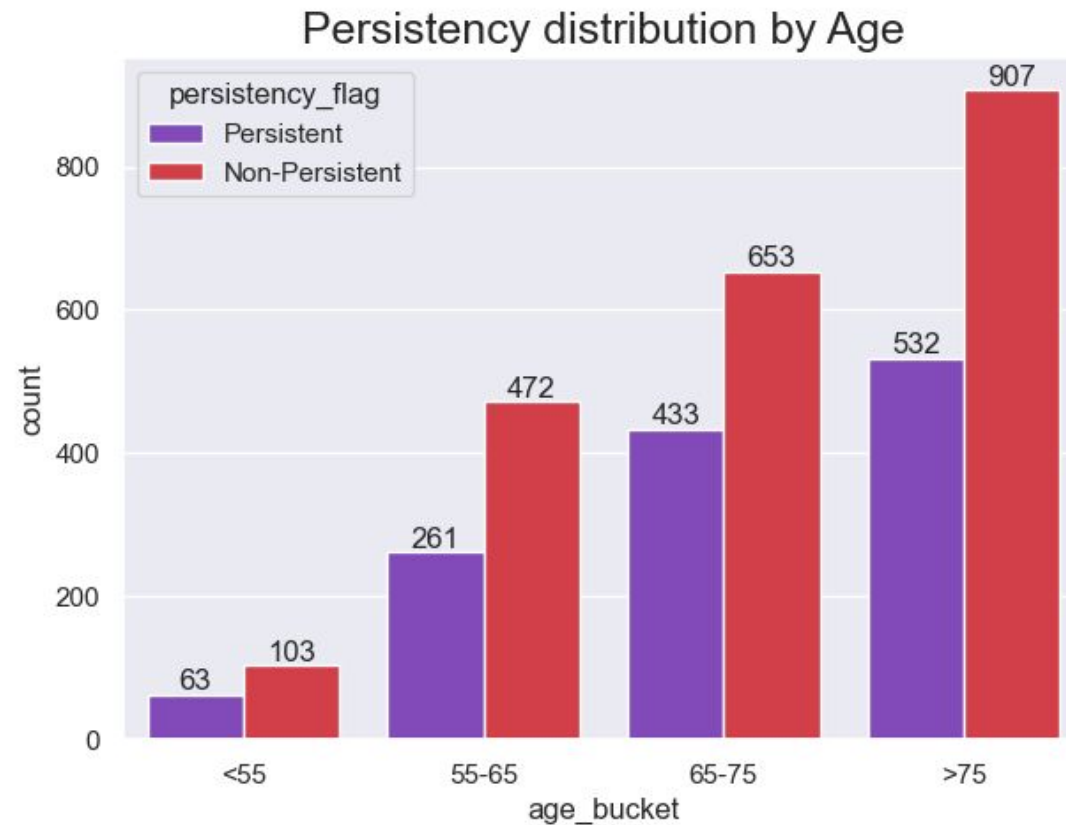
Demographics

- Features like 'Gender', 'Race', and 'Ethnicity' have been dropped as they tend to induce bias in the data and don't provide much information regarding impact of persistent of a patient.



