Patient Persistency on a Bone Density Treatment

Final Report

Data Glacier

Your Deep Learning Partner

By the Data Dynamos Team:

Jen Turley, Queen Echerenwa and Ateeb Aqil

December 18, 2021

Table of Contents

- 1. Overview of the project including information on the Exploratory Data Analysis for the provided dataset.

- 2. Detailed information on the classification methods utilized and the results of the models

- 3. Conclusions

Section 1:

Overview of the Project including Exploratory Data Analysis

Insights into Drug Persistency

Overview:

A classification model for drug persistency was created using a health database file containing clinical and demographic data for patients on a bone density treatment. The ultimate goal of this project is to understand better the patient and provider factors associated with patients’ drug persistency on this treatment.

Utilizing the Microsoft Excel Spreadsheet file **healthcare_dataset.xlsx** which contains data for 3424 patients prescribed a bone density treatment, exploratory data analysis and machine learning classification models were applied.

The Persistency_Flag column within this data set indicates for each patient whether the patient continues taking the medication (is persistent) or if the patient has discontinued the prescribed treatment (is non-persistent).

When a patient fails to be persistent with a medical treatment, there may be adverse health consequences for the patient. When many patients fail to be persistent, there are adverse consequences for the medical providers and the pharmaceutical company that produces the drug. Therefore, it is in the interest of medical providers and pharmaceutical companies to understand the degree of drug persistency for a medication and what are the factors possibly influencing drug persistency.

Data Exploration:

Exploration of the data revealed an imbalance between persistency and non-persistency: of 3424 patients total, only 1289 were persistent while 2135 were non-persistent.

Alarmingly, a majority of patients were non-persistent!

Can drug persistency be accurately predicted? If so, are there identifiable factors associated with non-persistence? Which of these factors might be modifiable by the medical provider?

Data Cleansing:

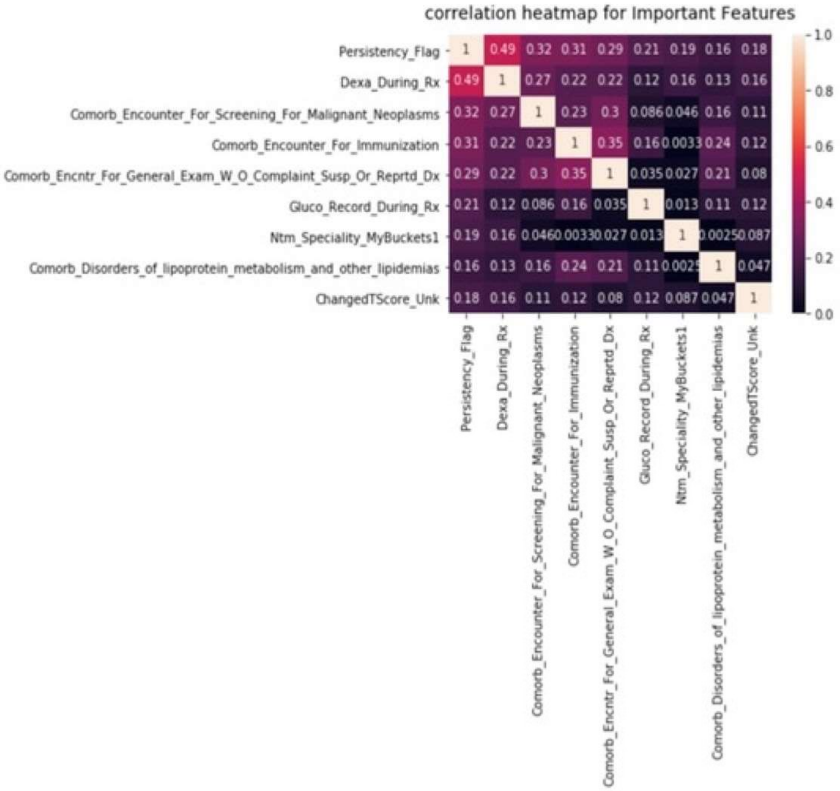
- The data cleansing required on this dataset was minimal, and there were no missing values.
- One issue was the majority of data columns were of type categorical; these were transformed to numerical (flag) format, with some dummy variables added as required.

