Data Glacier Internship: Presentation & Report for the Final Project

Obiri Vincent

18th March 2022

Business Problem

- One of the challenges faced by Pharmaceutical companies is to understand the persistency of drug use as prescribed by the physician.
- To solve this problem ABC Pharma company approached us to automate this process of identification.
- We were required to gather insights into the factors that influence persistency and build a classifier to help automate the process of identifying each patient’s persistency.

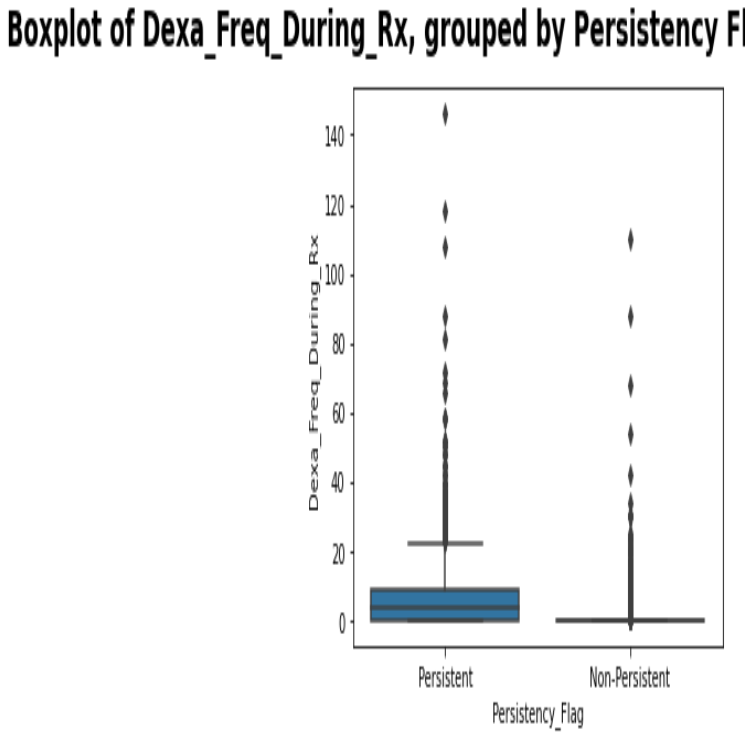
Weeks 8-9

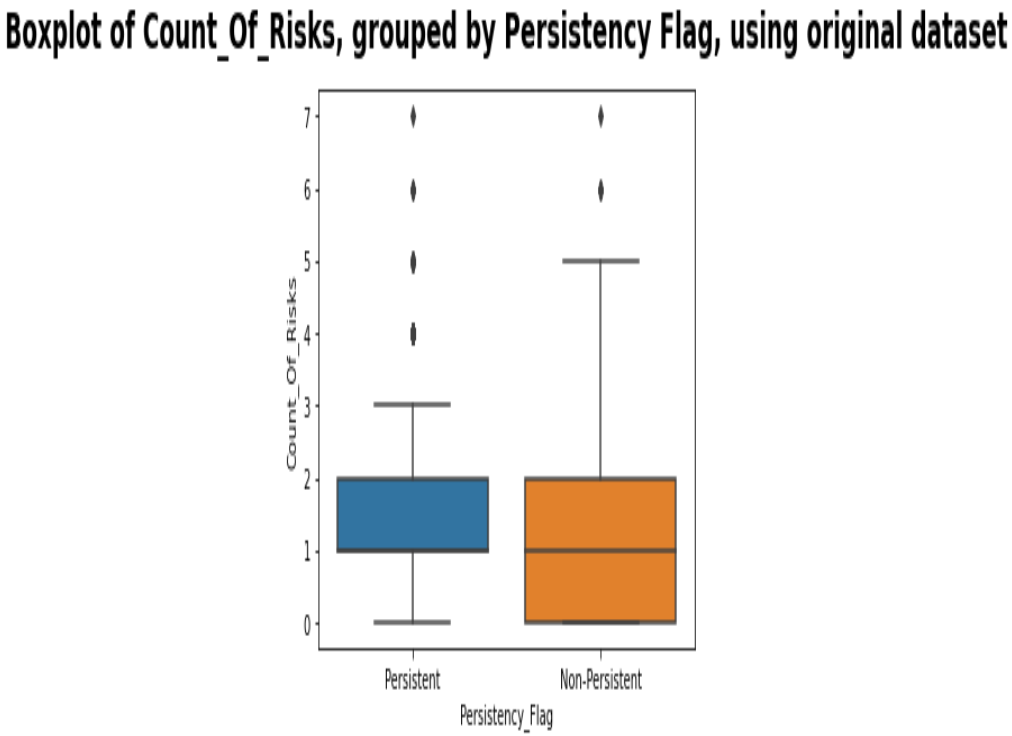
Weeks 8-9 involved:

- Understanding the dataset
- Identifying any outliers, duplicates and NA / nan values
- Transforming the data such that it can be used effectively for modelling
- Identifying the K best features (we used K=6) and some brief experimentation on a simple model

Weeks 8-9 (cont.)

Below we see a small sample of our work: boxplots for the numerical variables, grouped by persistency flag. The dots represent outliers.





Weeks 8-9 (cont.)

The 6 best features found in weeks 8-9 were:

- Dexa_During_Rx_Y" ,
- "Comorb_Encounter_For_Screening_For_Malignant_Neoplasms_Y",
- "Comorb_Encounter_For_Immunization_Y“
- "Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx_Y" ,
- "Comorb_Long_Term_Current_Drug_Therapy_Y" ,
- "Concom_Viral_Vaccines_Y",

Weeks 10-11 (EDA)

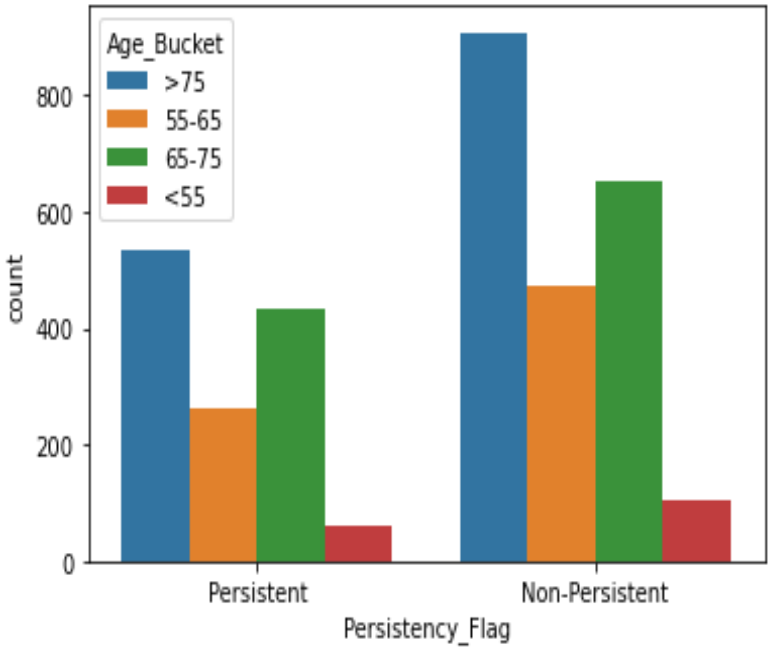
Weeks 10-11 involved:

- Understanding the dataset
- Investigating patient demographics
- Identifying relationships between the numerical variables and persistency flag
- Some comparisons of the data before and after transformation

Weeks 10-11 (EDA) (cont.)