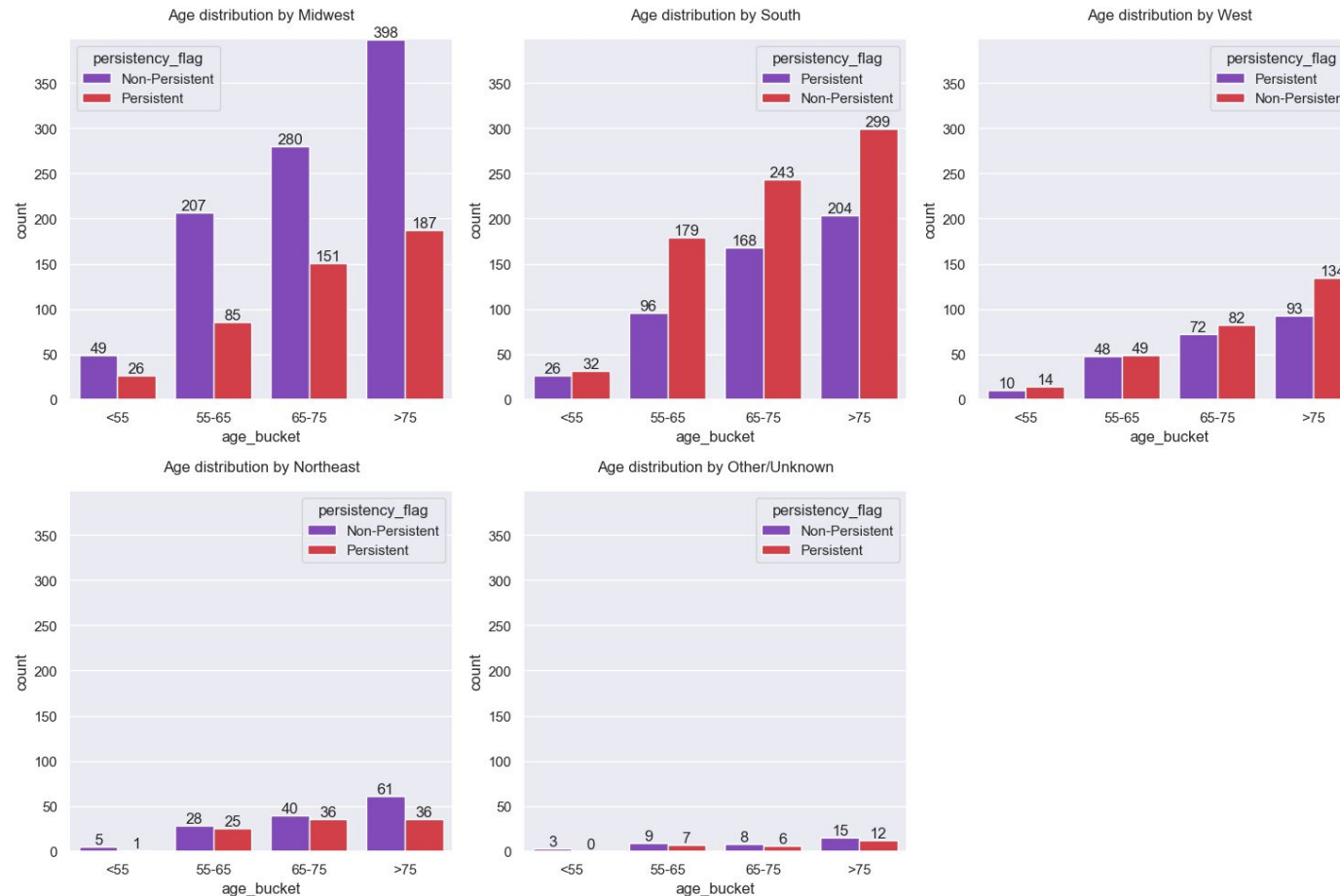
Demographics

- We can observe that majority of the patients are aged above *55 years* and majority **Non-Persistent** patients fall in the age group of more than *75 years* of age.