Yes, Drug Persistency Can Be Predicted:

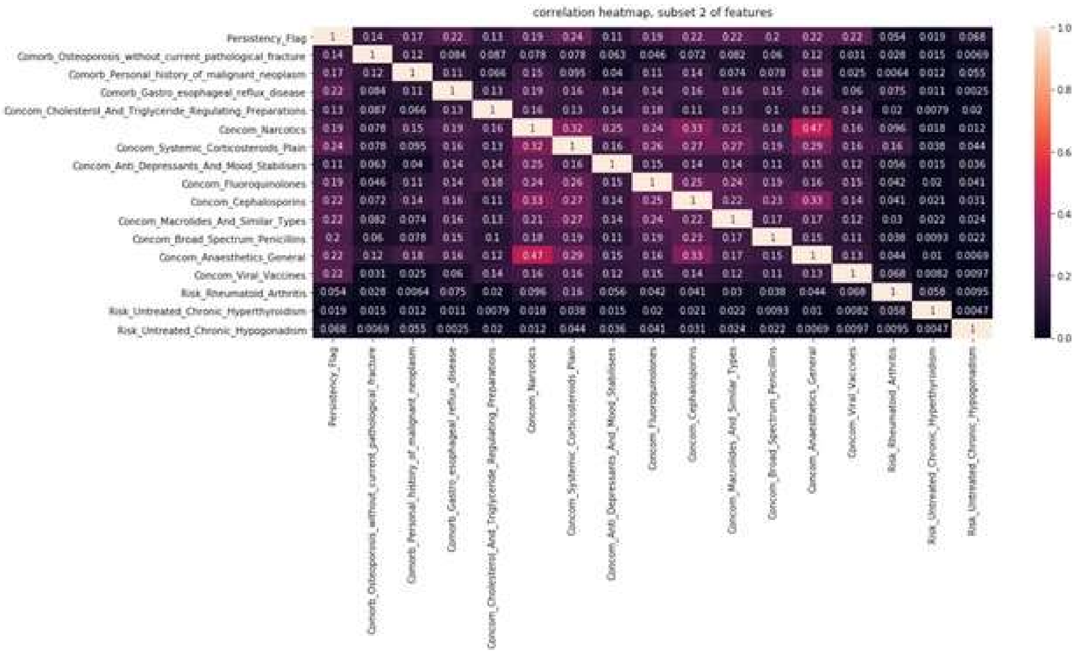
– Using machine learning methods for classification, patients' drug persistency could be predicted from the clinical and demographic data provided with greater than 80% accuracy.

– Which demographic or clinical factors are the most influential for drug persistency?

Heatmap of variable correlations for a subset of important variables in the dataset

