Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis and Model Recommendation

Project: Persistency of a Drug-Gathering insights for a Pharma company

Name: Shruthi Madgi

Email:Shruthi.madgi05@gmail.com

Country: USA

University: Visvesvaraya Technological University

Specialization: Data Science

Batch no:LISUM17

Reviewer: Data Glacier

Agenda

Problem Statement:

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription whether a patient will be persistent in completing his/her dose. To solve this problem, ABC pharma company has approached an analytics company to automate the process of identification.

Objective is to build a classification model for drug persistency identification.

Data and Approach

The dataset is in .xlsx format and has 3424 points and feature vectors have 69 variables. It has 2 discrete value features/columns,1 continuous value feature and 66 categorical value columns/features including the target which is the Persistency flag.

The goal of using this dataset is to predict whether the patient will be persistent in completing his/her treatment.

Apart from individual identificators and the target variables, there are four other buckets:

- Demographics
- Provider attributes
- Clinical factors
- Disease and treatment factors

Data provided was cleaned of outliers, missing values and skewed data and was converted into a csv file.

Five models have been discussed on the basis of EDA.

Risk, Comorbidity and Concomitancy factors

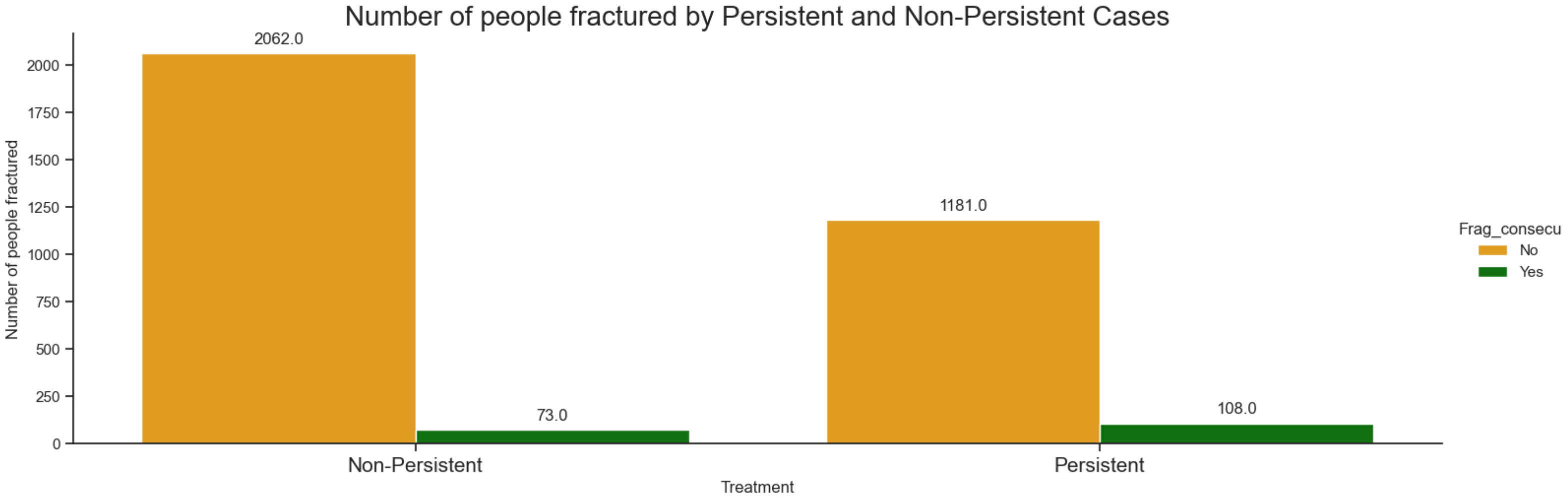
Most of the patients already hold comorbidity factors, while having risk factors is much less common.

- The main comorbidity factor is related to lipoproteins, metabolism and cholesterol.
- The main risk factor is deficiency in Vitamin D.
- More than one third have been found to be taken narcotics.

There are some significant differences between genders:

- Women seem to be more affected by vitamin D deficiency.
- There is more than a twofold difference in the number of women and men who have passed a screening for malignant neoplasms, with women passing in greater numbers.
- Four times as many men than women suffer from Hypogonadism (untreated).
- Patients older than 65 are affected by the mentioned factors in a higher proportion.
- There are some risks and other factors that seem to be significantly higher in Southern and West regions.
- There seem to be some remarkable differences between Asian and other races(due to either culture or race).