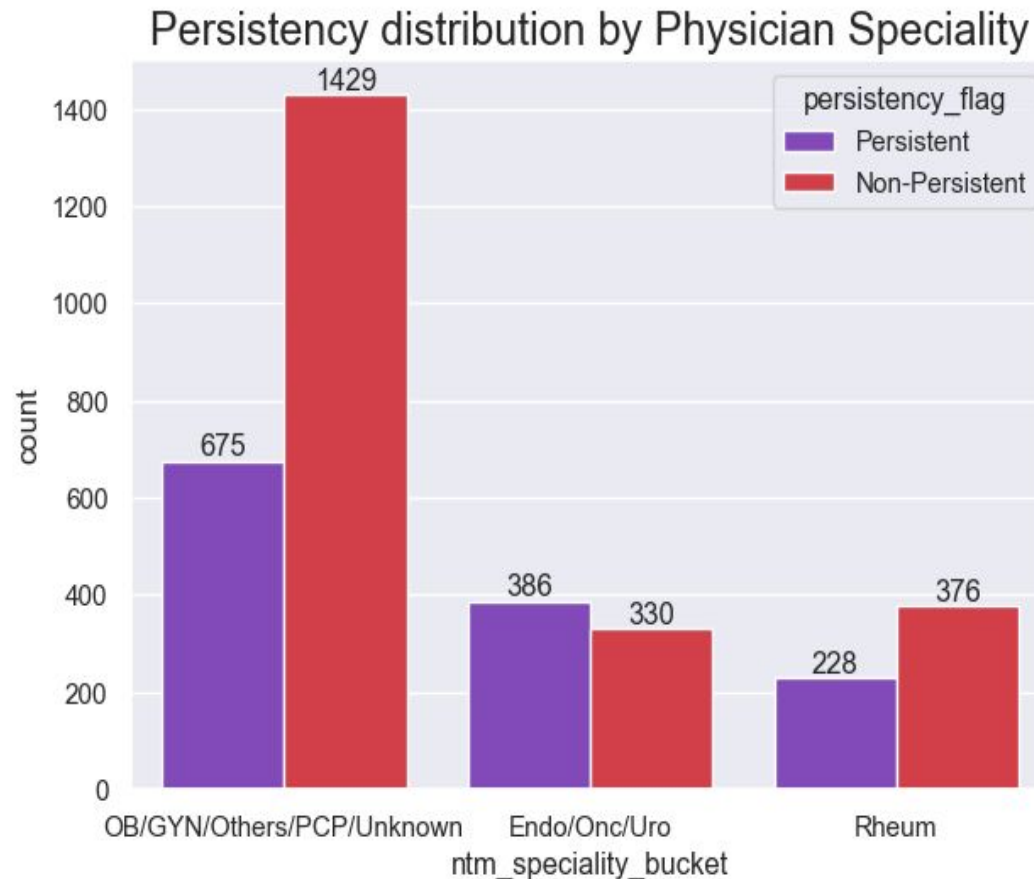
Demographics

- Most patients fall in the *'Midwest'* and *'South'* regions as observed from the following graphs.

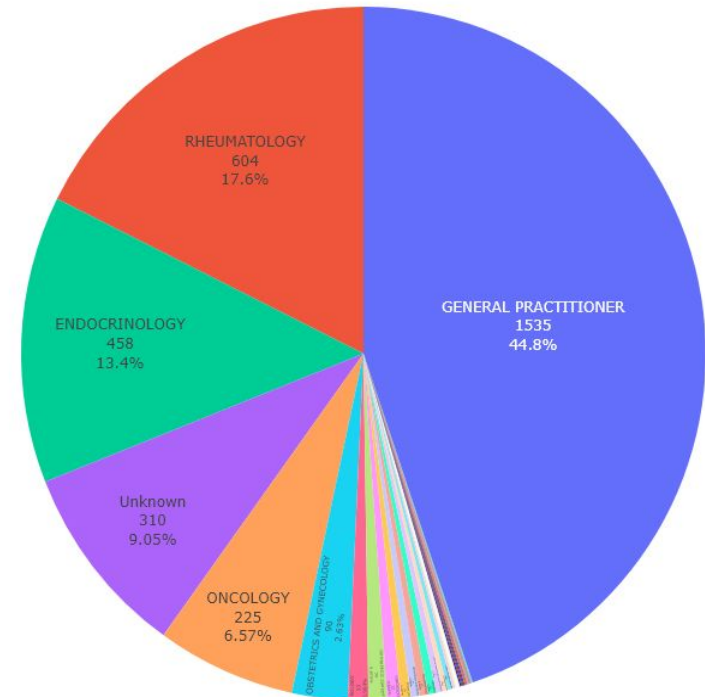


Physician Attributes

- Around **45%** of Physicians who have prescribed new medication to the patients are '*General Practitioners*'.