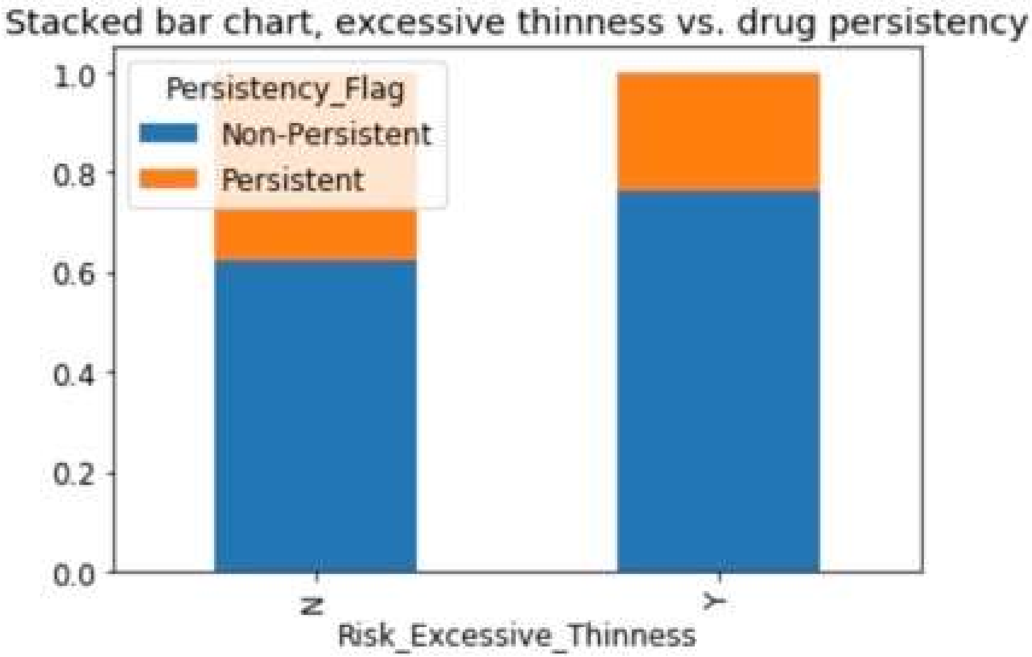


There were many more variables less strongly correlated with persistency

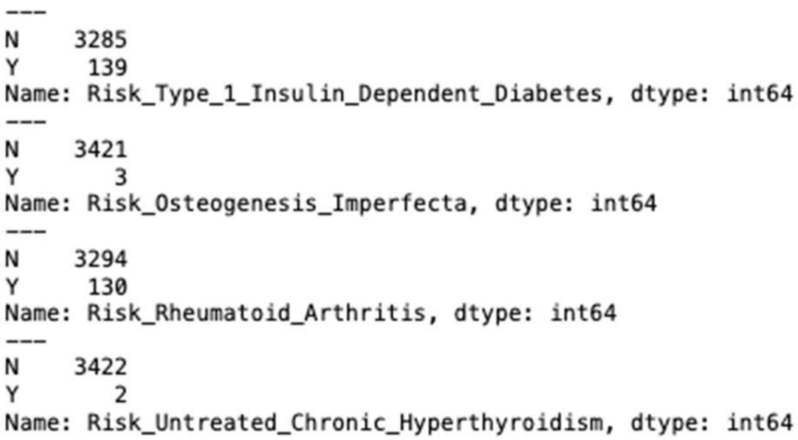


Bar Plots were created plotting individual variables against persistency

– Many of these plots indicated minimal correlation between the variable and drug persistency, such as with Risk of Excessive Thinness

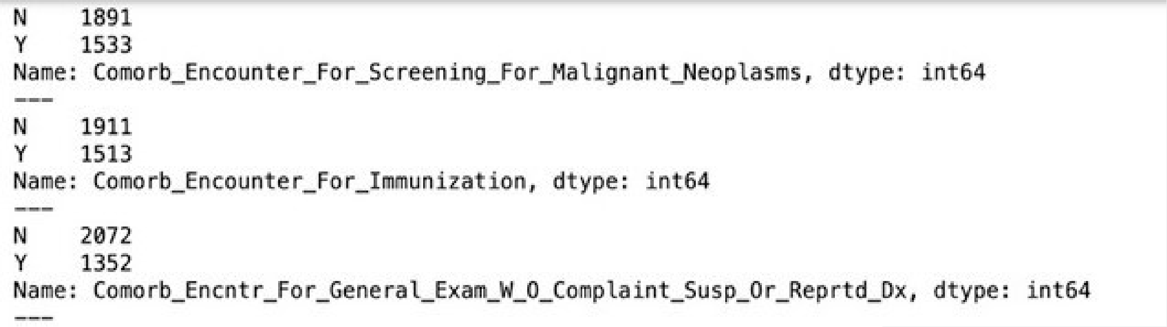


Some variables were excessively skewed, with almost all observations in one group and only a small number in the other, such those shown below:



```
---
N    3285
Y     139
Name: Risk_Type_1_Insulin_Dependent_Diabetes, dtype: int64
---
N    3421
Y       3
Name: Risk_Osteogenesis_Imperfecta, dtype: int64
---
N    3294
Y     130
Name: Risk_Rheumatoid_Arthritis, dtype: int64
---
N    3422
Y       2
Name: Risk_Untreated_Chronic_Hyperthyroidism, dtype: int64
```

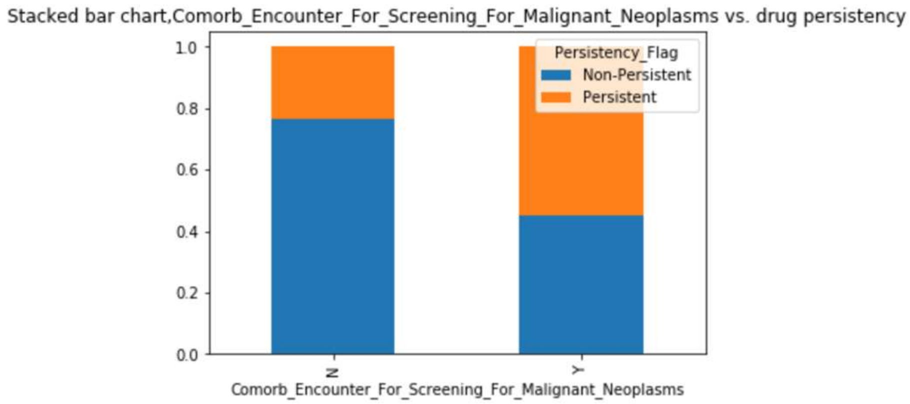
Some potential predictor variables were more promising as they were less skewed ...



