Data Glacier

Your Deep Learning Partner

Healthcare – Persistency of a Drug

The Insights Team

April, 2024

Week 8: Problem Description and Data Understanding

Team Members:

Name	Email	Country	College	Specialization
Tomisin Abimbola Adeniyi	tomisin_adeniyi11@yahoo.com	Nigeria		Data Science
Fabio Pontecchiani	pontecchianifabio@gmail.com	Belgium	University of Sheffield	Data Science
Bilikis Omolara Alayo	berlykis@gmail.com	United Kingdom		Data Science

Problem description

Accurately determining the duration of medication use as advised by doctors presents a substantial difficulty for pharmaceutical companies. This makes it more difficult to understand patient adherence trends and optimise treatment outcomes. To tackle this problem, ABC Pharma Company has enlisted the help of an analytics company to create an automated system that tracks and identifies drug persistency based on doctor prescriptions.

Data understanding

The dataset we will utilise for the project is named Healthcare_dataset.xlsx, which is available through the link.
https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view.

It consists of

- 3424 observations
- 67 Features (2 numerical features and 67 categorical features)
- 1 target class which is the Persostency_Flag

The initial exploration of the dataset shows the following problem:

Missing values - The dataset has a significant number of missing values from eight features, with Change_Risk_Segment having the largest proportion of 65% and Region having the lowest amount of 2%. The proportion of missing observations is represented in the bar chart below.

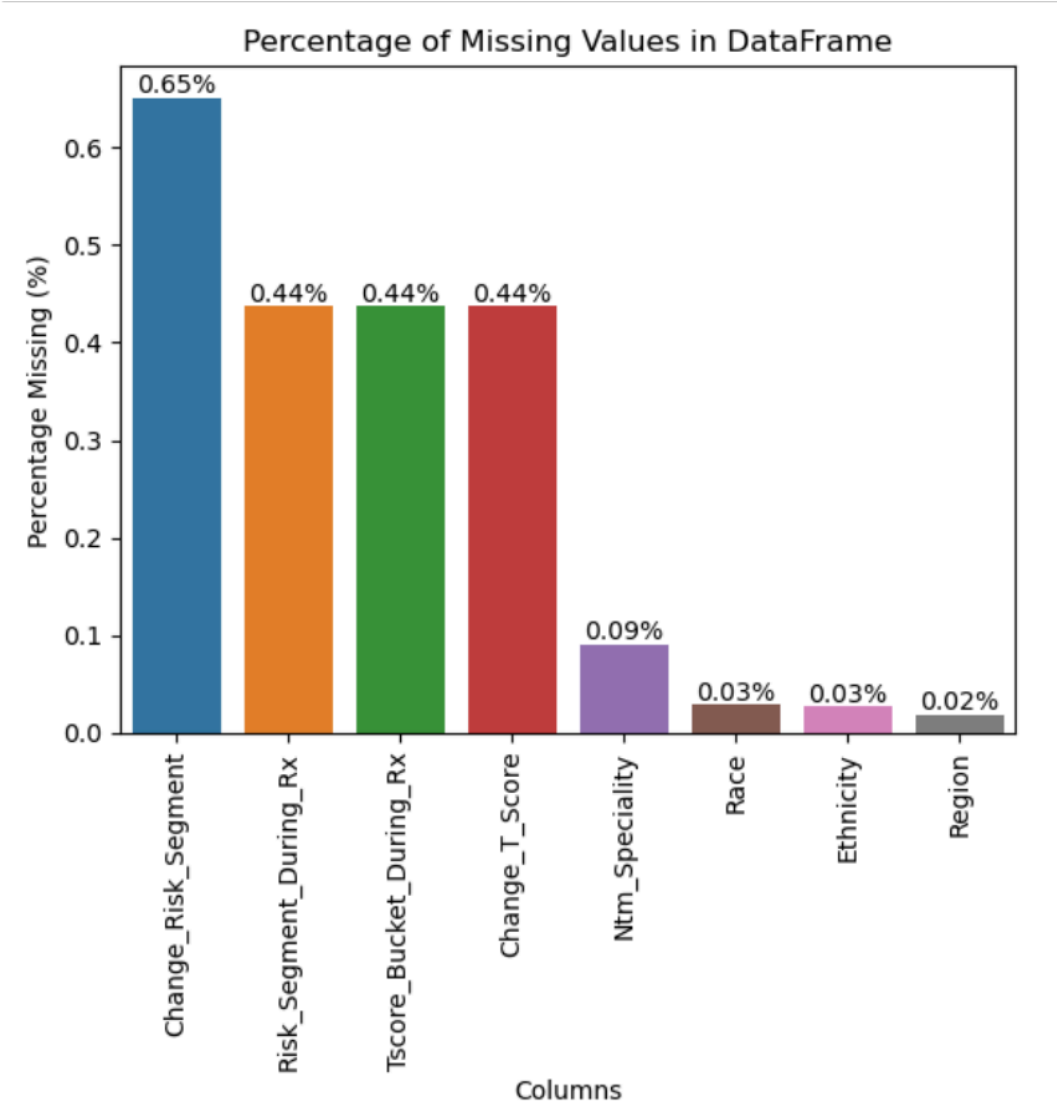

Figure 1: Percentage of missing values per feature

Class imbalance - The target class of the dataset, persistence_FLAG, has an imbalance class. The distribution of the target class categories (Persistent and Non_persistent) in the dataset is shown in the picture below.

