

Week 8 Deliverables

Group Name: The Insights Team

April 26, 2024

Team members:

|  |  |
| --- | --- |
| Name: Tomisin Abimbola Adeniyi  E-mail: tomisin\_adeniyi11@yahoo.com  Country: Nigeria  College/Company:  Specialization: Data Science | Name: Fabio Pontecchiani  E-mail: pontecchianifabio@gmail.com  Country: Belgium  College/Company: University of Sheffield  Specialization: Data Science |
| Name: Bilikis Omolara Alayo  E-mail: berlykis@gmail.com  Country: United Kingdom  College/Company:  Specialization: Data Science |  |

**Problem Description**

ABC Pharma Company faces the challenge of understanding drug persistence based on physician prescriptions. They want to automate the identification process to gain insights into the factors affecting persistence. The goal is to build a classification model using a given dataset. The target variable we are interested in is *Persistency Flag*.

**Data Understanding**

1. The dataset contains a total of 69 features that are sub-divided into the following five categories:

* 1 Target variable: Persistency\_Flag
* 1 Unique Identifier: Ptid6 Demographic Variables: Age\_Bucket, Gender, Race, Ethnicity, Region, Idn\_Indicator
* 3 Physician Specialist: 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

1. There are 68 independent features (66 categorical features and 2 numerical features), and one dependent feature
2. The dataset consists of 3242 records
3. There are no missing values in the dataset but there are *“unknown”* values which can be regarded as missing values

**Data Type**

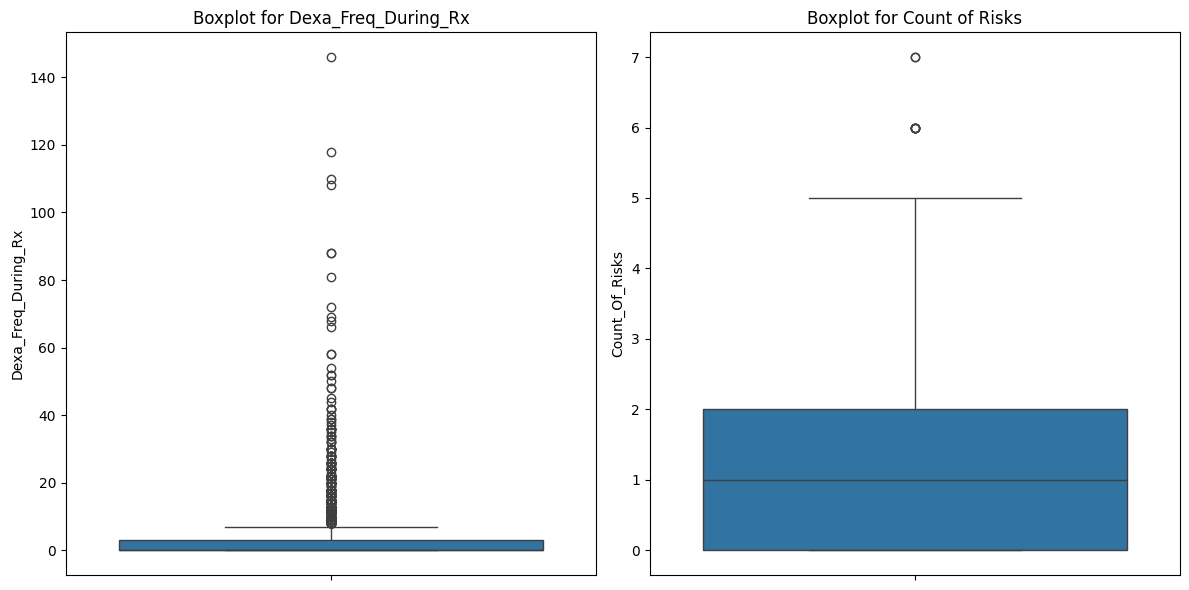
1. The dataset contains both categorical data and numerical data.

* Categorical features: 66
* Numerical features: 2

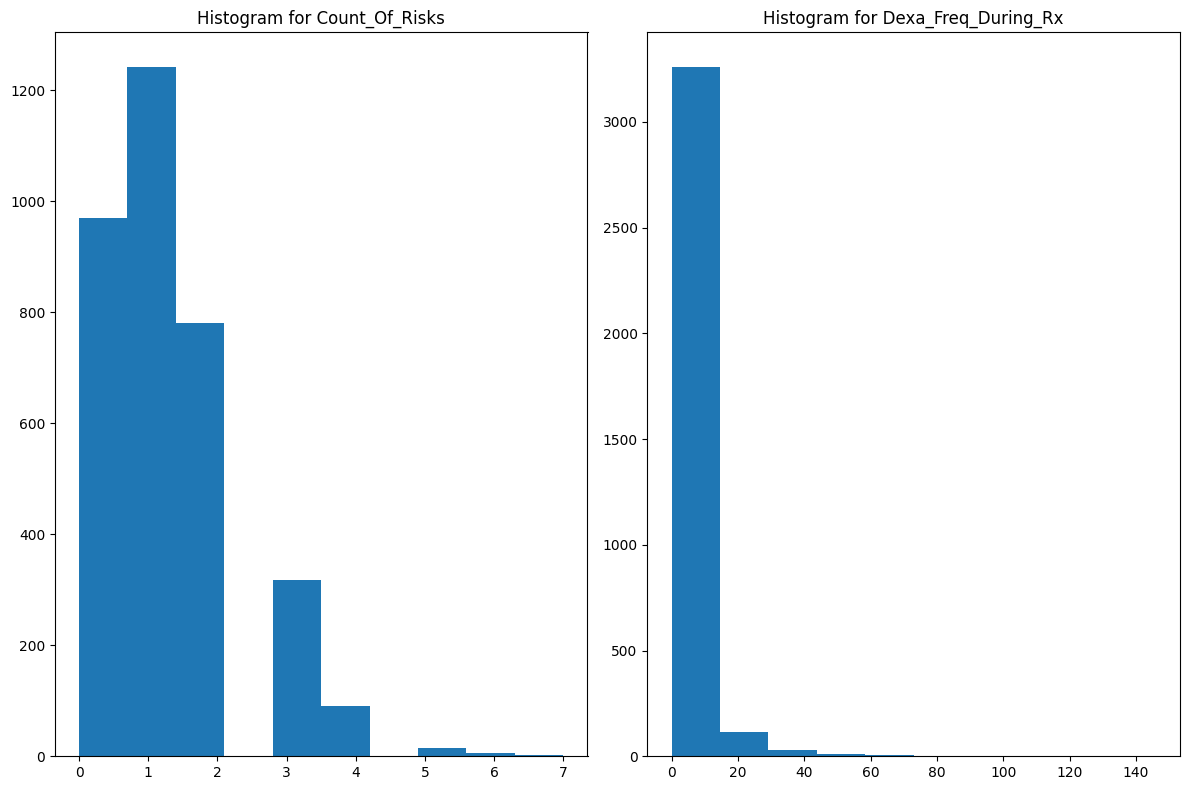
1. The dataset contains 68 independent variables and 1 target variable which is the Persistency\_Flag.