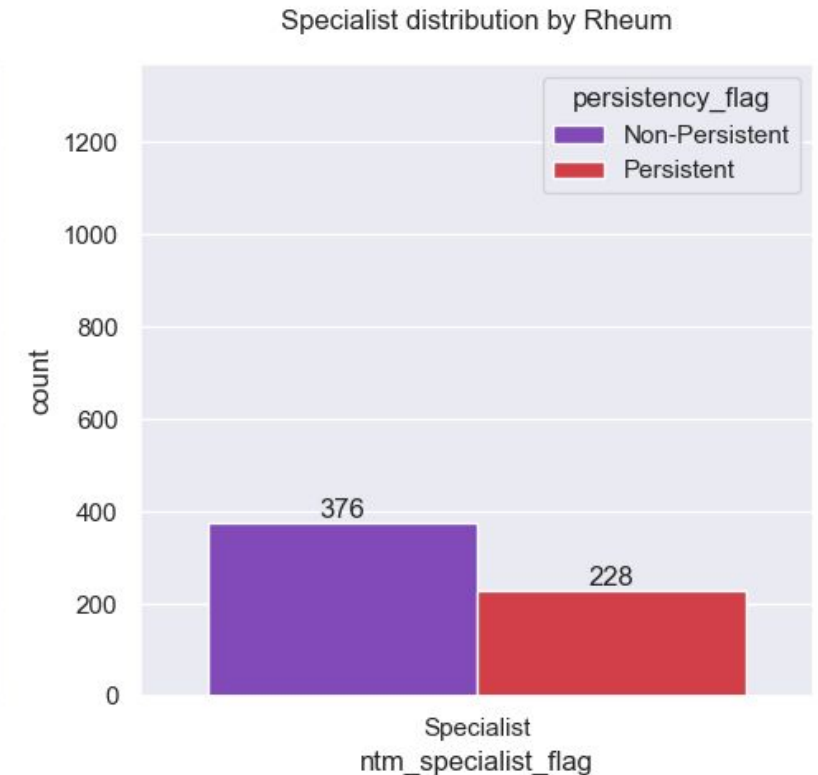
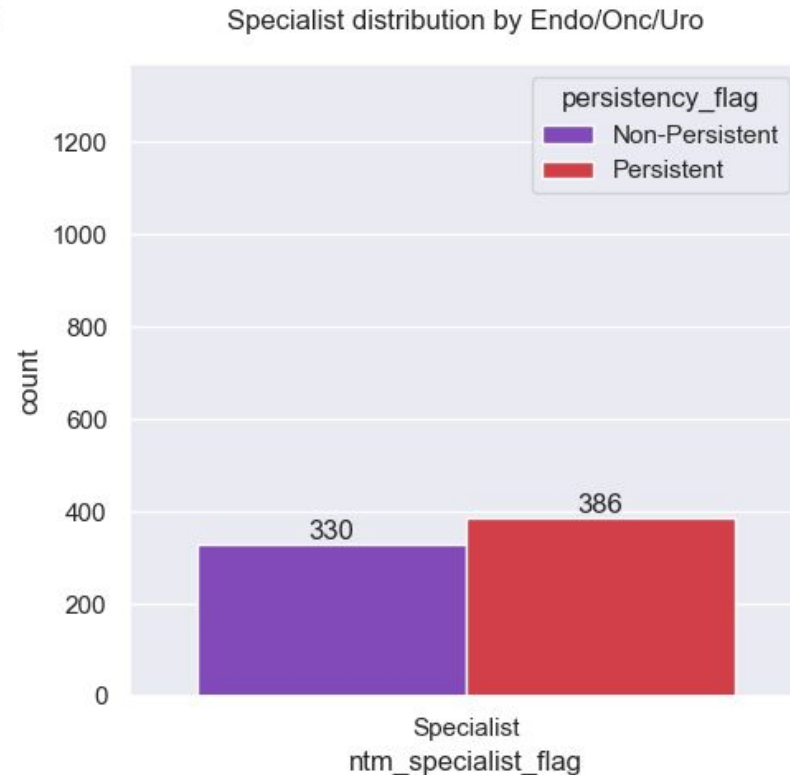
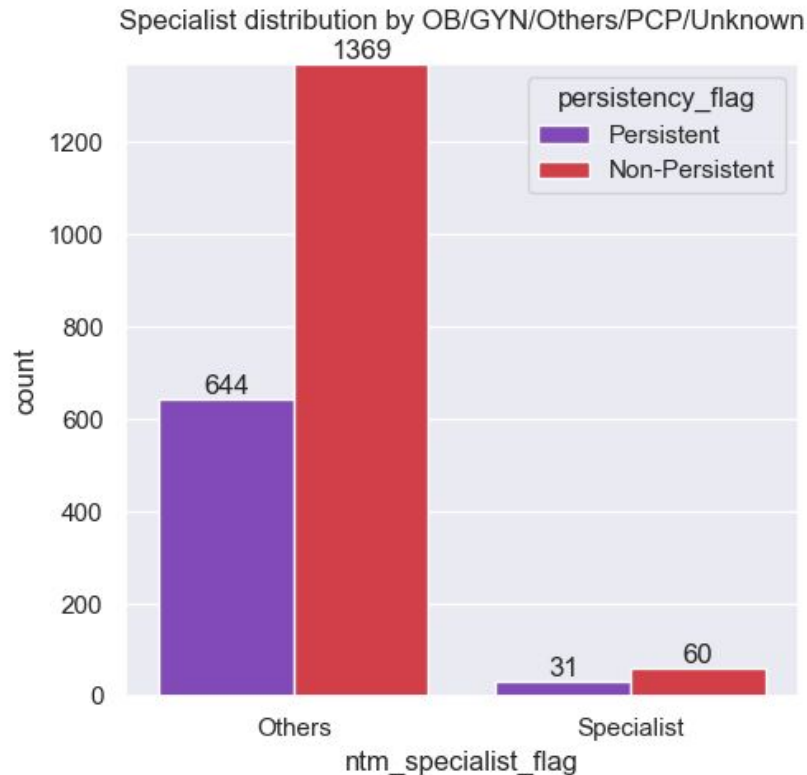