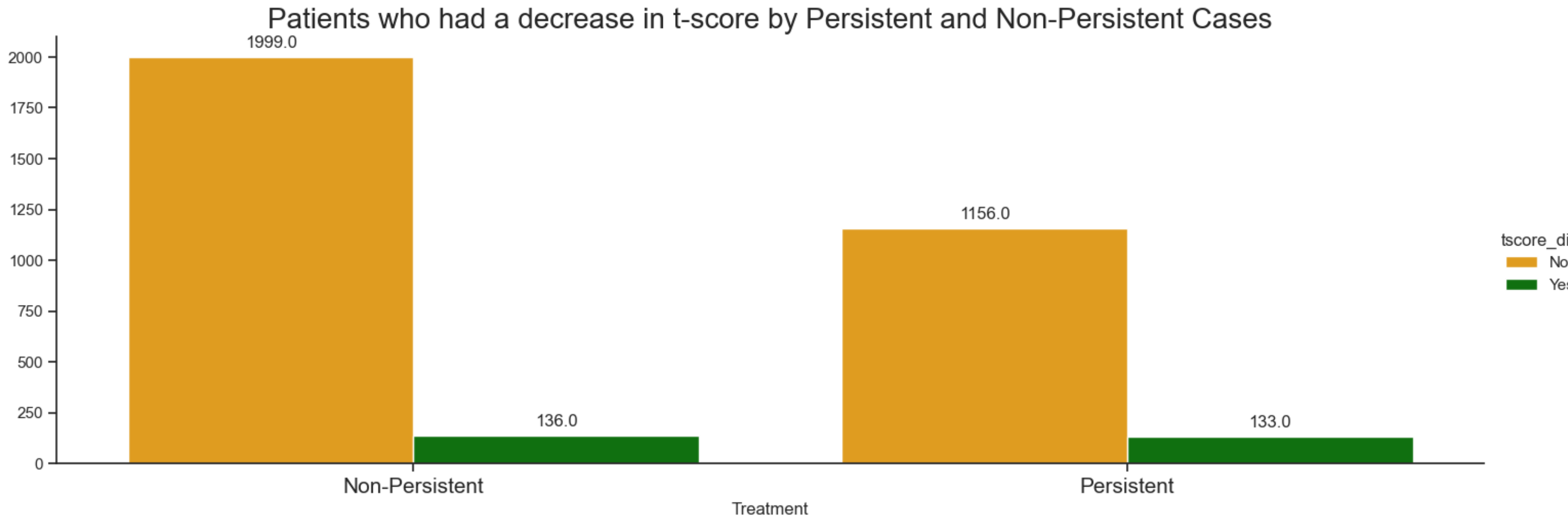
Fracture variable



Of the total number of patients, 8.38% of people were affected by the treatment, weakening their bones.

The percentage of people affected by treatment is significantly low with the sample size provided, i.e, out of 1280 who were persistent with the treatment, only 108 people were affected by the treatment. Thus, it can be inferred that the drug does not effect the considerable number of patients who adhere to the routine.

