Healthcare: Persistency of a Drug

EDA Presentation

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

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

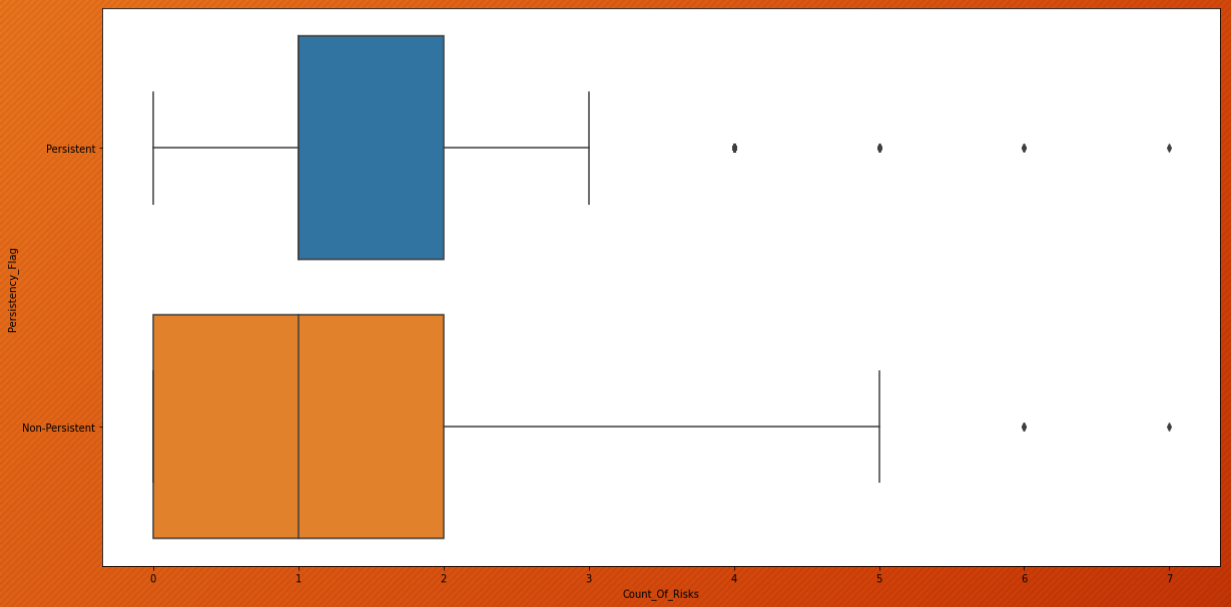
EDA Overview

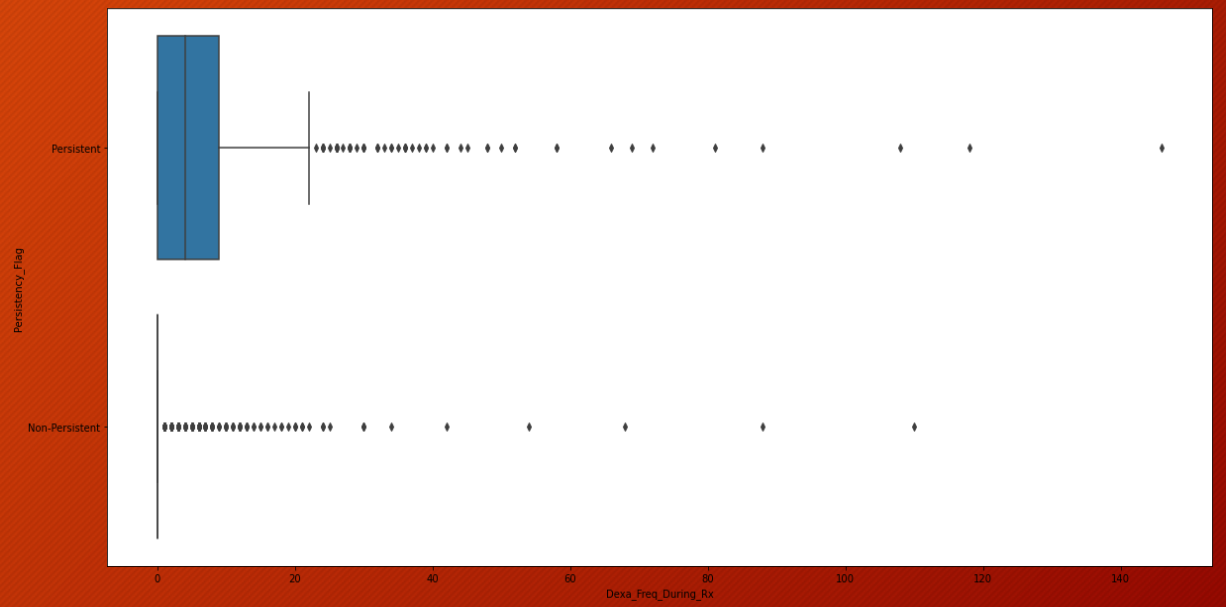
After performing Exploratory Data analysis on the dataset, the results show that most of the columns are of the Boolean data type and have the values of “Y” and “N”. These values will contribute to the model training in their current type and hence were mapped to the values of 1 and 0. Further analysis 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.

