Data Glacier

Your Deep Learning Partner

Data Science Intern at Data Glacier

Week 13 Deliverables

Healthcare - Persistency of a drug

LISUM12

Info

Data Glacier

Your Deep Learning Partner

Name: Mariam Ali Bashandy

Individual: Solo
Email: mariam.a.bashandy.1@gmail.com

Country: Egypt
College: German International University of Applied Science

Specialization: Data Science
Date : 19/9/2022
Batch code: LISUM12

Table of Contents

Data Glacier

Your Deep Learning Partner

Project Lifecycle along with deadlines

Data Glacier
Your Deep Learning Partner

Weeks	Date	Plan
Week 07	19/9/2022	Problem statement, Data Collection and Data Report
Week 08	26/9/2022	Data preprocessing
Week 09	2/10/2022	Feature engineering and data cleaning
Week 10	9/10/2022	EDA
Week 11	16/10/2022	Build Model and Model Result Evaluation (EDA Presentation)
Week 12	23/10/2022	Flask deployment
Week 13	20/10/2022	Final submission

Problem statement & Business Understanding

Data Glacier

Your Deep Learning Partner

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription and ABC pharma company has the same challenge and to solve this problem wants to automate the process of identification.

Objective to gather insights on the factors that are impacting the persistency and build a classification for the given dataset to automate the process.

For Business Understanding needs:

- Know drug persistency and how it calculated
- Define the business problem
- Define objective and criteria to solve it

Data Collection

The Data is about Healthcare which contain 69 number of features and 3424 number of observations and it was used to detect Persistent vs Non-Persistent

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	898KB

Data understanding

In this part, we will explain data understanding

Types of data

The dataset contains 69 features, we found 67 out of 69 data type are categorical and only two features were numerical (names of numeric: Dexa_Freq_During_RX and Count_Of_Risks)

Missing Values

No missing values found in this dataset

Outliers

We found outliers in both two numerical attributes and may have solve them by IQR or Simple Imputer in the next step.

Skew

We found that both numeric have skew greater than zero so will own more weight in the left

Data Cleaning and feature engineering

Data Glacier
Your Deep Learning Partner

In this part, we will explain data cleaning and feature engineering

Handling Outliers

We removed outliers by two ways which are Z-score and IQR and they did will and remove outliers and the one that choose from both wad IQR since when done handling outliers by it doesn’t remove too much data, we didn’t do this to categorical after doing encoding since this would results in large loss of data.

Skew

We solved the problem by using Power Transformer and it distribute the data.

Encoding

We encode the categorical attributes to not led to problem in ML part.

EDA

In this part, we will explain EDA

The columns Risk_Type_1_Insulin_Dependent_Diabetes, Risk_Osteogenesis_Imperfecta, Risk_Rheumatoid_Arthritis, Risk_Untreated_Chronic_Hyperthyroidism, Risk_Untreated_Chronic_Hypogonadism, Risk_Untreated_Early_Menopause, Risk_Patient_Parent_Fractured_Their_Hip ,Risk_Smoking_Tobacco, Risk_Chronic_Malnutrition_Or_Malabsorption, Risk_Chronic_Liver_Disease, Risk_Family_History_Of_Osteoporosis ,Risk_Low_Calcium_Intake, Risk_Vitamin_D_Insufficiency, Risk_Poor_Health_Frailty, Risk_Excessive_Thinness, Risk_Hysterectomy_Oophorectomy, Risk_Estrogen_Deficiency, Risk_Immobilization Risk_Recurring_Falls

So will be reduced due to Count_of_Risks includes summation of them, reached to 50 by that and also in cleaning phase we drop patient ID

