Data Glacier

Your Deep Learning Partner

Week 11 Deliverables

Group Name: The Data Doctors

16-Dec-2023

Team Details

Name: Noah Gallego Email: noahgallego394@gmail.com Country: United States College/Company: California State University Bakersfield Specialization: Data Science	Name: Tomisin Abimbola Adeniyi Email: tomisin_adeniyi11@yahoo.com Country: Nigeria College/Company: N/A Specialization: Data Science
Name: Mohammad Shehzar Khan Email: mshehzarkhan@gmail.com Country: Turkey College/Company: Koç University Specialization: Data Science	Name: Ashish Sasanapuri Email: sashrao21@gmail.com Country: India College/Company: N/A Specialization: Data Science

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.

Data Description

Data Glacier

Your Deep Learning Partner

Data Understanding

- The dataset provides the factors impacting the patient’s persistence to New Therapy Medication (NTM) by ABC pharmaceutical company prescribed by various physicians.

- The aim is to build a machine-learning model that classifies the patient into **Persistent** (Compliant) and **Non-persistent** (Non-Compliant).

- The dataset consists of 3242 records and is a an imbalanced dataset due to low number of **Persistent** records as compared to **Non-persistent**.

Data Understanding

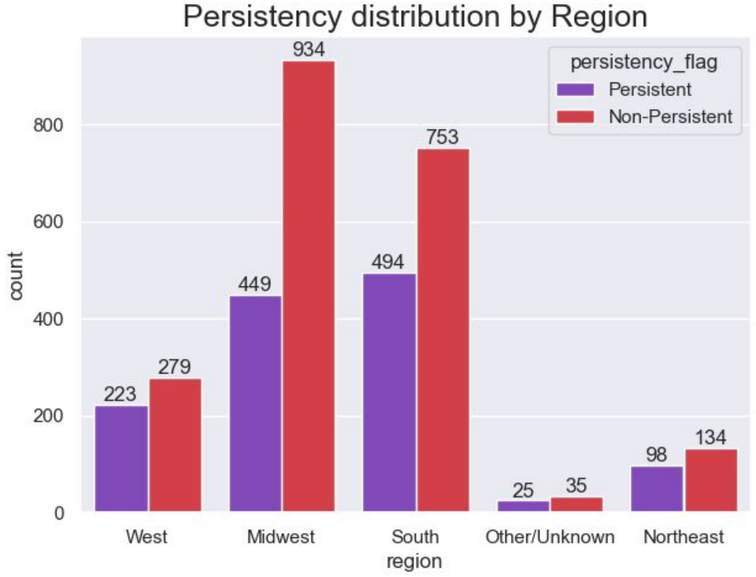
- The dataset contains a total of 69 features that are divided into multiple categories -
    - 1 Target variable: Persistency_Flag
    - 1 Unique identifier for each patient: Ptid
    - 6 Demographic variables of the each patient: Age_Bucket, Gender, Race, Ethnicity, Region, Idn_Indicator
    - 3 Physician Specialist attributes: Ntm_Speciality, Ntm_Specialist_Flag, Ntm_Specialist_Bucket
    - 13 Clinical factors: T-Score details, Risk_Segment details, Multiple risk factors count, DEXA details, Fragility fracture details, Glucocorticoid details
    - 45 Disease/Treatment factors: Injectable drugs, Risk factors, Comorbidities, Concomitancies, Adherence to therapy

