WEEK 8: DELIVERABLES

Team Details

Name	Email	Country	College/Company	Specialization
Fabian Umeh	Fabianumeh335@gmail.com	UK	Teesside University	Data Science
Rukevwe Ovuowo	rukevwe10@gmail.com	Nigeria	GBG Data science Academy	Data Science
Olutayo Oladeinbo	oladeinboolutayo@yahoo.com	UK	Teesside University	Data Science

PROBLEM STATEMENT

One of the challenges for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification. With an objective to gather insights on the factors that are impacting the persistency, it is necessary to build a classification for the given dataset, using the variable 'Persistency_Flag' as target variable and other attributes as prediction variables.

DATA UNDERSTANDING

Variables Description:

Here is a description of the columns in details

Unique Row Id:

- Patient ID: Unique ID of each patient;

Target Variable:

- Persistency_Flag: Flag indicating if a patient was persistent or not;
- Age: Age of the patient during their therapy;
- Race: Race of the patient from the patient table;
- Region: Region of the patient from the patient table;

Demographics:

- Ethnicity: Ethnicity of the patient from the patient table;
- Gender: Gender of the patient from the patient table;
- IDN Indicator: Flag indicating patients mapped to IDN;

Provider Attributes:

- NTM - Physician Specialty: Specialty of the HCP that prescribed the NTM Rx;
- NTM - T-Score: T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate);
- Change in T Score: Change in Tscore before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
- NTM - Risk Segment: Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate);
- Change in Risk Segment: Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
- NTM - Multiple Risk Factors: Flag indicating if patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate);

Clinical Factors:

- NTM - Dexa Scan Frequency: Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate);
- NTM - Dexa Scan Recency: Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable);
- Dexa During Therapy: Flag indicating if the patient had a Dexa Scan during their first continuous therapy;
- NTM - Fragility Fracture Recency: Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate);
- Fragility Fracture During Therapy: Flag indicating if the patient had fragility fracture during their first continuous therapy;
- NTM - Glucocorticoid Recency: Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx;
- Glucocorticoid During Therapy: Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy;
- NTM - Injectable Experience: Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx;
- NTM - Risk Factors: Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx;

Disease/Treatment Factors:

- NTM - Comorbidity: Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied;
- NTM - Concomitancy: Concomitant drugs recorded prior to starting with a therapy (within 365 days prior from first rxdate) Adherence: Adherence for the therapies.
- Adherence : Adherence for the therapies

WHAT TYPE OF DATA HAVE YOU GOT FOR ANALYSIS

The data is a csv file and also a classification problem

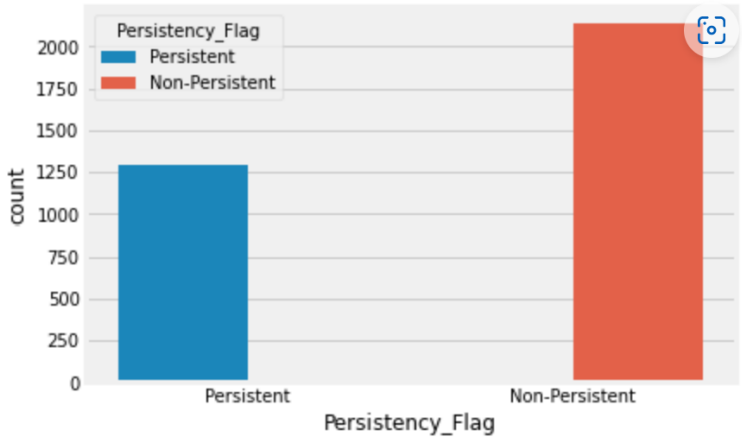
WHAT ARE THE PROBLEMS IN THE DATA (NUMBER OF NA VALUES, OUTLIERS, SKEWED ETC)

- Gather insights on the factors that are impacting the persistency of a drug during treatment.
- There are no NA values

WHAT APPROACHES YOU ARE TRYING TO APPLY ON YOUR DATA SET TO OVERCOME PROBLEMS LIKE NA VALUE, OUTLIER ETC AND WHY?

