

Healthcare: Persistency of a Drug

Final Project Presentation



Data Glacier

Your Deep Learning Partner

Intern Details

- **Group:** Single Member Group
- **Name:** Hassan Faheem
- **Batch:** LISUM06
- **Email:** Hassan_hsn9@hotmail.com
- **Country:** UAE
- **University:** Heriot-Watt University
- **Specialization:** Data Science

Agenda

- Executive Summary
- Overview
 - Problem Description
 - Business Understanding
 - Data Understanding
 - Data Intake Report
- Exploratory Data Analysis
 - EDA Overview
 - Problems in the Dataset
 - Analyzing Outliers
 - Skewness & Kurtosis
 - Correlation Between Target & Features
 - Data Transformation
- Model Training & Testing
 - Introduction
 - Classifiers Used
 - Evaluation of Performance of Classifiers
 - Classifiers Train & Test Results
 - Final Recommendation & Conclusion

Executive Summary

- The Pharmaceutical Company wants to understand the persistency of a drug. The company has come to the conclusion that the best way to understand the persistency is to automate the whole process. In order to achieve this goal, the company has approached a Company that excels in Analytics. This Analytics Company will automate the process of identifying the persistency in an efficient way.
- The goal is to understand the data given, perform the necessary EDA and transformations and then build a classification model that is trained and tested on the given data.

Overview

Problem Description

The problem given here is that the Pharmaceutical Company, ABC is in need to understand the persistency of drug as per the physician prescription. The company ABC has thus approached a company that specializes in Analytics, to get this process of identification to be automated. The company has assigned the case to the relevant member to figure out the solution for the automation of persistency of drug for the company ABC.

Business Understanding

The objective of Pharmaceutical Company, ABC is to understand the persistency of a drug for patients. The data obtained shows a large amount of NTM or Non-Tuberculous Mycobacterial infection. The Company hence wants to verify the persistency of the drug that is being prescribed and so the Company would in turn manufacture more those drugs in demand for a more successful business.

Data Understanding

The Healthcare Dataset provided has 69 columns and 3424 number of observations. The target variable is Persistency_Flag. This variable is of Boolean data type with values that are either True or False. After understanding and analyzing the data, it's been found that there are few columns that are of numerical data type. Most of the columns are of either Boolean data type or String data type. The column of "Ptid" which refers to Patient ID has no value in terms of model training and thus will be removed the dataset.

Data Intake Report

Name: Healthcare – Persistency of a Drug

Report date: 18th March 2022

Internship Batch: LISUM06

Version: 1.0

Data intake by: Hassan Faheem

Data intake reviewer:

Data storage location: <https://github.com/hf904/Data-Glacier-Internship/tree/main/Week%207>

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	26
Base format of the file	.xlsx
Size of the data	899 KB