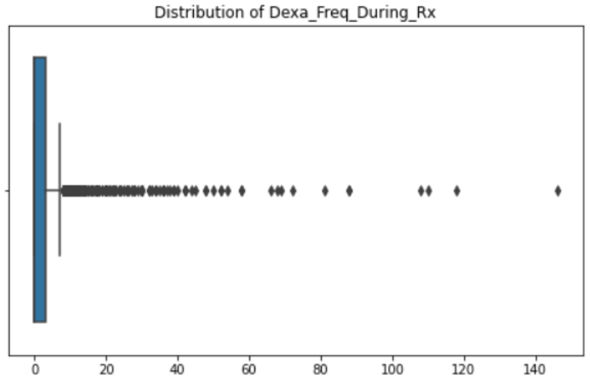
The Columns

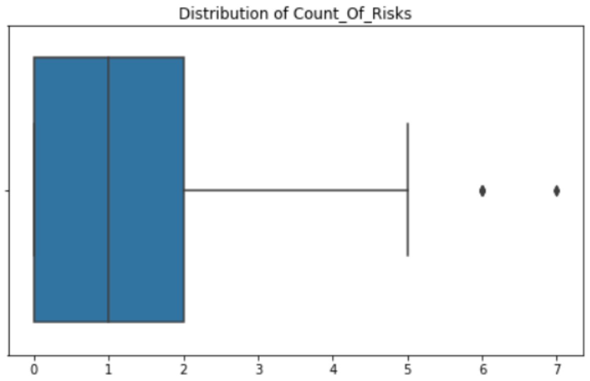
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms,Comorb_Encounter_For_Immunization,Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx,Comorb_Vitamin_D_Deficiency,Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified',Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx,Comorb_Long_Term_Current_Drug_Therapy,Comorb_Dorsalgia,Comorb_Personal_History_Of_Other_Diseases_And_Conditions,Comorb_Other_Disorders_Of_Bone_Density_And_Structure,Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias,Comorb_Osteoporosis_without_current_pathological_fracture,Comorb_Personal_history_of_malignant_neoplasm,Comorb_Gastro_esophageal_reflux_disease,Concom_Cholesterol_And_Triglyceride_Regulating_Preparations,Concom_Narcotics,Concom_Systemic_Corticosteroids_Plain,Concom_Anti_Depressants_And_Mood_Stabilisers,Concom_Fluoroquinolones, Concom_Cephalosporins,Concom_Macrolides_And_Similar_Types,Concom_Broad_Spectrum_Penicillins, Concom_Anaesthetics_General,Concom_Viral_Vaccines

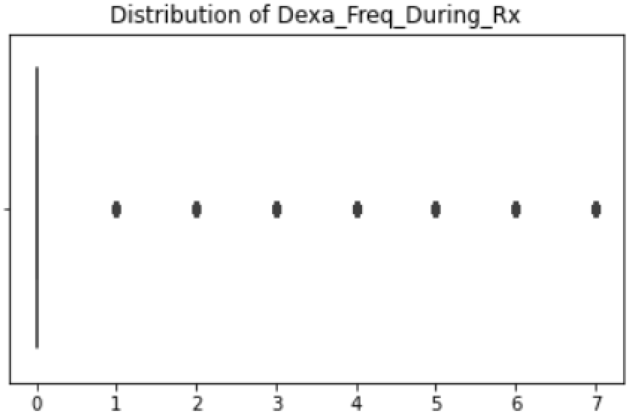
Will be two columns have count of them one for columns that start with comorb and one for concom and reached 28 features only

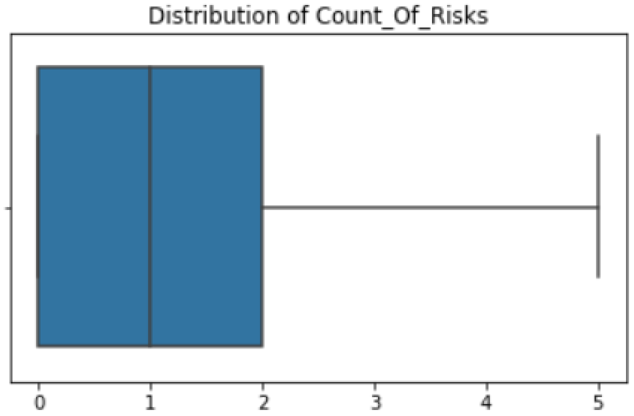
EDA

Handling outliers the first graph was before handling and the second one after handling outliers in numeric attributes