The solution required in generating business insights and create a machine classification model to solve the proposed problem include:

- Exploratory data analysis: With exploratory data, it is possible to gain business experience, to understand how the business works and to also understand the behavior that the business has through the data
    - ✓ Numerical variable
    - ✓ Target variable



The quantity of patient that did not persist in the use of treatment is more than the quantity of patient that persist.

- Data Description
  - ✓ Data Dimension

    The data contain 3424 rows and 69 features. Informing the quality of patient that are persistent of drugs and patient that are not persistent of drugs

  - ✓ Descriptive data

    This is use to gain business knowledge and to be able to detect error. There are two types of descriptive statistics.

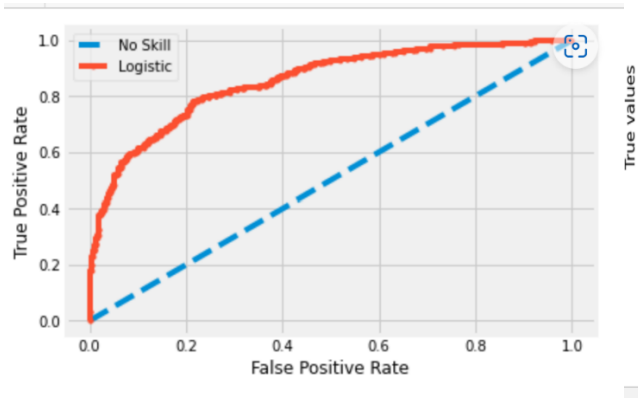
    1. Central Tendency: summaries data into a single number. E.g., Mean
    2. Dispersion metrics: which tells if the data is too close to the mean or not. E.g., std, min, max

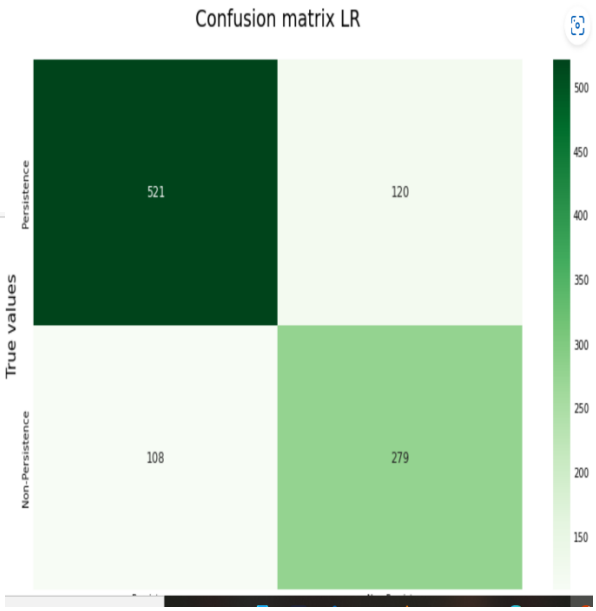
       Example

	Dexa_Freq_During_Rx	Count_Of_Risks
count	3424.000000	3424.000000
mean	3.016063	1.239486
std	8.136545	1.094914
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	1.000000
75%	3.000000	2.000000
max	146.000000	7.000000

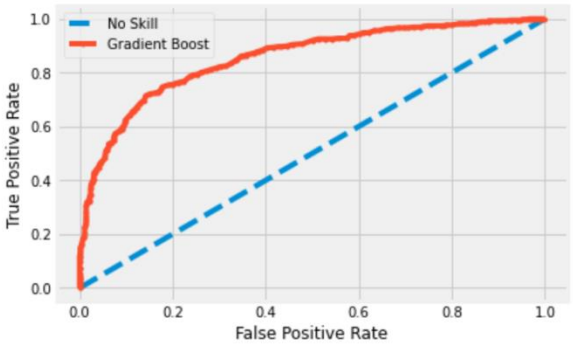
From the chart we can see that Dexa_Freq_During_Rx has a high value, which means that it can cause aa problem at the model training and has to be fixed

