Healthcare - Persistence of a drug

Presented by : Rebhi Wiem

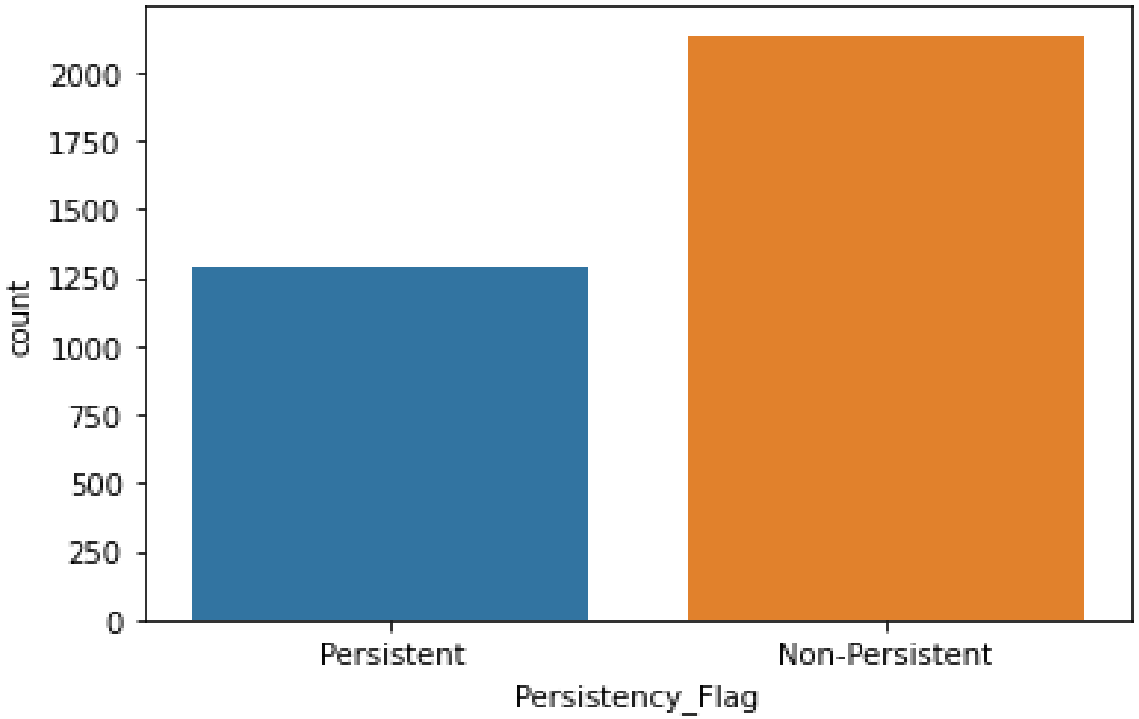
Problem understanding:

The treatment of chronic illnesses commonly includes the long-term use of pharmacotherapy. Although these medications are effective in combating disease, their full benefits are often not realized because approximately 50% of patients do not take their medications as prescribed.in order to gather insights on the factors that are impacting the persistence,we are going to build a classification that will help us to predict the treatment persistence and avoid a lot of problem that can happen to patients in the future.

Data Understanding :

- The dataset contains 3424 samples of patient's details
- Where the details are classified in six categories as it shown below :
    - **Unique Row Id :** Unique ID of each patient (**Patient ID**)
    - **Target Variable :** Flag indicating if a patient was persistent or not (**Persistency_Flag**)
    - **Demographics :** Age,Race,Region,Ethnicity,Gender,IDN Indicator
    - **Provider Attributes :** NTM - Physician Specialty
    - **Clinical Factors :** NTM - T-Score ,Change in T Score ,NTM - Risk Segment,Change in Risk Segment,NTM - Multiple Risk Factors,NTM - Dexa Scan Frequency,NTM - Dexa Scan Recency,Dexa During Therapy,NTM - Fragility Fracture Recency,Fragility Fracture During Therapy,NTM - Glucocorticoid Recency,Glucocorticoid Usage During Therapy
    - **Disease/Treatment Factor :** NTM - Injectable Experience,NTM - Risk Factors,NTM - Comorbidity ,NTM – Concomitancy,Adherence

Data Understanding :



- The diagramme plot represent the number of patient of each category and we can clearly see that we have inbalanced data that could later affect the performance of models

Data Cleaning and Feature engineering:

The process of data cleaning consist of :

- Remove duplicate or irrelevant observations
- Handle missing data
- Remove NAN values

The dataset contains a lots of categorical values that need to be converted to numerical values and to do that we use **label encoding** where :

- 0 represent the following value of target label : Non_Persistent
- 1 represent the following value of target label : Persistent

for the other columns the values vary according to the number of categories

Machine learning models for binary classification :

In machine learning, there are many methods used for binary classification. The most common are :

- ✓ Logistic Regression
- ✓ Support Vector Machines
- ✓ Naive Bayes
- ✓ K-nearest Neighbor
- ✓ Decision Tree

Model Selection:

Model	Score	Test Score
K-Nearest Neighbours	0.808324	0.678832
Support Vector Machine	0.707558	0.700730
Naive Bayes	0.757576	0.745985
Decision Tree	1.000000	0.748905
logistic regression	0.826214	0.795620