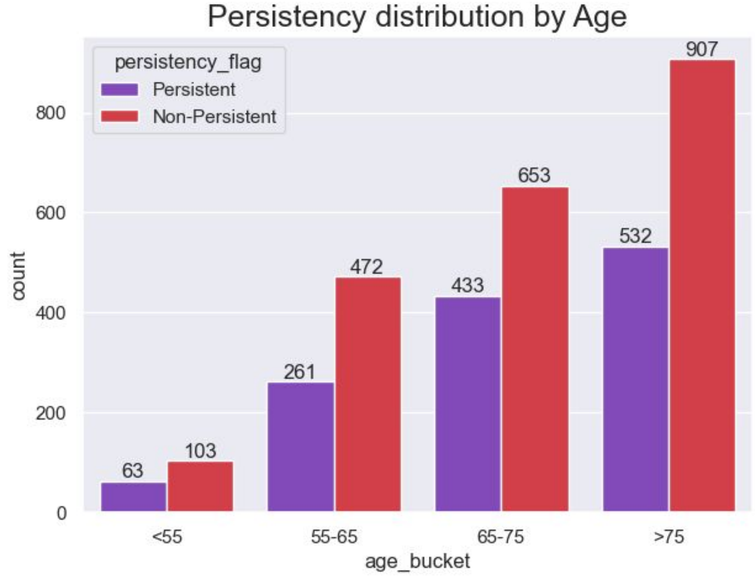
Data Analysis

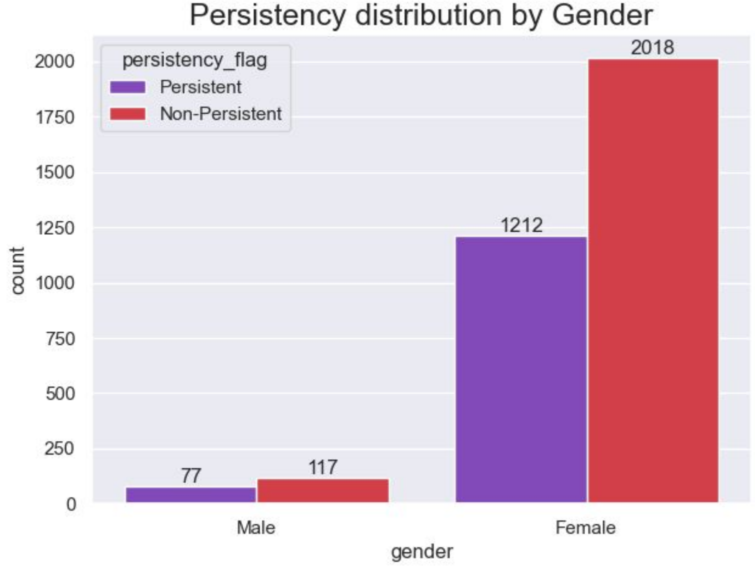
Data Glacier

Your Deep Learning Partner

Demographic Data

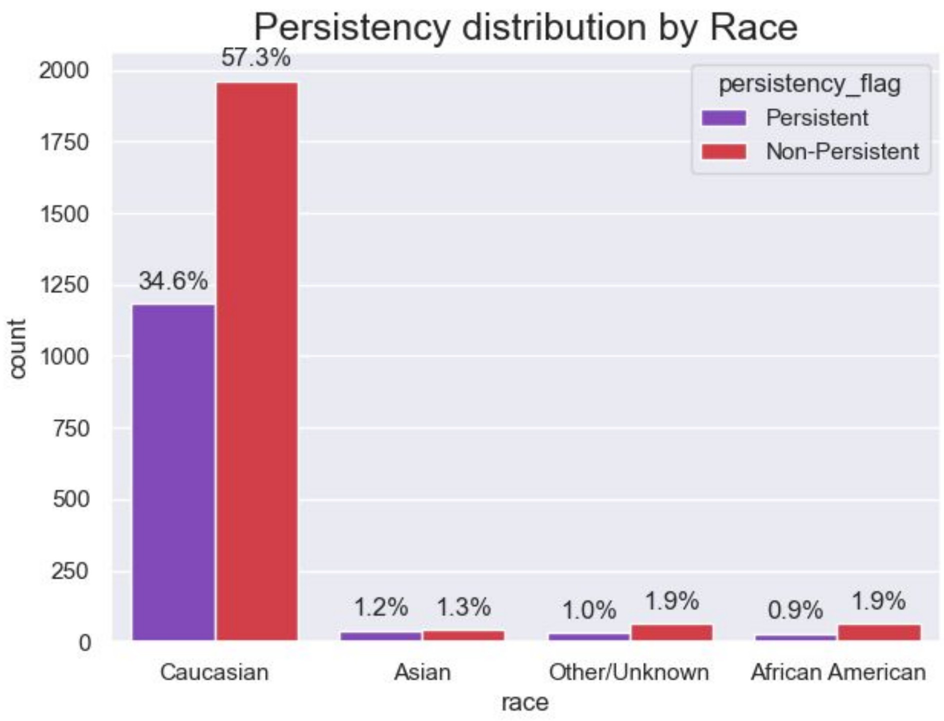

Persistency distribution by Region

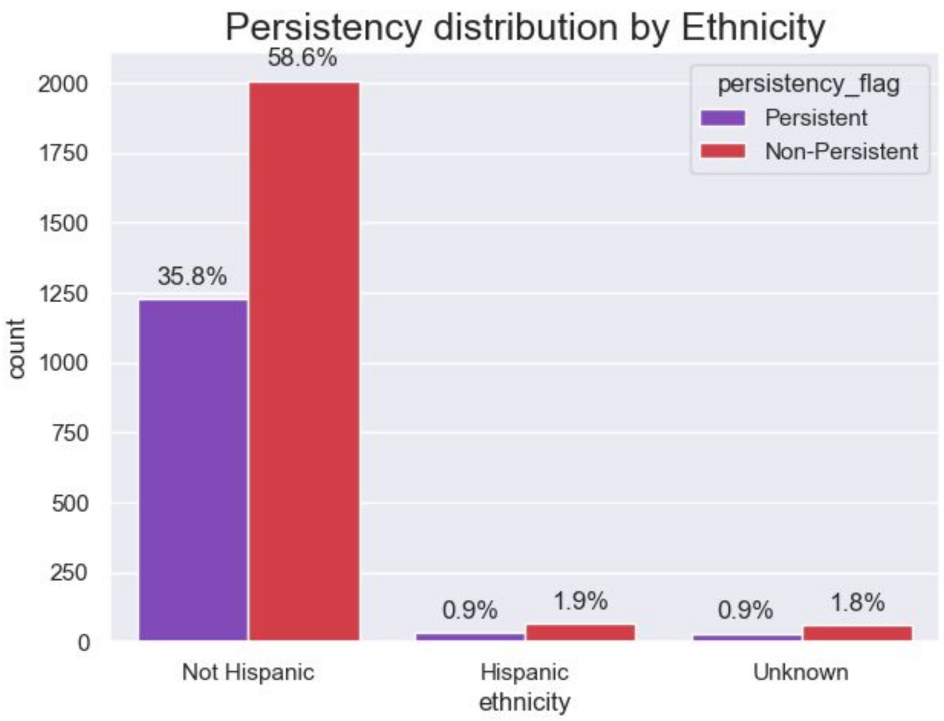

Persistency distribution by Age


Persistency distribution by Gender

- Majority of the patients recorded are **Females** and most of them are **Non-Persistent** to NTM therapies.
- 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.
- *Midwest*, *South*, and *West* regions display majority of the patients recorded.

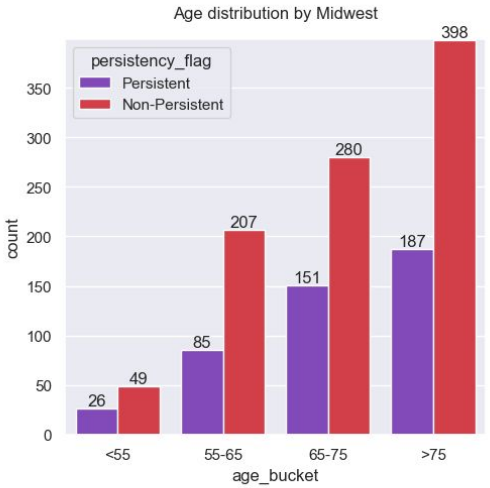
Demographic Data

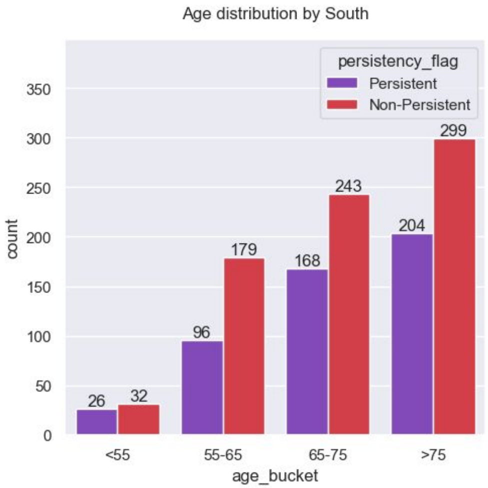


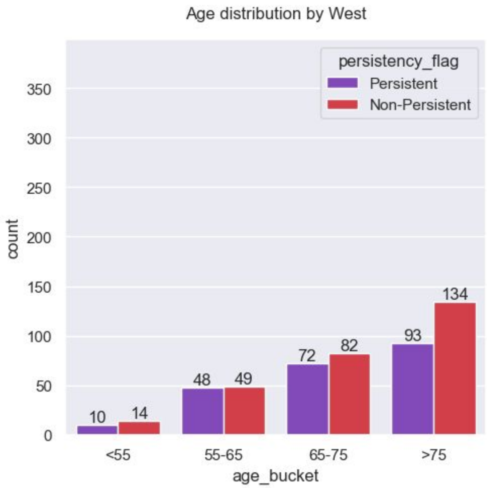


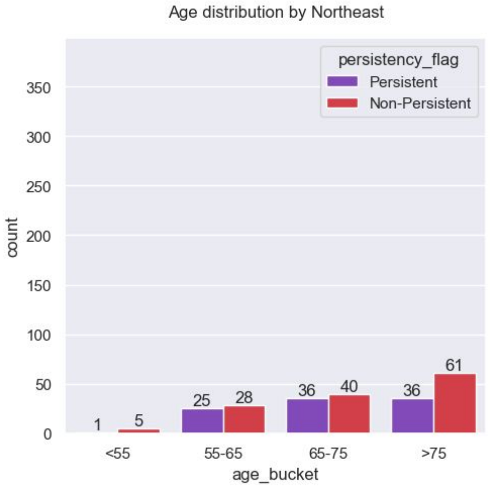
- We can see that majority of the patients are **Caucasian** and **Non-Hispanic**.