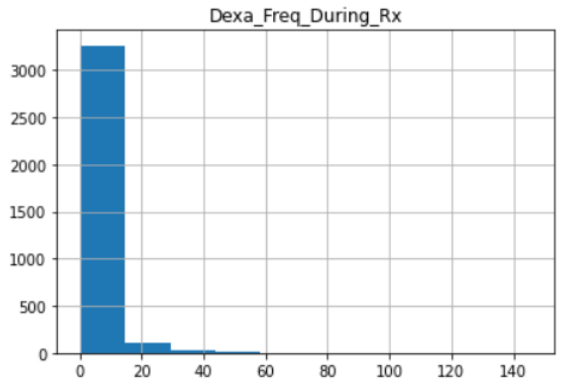


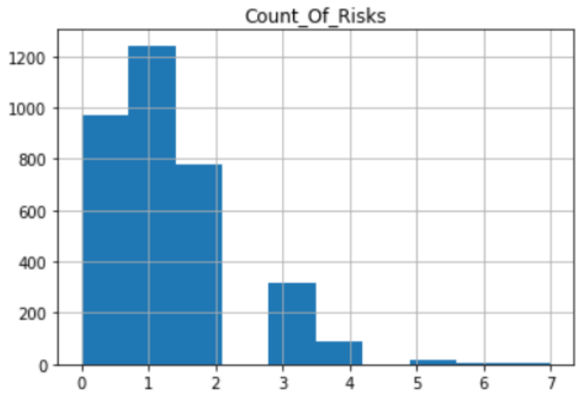




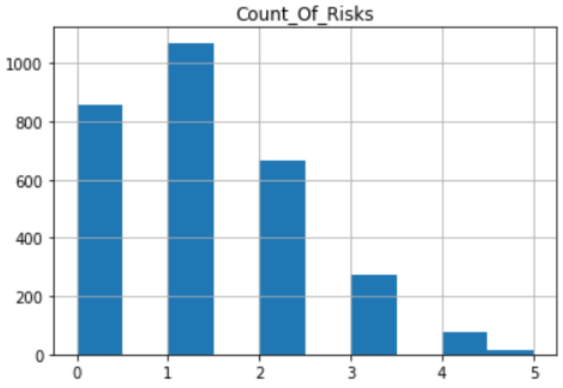
EDA

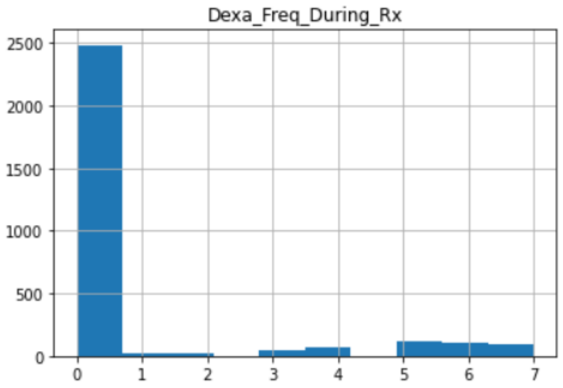
Skew graph before doing transformation





Skew graph after doing transformation

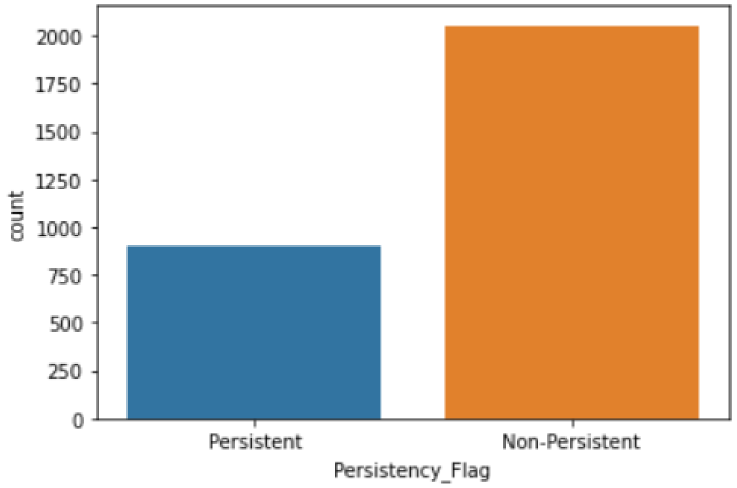




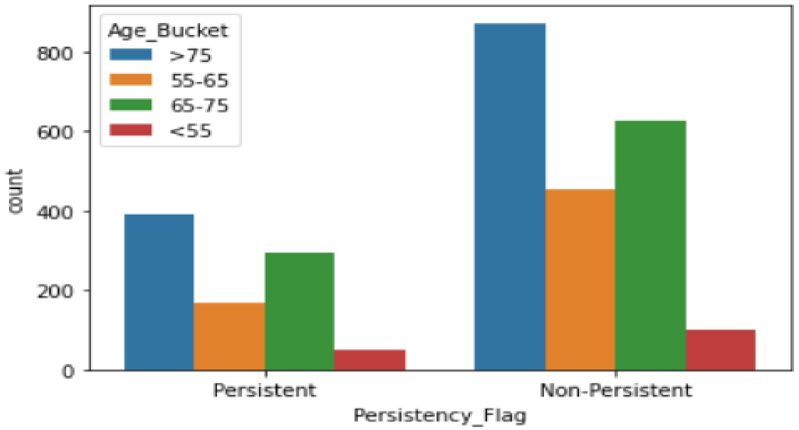
EDA

Data Glacier
Your Deep Learning Partner

Non-Persistent for drug is higher in our data

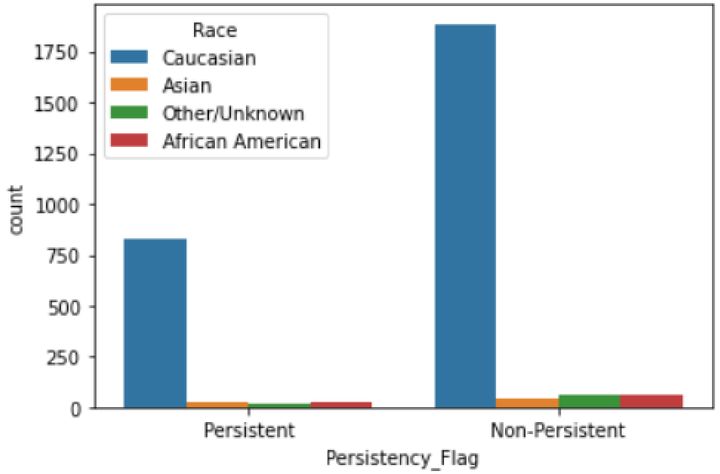


The age_Bucker >75 in Persistent and non-Persistent have higher value so they people greater than 75 have higher chance to persistent by drug