Exploratory Data Analysis

EDA Overview

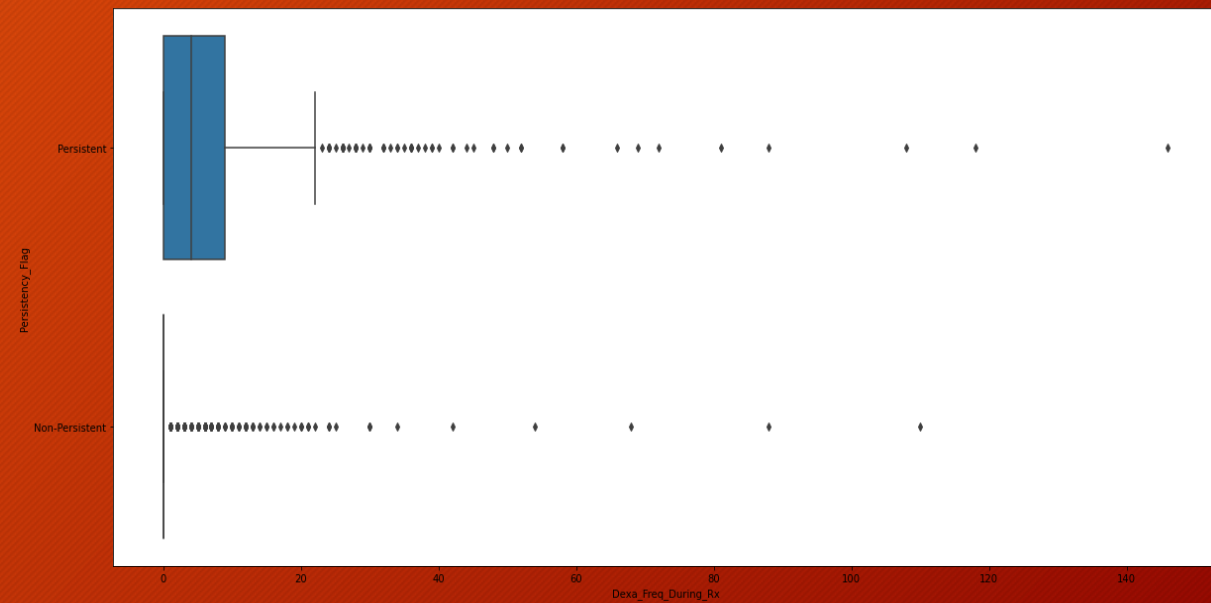
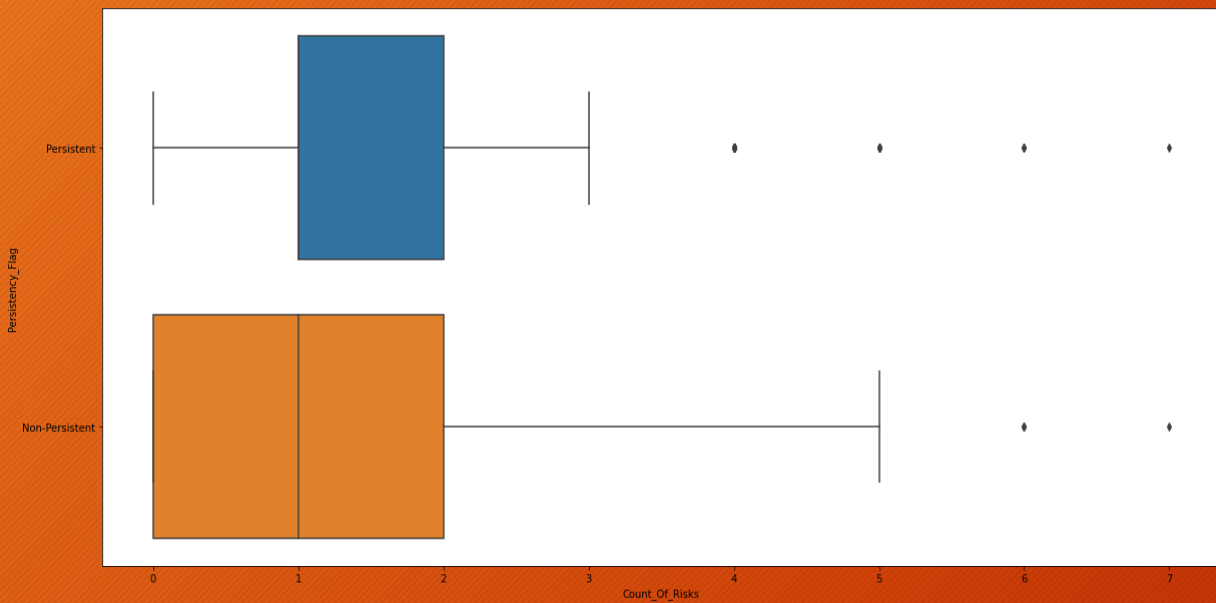
After performing Exploratory Data analysis on the dataset, the results show that most of the columns are of the Boolean data type and have the values of “Y” and “N”. These values will contribute to the model training in their current type and hence were mapped to the values of 1 and 0. Further analysis shows that no Null values were found in the dataset and so did not require any sort of data handling. The analysis show that a certain feature has some outliers and needed to be handled. To fix this, log transformation was performed on this feature to handle the outliers.

