Data Glacier

Your Deep Learning Partner

Data Science Project

ABC Pharma Case Study of Drug Persistency

15th August, 2021

Drug Persistence Case Study

- One of the challenges for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription.
- Medical Persistence refers to the act of continuing the treatment for the prescribed duration, or the duration of time from initiation to discontinuation of therapy.
- Numerous studies have demonstrated that inadequate compliance and non persistence with prescribed medication regimens result in increased morbidity and mortality from a wide variety of illnesses, as well as increased healthcare costs.
- This source of wasted healthcare spending every year had an estimate in the US of the potential to reach even $300 billion, while also affecting pharmaceutical companies.

Drug Persistence Case Study

- A lot of factors could contribute to a patient stopping, or altering their medication regimen, to name a couple:
  - Competing priorities for patients: their lifestyle, their finances as well as emotional factors “Will the medicine make them feel sick?”, “Will it be a daily reminder of their illness?”
  - Competing priorities for doctors, such as lack of time to properly talk with the patients and explain them their situation and the need for medicatio. At times, less than a minute is given to the “what and why” about prescribed medication therapies, including side effects. Failure to adequately explain what the medication is and why it is important is a massive barrier to persistence.

ABC Pharma’s Case and Dataset

- Our particular case involves patients treated for an affliction which originates from a family of common organisms found in water and soil, which is Nontuberculous Mycobacterial.
- It is a rare affliction and can affect people with damaged lungs, or with a weakened immune system.
- If diagnosed, a patient might need up to two years of treatment and could get infected again in the future.
- This makes it very important for patients to remain persistent with their medication, since therapy could take a long time.

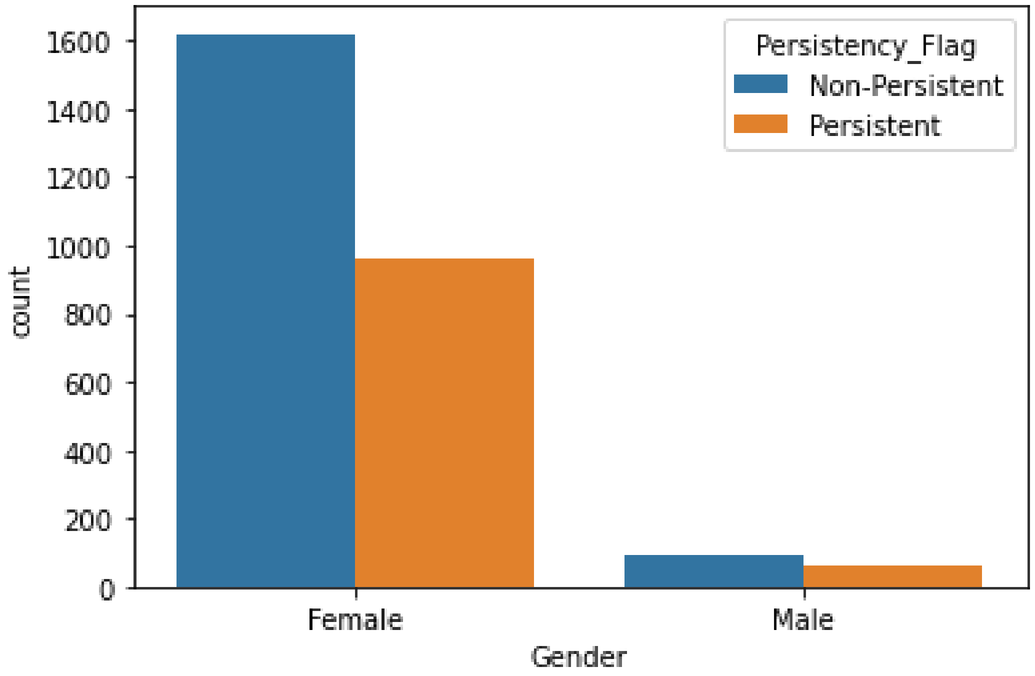
ABC Pharma’s Case and Dataset

- The dataset provided to us had 3424 patients with 67 features for each patient. The features could be grouped in four buckets:
  - **Demographics**: such as Age, Race, Gender, etc. for each patient.
  - **Provider Attributes**: some information about the provider that prescribed the medication to the patient, with variables such as the Specialty of the Physician.
  - **Clinical Factors**: certain physiological attributes which could be associated with the disease, with variables such as Frequency of a Dexa Scan.
  - **Disease/Treatment Factors**: such as a Comorbidity factor, divided into two categories – Acute and Chronic and a Concomitancy factor, i.e. concomitant drugs recorded prior to starting with the therapy.