Distribution of Physician's Speciality



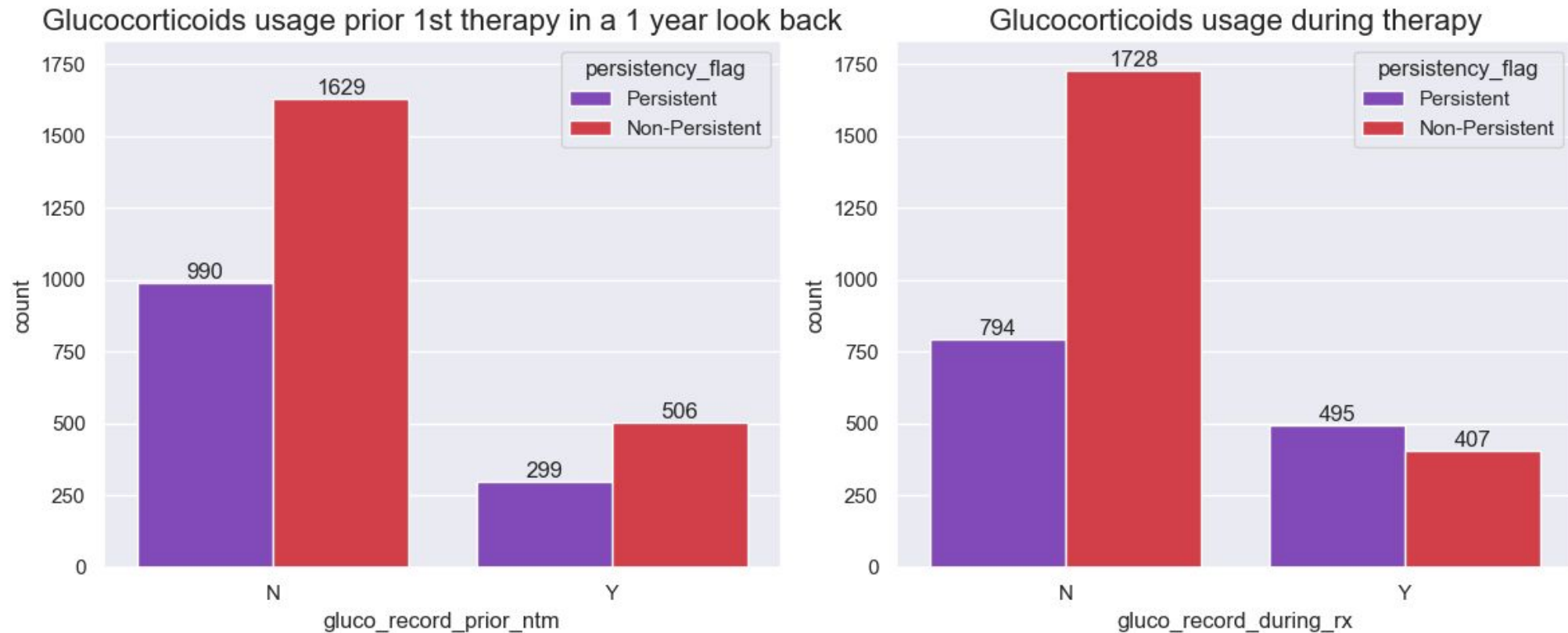
Physician Attributes

- Majority of the **Non-Persistent** patients have been prescribed the new medication by Physicians who are not *Specialists*.