Problems in the Dataset

After analysis, it was found that the dataset has no null values. The function “isnull” was used along with sum function to check and verify the null values in the dataset. The screenshot shows that no null values were found. There were some outliers present in the numerical columns of the dataset.

```
1 Ptid 0
2 Persistence_Flag 0
3 Gender 0
4 Race 0
5 Ethnicity 0
6 Region 0
7 Age_Bucket 0
8 Ntm_Speciality 0
9 Ntm_Specialist_Flag 0
10 Ntm_Speciality_Bucket 0
11 Gluco_Record_Prior_Ntm 0
12 Gluco_Record_During_Rx 0
13 Dexa_Freq_During_Rx 0
14 Dexa_During_Rx 0
15 Frag_Frac_Prior_Ntm 0
16 Frag_Frac_During_Rx 0
17 Risk_Segment_Prior_Ntm 0
18 Tscore_Bucket_Prior_Ntm 0
19 Risk_Segment_During_Rx 0
20 Tscore_Bucket_During_Rx 0
21 Change_T_Score 0
22 Change_Risk_Segment 0
23 Adherent_Flag 0
24 Ids_Indicator 0
25 Injectable_Experience_During_Rx 0
26 Comorb_Encounter_For_Screening_For_Malignant_Neoplasms 0
27 Comorb_Encounter_For_Immunization 0
28 Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx 0
29 Comorb_Vitamin_D_Deficiency 0
30 Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified 0
31 Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx 0
32 Comorb_Long_Term_Current_Drug_Therapy 0
33 Comorb_Dorsalgia 0
34 Comorb_Personal_History_Of_Other_Diseases_And_Conditions 0
35 Comorb_Other_Disorders_Of_Bone_Density_And_Structure 0
36 Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias 0
37 Comorb_Osteoporosis_without_current_pathological_fracture 0
38 Comorb_Personal_history_of_malignant_neoplasm 0
39 Comorb_Gastro_esophageal_reflux_disease 0
40 Concom_Cholesterol_And_Triglyceride_Regulating_Preparations 0
41 Concom_Narcotics 0
42 Concom_Systemic_Corticosteroids_Plain 0
43 Concom_Anti_Depressants_And_Mood_Stabilisers 0
44 Concom_Fluoroquinolones 0
45 Concom_Cephalosporins 0
46 Concom_Macrolides_And_Similar_Types 0
47 Concom_Broad_Spectrum_Penicillins 0
48 Concom_Anesthetics_General 0
49 Concom_Viral_Vaccines 0
50 Risk_Type_1_Insulin_Dependent_Diabetes 0
51 Risk_Osteogenesis_Imperfecta 0
52 Risk_Rheumatoid_Arthritis 0
53 Risk_Untreated_Chronic_Hyperthyroidism 0
54 Risk_Untreated_Chronic_Hypogonadism 0
55 Risk_Untreated_Early_Menopause 0
56 Risk_Patient_Parent_Fractured_Their_Hip 0
57 Risk_Smoking_Tobacco 0
58 Risk_Chronic_Malnutrition_Or_Malabsorption 0
59 Risk_Chronic_Liver_Disease 0
60 Risk_Family_History_Of_Osteoporosis 0
61 Risk_Low_Calcium_Intake 0
62 Risk_Vitamin_D_Insufficiency 0
63 Risk_Poor_Health_Frailty 0
64 Risk_Excessive_Thinness 0
65 Risk_Hysterectomy_Oophorectomy 0
66 Risk_Nitrogen_Deficiency 0
67 Risk_Immobilization 0
68 Risk_Recurring_Falls 0
69 Count_Of_Risks 0
70 dtype: int64
```

Analyzing Outliers



Skewness & Kurtosis

The figures on the right shows the difference in outliers through skewness and kurtosis in fields of Count_Of_Risks and Dexa_Freq_During_Rx. It can be seen that Dexa_Freq_During_Rx have more skewness and kurtosis which shows that it has more outliers.

```
Count of risks skweness: 0.8797905232898707  
Count of risks Kurtosis: 0.9004859968892842
```

```
dexa_freq_during_rx skweness: 6.8087302112992285  
dexa_freq_during_rx Kurtosis: 74.75837754795428
```