T-score variable

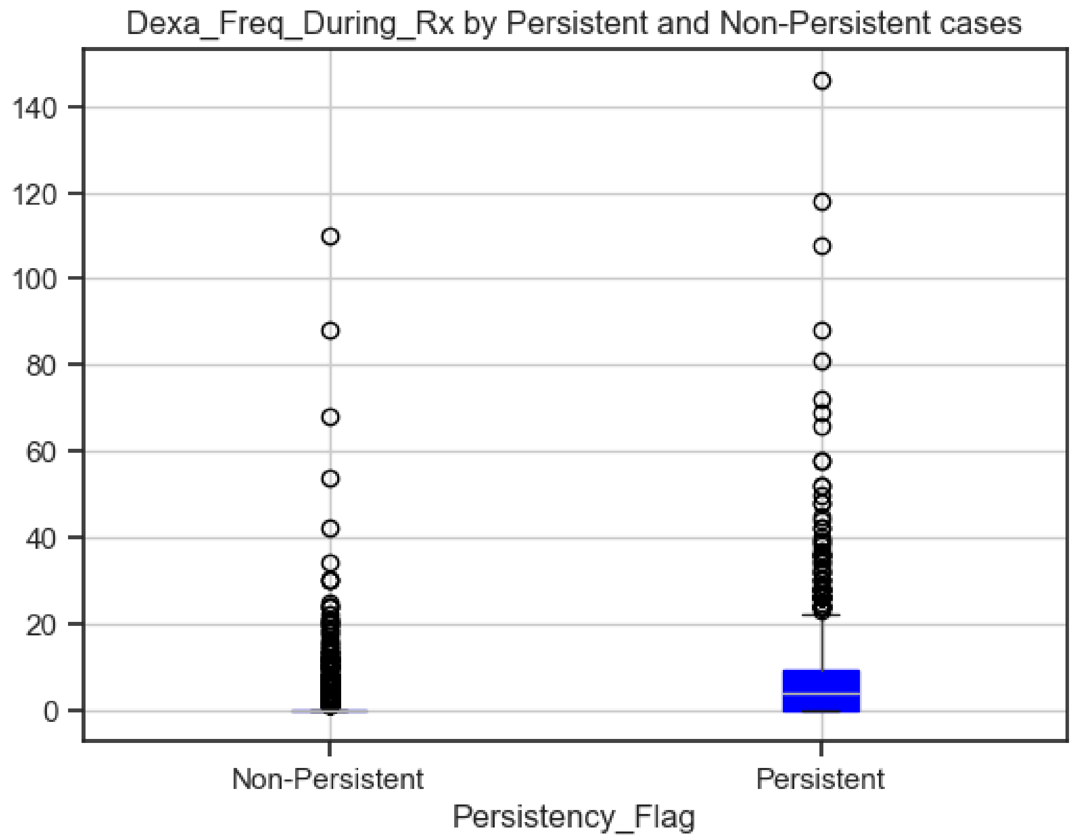


There are 10.31% of people with treatment who had a decrease in their T-score.

This indicates that T score is improving with the persistent routine.