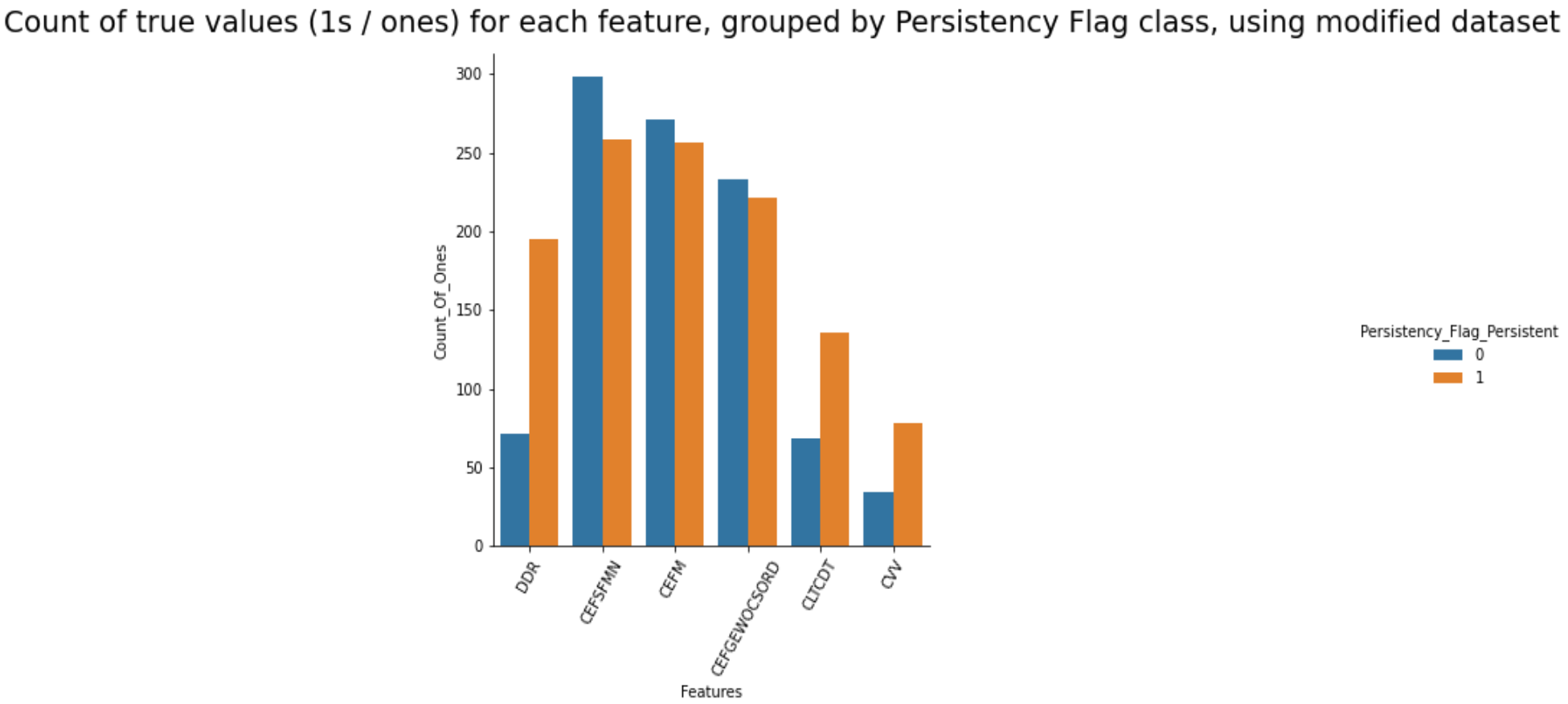
The two sets of plots for persistent / non-persistent patients have similar forms which implies that age isn't a useful indicator of
whether or not a patient is persistent, and also some age ranges are heavily underrepresented (namely <55)

Count of people by age group, grouped by persistency flag, using original dataset



