Data Glacier

Your Deep Learning Partner

Week 9 Deliverables

Group Name: The Insights Team

May 2, 2024

Week 9: Data Cleaning and Transformation

Team Members:

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

Data cleaning and Transformation

The data cleaning and transformation done on the data entails:

1. Checking and handling of missing values

The dataset contains a substantial number of missing values from 8 columns and this was handled using:

- Mode imputation – Since the missing values were categorical, mode imputation was used as a single imputation method for addressing the missingness.

- Model-based imputation - k-Nearest Neighbours imputation was used for completing missing values. The data was divided into two sets and each sample's missing values were imputed using the mean value from n_neighbors nearest neighbours found in the training set.

The mode imputation method appears suitable as it is straightforward and does not require categorical label encoding while the model-based imputation does require label encoding.

2. Skewness handling

   Numerical variables in the dataset are rightly skewed and this was addressed using log transformation which significantly reduced the skewness.

   Skewness before transformation:

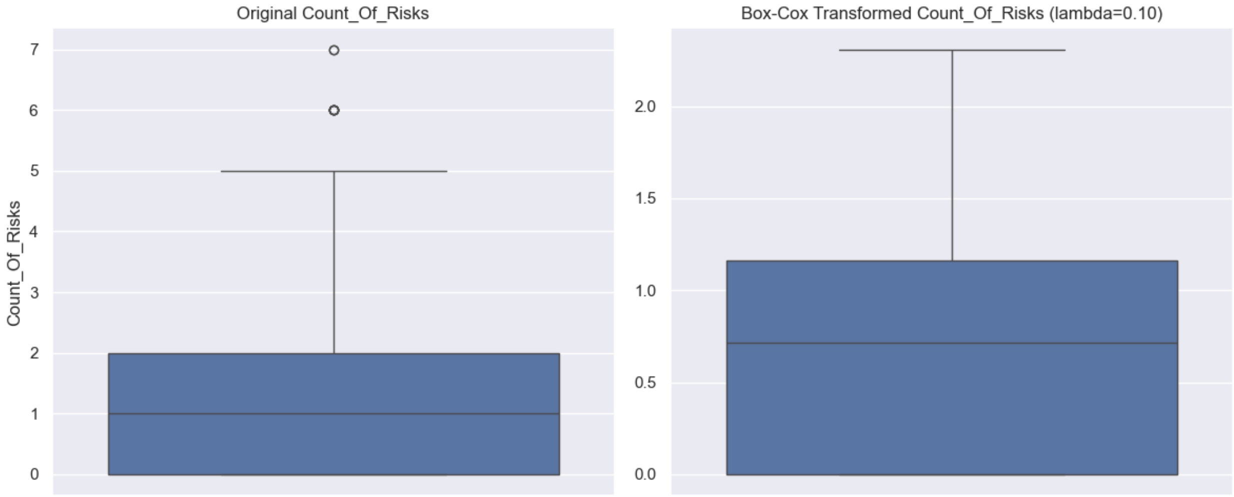
```
Skewness of each numerical variable before transformation:
Dexa_Freq_During_Rx    6.808730
Count_Of_Risks         0.879791
dtype: float64
```

   Skewness after transformation

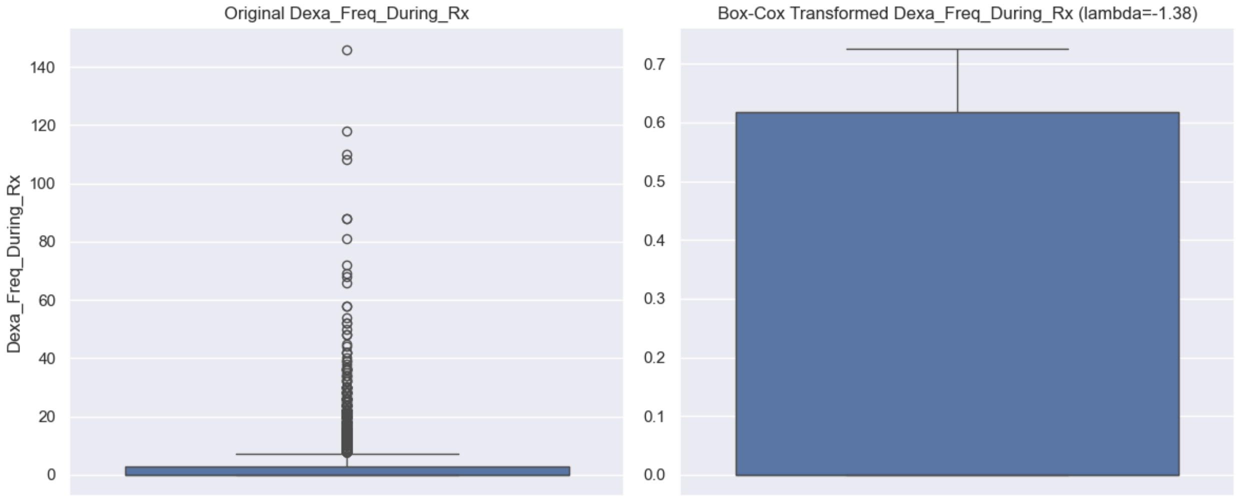
```
Skewness of each numerical variable after transformation:
Dexa_Freq_During_Rx    1.048950
Count_Of_Risks        -0.803041
dtype: float64
```

3. Outlier handling

   Outliers in the numerical variables of the dataset were addressed through Box-Cox transformation. The `Count_of_Risks` feature was transformed with an estimated lambda function of 0.09947523122712457, while the `Dexa_Freq_During_Rx` feature was transformed with an estimated lambda function of -1.3784802096724922



Boxplot of `Count_Of_Risks` before and after Box-Cox transformation



Boxplot of `Dexa_Freq_During_Rx` before and after Box-Cox transformation