Data Glacier

Your Deep Learning Partner

Week 8 Deliverables

Group Name: The Data Doctors

25-Nov-2023

Team Details

Name: Noah Gallego Email: noahgallego394@gmail.com Country: United States College/Company: California State University Bakersfield Specialization: Data Science	Name: Tomisin Abimbola Adeniyi Email: tomisin_adeniyi11@yahoo.com Country: Nigeria College/Company: N/A Specialization: Data Science
Name: Mohammad Shehzar Khan Email: mshehzarkhan@gmail.com Country: Turkey College/Company: Koç University Specialization: Data Science	Name: Ashish Sasanapuri Email: sashrao21@gmail.com Country: India College/Company: N/A Specialization: Data Science

Problem Description

One challenge for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription. To solve this problem ABC Pharma company approached an analytics company to automate this process of identification.

Data Understanding

- The dataset provides the factors impacting the patient’s persistence to ABC pharmaceutical company’s drug prescribed by various Nontuberculous mycobacteria (NTM) specialists.

- The aim is to build a machine-learning model that classifies the patient into **Persistent** (Compliant) and **Non-persistent** (Non-Compliant).

- The dataset consists of 3242 records and is a an imbalanced dataset due to low number of **Persistent** records as compared to **Non-persistent**.

- There are no missing values in the dataset, other than *‘unknown’* values.

- Among the independent features, there are 2 features - *Dexa_Freq_During_Rx* and *Count_Of_Risks*, that have outliers.

Data Understanding

- The dataset contains a total of 69 features that are divided into multiple categories -
    - 1 Target variable: Persistency_Flag
    - 1 Unique identifier for each patient: Ptid
    - 6 Demographic variables of the each patient: Age_Bucket, Gender, Race, Ethnicity, Region, Idn_Indicator
    - 3 Physician Specialist attributes: Ntm_Speciality, Ntm_Specialist_Flag, Ntm_Specialist_Bucket
    - 13 Clinical factors: T-Score details, Risk_Segment details, Multiple risk factors count, DEXA details, Fragility fracture details, Glucocorticoid details
    - 45 Disease/Treatment factors: Injectable drugs, Risk factors, Comorbidities, Concomitancies, Adherence to therapy

Type of Data

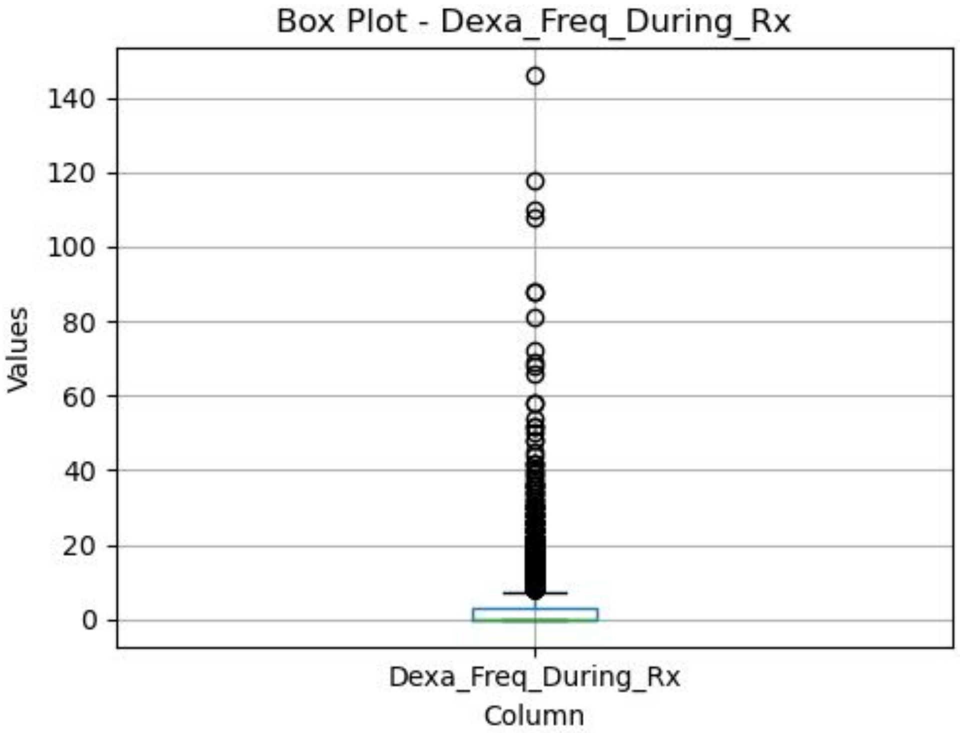
- The dataset contains high majority of categorical data rather than numerical data.
    - Categorical features: 66
    - Numerical features: 2