Dataset Exploration

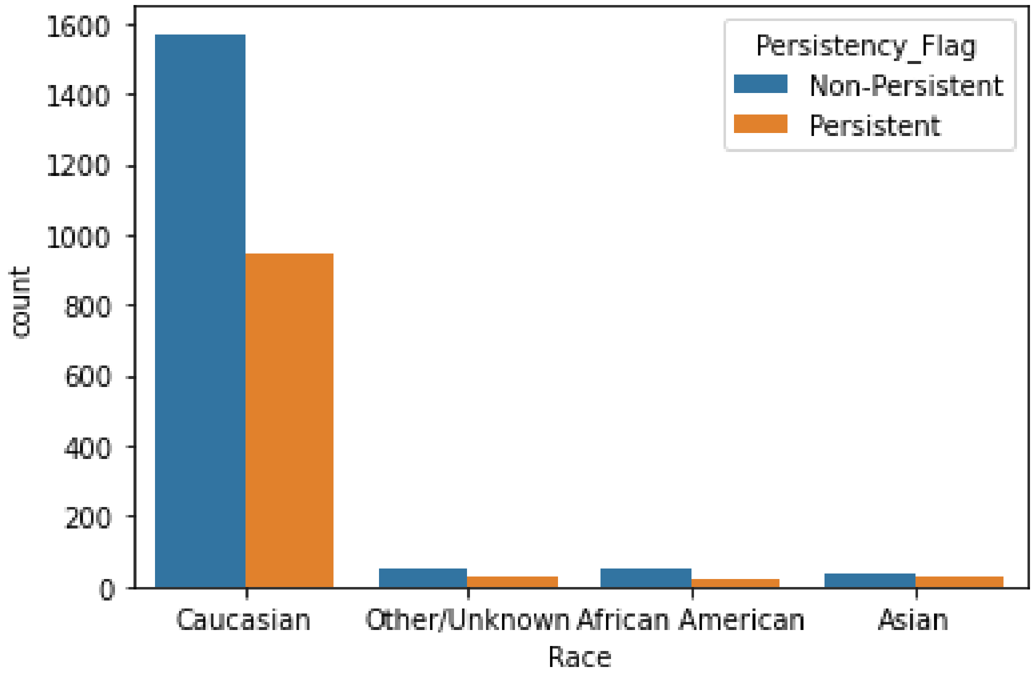
- Out of the 67 features, only 2 can be considered numerical, the rest are categorical, such as Race, or Physician’s Speciality.
- Out of the 2 numerical, 1 is removed ‘Count of Risks’, since it is a linear combination of all other Risk Factors.
- No missing values were found (NAs) but 5 features had ‘Unknown’ categories, which were considered as missing.
- One of these 5 was removed, as over 65% of its data points were missing, the other variables’ values imputed from all other features of the dataset.

Demographics Analysis



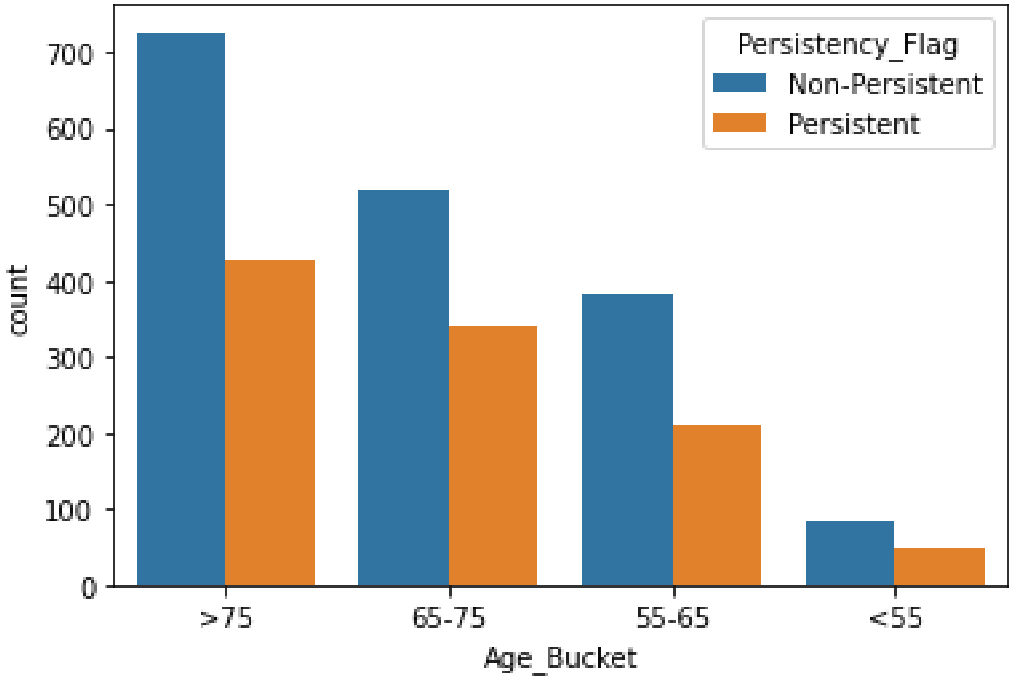
- The overwhelming majority of patients were women, 95.3%. Could be due to the fact that this affliction tends to affect women more frequently.
- Both gender appear to have more Non-Persistent cases, than Persistent ones.