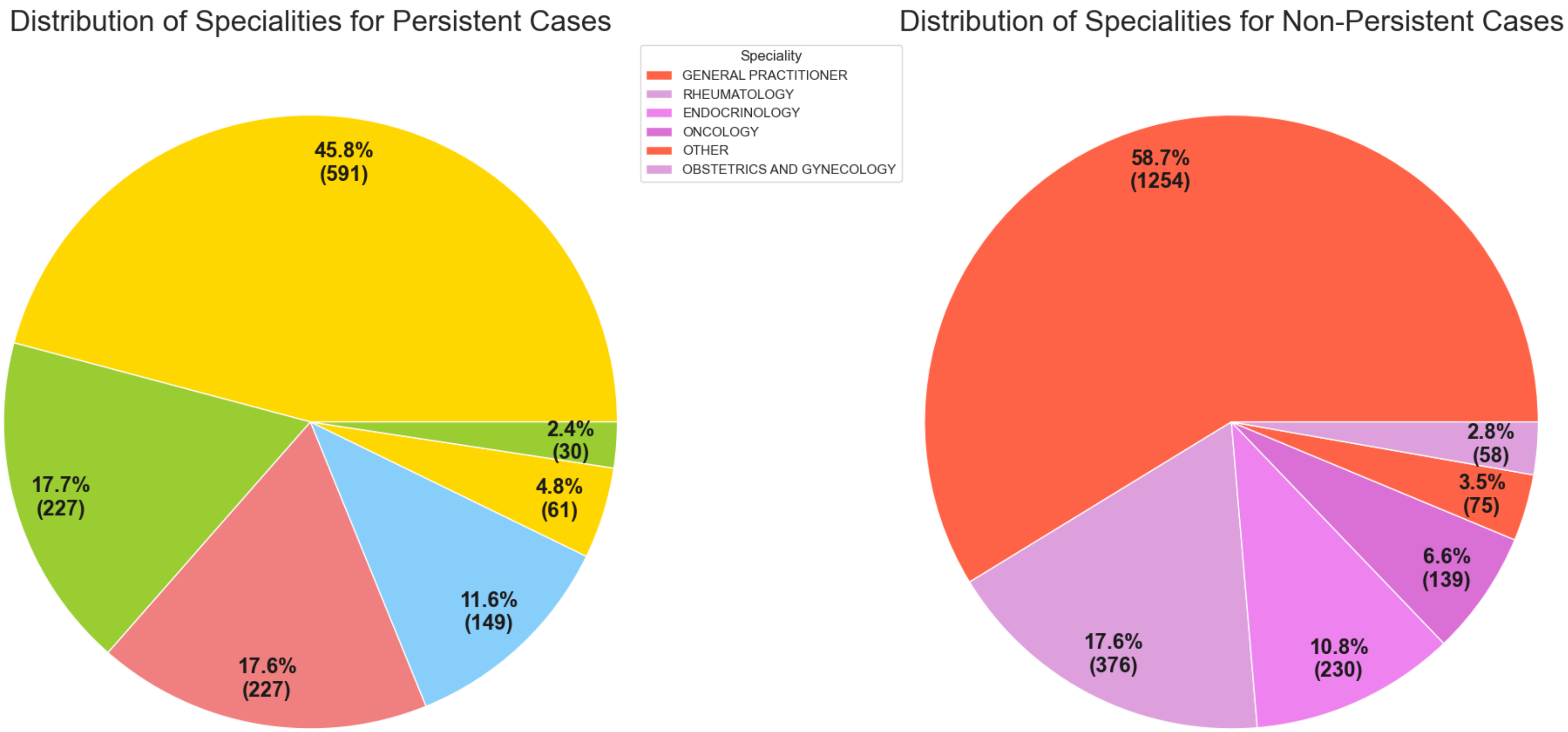
Clinical Factors

The distribution of Dexa_Freq_During_Rx numbers is higher in the Persistent patients.



The distributions of frequency for the target variable by specialty are pretty similar. There does not seem to be any effect of specialty.

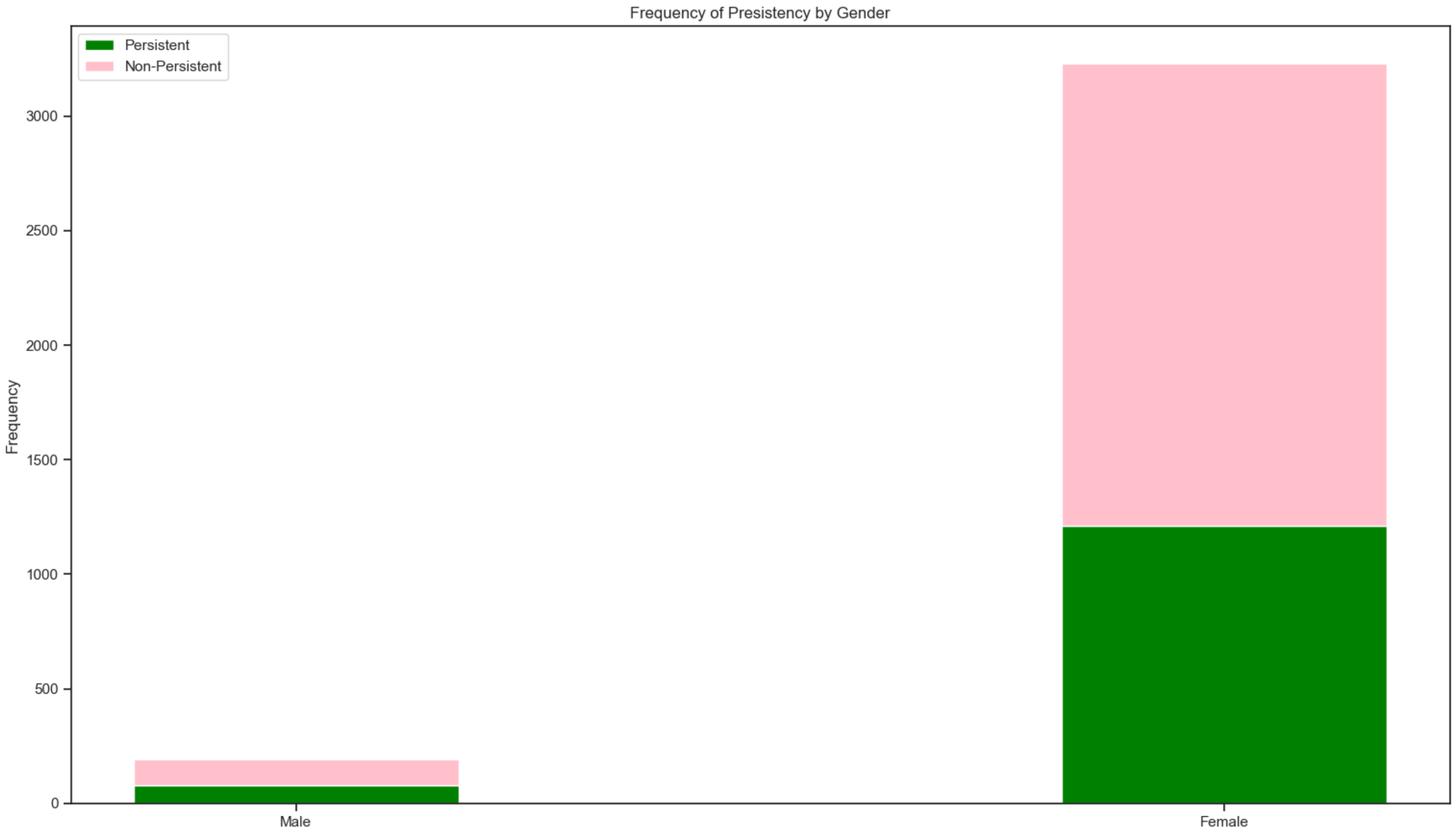
Thus, it can be inferred that specialty of the physician does not have a bearing on the adherence of the drug.