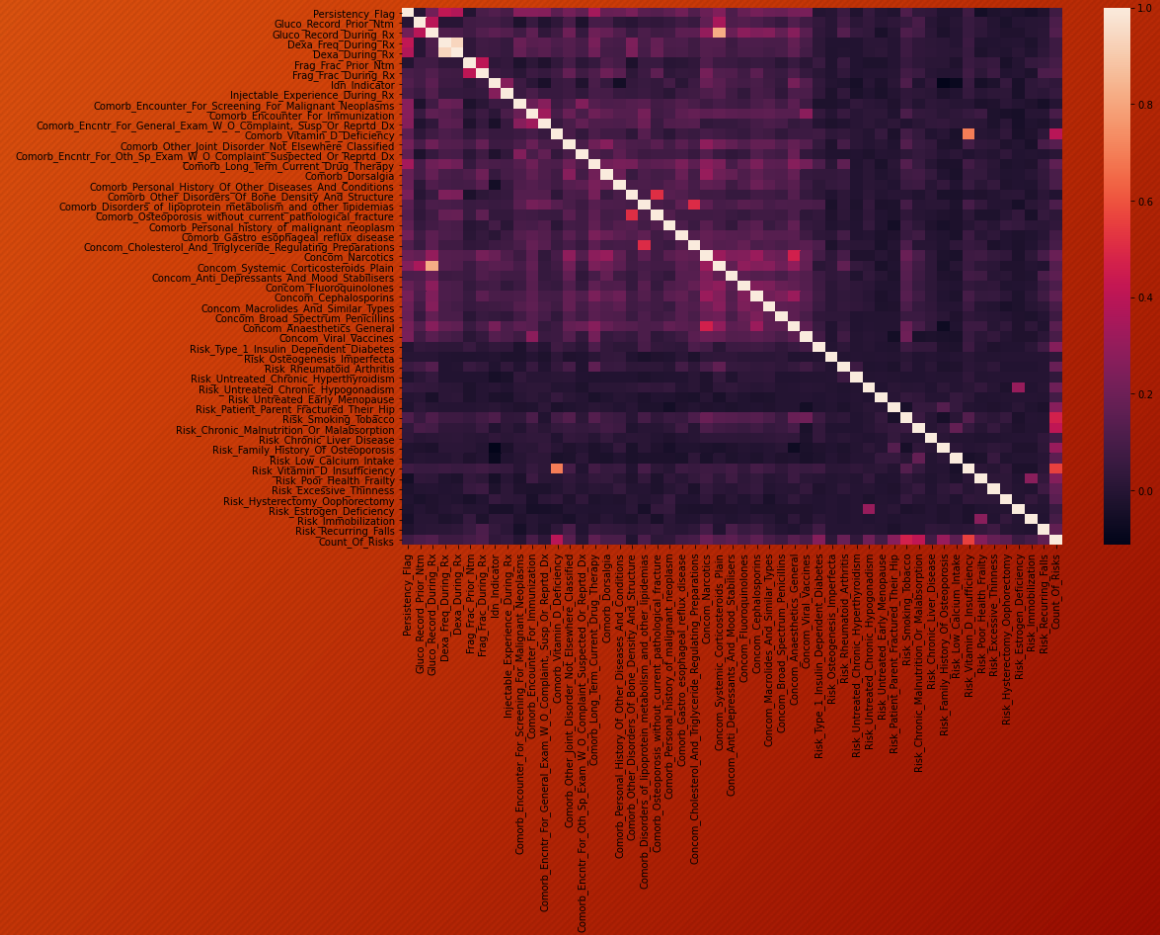

Correlation Between Target & Features

Figure on the right shows correlation between the target value and the features. It can be seen that there are not many features that are highly correlated with the target value.

	persistence flag
persistence flag	1.000000
dexa during rx	0.491823
dexa freq during rx	0.395247
comorb long term current drug therapy	0.352760
comorb encounter for screening for malignant neoplasms	0.322320
comorb encounter for immunization	0.314887
comorb encntr for general exam w o complaint, susp or rptd dx	0.289828
comorb other disorders of bone density and structure	0.247283
concom systemic corticosteroids plain	0.242854
comorb other joint disorder not elsewhere classified	0.233279
concom anaesthetics general	0.222293
concom viral vaccines	0.222241
concom macrolides and similar types	0.221611
concom cephalosporins	0.221543
comorb gastro esophageal reflux disease	0.220644
comorb personal history of other diseases and conditions	0.219665
comorb dorsalgia	0.215307
comorb encntr for oth sp exam w o complaint suspected or rptd dx	0.213413
gluco record during rx	0.212704
concom broad spectrum penicillins	0.197854
concom narcotics	0.191910
concom fluoroquinolones	0.186190
comorb personal history of malignant neoplasm	0.174835
comorb vitamin d deficiency	0.172664
comorb disorders of lipoprotein metabolism and other lipidemias	0.163495
comorb osteoporosis without current pathological fracture	0.139920
ntm specialist flag	0.139387
concom cholesterol and triglyceride regulating preparations	0.125552
adherent flag	0.112488
idn indicator	0.111440
concom anti depressants and mood stabilisers	0.110045
frag frac during rx	0.106935
change risk segment	0.106185
injectable experience during rx	0.098360
risk smoking tobacco	0.098045
ntm speciality bucket	0.091667
risk vitamin d insufficiency	0.079782
count of risks	0.071562
risk untreated chronic hypogonadism	0.067588
risk rheumatoid arthritis	0.053809

Dependency of Data Features

The **Figure** on the right shows the correlation between all the features. It can be seen from the figure that the features that are less correlated are in darker color while the features that are highly correlated are in lighter color.