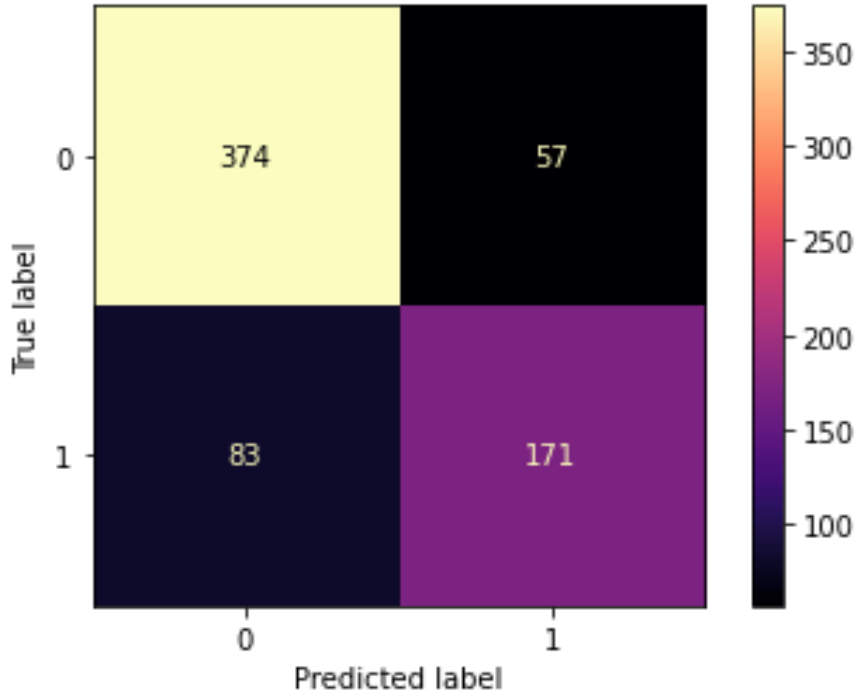
- The table above shows that the logistic regression gives the highest accuracy for the dataset
- We used Grid search to get the best parameters for each model

Model Evaluation:

	precision	recall	f1-score	support
0	0.82	0.87	0.84	431
1	0.75	0.67	0.71	254
accuracy			0.80	685
macro avg	0.78	0.77	0.78	685
weighted avg	0.79	0.80	0.79	685

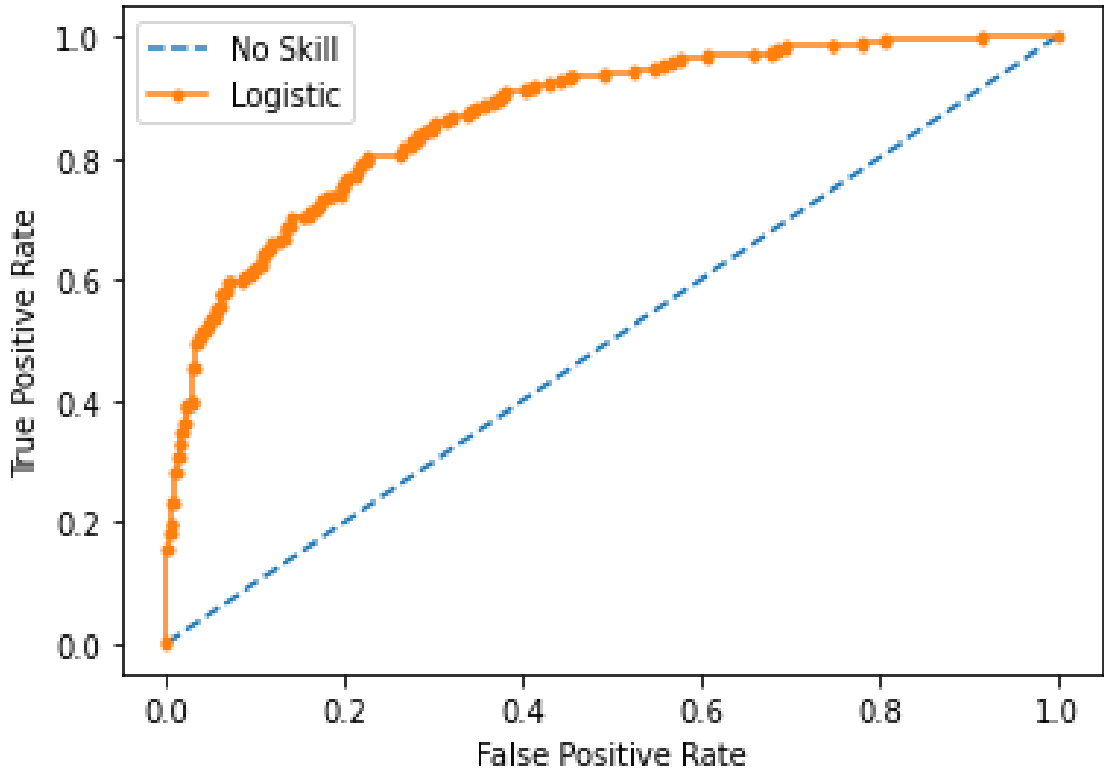
- The model predicted better the Non_Persistent than the presistent because we have imbalanced dataset but in general the results are good

Model Evaluation:



- The confusion matrix above shows that there is a lot of presistent patients predicted as non-presistent
- One of the solutions for this problem is adding more data

Model Evaluation:



- a good classifier, is a classifier with the ROC curve closer to the axes and the “elbow” close to the coordinate (0,1) which mean that the logistic regression is a good enough classifier, but we still can make it better

Conclusion:

- The logistic regression gave us a good result for a small inbalanced dataset
- Adding more data can help the model to performe better

Thank you