Demographics Analysis



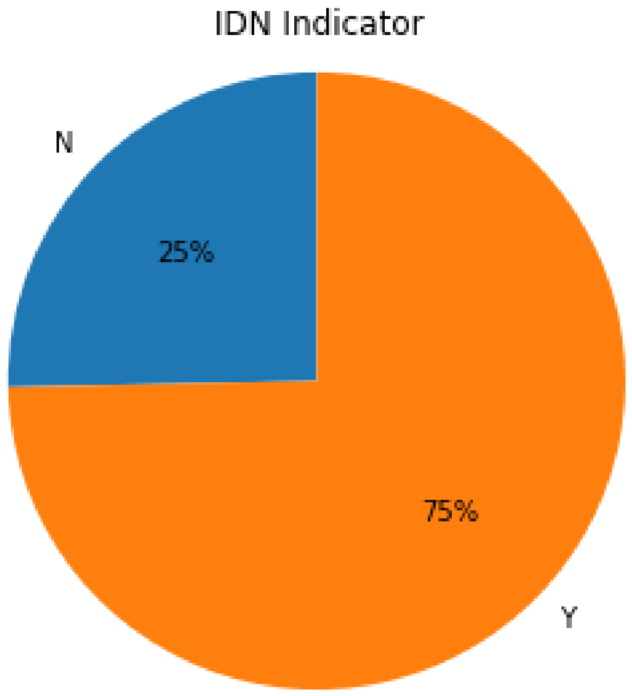
- Most of the patients, 91.75%, were Caucasian in Race, followed by Other, then African American and then Asian.
- Again, in all categories, we have more Non Persistent than Persistent cases.

Demographics Analysis



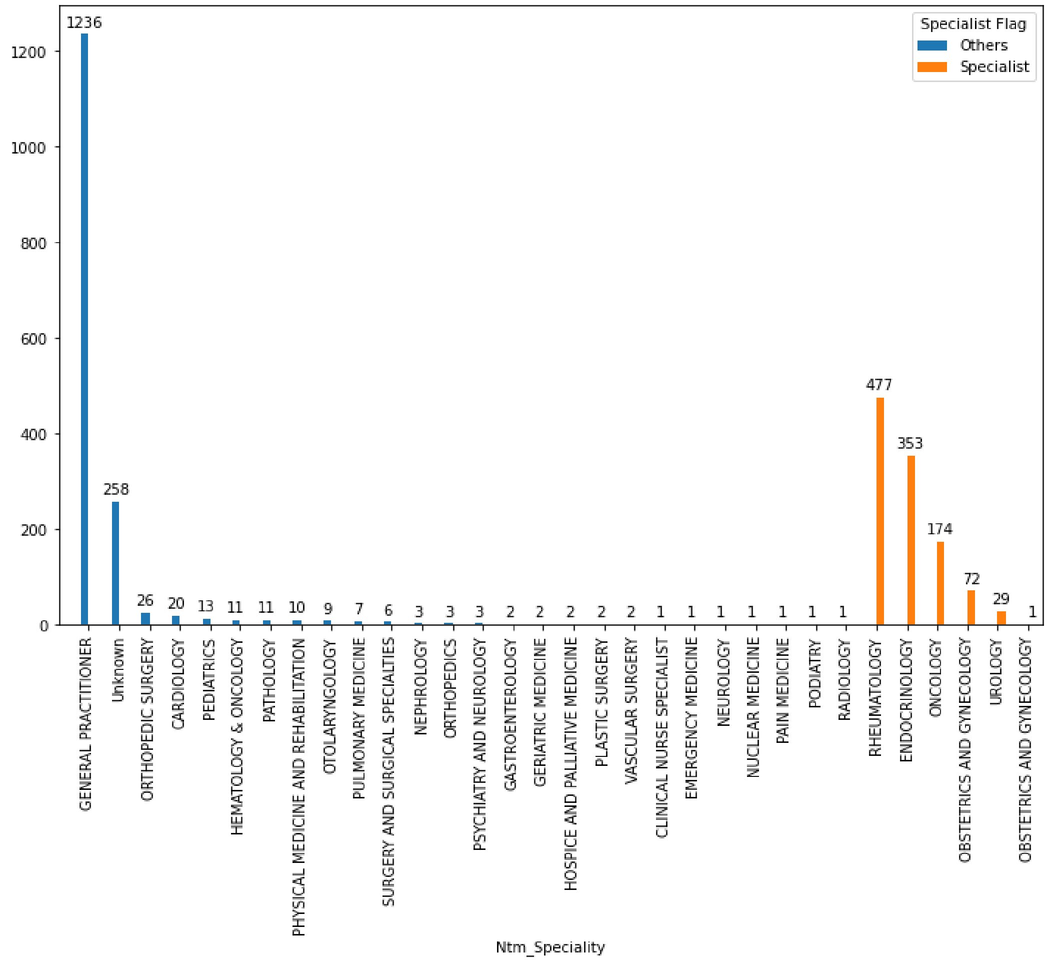
- The largerst percentage of patients belongs to the [>75] years of age bucket, with 42.17%, followed by the [65-75] bucket with 31.4%.
- Then [55-65] with 21.58% and [<55] years old with 4.9%.
- This categorisation seems to represent the population, as NTM tends to affect older people more.