.

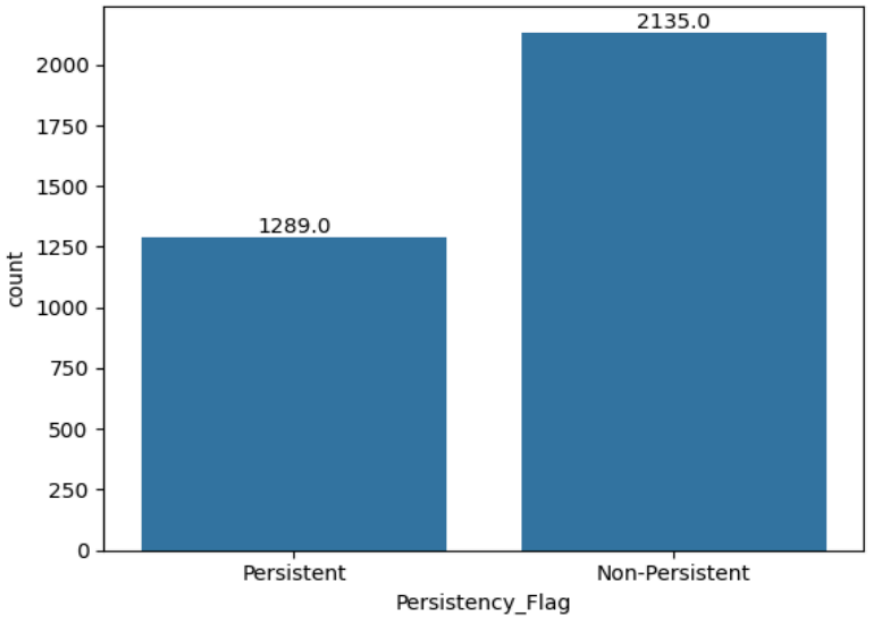

Fig 2: Target class categories distributions

Outliers – The numerical features Dexa_Freq_During_Rx and Counts_of_Risks appear to contain some outliers during data exploration; however, they will not be treated as such because they represent the "Number of DEXA scans performed prior to the first NTM Rx date (within 365 days of the rxdate)" and "the count of risk factors that the patient is falling into," respectively.

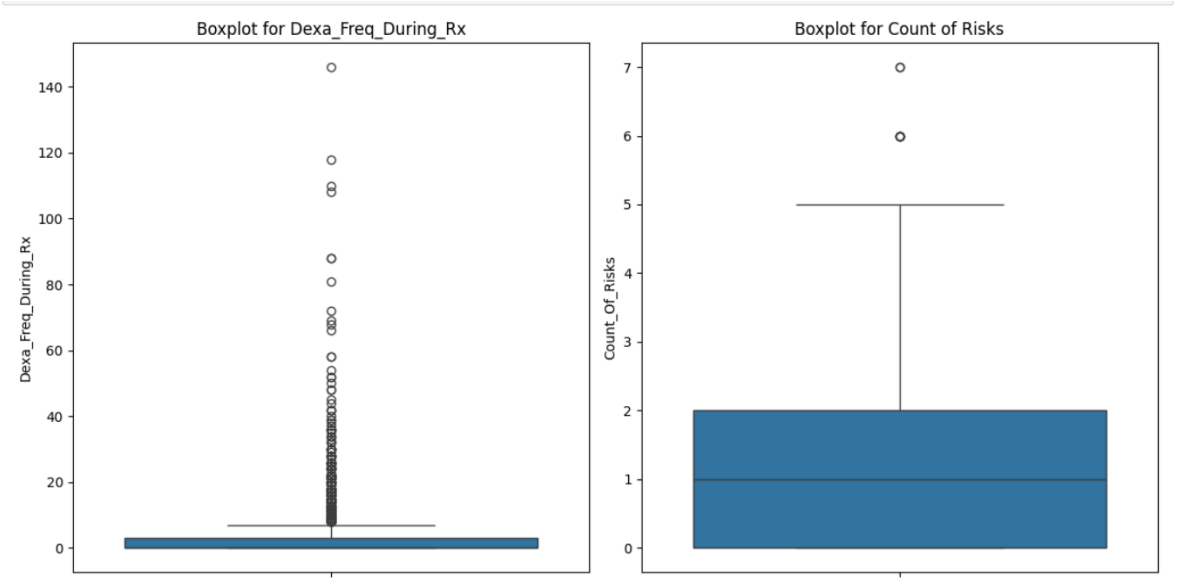

Fig 3: Boxplot showing outliers in the numerical features

Skewness – During exploration, it was observed that the numerical features appear to have a rightly skewed distribution.

```
Skewness of each numerical variable:
Dexa_Freq_During_Rx    6.808730
Count_Of_Risks         0.879791
dtype: float64
```

The Dexa_Freq_During_Rx feature has a skewness of 6.808730, indicating a highly right-skewed distribution. The majority of data points are on the left side, with a long tail towards the right. This indicates few data points with extremely high values, which can significantly impact the mean and potentially the overall analysis.