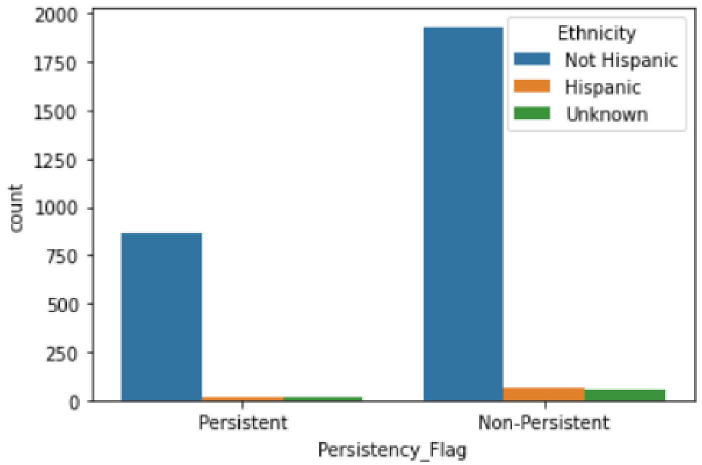


EDA

The Caucasian Race in both persistent and non-Persistent have highest count



Not Hispanic is dominant in Persistency Flag of drugs

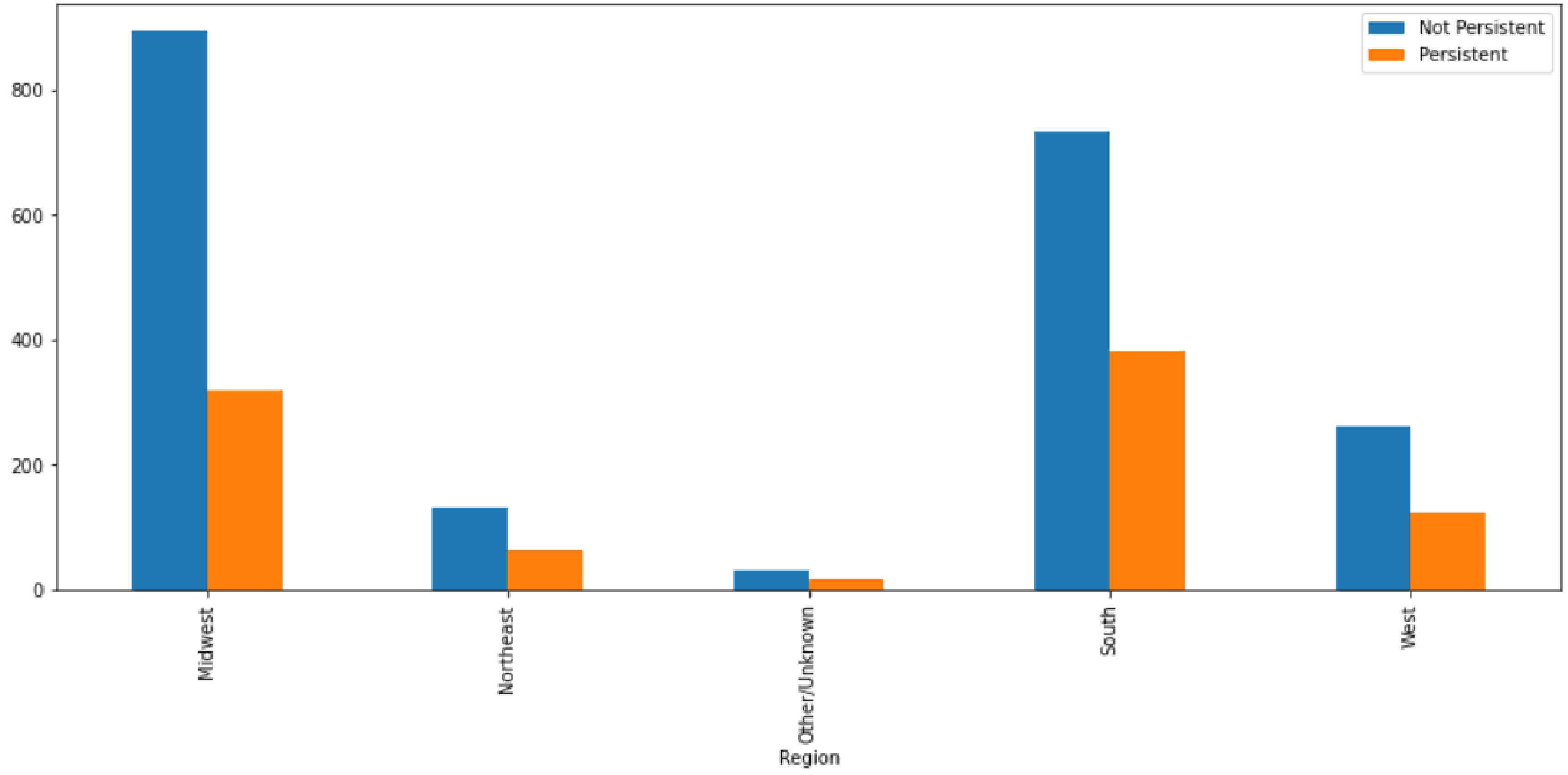


EDA

Data Glacier