Demographics Analysis



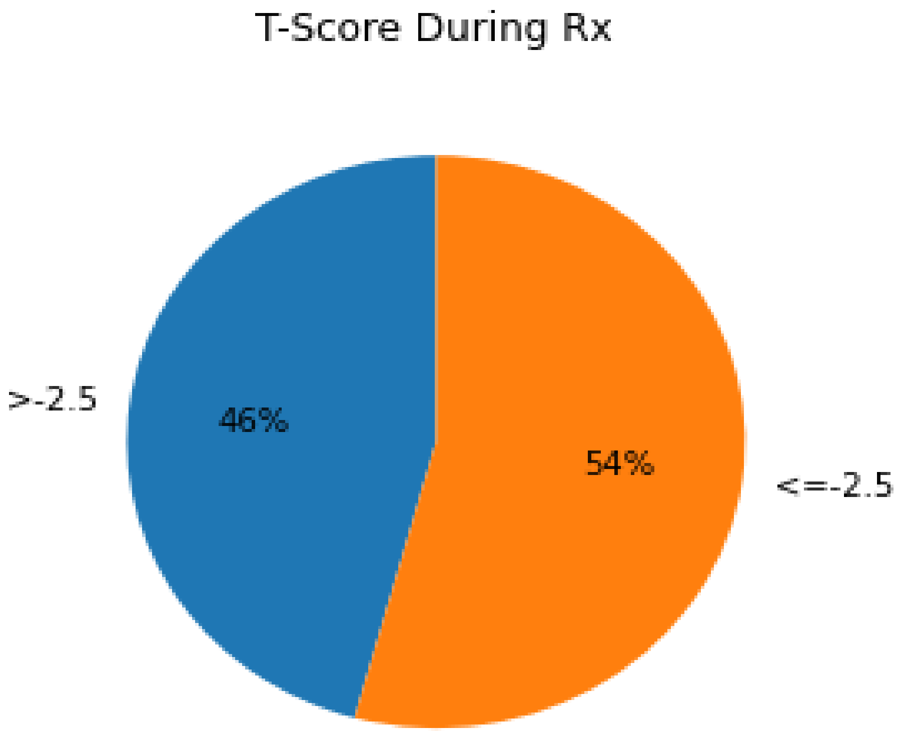
- This is the percentage of patients that were mapped to an Integrated Delivery Network (IDN), at 75%.
- An estimation of 76% of patients in the US are mapped to an IDN, so our dataset seems to follow that trend.

Physician Attributes



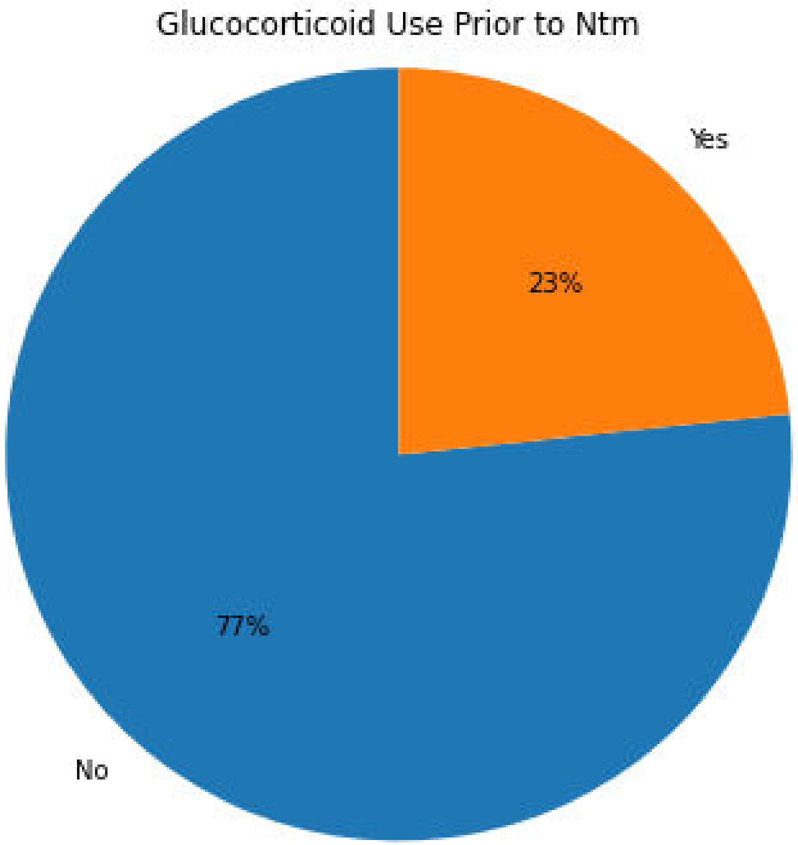
- There are 1106 Specialists in total and 1633 Others.
- Almost half of all our patients, 45%, received their prescription from a General Practitioner. This could indicate how people are being informed on whether they have the NTM infection or not, and how they are then given treatment.
- There are 258 Unknown values, which will be imputed from the rest of the categories.