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
```

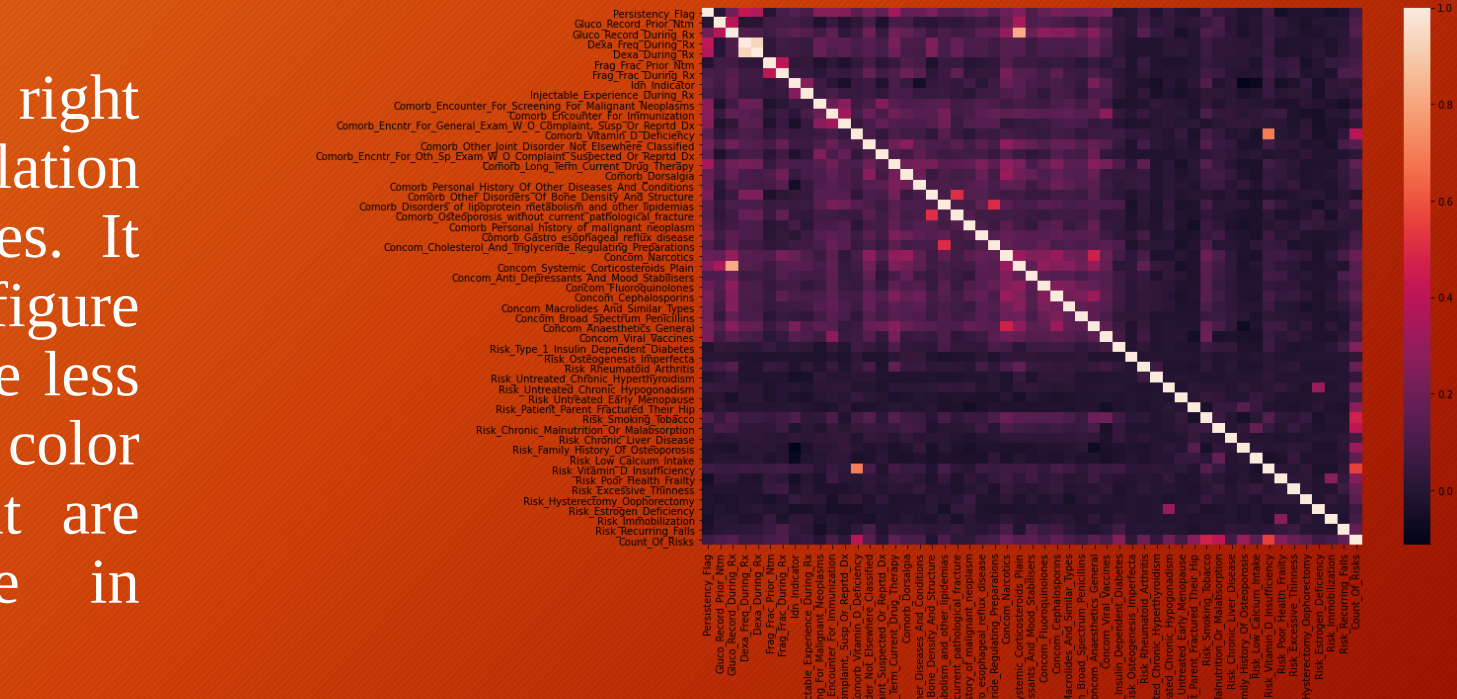
Correlation Between Target & Features

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.

	persistency_flag
persistency_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_reprtd_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_reprtd_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.

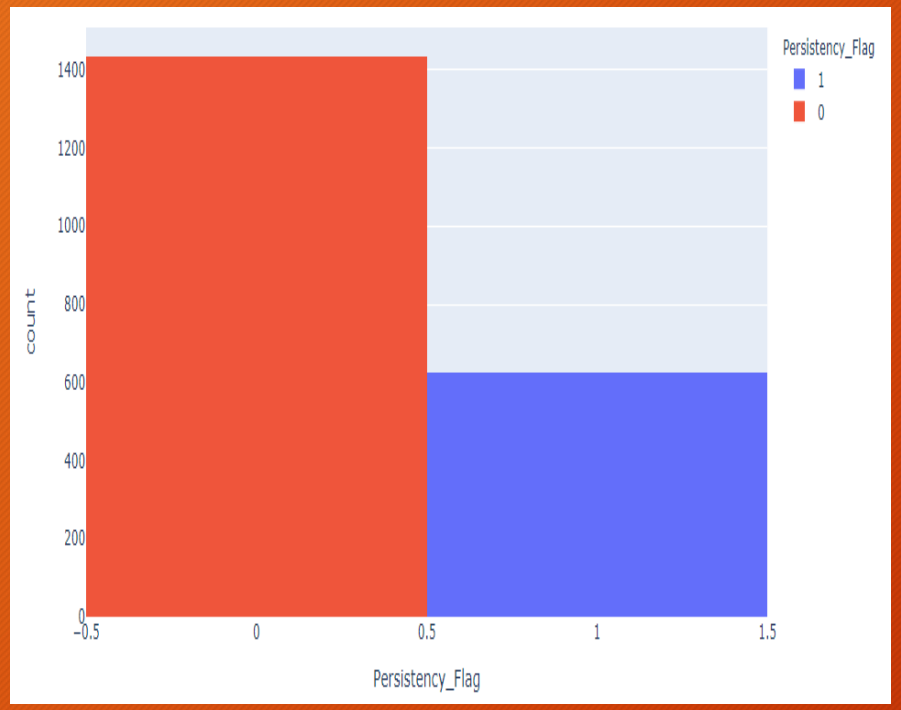