Your Deep Learning Partner

In all the regions the dominant was Not-Persistent

The Midwest was higher region to not persistent from drug than South

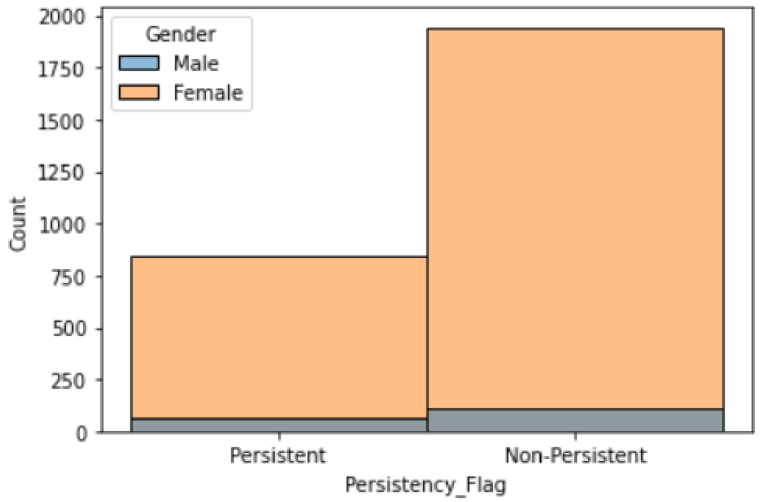


EDA

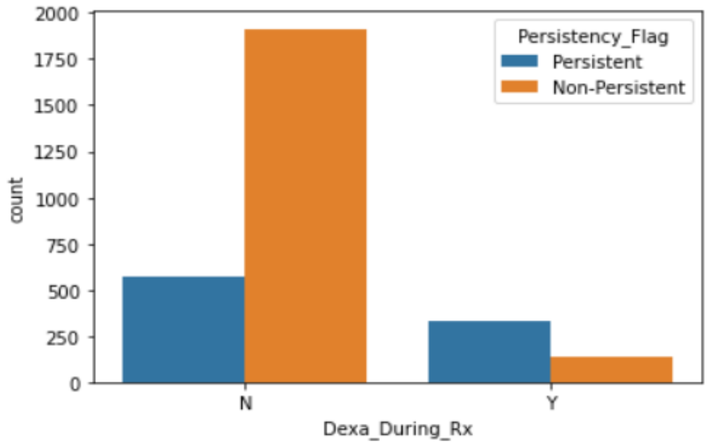