Clinical Factors Analysis



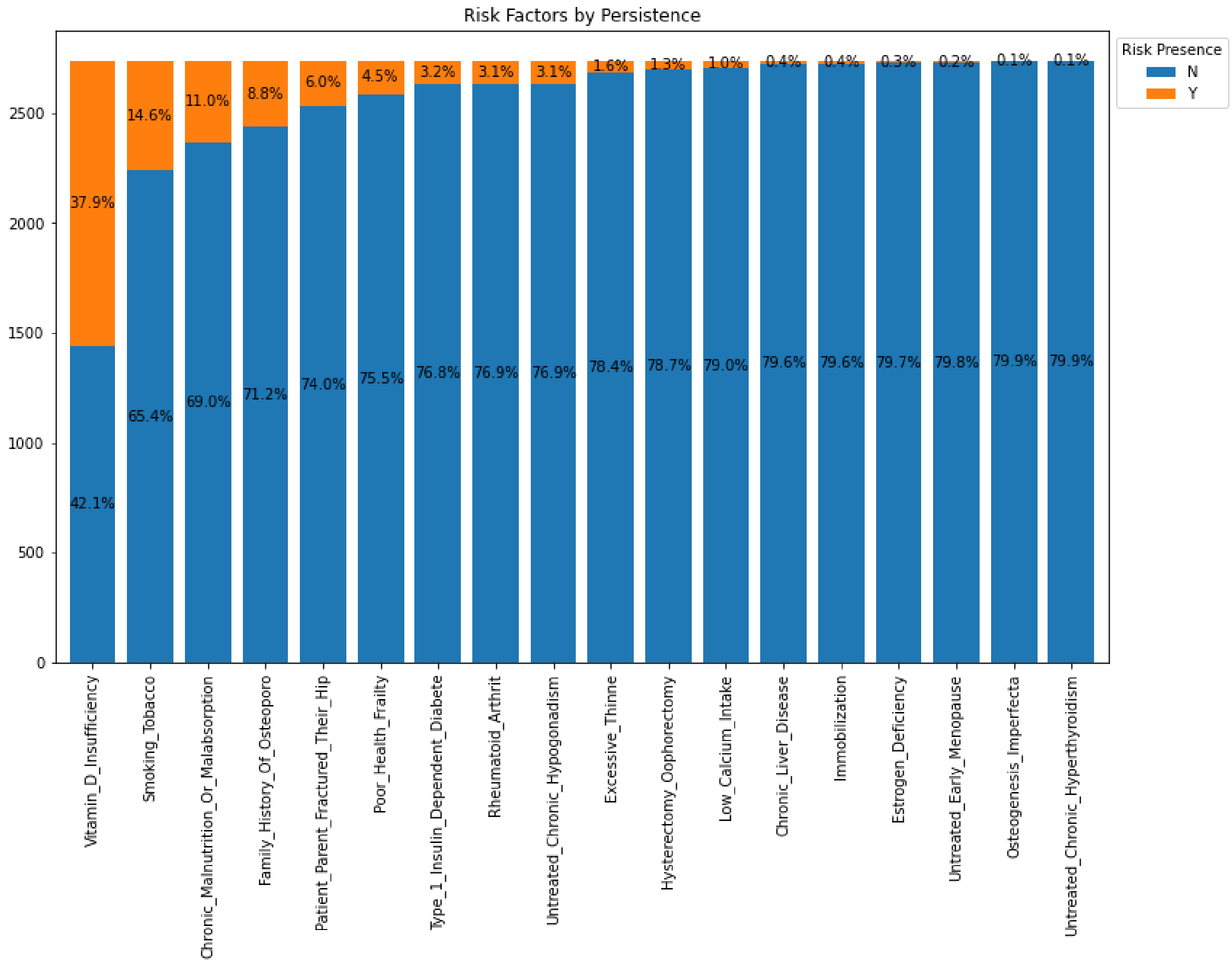
- A T-Score is produced from a bone densitometry scan (DEXA) and it determines bone mineral density. These values are during the patients' treatment period.
- Values below -2.5 indicate presence of osteoporosis, while values above -2.5 indicate low bone density.
- T-Scores before and during treatment for NTM showed an 88% of 'No Change'.