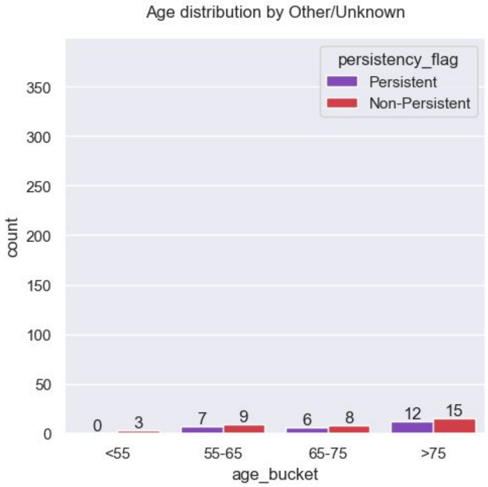
Demographic Data


Age distribution by Midwest


Age distribution by South


Age distribution by West

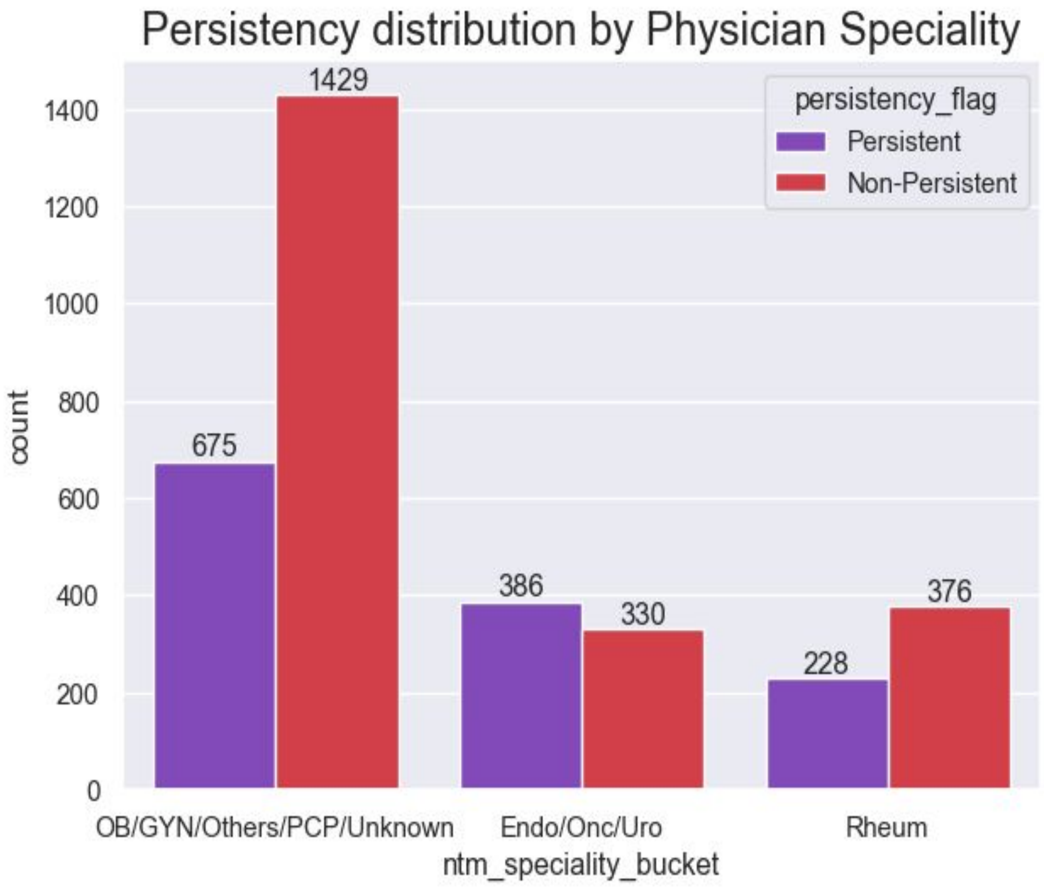

Age distribution by Northeast

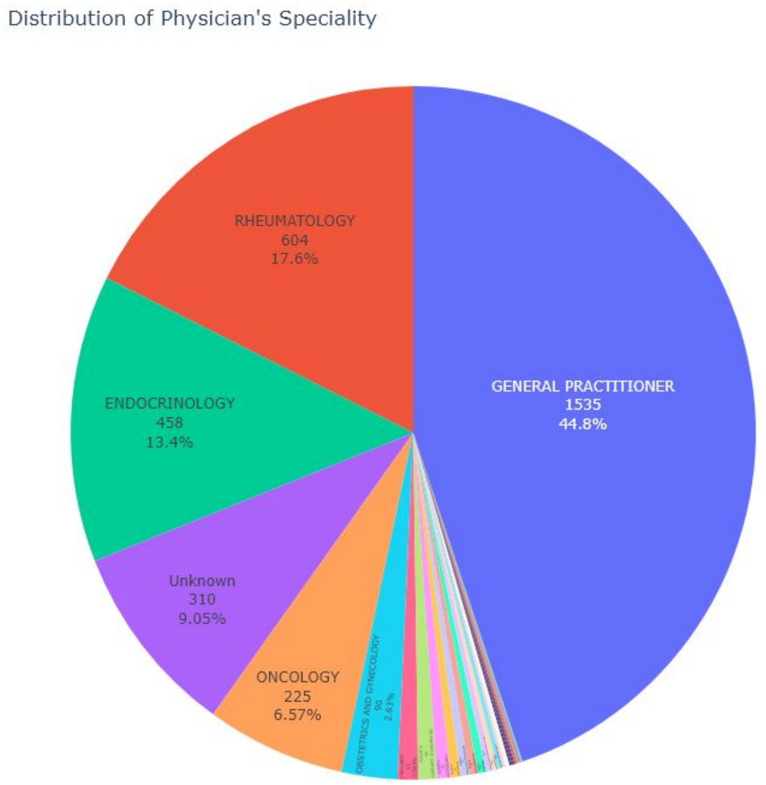

Age distribution by Other/Unknown

- Majority of **Non-Persistent** patients belong to the age group above *75 years* in the **Midwest** region.