Data Transformation

- **Null Values:** The dataset did not have any Null values present after the analysis and thus no step was taken in this transformation step.
- **Outliers:** In the numerical features of the dataset, there were outliers present which were shown by the skewness and kurtosis. The function RobustScaler was used to scale the values and the next step is to remove the outliers present and this is done by calculation the inter-quartile range and removing the values which lie outside the whiskers. This step changes and decreases the shape of the data from (3424, 69) to (2964, 69).
- **Changing data type:** The dataset had a lot of columns with the Boolean values of “Y” and “N”. For the purpose of model training, all the values of “Y”, “N” and of the target feature “Persistent”, “Non-Persistent” were changed to [1,0].
- **Unbalanced Dataset:** Unbalanced dataset was the next issue faced and this unbalanced dataset will in turn affect the prediction results. Hence to counter this unbalancing issue, Up Sampling method was used. This method will bump up the records of the class with minority and thus will make all the records equal in count.
- **One-Hot Encoding:** The final step in the data transformation was the implementation of the function get_dummies which was used for the purpose of One-Hot Encoding. The numerical values are needed for the classifiers to work on and so by this method, the values are transformed into numerical values which can be used by the classifiers.

Model Training & Testing

Introduction to Model Testing & Training

As seen from the figures of the previous section, its clear that not many features are highly correlated with the target value. Therefore, to avoid any overfitting, it would be in the best interest to ignore the less correlated features during the model training section of the project which comes after this. In the model training section, the dataset is divided into two sections with 70% data for training the model and 30% data is given for testing the model.

Classifiers Used

There are Classifiers used from each of the family of Models which include Linear Models, Ensemble & Boosting Models and Neural Network. The following are the classifiers that were trained and tested on:

- Ensemble & Boosting Models
 - Bagging Classifier
 - Gradient Boosting Classifier
 - Random Forest Classifier
 - ExtraTrees Classifier
 - AdaBoost Classifier
 - XGBoost Classifier
 - Stacking Classifier

Classifiers Used - Continued

- Linear Models
 - Ridge Classifier
 - SGD Classifier
 - Logistic Regression Classifier
- Neural Network
 - Multi-Layer Neural Network
 - Multi-Layer Perceptron

Evaluation of Performance of Classifiers

To evaluate the classifier models and select the best classifier, there are various methods that can be used. For this project, the following matrices are used to evaluate the classifiers:

- Accuracy
- Recall
- Precision
- F1-Score
- Confusion Matrix
- ROC

Classifiers Train & Test Results

The following table shows the results of Ensemble & Boosting Models

Classifier	Accuracy
Bagging Classifier	0.80
Gradient Boosting Classifier	0.73
Random Forest Classifier	0.79
ExtraTrees Classifier	0.79
AdaBoost Classifier	0.78
XGBoost Classifier	0.77
Stacking Classifier	0.80

Classifiers Train & Test Results- Continued

The following table shows the results of Linear Models

Classifier	Accuracy
Ridge Classifier	0.79
SGD Classifier	0.78
Logistic Regression Classifier	0.78

The following table shows the results of Neural Network

Classifier	Accuracy
Multi-Layer Neural Network	0.79
Multi-Layer Perceptron	0.76

Final Recommendation & Conclusion

From the previous slides it is clear that the following classifiers gave out the best results:

- Bagging Classifier
- Stacking Classifier
- Ridge Classifier

By taking into consideration all the mentioned evaluation metrics in the previous slides, it can be seen that Stacking Classifier gives slightly better result than Ridge Classifier and Bagging Classifier.

Final Recommendation & Conclusion- Continued

The figures below shows the evaluation metrics of the Classifier and the Confusion Matrix of the classifier selected.

	precision	recall	f1-score	support
Non-Persistent	0.81	0.93	0.87	614
Persistent	0.76	0.51	0.61	269
accuracy			0.80	883
macro avg	0.79	0.72	0.74	883
weighted avg	0.80	0.80	0.79	883