Clinical Factors Analysis



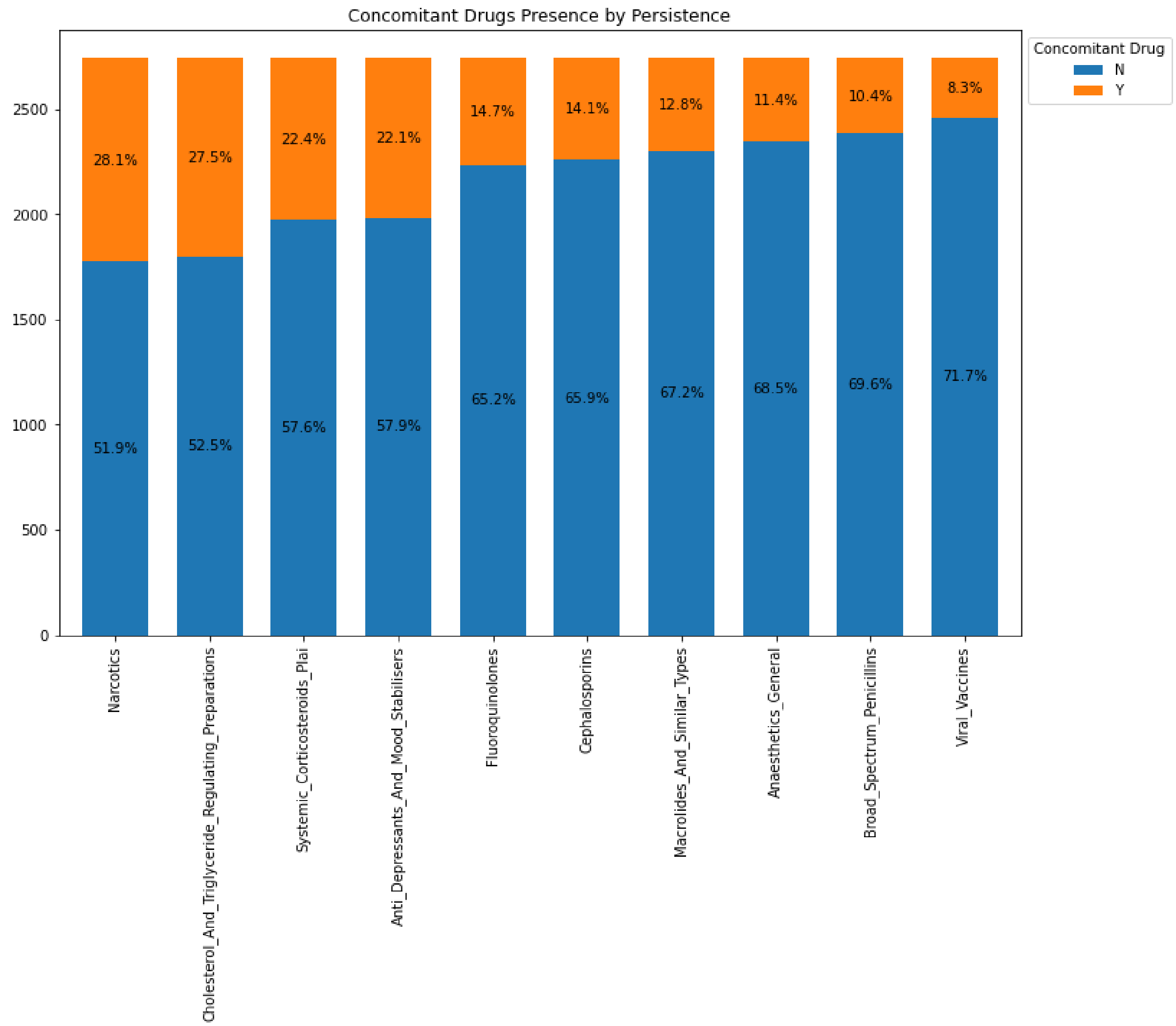
- Glucocorticoids are powerful medicines that fight inflammation and work with the patient's immune system to treat a wide range of health problems.
- We can see that most patients, 77%, did not use any Glucocorticoids before their treatment and a similar picture could be seen during their treatment, with 74%.