Physician Attributes

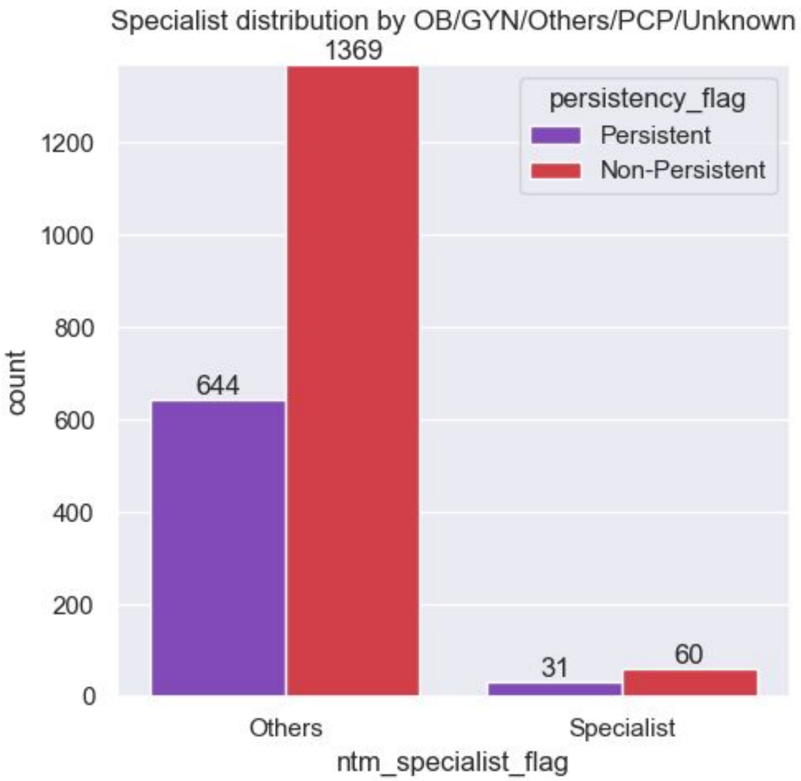
- Around **45%** of Physicians who have prescribed new medication to the patients are *'General Practitioners'*.



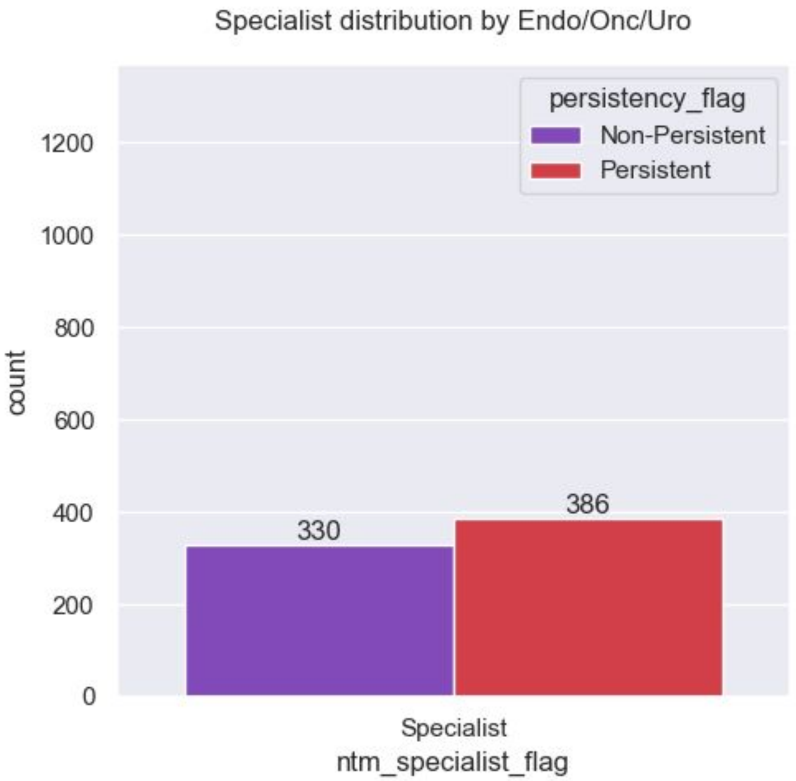


Physician Attributes

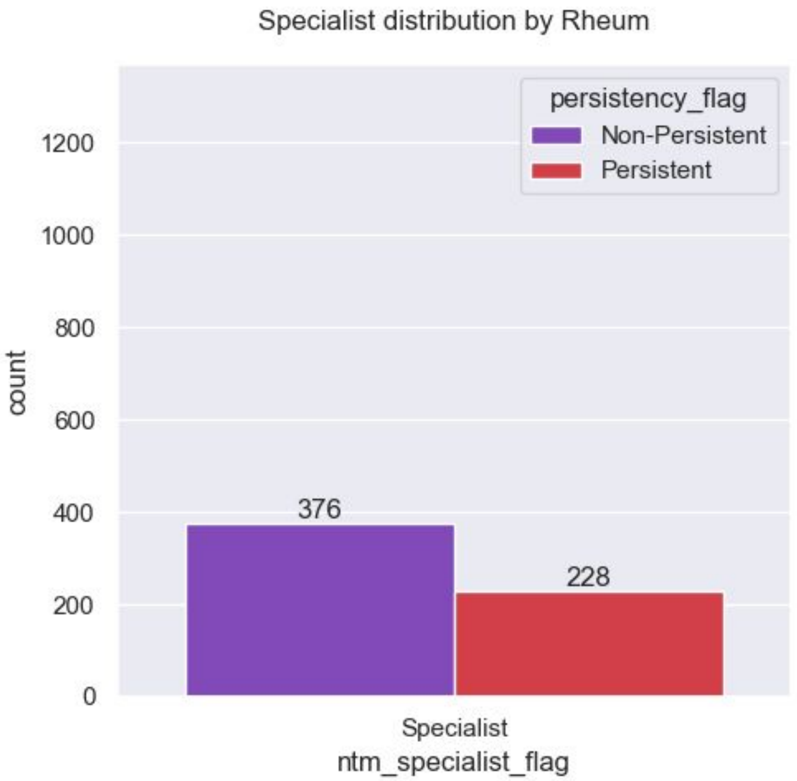
- Majority of the **Non-Persistent** patients have been prescribed the new medication by Physicians who are not *Specialists*.