**Dataset Problem**

1. The two numerical features in the dataset, *Dexa\_Freq\_During\_Rx* and *Count\_Of\_Risks*, have **outliers**

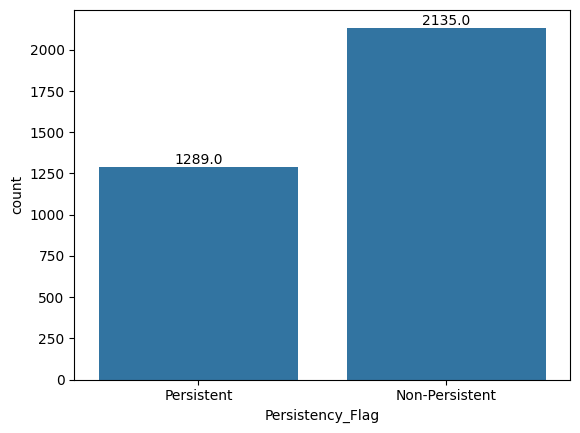


1. The dataset is **positively skewed**



**Dataset Problem Cont’d**

1. The target variable, *Persistency\_Flag,* is imbalanced (Persistent-1289, and Non-persistent-2135)



1. **Too many unique values:**

* *NTM\_Speciality* feature has 36 unique values
* *Count\_Of\_Risks* feature has 8 unique values
* *Dexa\_Freq\_During\_Rx* has 58 unique values

1. **Missing values**: the following features have missing values

* Race
* Ethnicity
* Region
* Ntm\_Speciality
* Risk\_Segment\_During\_Rx
* Tscore\_Bucket\_During\_Rx
* Change\_T\_Score
* Change\_Risk\_Segment

**Handling Dataset Problem**

1. **Outliers**: various methods can be employed to manage outliers within a dataset:

* Winsorize: This technique involves capping the lower and upper bounds of the outliers, effectively limiting their impact on statistical analysis.
* Log Transformation: By applying a logarithmic function to the data points, values can be altered while simultaneously reducing skewness and promoting a more normalized distribution.
* Median Absolute Deviation (MAD): Similar to the Z-score approach, MAD employs the median and Median Absolute Deviation statistics instead of the mean and standard deviation, offering robustness against outliers.
* Box-Cox Transformation: Utilizing a power transformation method, Box-Cox aims to stabilize variance and enhance the normality of the data. In Python, the boxcox() method can be used for this purpose.
* Square-root Transformation: This method involves taking the square root of each data point. It is commonly utilized to mitigate right-skewed data distributions and facilitate a more normalized shape.
* Inverse Transformation: The inverse transformation serves to reverse a previous transformation, effectively restoring data to its original scale. This can be valuable for reverting data back to its initial representation post-analysis.

1. **Skewness: ##Fabio##**
2. **Imbalance:** Addressing imbalanced datasets can be achieved through several techniques:
   * **Resampling Techniques**:
     1. **Oversampling and Under-sampling**: Equilibrate class distribution by oversampling the minority class or under-sampling the majority class.
     2. **Synthetic Minority Oversampling Technique (SMOTE):** Employ SMOTE to generate synthetic instances of the minority class, thereby balancing the dataset. SMOTE helps alleviate overfitting risks associated with the original minority class.
   * **Weighted Sampling**: Incorporate weighted sampling during training. Assign higher weights to samples from the minority class to enhance their influence on model training. This ensures the model adequately addresses the minority class.
3. **High number of unique values:**

* To simplify the complexity of the "Count\_Of\_Risks" feature, its various categories can be condensed into five groups: 0, 1, 2, 3, and >3.
* The "Dexa\_Freq\_During\_Rx" feature can be reorganized into distinct intervals as follows: (0 - 6], (6 - 12], (12 - 18], (18 - 24], (24 - 30], and (>30).
* Considering the presence of two separate features for the Physician specialist attribute, it is advisable to eliminate the "Ntm\_Speciality" feature. Instead, we can rely on the "Ntm\_Speciality\_Bucket" feature, which encompasses three distinct categories, offering a generalized representation of the "Ntm\_Speciality" feature.
* The steps above will also reduce the skewness in the dataset

1. **Missing Values**: **##Bilkis##**