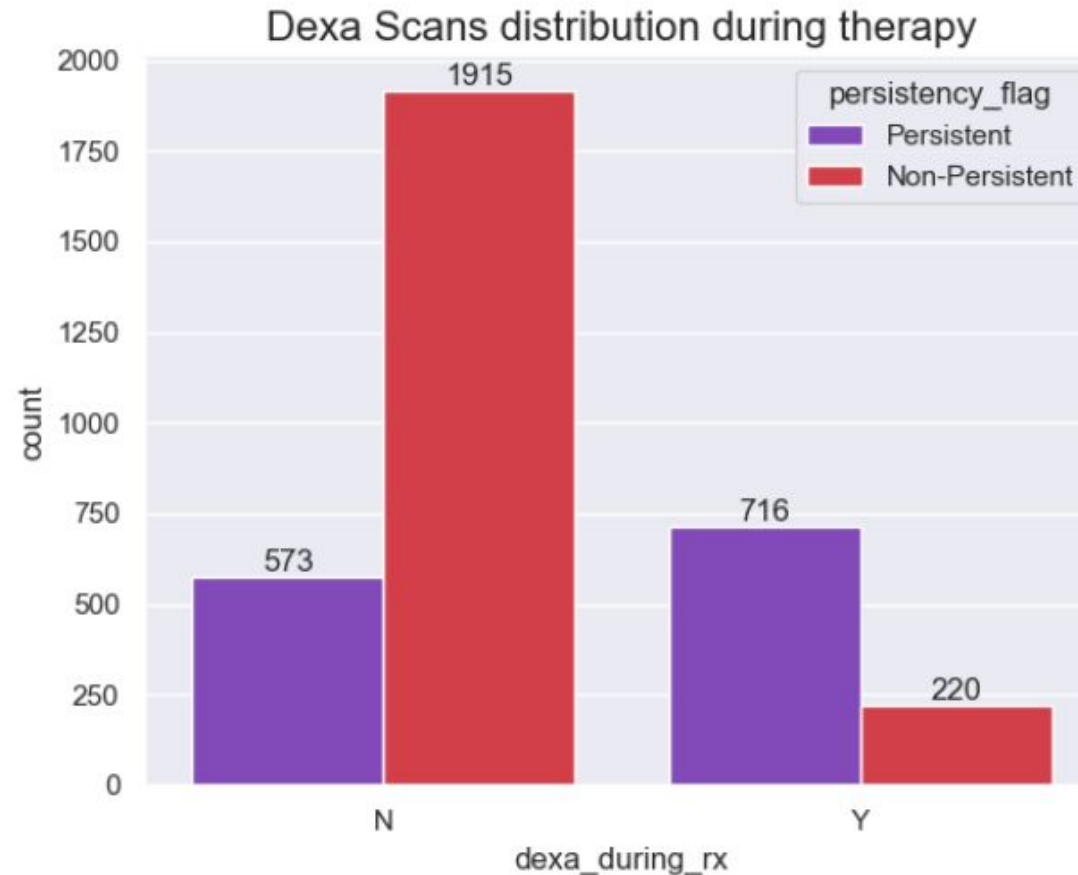
Clinical Factors

- When we consider the case of usage *Glucocorticoids*, the **Persistent** patients are less compared to **Non-Persistent** patients prior to therapy but vice-versa during therapy. There is increase in the number of patients using *Glucocorticoids* during therapy.



Clinical Factors

- Based on the below graph, the *Dexa Scans* is part of the therapy and majority of patients who haven't gone through *Dexa Scans* are **Non-Persistent**.



Clinical Factors

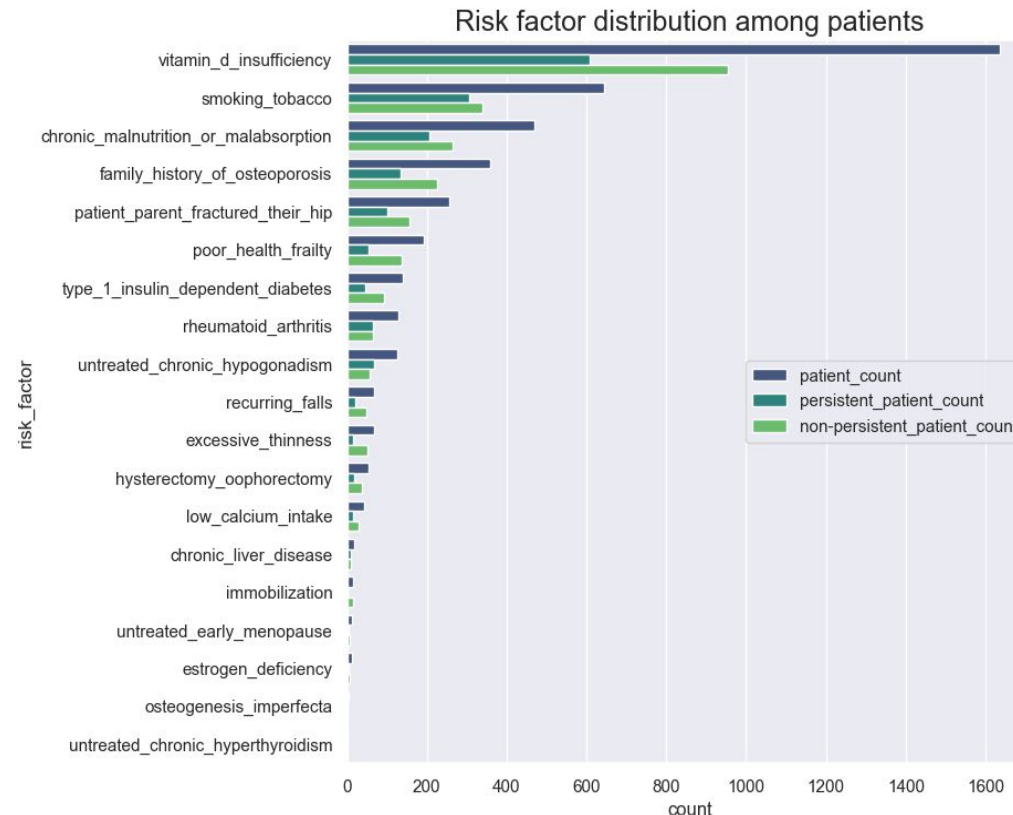
- Some of the features have '*Unknown*' values as one of their categories as given below -

```
risk_segment_during_rx: 1497  
tscore_bucket_during_rx: 1497  
change_t_score: 1497  
change_risk_segment: 2229
```