Specialist distribution by OB/GYN/Others/PCP/Unknown



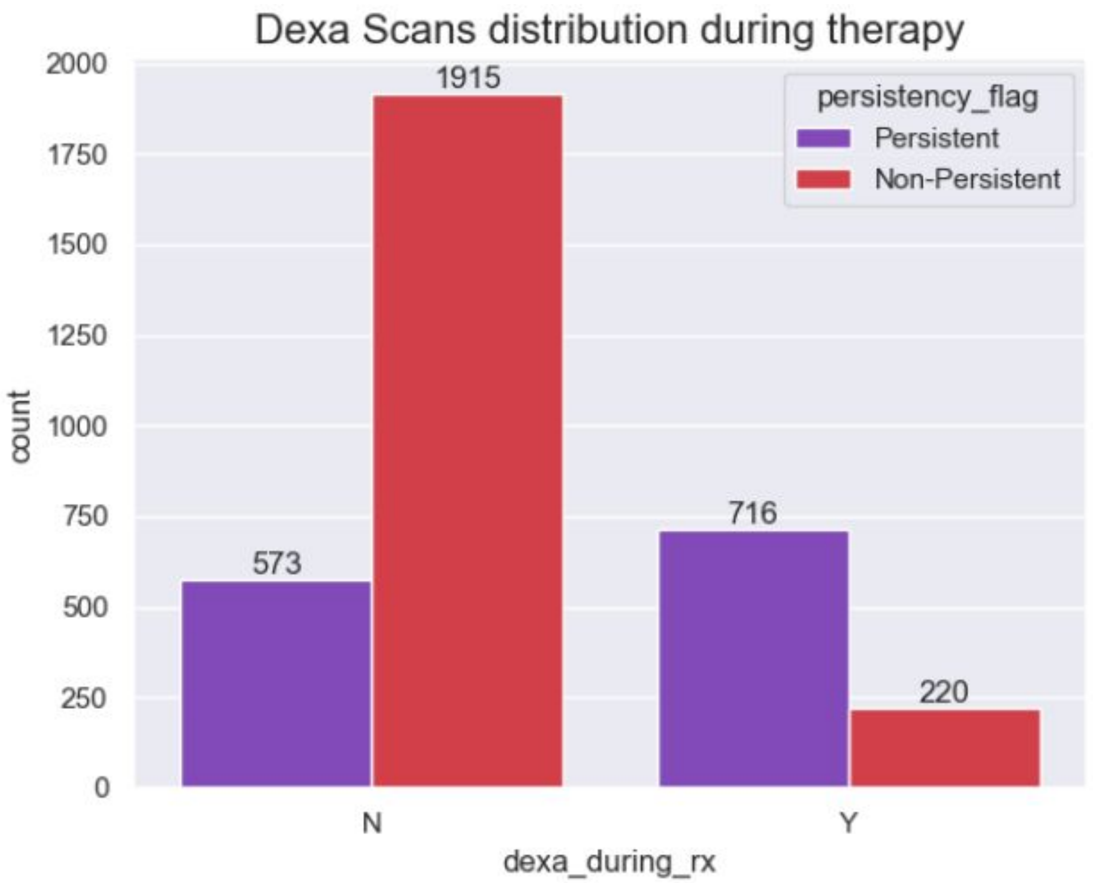
Specialist distribution by Endo/Onc/Uro



Specialist distribution by Rheum

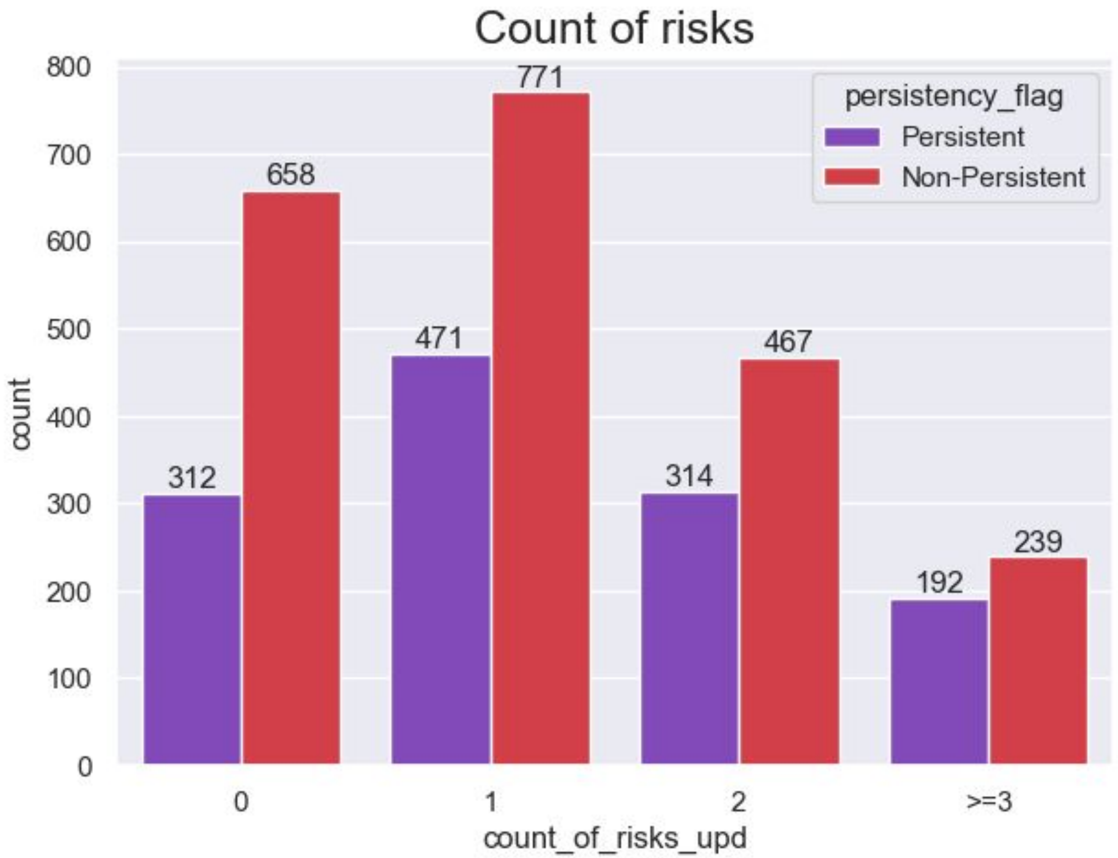
Clinical Factors

- 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**.