Disease / Treatment Factors



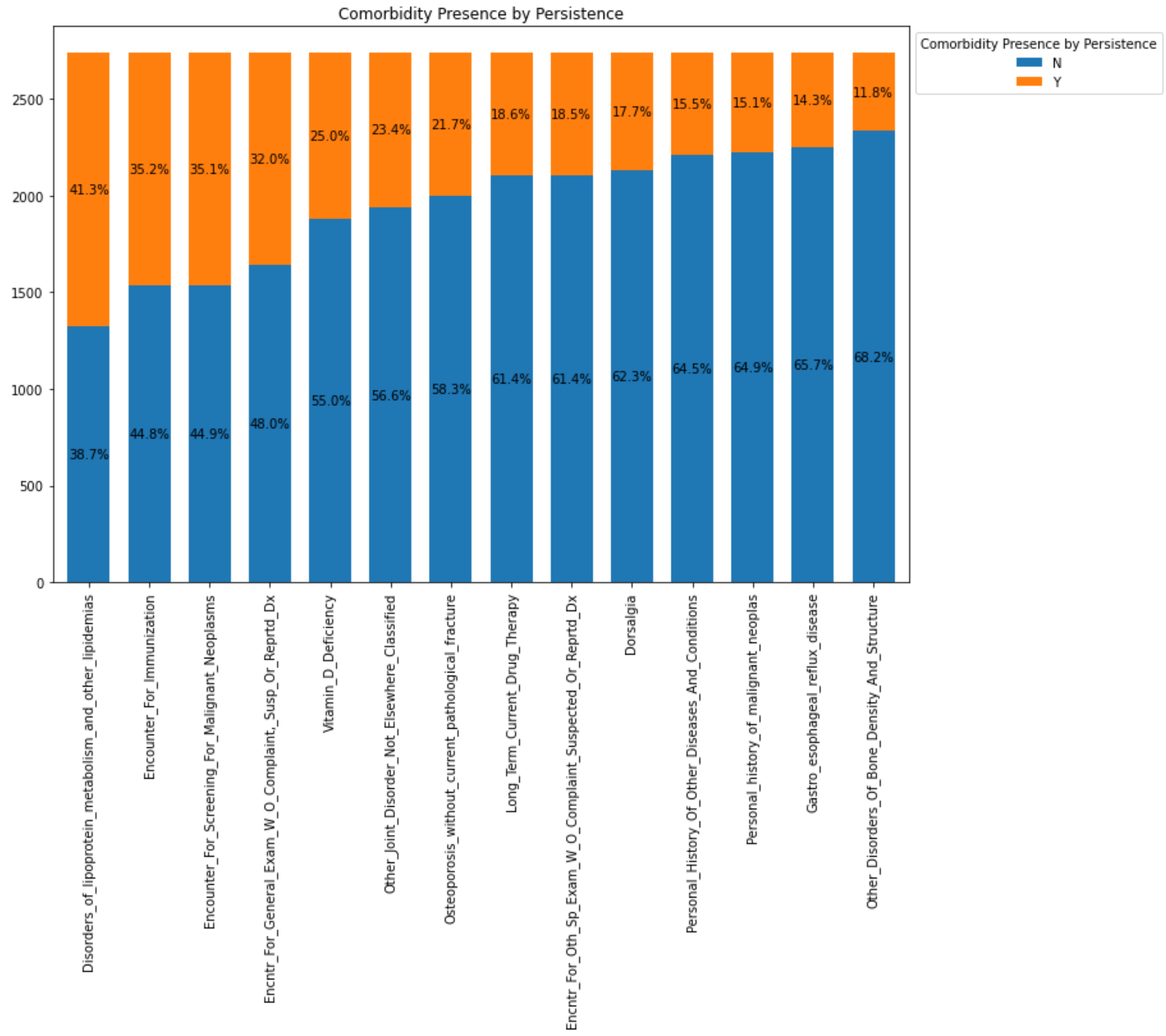
- Almost for each Risk Factor presence, patients were Non Persistent.
- An exception is seen where 37.9% of patients with ‘Vitamin D Insufficiency’ were Persistent.
- Followed by 14.5% by those who smoke tobacco.

Disease / Treatment Factors



- Concomitant drugs are two or more drugs used/given within 365 days prior to 1st prescription date.
- The cases with Concomitant drugs and highest patient Persistence were for various ‘Narcotics’, ‘Cholesterol Regulators’, ‘Systemic Corticosteroids’ and ‘Anti Depressants’.

Disease / Treatment Factors



- Comorbidity is the presence of one or more additional conditions, often co-occurring, with a primary condition.
- Patients with ‘Lipoprotein Metabolism Disorder’ tend to be more Persistent, with 41%.

Univariate Relationships

Variable	pvalue
Gender	0.426
Age_Bucket	0.416
Gluco_Record_Prior_Ntm	1.0
Frag_Frac_Prior_Ntm	0.845
Risk_Segment_Prior_Ntm	0.618
. . .	. . .
Risk_Hysterectomy	0.408
Risk_Estrogen_Deficiency	0.548
Risk_Recurring_Falls	0.699
Race	0.420
Ethnicity	0.463

- One-on-One relationships of variables considered to be independent from our target variable, 'Persistence'. In total there are 20 such variables.
- Pvalues are between 0 and 1, and anything above 0.05 indicates an independent relationship.
- 13 out of 19 Risk variables were on this table.

EDA Results

Following the Exploratory Data Analysis (EDA) we have noted:

- The majority of patients were Non Persistent than Persistent, about 63%.
- Most patients are Caucasian Females, over the ages of 65 and come from either the Midwest of South regions of the US.
- Almost half of all the patients, received their prescription from a General Practitioner.
- From the Persistent patients, there are people with high Vitamin D Defficiency, some with Lipoprotein Metabolism Disorder, Malignant Neoplasms.
- When it comes to one-on-one relationships with our target variable, almost all Risk Factors seem to be independent from it.

Recommendations

- Further analysis of all/almost all predictors and our target variable should be done in order to determine whether or not they can affect it.
- This can be achieved through Machine Learning (ML) models that will classify each data point according to the predictors' values.
- Some ML model could be Logistic Regression Classification, Trees Classification and ensemble methods, such as AdaBoost.

Feature Selection

- In this part of the analysis the features that seemed most important in the training of our model on the 'Train' dataset were chosen.
- An 87% accuracy was achieved with 31 features.
- An 86% accuracy was achieved with 20 features.
- We choose the 31 features since there is not much difference in computational speed.