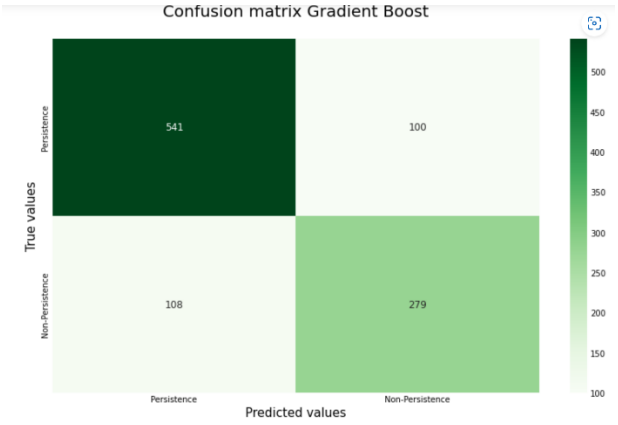
- Machine learning model performance
  - ✓ LogisticRegression



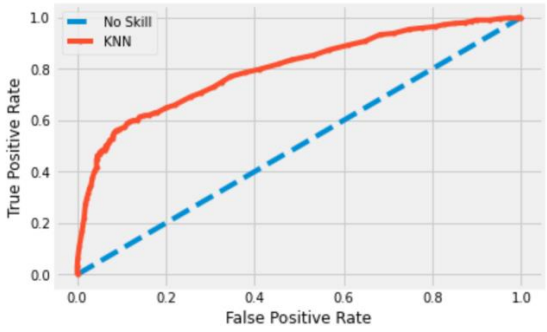


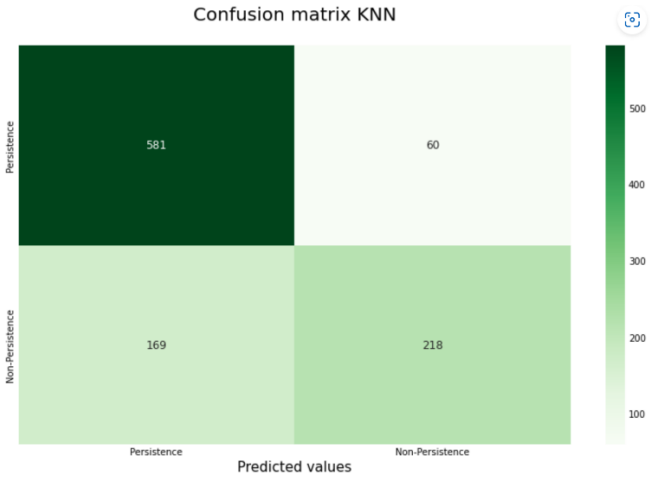
  - ✓ Gradient Boost



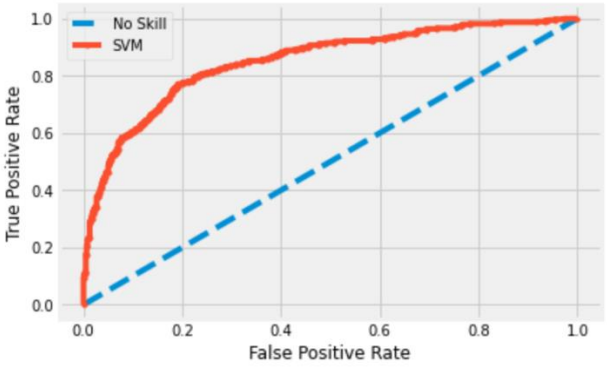


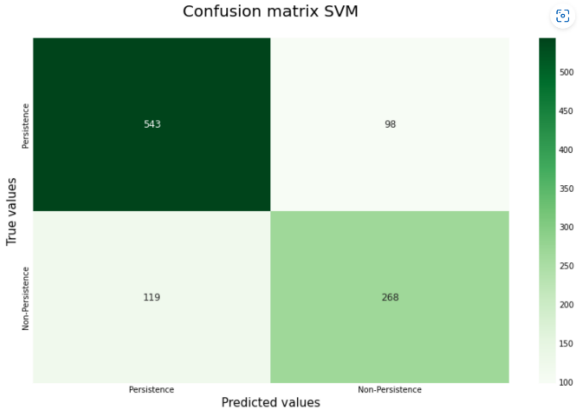
  - ✓ KNeighborsClassifier



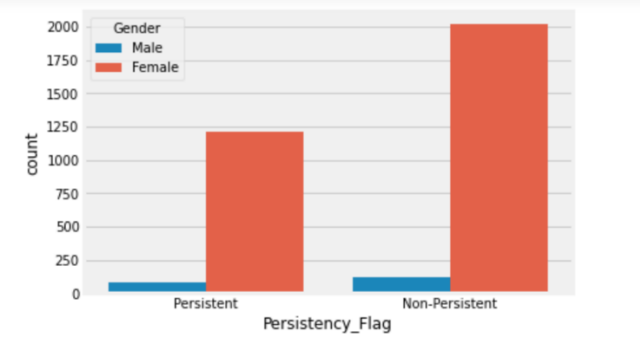


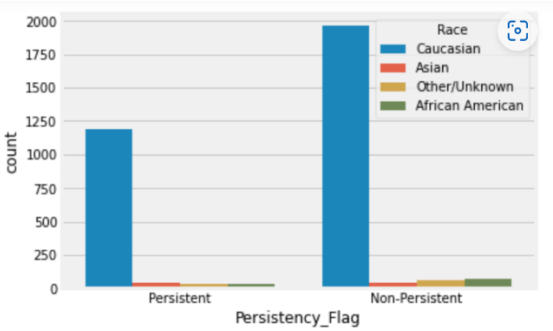
- ✓ SVC

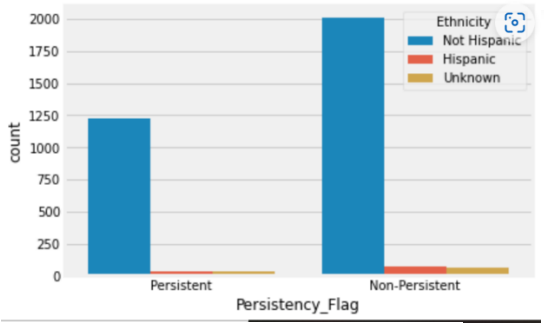


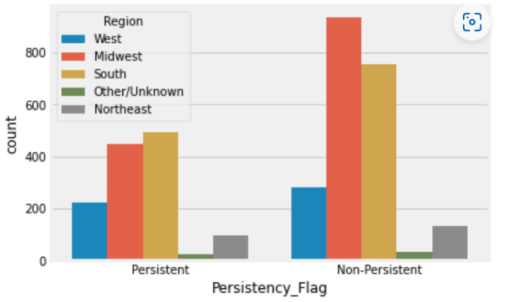


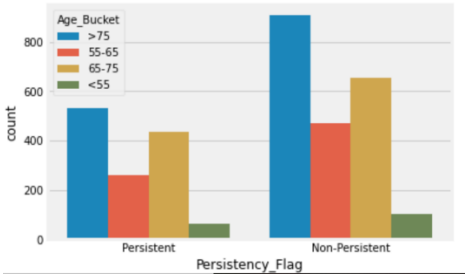
- Business performance
  - ✓ Classification_report
  - ✓ Confusion_matrix
  - ✓ roc_curve
  - ✓ roc_auc_score
- Hypothesis creation











REASONS FOR USING THIS APPROACH TO SOLVE THE ABOVE PROBLEM ARE

- Machine learning was used to classify future patients, informing if they will use the drugs during the entire treatment or if they won't.
- Creating of dashboard with several hypothesis and insights to help the company CEO with future decisions.
- Visualizing it if they will use the drugs during treatment or not.

Github Repo link

Data storage location: Github
Total number of observations: 3424
Total number of files: 1
Total number of features:69

Base format of the file: .csv
Size of the data: 898 KB