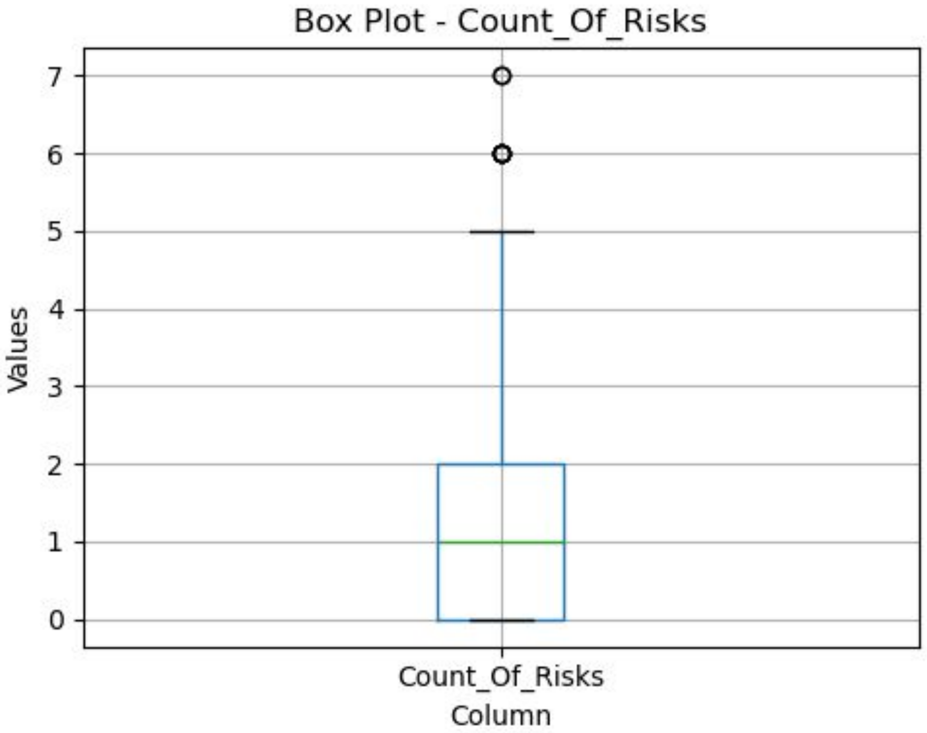
- Among the given features, 68 are independent variables and the target variable is the **Persistency_Flag**.