Feature Selection

Feature
Whether or not patient had a Dexa Scan during Rx
Comorbidity of Bone Density Disorders
Comorbidity for Malignant Neoplasms
Concomitancy of Fluoroquinolones
Physician Speciality_Oncology
Comorbidity of Long Term Drug Therapy
Comorbidity of Systemic Corticosteoids Plain
Dexa Scan Frequency During Rx
Gluco Record During Rx
Concomitancy of Narcotics
Comorbidity of Vitamin D Defficiency

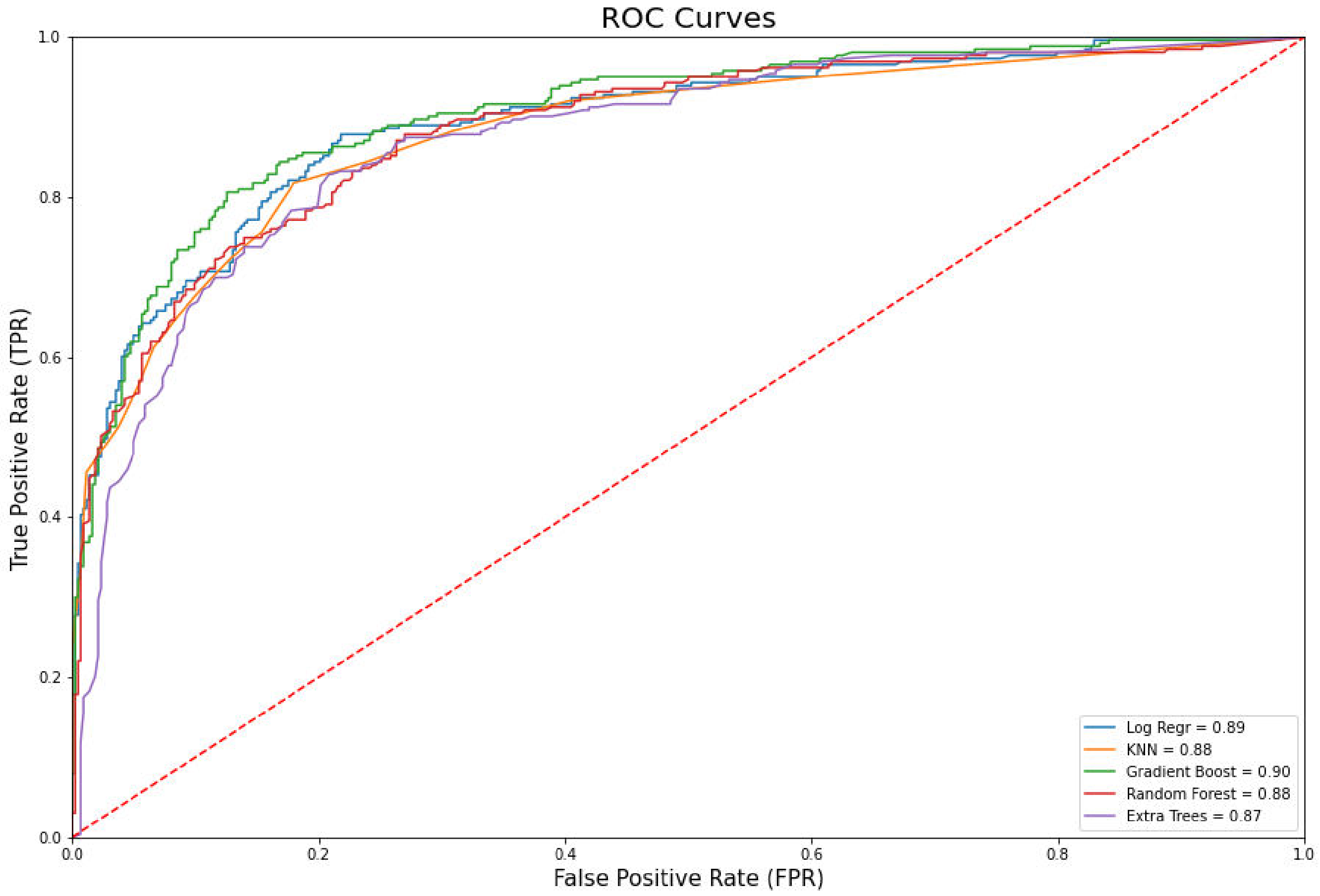
- Some of the features chosen are presented here.
- Dexa Scans are important features.
- Some comorbidities and concomitancies as well.
- As far as Risk factors are concerned, none were picked up as features.

Machine Learning Models Results

Classification Method	Accuracy	Precision	AUC
Logistic Regression	80.44	80.57	88.55
RandomForest	81.31	81.26	88.44
K Nearest Neighbors	81.31	81.47	88.18
Gradient Boosting	84.10	84.00	90.05
ExtraTrees	80.09	80.08	87.40

- The Gradient Boosting ensemble method seems to be most accurate, with Random Forest following.
- The resulting models do not differ by much, which can also be observed from their ROC curves plotted below.

ROC Curves of ML Models



Thank You

Group Name: BetterHealth Analytics
Name: Enias Vontas Country: Greece Email: vondas100@gmail.com Specialization: Data Science

Data Glacier
Your Deep Learning Partner