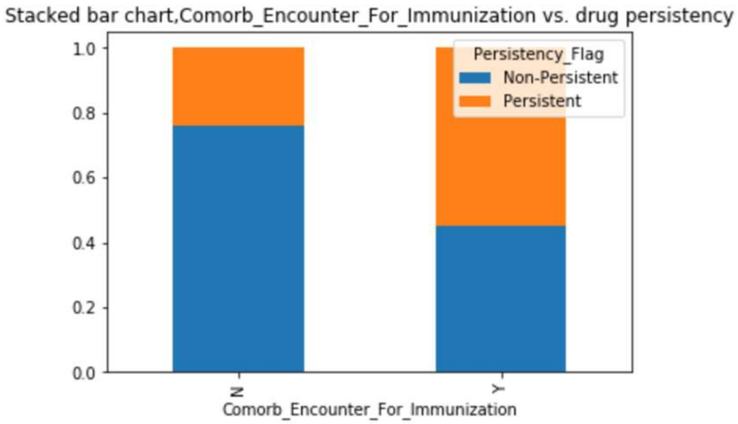
```
N    1891
Y    1533
Name: Comorb_Encounter_For_Screening_For_Malignant_Neoplasms, dtype: int64
---
N    1911
Y    1513
Name: Comorb_Encounter_For_Immunization, dtype: int64
---
N    2072
Y    1352
Name: Comorb_Encntr_For_General_Exam_W_O_Complaint_Susp_Or_Reprtd_Dx, dtype: int64
---
```

… and had Bar Charts indicating a correlation with Drug Persistency:

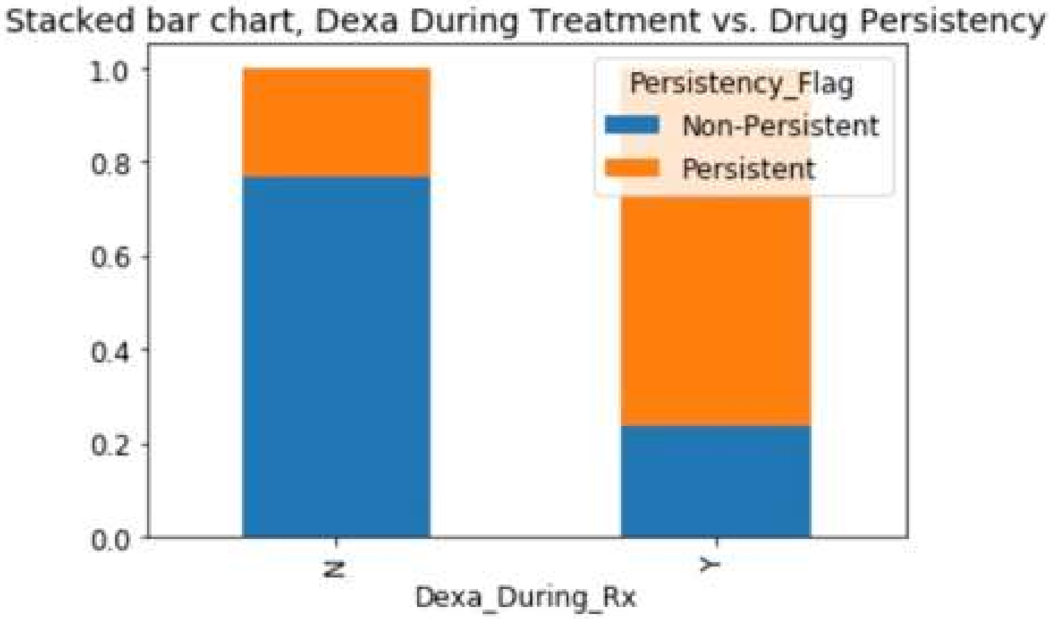
A medical visit to screen for malignant neoplasms appears positively correlated with being Persistent.