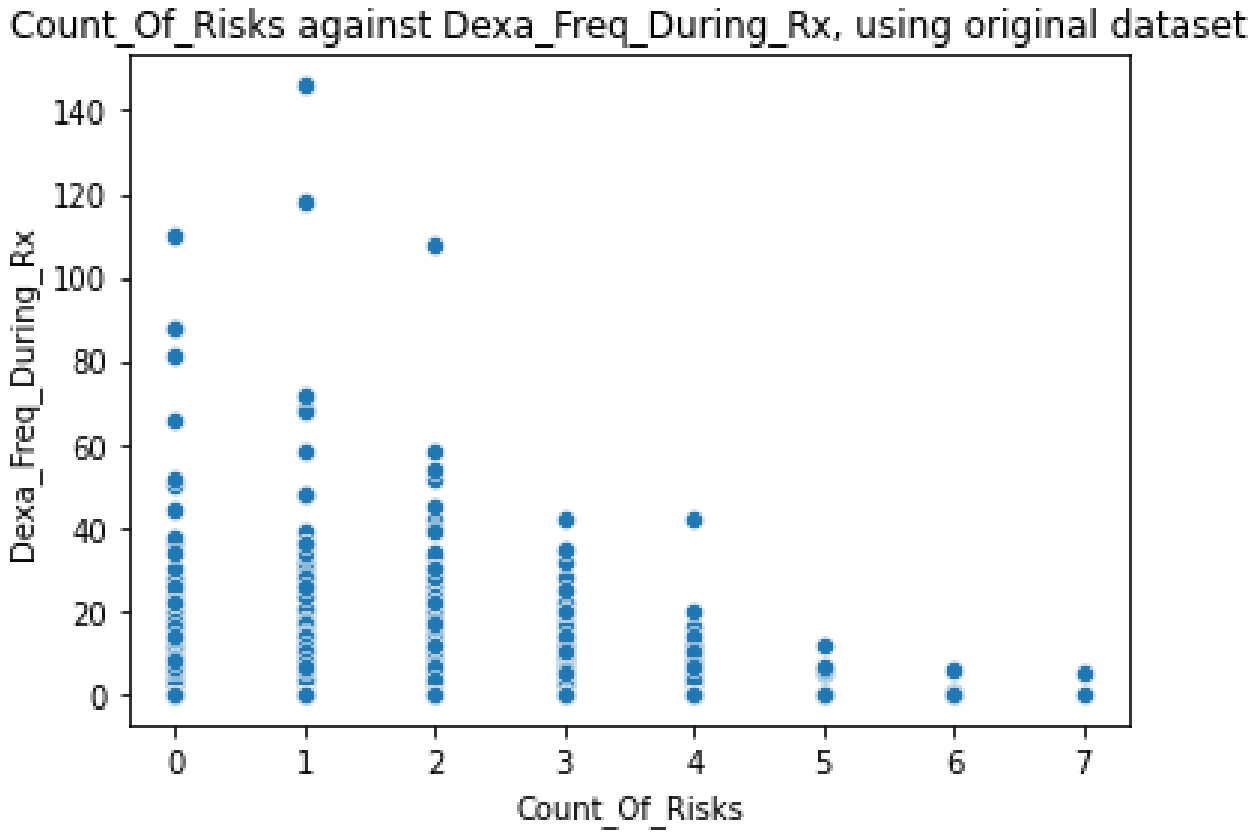
Weeks 10-11 (EDA) (cont.)

The following sample of our work corresponds to the 6 best features (some feature names are abbreviated). Dexa_During_Rx (DDR) was found to be a strong predictor from SelectKBest, and it does look good here also (many patients with DDR=1 were persistent):



Weeks 10-11 (EDA) (cont.)

We clearly see that as the Count_Of_Risks increases, Dexa_Freq_During_Rx tends to decrease (negative correlation):



Weeks 12-13 (modelling, evaluation and deployment)

Weeks 12-13 involved:

- Importing, splitting and manipulating data
- Defining and training appropriate classifiers
- Deciding on the best dataset / data frame to use
- Evaluating the performance of the classifiers and using grid search to find optimal parameters
- Comparing the models and choosing the best one for deployment
- Deploying the model

Weeks 12-13 (cont.)

Challenges / problems:

- The main challenges were with syntax errors and various errors when compiling. These were resolved with stack overflow and by reading relevant documentation. I also needed to refamiliarize myself with various libraries

- There was great difficulty with getting Heroku to work due to package incompatibilities

- Choosing which modified dataset / data frame to use was challenging, since there were a few good options. Eventually we decided to use the dataframe with the best 6 predictors and outliers removed, due to good evaluation scores and computational efficiency.