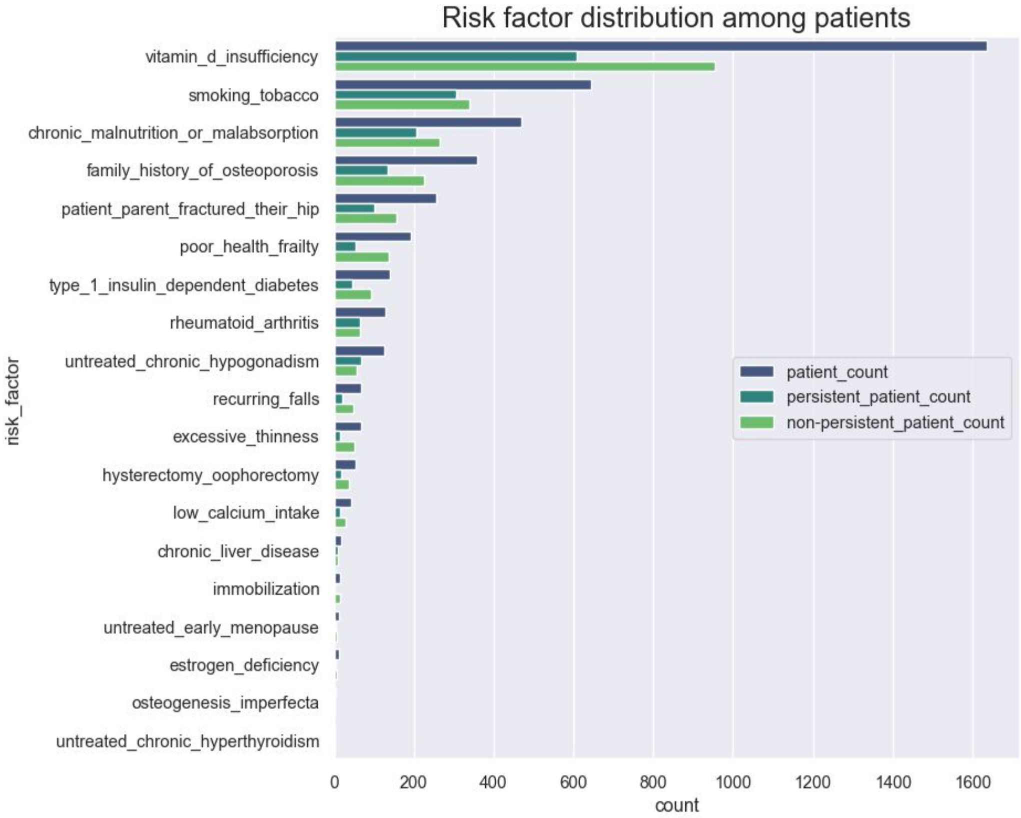
Risk Factors

- As the number of risks per patient increases, the number of **Non-Persistent** patients decreases.



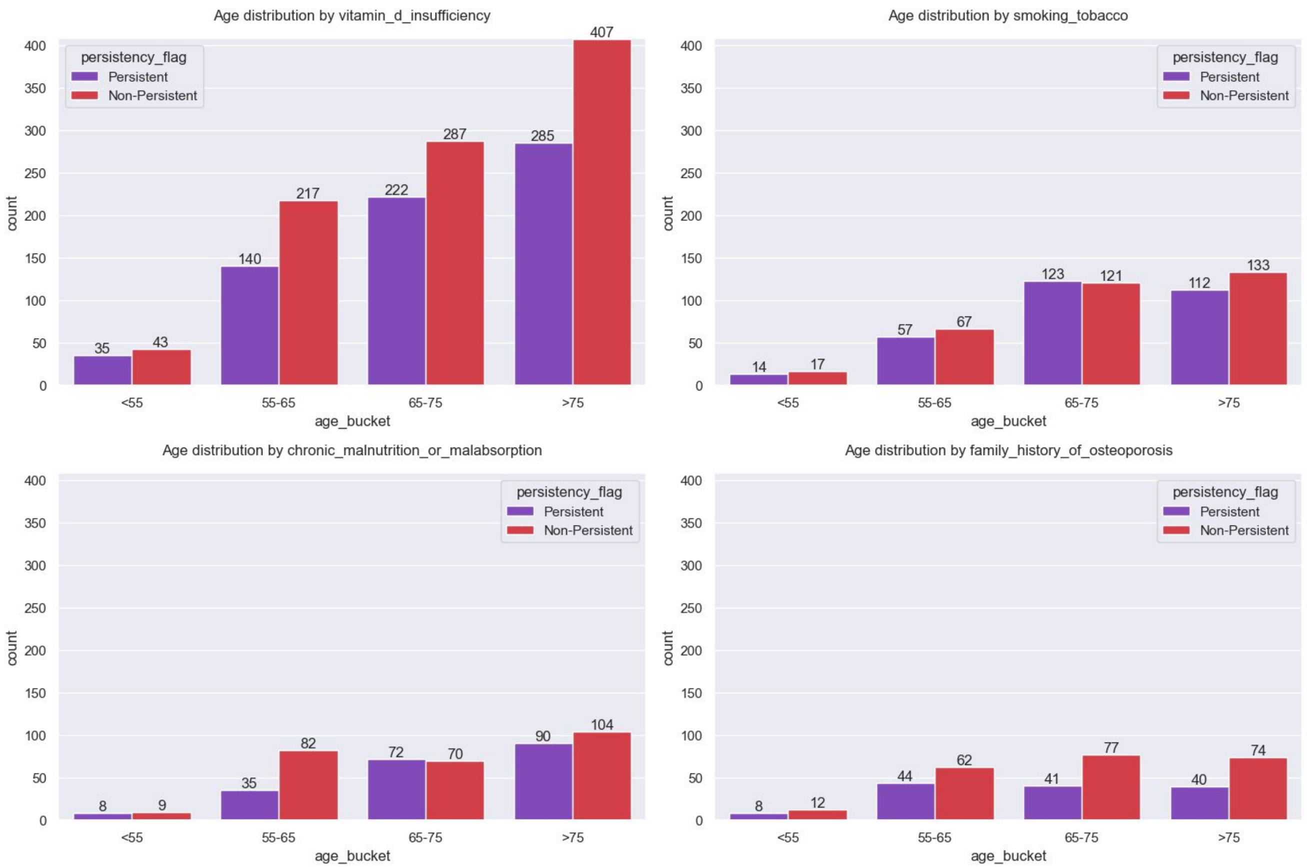
Risk Factors

- 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.