- As these features have a large number of '*Unknown*' category values, they don't provide much information. Hence, dropping these features will also not impact the outcome of the model.

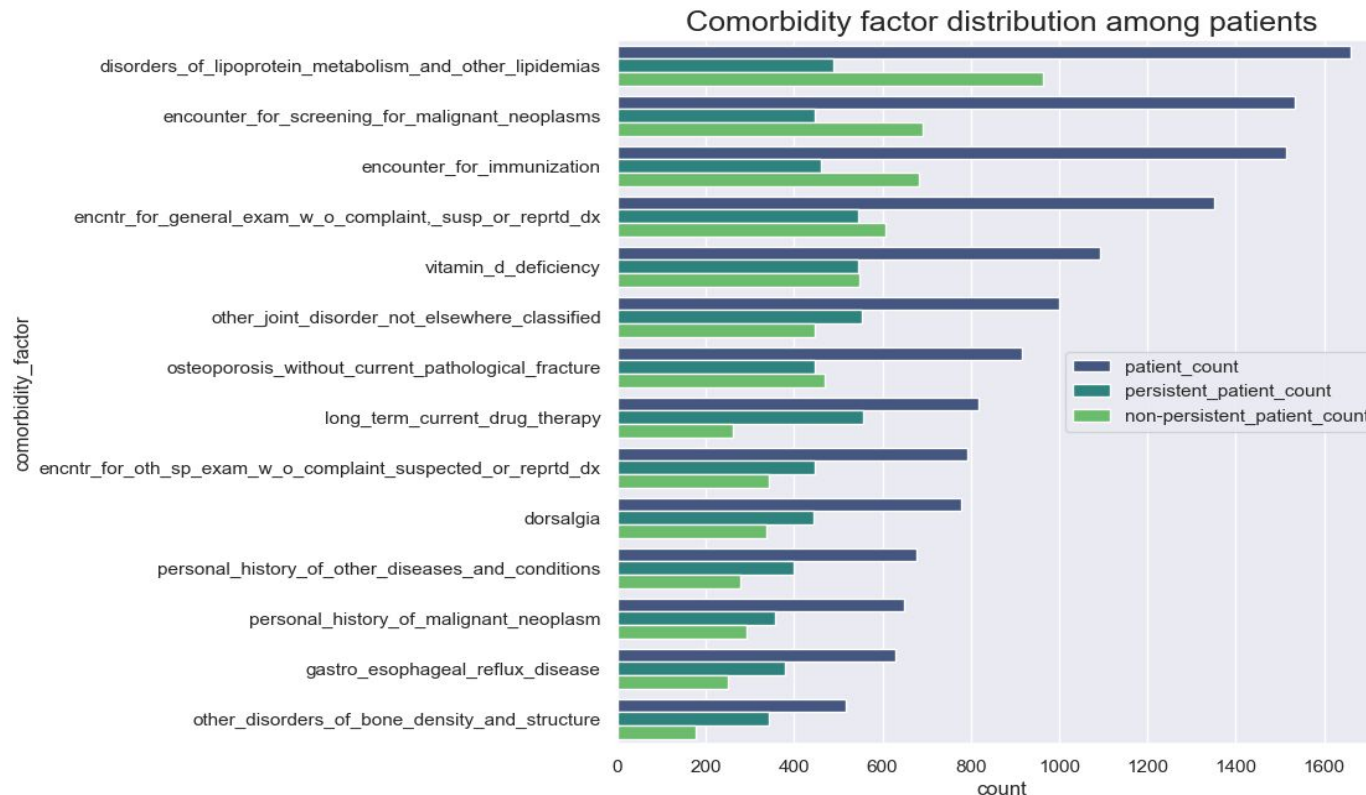
Disease/Treatment Factors

- Majority of the patients have been susceptible to **Risk Factors** such as '*Vitamin D insufficiency*', '*smoking tobacco*', '*chronic malnutrition or malabsorption*' and have a '*family history of osteoporosis*'.
- Due to heavy imbalance of data in **Risk Factor** categories, we can reduce dimensionality by reducing the categories capturing less data into a single category.