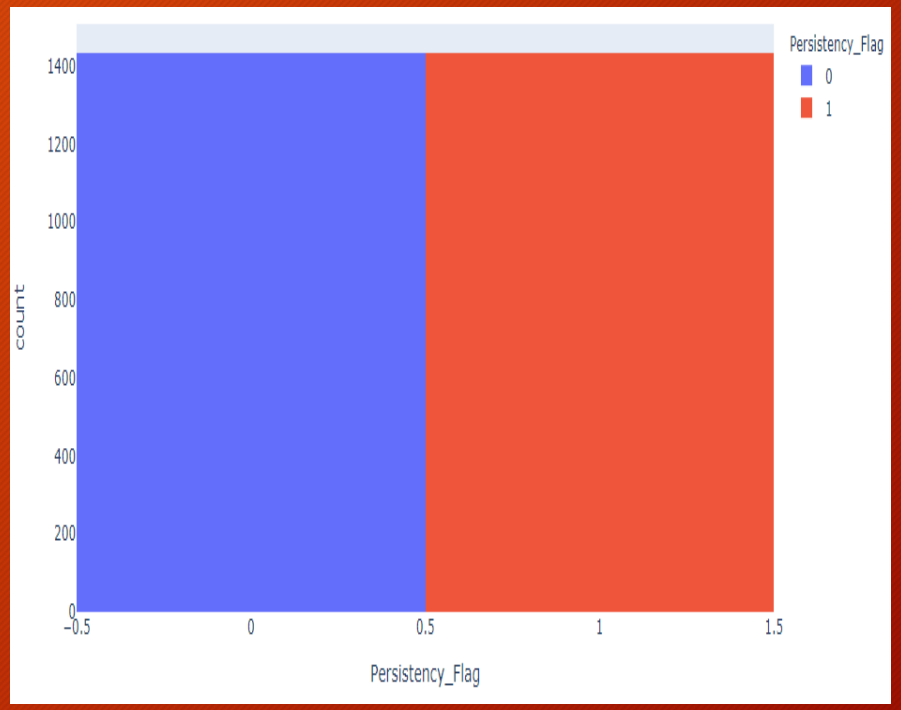
Data Transformation

- **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 with 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.

Up Sampling Unbalanced Dataset

The figures below show before and after the use of Up Sampling method





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 in divided into two section 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