A visit for immunizations appears positively correlated with being Persistent.



The variable indicating whether the patient had a Dexa Scan during the bone density treatment appeared especially promising in terms of predicting Drug Persistency.



Section 2

Models

After the data was converted to numerical format (1 / 0 replacing Y/N, with additional dummy variables as needed), and the transformed dataset was split into training and test sets, our team fitted models from the SciKit-Learn Package: Logistic Regression, Random Forest, basic Decision Tree, and the ensemble methods AdaBoost and Gradient Boosting Classifier. We also applied Logistic Regression from the Statsmodels package and examined the output from fitting that model.

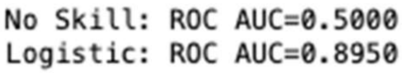
The classifiers built using these very different methods produced surprisingly comparable results when applied to the test set, with the notable exception of the Decision Tree classifier which performed worse than the others.

Model Evaluation and Final Selection

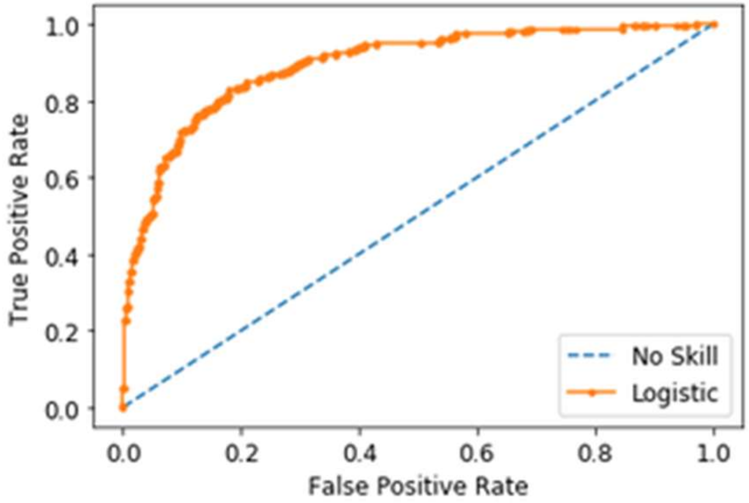
We evaluated our classifiers initially by comparing their accuracy (the proportion of the test observations correctly classified) as well as their Area Under the Curve for the Receiver Operating Characteristics curve or 'AUC ROC' value. While accuracy is perhaps the most intuitive metric for assessing classifier performance, the AUC ROC overcomes some of the weaknesses of the accuracy metric. The AUC ROC provides slightly different information and works especially well as a metric for evaluating classifiers when the target variable is skewed (as is the case with Drug Persistency in our data set, with many more patients non-persistent than persistent).

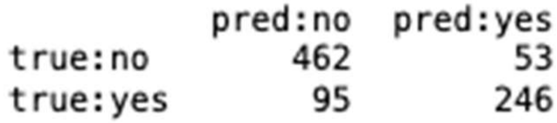
We ultimately selected **Logistic Regression** for our model, based on its excellent performance in terms of accuracy and the other measures (including AUC ROC) and because this model is well understood and provides especially clear information about the contribution of each of the predictor variables. On our chosen model we then calculated the Precision, Recall, and Specificity.

Logistic Regression



```
No Skill: ROC AUC=0.5000
Logistic: ROC AUC=0.8950
```





```
             pred:no  pred:yes
true:no          462        53
true:yes          95       246
```

Accuracy: 0.83
Precision: 0.82
Recall/Sensitivity: 0.72
Specificity: 0.90

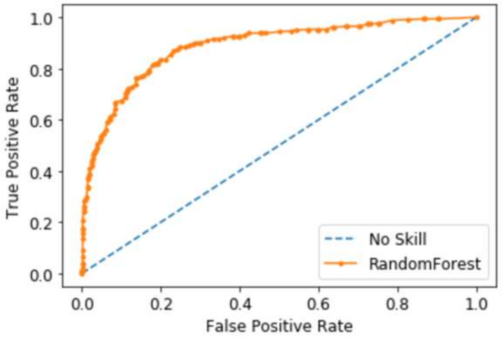
e

Runners Up

Random Forest

Accuracy: 0.81

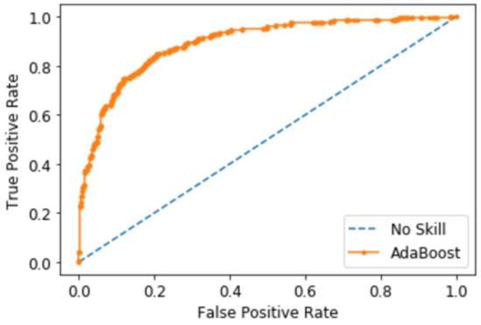
```
No Skill: ROC AUC=0.5000
RandomForest: ROC AUC=0.8880
```