Problems in the Data - Outliers

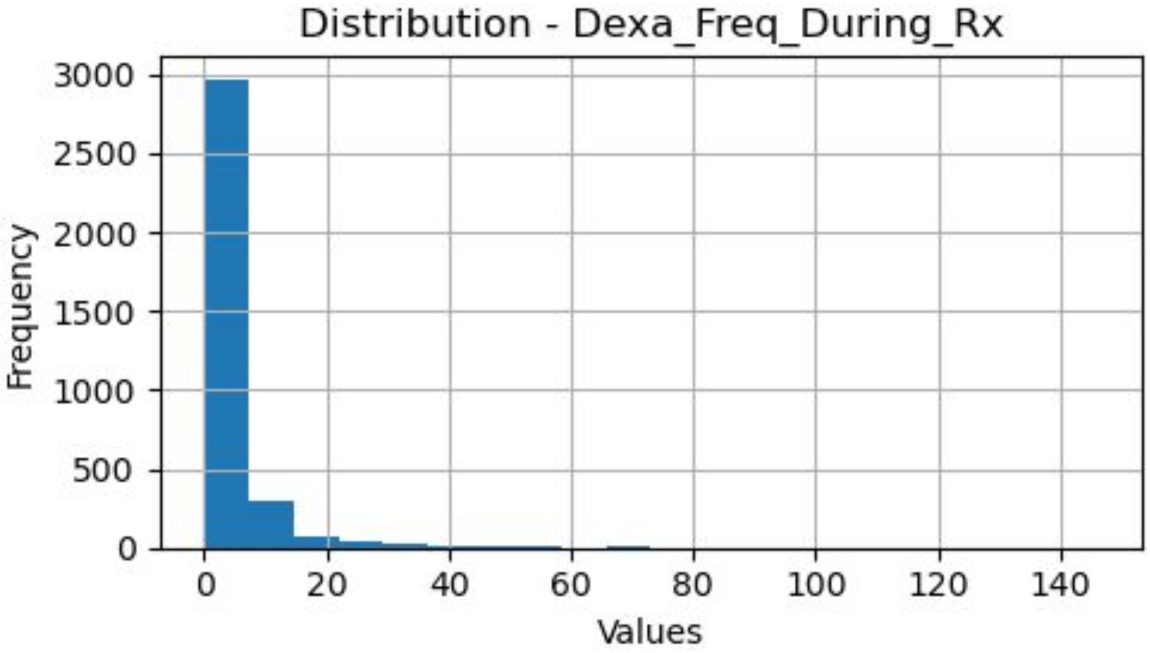
- 2 of the features in the dataset contain outliers - *Dexa_Freq_During_Rx* and *Count_Of_Risks*