Data Glacier

Your Deep Learning Partner

Female patients are more persistent of a drug than male



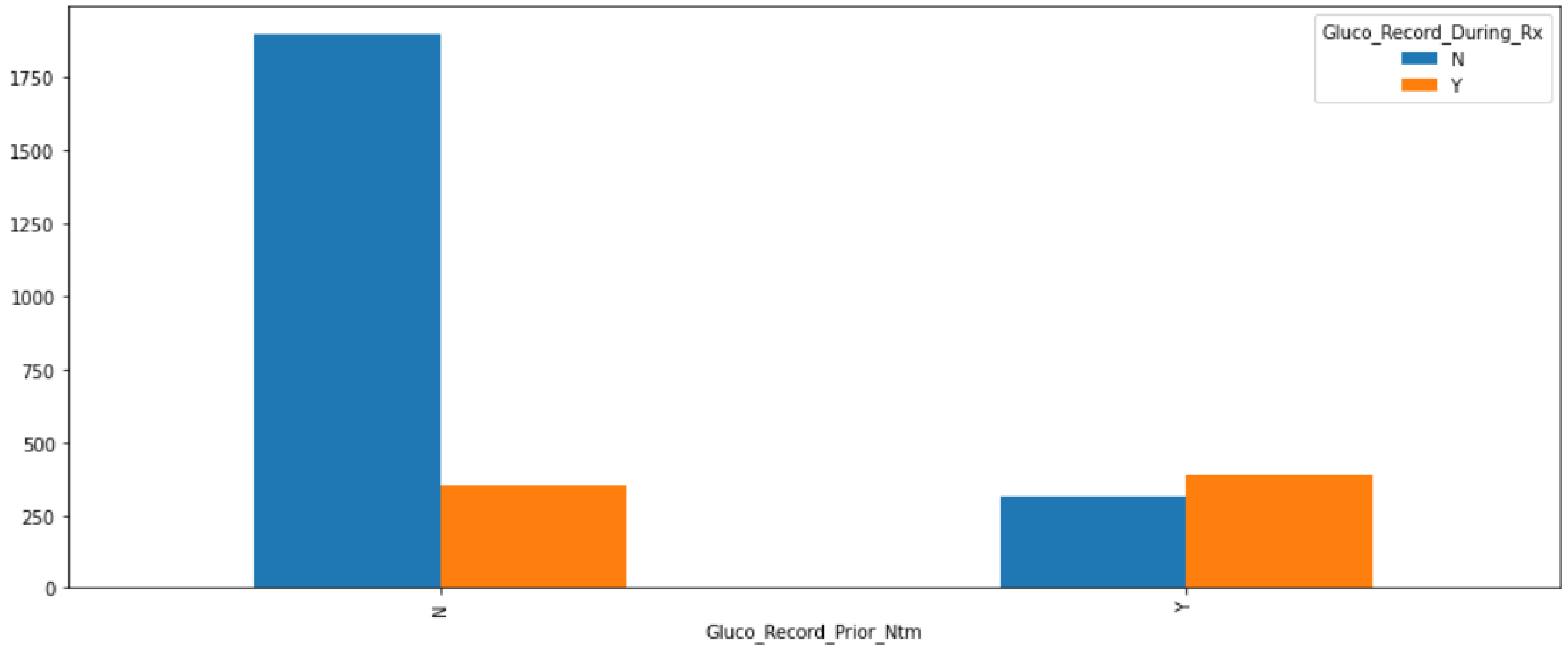
The number of patients without having Dexa scan is higher to be non persistent by drugs