AdaBoost

Accuracy: 0.82

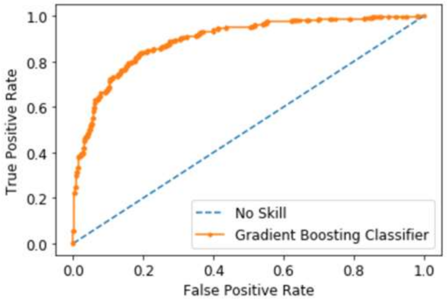
```
No Skill: ROC AUC=0.5000
AdaBoost: ROC AUC=0.8938
```



Gradient Boosting Classifier

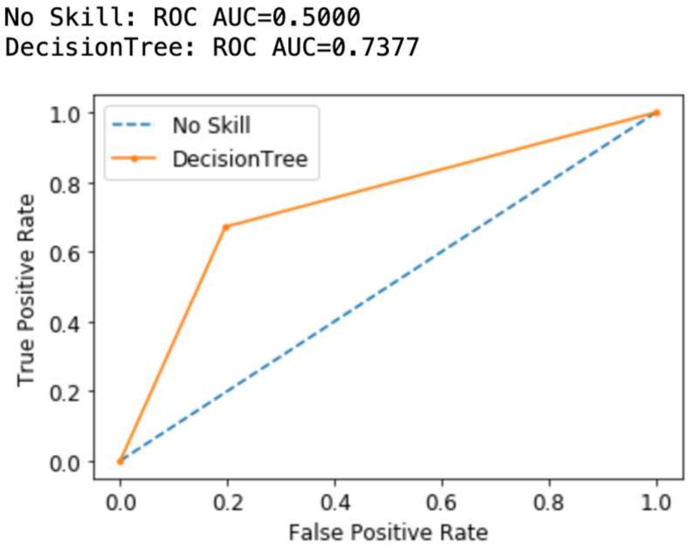
Accuracy: 0.82

```
No Skill: ROC AUC=0.5000
Gradient Boosting Classifier: ROC AUC=0.8940
```



Less Successful Classifier: Decision Tree

Accuracy: 0.75



```
No Skill: ROC AUC=0.5000
DecisionTree: ROC AUC=0.7377
```

Section 3

Conclusions and Recommendations

Drug Persistency can be predicted with very good accuracy (greater than 0.80) from the predictor variables in this data set.

As the ultimate goal is to understand drug persistency in terms of these predictors, we will now look at the variables most strongly correlated with Persistency.

The variable indicating whether the patient had a Dexa scan during treatment made the greatest contribution to predicting Drug Persistency.

Having had the scan during the treatment was associated with a 22% increase in the odds of being persistent, holding all other variables constant, based on our Logistic Regression model.

The next most influential factors

- If during the bone density treatment there were medical encounters such as
  - Screenings for malignancies (9%)
  - Visits relating to other long-term therapies (16%)
  - immunization (6%)
  - General medical check-ups (8%)

  the odds of drug persistency were increased.

  The percentages in parentheses next to the above visit types give the increase in odds for positive values on these flags, where in each individual case all other variables are held constant.

Visits with their medical providers appear to make a difference for patients, increasing the odds of drug persistency for those patients.

Other factors associated with persistency, but more modestly:

- The medical specialty of the prescribing provider
- Whether the patient lives in the Midwest
- Whether the patient has certain other medical conditions including gastro-esophageal reflux disease, disorders of lipoprotein metabolism or other lipidemias, or certain joint disorders, among others.

Fortunately, the factors most strongly correlated with drug persistency are potentially modifiable by the treating medical provider.

Recommendations to prescribing medical providers:
- order Dexa scans while the patient’s bone density treatment is ongoing
-encourage the patient to attend regularly for routine check-ups and any recommended immunizations

Thank you!

Data Glacier

Your Deep Learning Partner