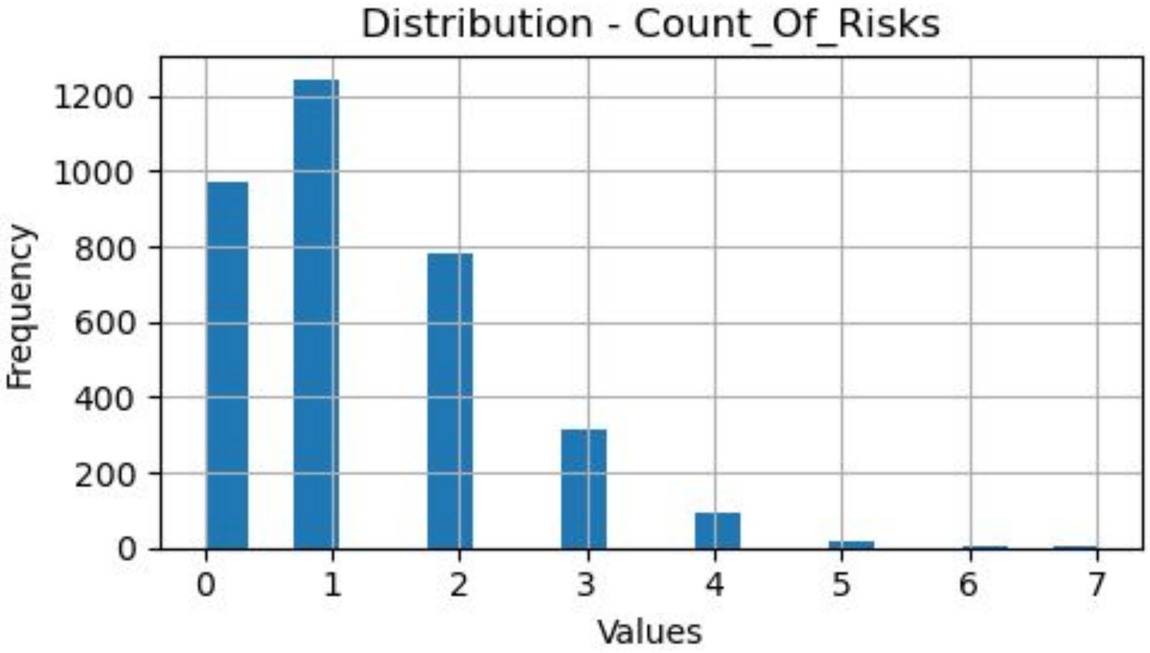
- Data is positively skewed





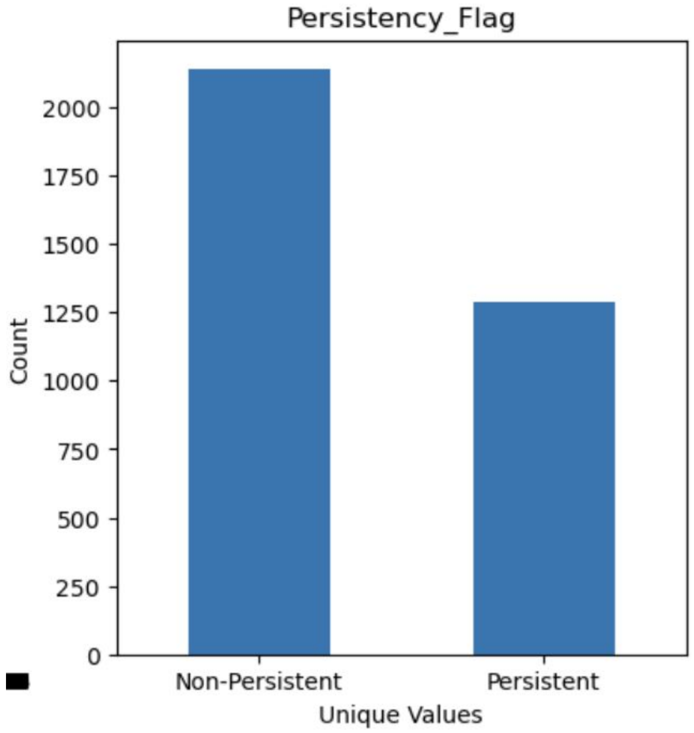
Problems in the Data - Outliers


Distribution - Dexa_Freq_During_Rx


Distribution - Count_Of_Risks

Problems in the Data - Imbalance

- The target feature - *Persistency_Flag* is imbalanced with 2135 records **Persistent** and 1289 records **Non-persistent**.



Problems in the Data - Others

- *NTM_Speciality* feature has 36 unique categories.
- *Count_Of_Risks* feature has 8 unique values.
- *Dexa_Freq_During_Rx* has 58 unique values.

Handling Problems in Data

- Handling of outliers can be performed using below methods:
    - **Winsorize** - This can be used to cap the lower and upper bound of the outliers.
    - **Log Transformation** - Applying log to the data points will change the values but also help reduce skewness and make the data more normal.
    - **Median Absolute Deviation**(**MAD**) - Approach is similar to *Z-score* but uses *Median* and *Median Absolute Deviation* statistics instead of *Mean* and *Standard Deviation*.
    - **Box-Cox Transformation** - Power transformation method that tries to stabilize variance and make the data more normally distributed. We can use the boxcox() method in python.
    - **Square-root Transformation** - Takess the square root of each data point. It's often used to moderate right-skewed data and can make a more-normal distribution.
    - **Inverse Transformation** - The inverse transformation is the reversal of a previous transformation, typically used to revert data to its original scale.

Handling Problems in Data (*Continued*)

- Since the dataset contains 2 different features for Physician specialist attribute, we can drop the *Ntm_Speciality* feature. We can use the *Ntm_Speciality_Bucket* feature as it consists of 3 different categories generalising the *Ntm_Speciality* feature.

Handling Problems in Data (*Continued*)

- The different categories for *Count_Of_Risks* feature can be reduced to (0, 1, 2, 3 and >3) to reduce complexity.

- The values *Dexa_Freq_During_Rx* feature can be updated into different buckets - (0 - 6], (6 - 12], (12 - 18], (18 - 24], (24 - 30] and (>30).

- The above 2 methods can also be considered as outlier handling methods.

- Furthermore, to combat missing values in the Risk Segment during Rx column, we can take data from the Risk Segment Prior to Rx column. This is because, the taking the prescription and not taking it resulted in no change the majority (86%) of the time.

Data Glacier

Your Deep Learning Partner

Thank You