Distribution of Specialities for Persistent Cases

Distribution of Specialities for Non-Persistent Cases

Drug persistency by Gender



60.31% of males were flagged as non-persistent.62.48% of females were flagged as non-persistent.
Both the genders seem to experience same result with the drug persistency.

Model recommendations

1.Decision Tree algorithm: Decision trees are a type of machine learning algorithm that work by recursively splitting the dataset based on the values of the input features, with the goal of creating subgroups that are more homogeneous in terms of the target variable (in this case, drug persistence). The max depth parameter determines the maximum number of levels or splits that the tree can have. In this case, the best results were obtained with a max depth of 1, which suggests that one of the input features had a strong predictive power in determining drug persistence. One advantage of decision trees is their interpretability, as they allow for visual inspection of the decision-making process.

2.Random Forest algorithm: Random Forest is an ensemble learning method that works by combining multiple decision trees, with the goal of reducing overfitting and increasing prediction accuracy. The number of estimators parameter determines the number of decision trees that will be trained. The max depth parameter sets a limit on the depth of each individual decision tree to prevent overfitting. The accuracy of 81% on the test data suggests that the model was able to accurately classify drug outcomes. The AUC of 89% suggests that the model has good discriminative power, i.e. it is able to differentiate well between positive and negative drug outcomes.

3.Support Vector Machines algorithm: Support Vector Machines (SVM) is a popular machine learning algorithm that can be used for classification and regression problems. The linear kernel is one type of SVM kernel function that works by finding the hyperplane that best separates the data into two classes (in this case, positive and negative outcomes). The accuracy of 83.5% on the test data suggests that the model was able to accurately classify drug outcomes. One advantage of SVMs is their ability to handle high-dimensional data, such as the 83-dimensional feature vectors used in this study. The output labels of 1 and -1 are a common convention in SVMs, with positive outcomes assigned a label of 1 and negative outcomes assigned a label of -1.

4.K-Nearest Neighbors (KNN): With the results from Random Forest (RF) and Support Vector Machines (SVM), KNN can be used to predict probabilities or scores from RF and SVM as additional features, along with the original 64 features, to train a KNN model. This allows the KNN model to potentially capture additional patterns in the data that were not captured by the RF or SVM models alone. To implement this approach, first train the RF and SVM models on the dataset, then use their predicted probabilities or scores as additional features in a new feature matrix. Then train a KNN model on this new feature matrix, choosing an appropriate value of k (the number of nearest neighbors to consider) using cross-validation, and evaluate the performance of the KNN model on a separate test set. It's important to use appropriate regularization techniques and evaluate the model carefully to avoid overfitting.

5.Logistic Regression (LR): Model was trained on a dataset of drug persistence, using binary classification, where 0 represents Non-Persistent and 1 represents Persistent labels. The LR model was optimized using Bayesian optimization, which builds a probabilistic model of the objective function, and selects the next set of hyperparameters to evaluate. The Bayesian optimization algorithm searched for the optimal values of hyperparameters such as regularization strength, learning rate, or penalty type, and achieved an f1_score of 84%.Overall, the Bayesian optimization technique proved to be an efficient alternative for optimizing hyperparameters in the LR model, leading to better predictions of drug persistence.

As per EDA, there can be as many as five algorithms to develop a model that can predict the adherence of a drug by the patient. Among the five, the best algorithms to follow are Decision Tree, Support Vector Machine and Random Forest.

GitHub: https://github.com/shru0405/DataGlacierInternship/tree/main/Week_11