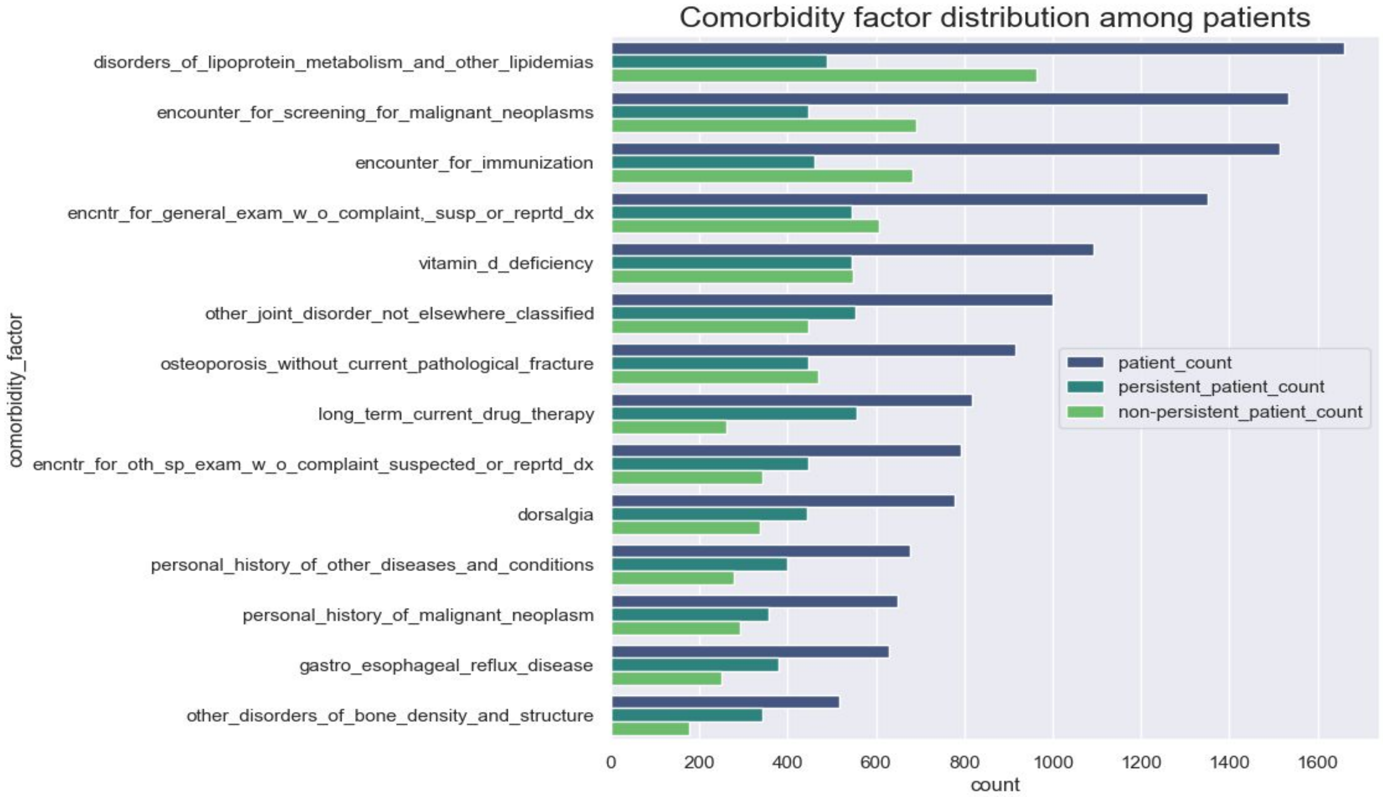
Risk Factors

- Below graph displays the distribution of top **Risks** between different **Age** groups 1 year prior starting NTM therapy.