Similarly, the Count_Of_Risks data has a moderately right-skewed distribution with a skewness of 0.879791, indicating a slight trend towards higher values and a moderate number of data points with values greater than the mean.
The figure below shows the histogram representation of these features in relation to their skewness

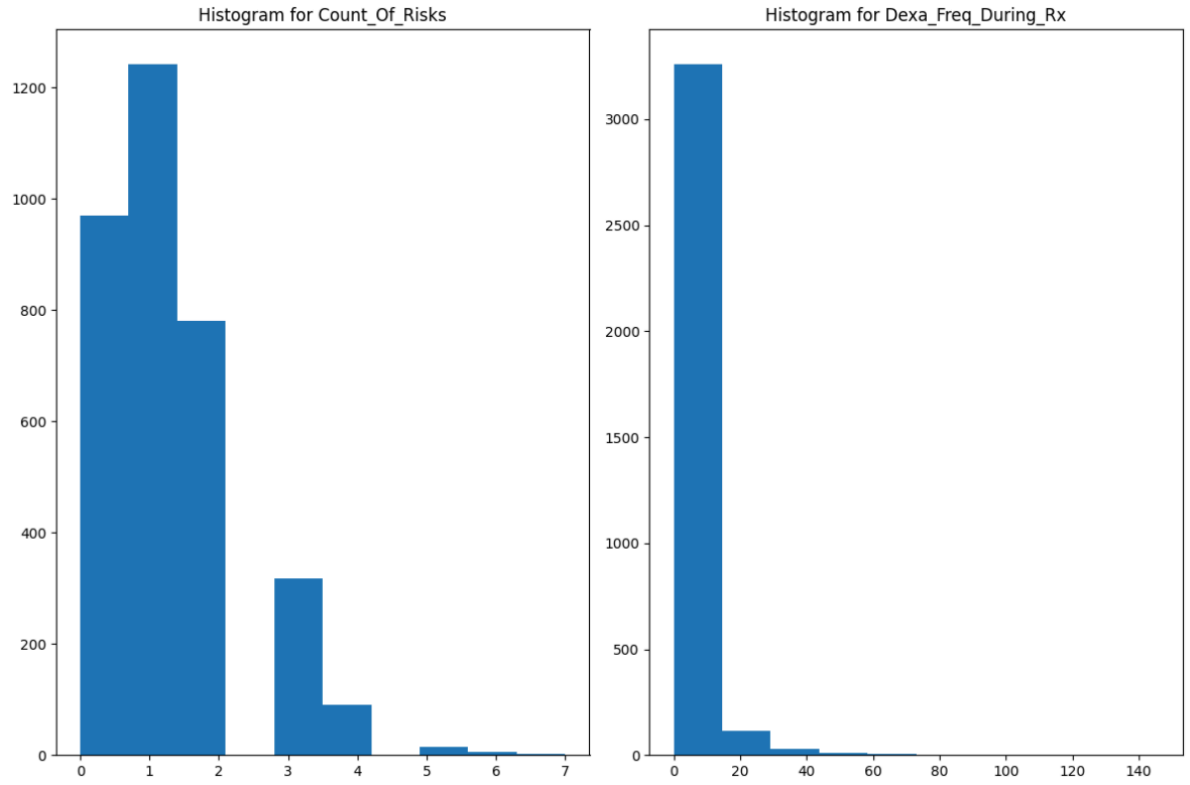

Fig 4: Histogram representation of the numerical features.

Approaches to solve identified problems

The identified problems will be addressed as follows:

Missing values - This can be overcome in a variety of ways. One option is to employ the whole case analysis deletion, which would result in the loss of a substantial amount of data because some features have more than 40% missingness. As a result, we want to use the imputation method to help prevent information loss, preserve sample size, and reduce bias.

Class imbalance: This will be addressed through the application of Synthetic Minority Oversampling Techniques (SMOTE), a specific type of oversampling method for addressing imbalance. This method creates synthetic examples for the minority class by interpolating existing instances. This strategy is more useful because it incorporates diversity into the synthetic sample, lowering the danger of overfitting. This makes it a popular and effective strategy.

Outliers: Outliers in numerical features will not be treated as such because we intend to deploy outlier-tolerant machine learning methods such as random forest, support vector machines, and K-nearest neighbours.

Skewness: if we take the logarithms of the positive values “Dexa_Freq_During_Rx”, the Skewness is highly reduced. The “Count of Risk” feature is not as skewed and it can be left as it is.