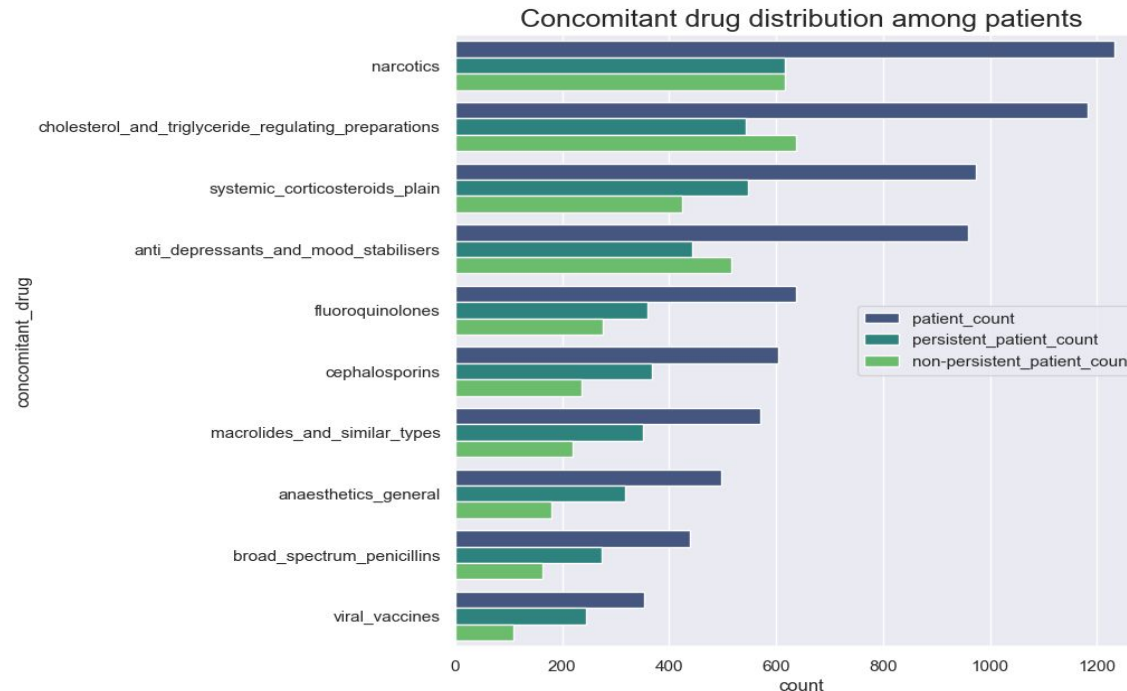
Disease/Treatment Factors

- There are total 14 **Comorbidity Factors** recorded for each patient.
- The top **Comorbidity Factors** include *disorders_of_lipoprotein_metabolism_and_other_lipidemias*, *encounter_for_screening_for_malignant_neoplasms*, *encounter_for_immunization*, and *encntr_for_general_exam_w_o_complaint,_susp_or_reprtd_dx*.



Clinical Factors

- We can see that the graph shows the distribution of patients who have received **Concomitant Drugs** 1 year prior to start therapy.
- The count for **Non-Persistent** patients who have been given **Concomitant Drugs** such as *Narcotics*, *cholesterol_and_triglyceride_regulating_preparations*, and *anti_depressants_and_mood_stabilisers* is greater compared to the other categories.



Recommendations

- We can drop features which contain large number of '*Unknown*' category as value.
- As the dataset heavily dimensional, we can apply different methods such as PCA or Attribute Analysis for feature selection as well as dimensionality reduction.
- As the dataset is highly imbalance, we can apply *SMOTE* or *Weighted sampling* techniques for balancing the dataset.
- We can try to train the data on simpler models and apply grid search and cross validation methods for hyper-parameter tuning. The test results will help in understanding the best model.

Thank You



Data Glacier

Your Deep Learning Partner