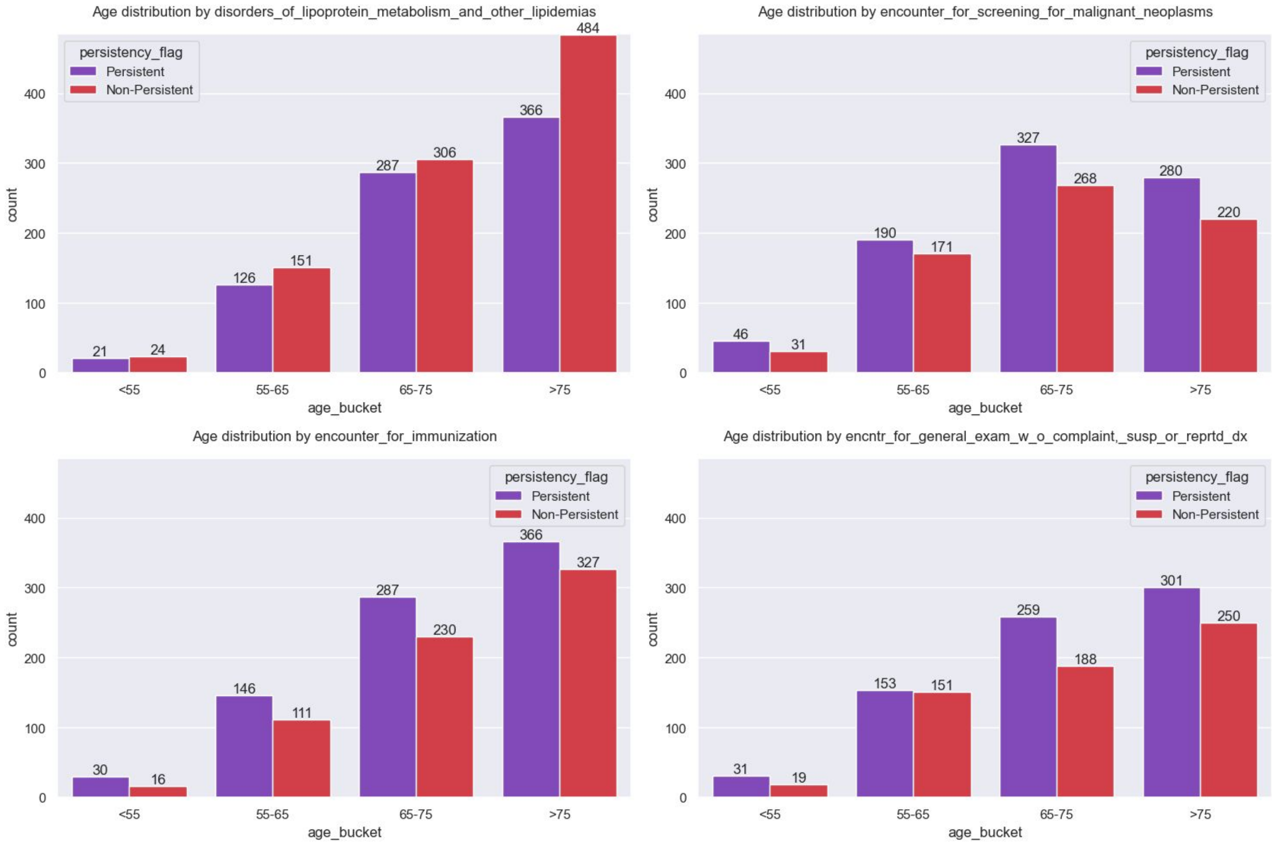
Comorbidity Factors

- 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.*



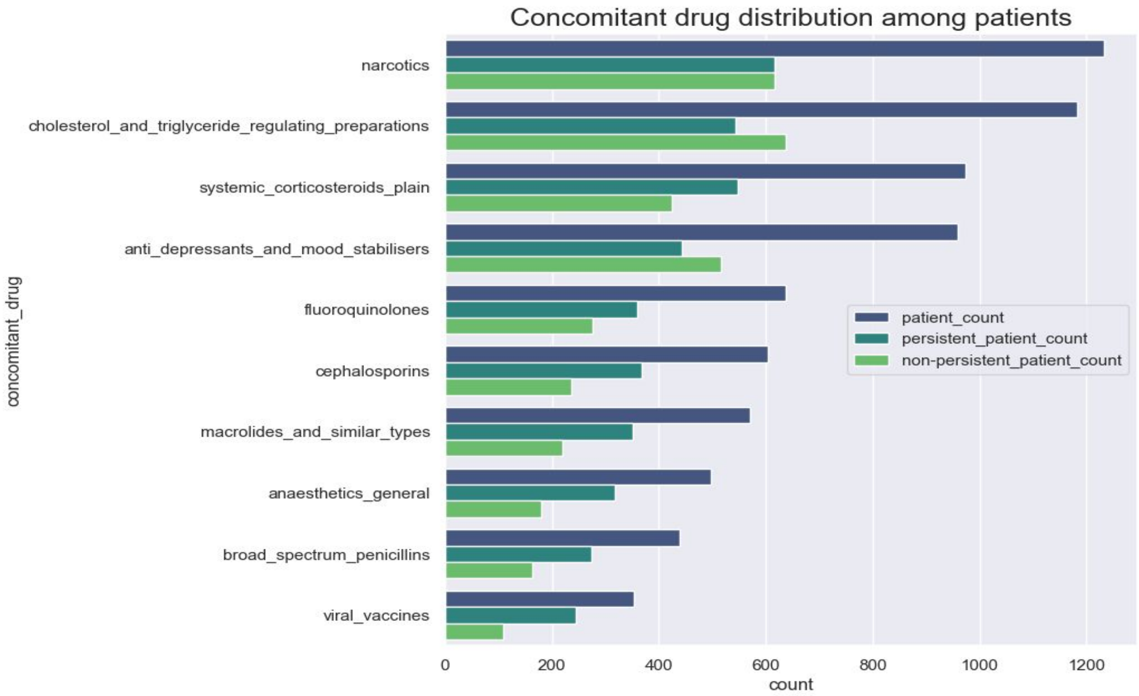
Comorbidity Factors

- Below graph displays the distribution of top **Comorbidities** between different **Age** groups 1 year prior to NTM OP therapy.



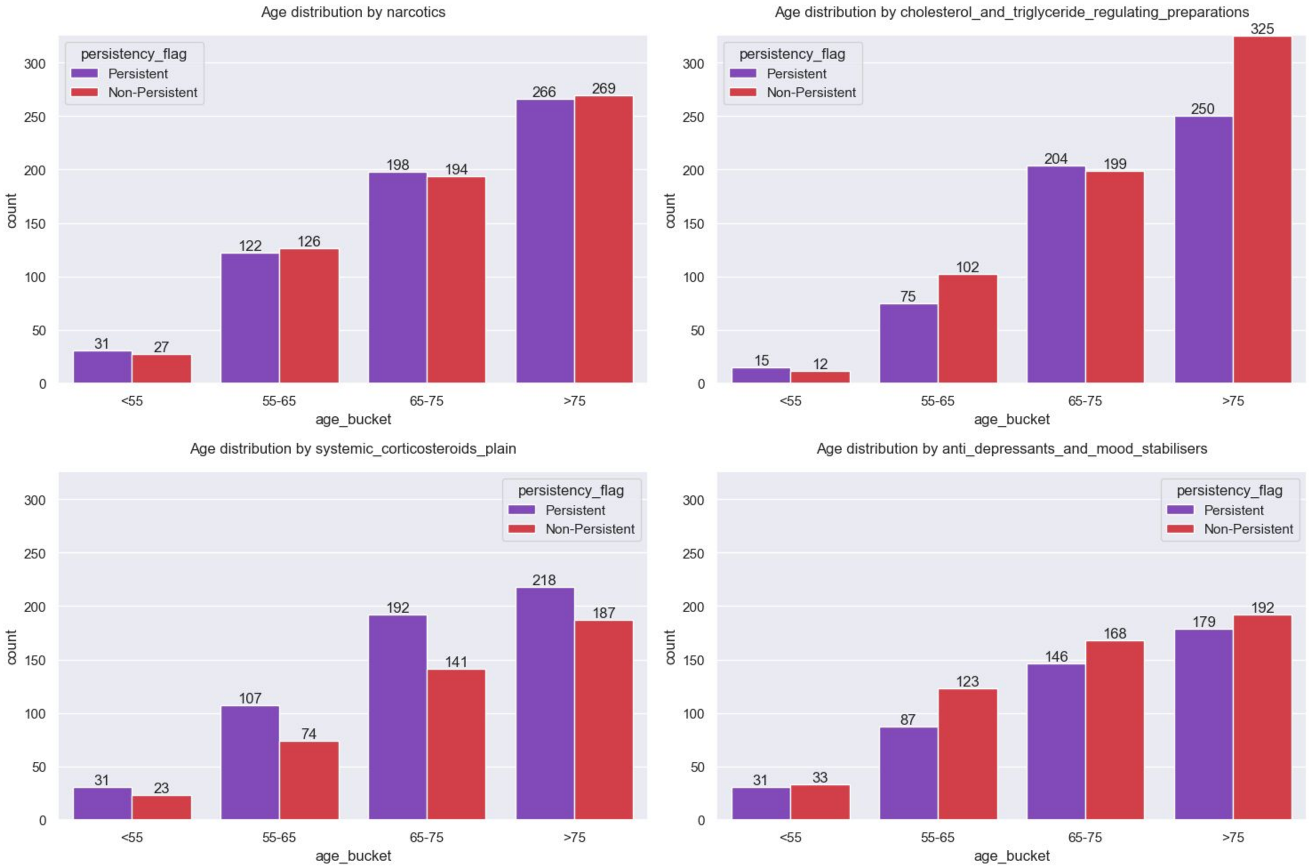
Concomitant Drugs

- 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.



Concomitant Drugs

- Below graph displays the distribution of top **Concomitant Drugs** administered to patients between different **Age** groups 1 year prior to NTM OP therapy.



Recommendations

Data Glacier

Your Deep Learning Partner

Recommended Models

- Majority of the attributes in the dataset are categorical in nature. **Label Encoding** will help in converting the data type from string to numerical.

- As the problem statement requires a classification model, following models are recommended - **Logistic Regression**, **Support Vector Classifier**, **Random Forest**, **Decision Tree Classifier** etc.

- Model optimisation can be carried out by applying **Grid Search** with **Cross Validation** as it helps in hyper-parameter tuning.

- Methods like **Recursive Feature Elimination**, **Attribute Relevance Analysis**, **Principal Component Analysis(PCA)** can be employed to handle dimensionality and complexity of dataset.

- Performance metrics such as **AUC-ROC curves**, **Accuracy**, and **F1-Score** will help us understand the performance of the models.

Data Glacier

Your Deep Learning Partner

Thank You