EDA

Data Glacier
Your Deep Learning Partner

The ratio of the patients which are stable is much more higher than of the ratio of the improved patients



Preparation for Model building and Model Selection

Data Glacier
Your Deep Learning Partner

We split data into train and test

Start preparing for model selection

Choose models: Random forest, Logistic Regression, Support Vector machine , KNN , Neural network , decision tree and Gradient Boost Model

The best model Gradient Boost Model was since have the highest accuracy by 79.05 which was near to the accuracy of logistic Regression

Accuracy

Data Glacier
Your Deep Learning Partner

The report of every model and the selection one

Logistic Regression

	precision	recall	f1-score	support
0	0.76	0.43	0.55	183
1	0.79	0.94	0.86	409
accuracy			0.78	592
macro avg	0.77	0.68	0.70	592
weighted avg	0.78	0.78	0.76	592

Support Vector Machine

	precision	recall	f1-score	support
0	0.00	0.00	0.00	183
1	0.69	1.00	0.82	409
accuracy			0.69	592
macro avg	0.35	0.50	0.41	592
weighted avg	0.48	0.69	0.56	592

Decision tree

	precision	recall	f1-score	support
0	0.52	0.53	0.52	183
1	0.79	0.78	0.78	409
accuracy			0.70	592
macro avg	0.65	0.65	0.65	592
weighted avg	0.70	0.70	0.70	592

KNN

	precision	recall	f1-score	support
0	0.66	0.27	0.38	183
1	0.74	0.94	0.83	409
accuracy			0.73	592
macro avg	0.70	0.60	0.60	592
weighted avg	0.72	0.73	0.69	592

Accuracy

Data Glacier

Your Deep Learning Partner

Rest

Neural Network

	precision	recall	f1-score	support
0	0.63	0.51	0.56	183
1	0.80	0.87	0.83	409
accuracy			0.76	592
macro avg	0.71	0.69	0.70	592
weighted avg	0.75	0.76	0.75	592

Gradient Boosting Model

	precision	recall	f1-score	support
0	0.79	0.44	0.56	183
1	0.79	0.95	0.86	409
accuracy			0.79	592
macro avg	0.79	0.69	0.71	592
weighted avg	0.79	0.79	0.77	592