Weeks 12-13 (cont.)

Model selection:

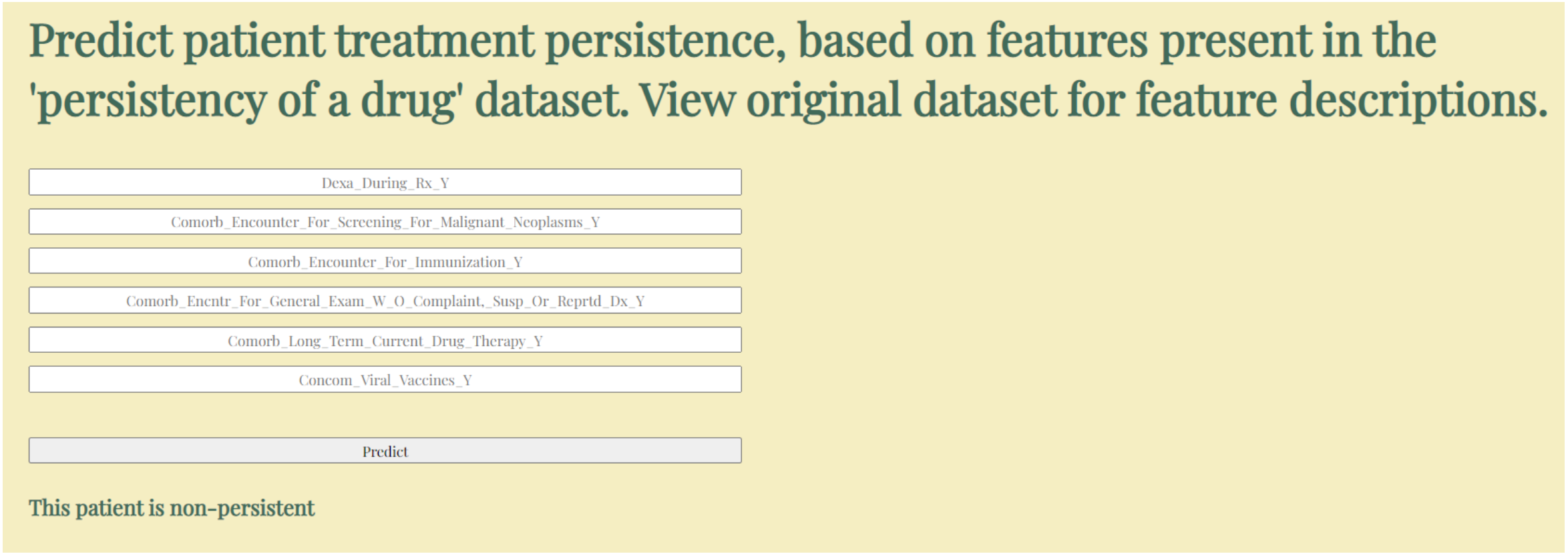
We used grid search to find optimal parameters for various classifiers. The neural network was chosen for deployment due to its outstanding evaluation scores relative to the other models. An XGB classifier also performed very well.

Evaluation scores for the neural network (with macro averaging due to class imbalance):

- Precision:0.848
- Recall:0.809
- F1:0.824
- Accuracy:0.858
- ROC-AUC:0.809

Weeks 12-13 (cont.)

The model was deployed successfully:



Weeks 12-13 (cont.)

Predict patient treatment persistence, based on features present in the 'persistency of a drug' dataset. View original dataset for feature descriptions.

Dexa_During_Rx_Y

Comorb_Encounter_For_Screening_For_Malignant_Neoplasms_Y

Comorb_Encounter_For_Immunization_Y

Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx_Y

Comorb_Long_Term_Current_Drug_Therapy_Y

Concom_Viral_Vaccines_Y

Predict

This patient is persistent

Thank you