Week 9 Submission

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.

Project Lifecycle

Project Name	Healthcare - Persistency of a drug
Start Date	16th March 2022 (Week 7)
Final Submission Date	20th April 2022
Project Duration	5 weeks
Deliverables Submission Dates	1. March 16th 2. March 23rd 3. March 30th 4. April 6th 5. April 13th 6. April 20th

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.

Exploratory Data Analysis (EDA)

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.

Data Types

After analyzing the data, it can be seen that the data has a dimension of (3424, 69). Most of the features here are of type “Object” and very few are of type “int64”. The Object type means that the data is of categorical in nature.

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. There were some outliers present in the numerical columns of the dataset. The **Figure 1** below shows the outliers present in Count_Of_Risks. The **Figure 2** shows outliers in Dexa_Freq_During_Rx.

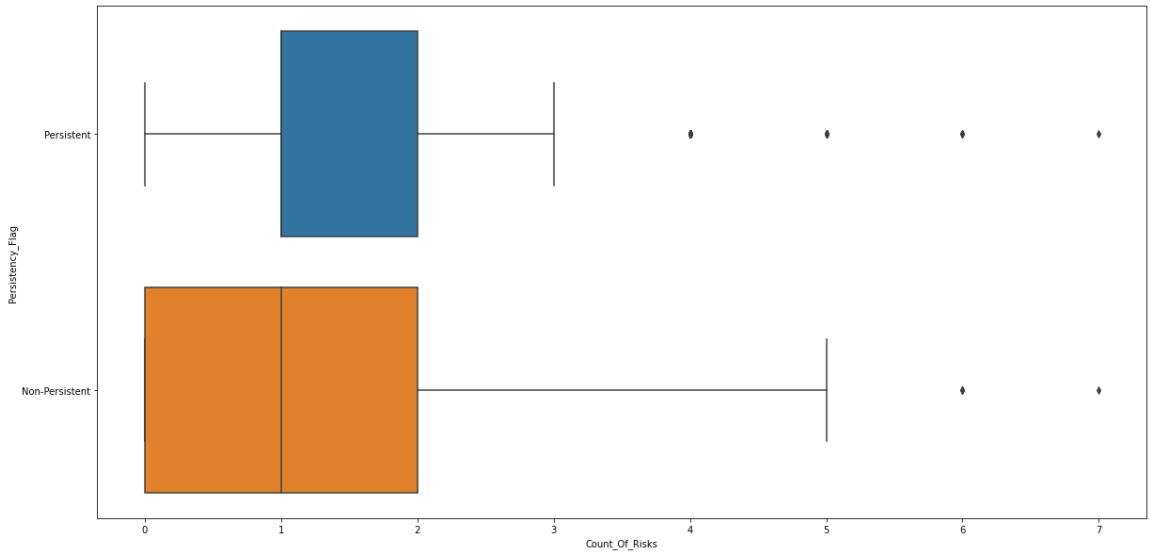

Figure 1: Outliers in Count_Of_Risks

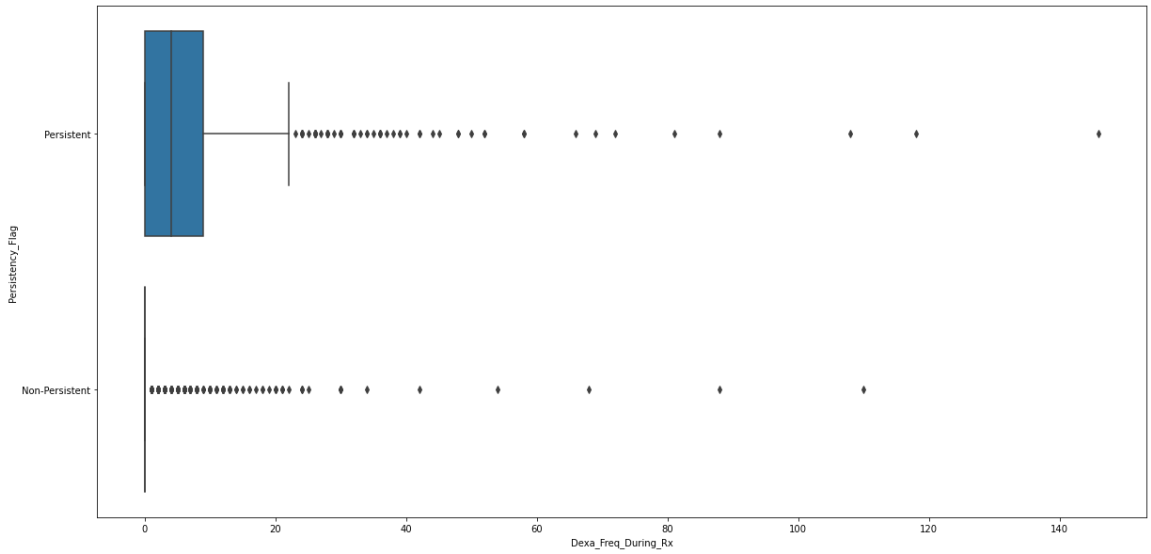

Figure 2: Outliers in Dexa_Freq_During_Rx

Figure 3 and **Figure 4** 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
```

Figure 3

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

Figure 4

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.