Random Forest Classifier

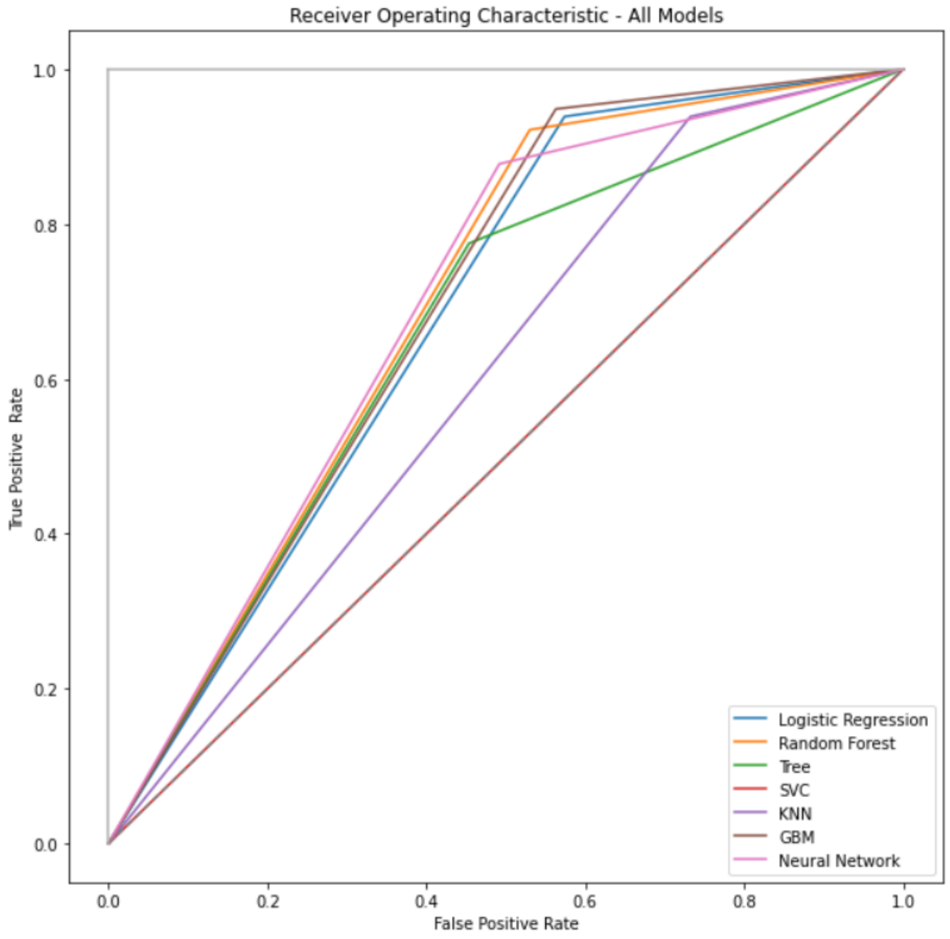
	precision	recall	f1-score	support
0	0.73	0.46	0.57	183
1	0.79	0.92	0.85	409
accuracy			0.78	592
macro avg	0.76	0.69	0.71	592
weighted avg	0.78	0.78	0.77	592

With the result of accuracies the best Model was Gaussian Boosting Model

ROC Plot

Data Glacier
Your Deep Learning Partner

The best model Gradient Boost Model was since have the highest accuracy by 79.05 which was near to the accuracy of logistic Regression



Flask Deployment

Data Glacier
Your Deep Learning Partner

We created the model.pkl file and html and style.css for styling the form of app

Then we create the app at port 5000

Make pickle file to the model

```python
import pickle
pickle.dump(gbm_select, open('E:/solo projects/Data_Glacier_virtual_internship/Data_Glacier_virtual_internship/Week13/model/model_gbm.pkl', 'wb'))
```

[78] ✓ 0.6s Pyth

Loading model

```python
model = pickle.load(open('E:/solo projects/Data_Glacier_virtual_internship/Data_Glacier_virtual_internship/Week13/model/model_gbm.pkl', 'rb'))
```

[79] ✓ 0.7s Pyth

```python
from unittest import result
from flask import Flask , request,render_template
import numpy as np
import pickle


app=Flask(__name__)
model = pickle.load(open('model/model_gbm.pkl','rb'))

@app.route('/')
def home():
    return render_template('template/index.html')


@app.route('/predict',methods=['POST'])
def predict():



    features = [float(x) for x in request.form.values()]
    final_features = [np.array(features)]
    prediction = model.predict(final_features)

    return render_template('template/index.html', prediction_text='The Presistent_Flag is {}'.format(prediction))

if __name__ == '__main__':
    # app.run(port =5000,debug=True)
    app.run(port =5000,debug=True, use_reloader=False)
```

[5] 4m 50.9s Python

The attachment screenshots for code of flask deployment

GitHub link

Data Glacier

Your Deep Learning Partner

https://github.com/Mariamali2001/Data_Glacier_virtual_internship

Thank You

Data Glacier

Your Deep Learning Partner