Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Drug Persistency

February 2023

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

In the medical field, persistence is defined as continuing a course of therapy prescribed by a medical professional. A pharmaceutical company that can’t keep its customers using its product loses money. It is a well-regarded fact that obtaining new customers is more than double the cost of keeping current ones.

As a society, patients that do not follow long term medical advice drain additional resources from an already overloaded medical system. Low levels of persistency lead to 125,000 deaths per year and cost the U.S. health care system $100 billion annually. (Rubin, 2006).

Approach

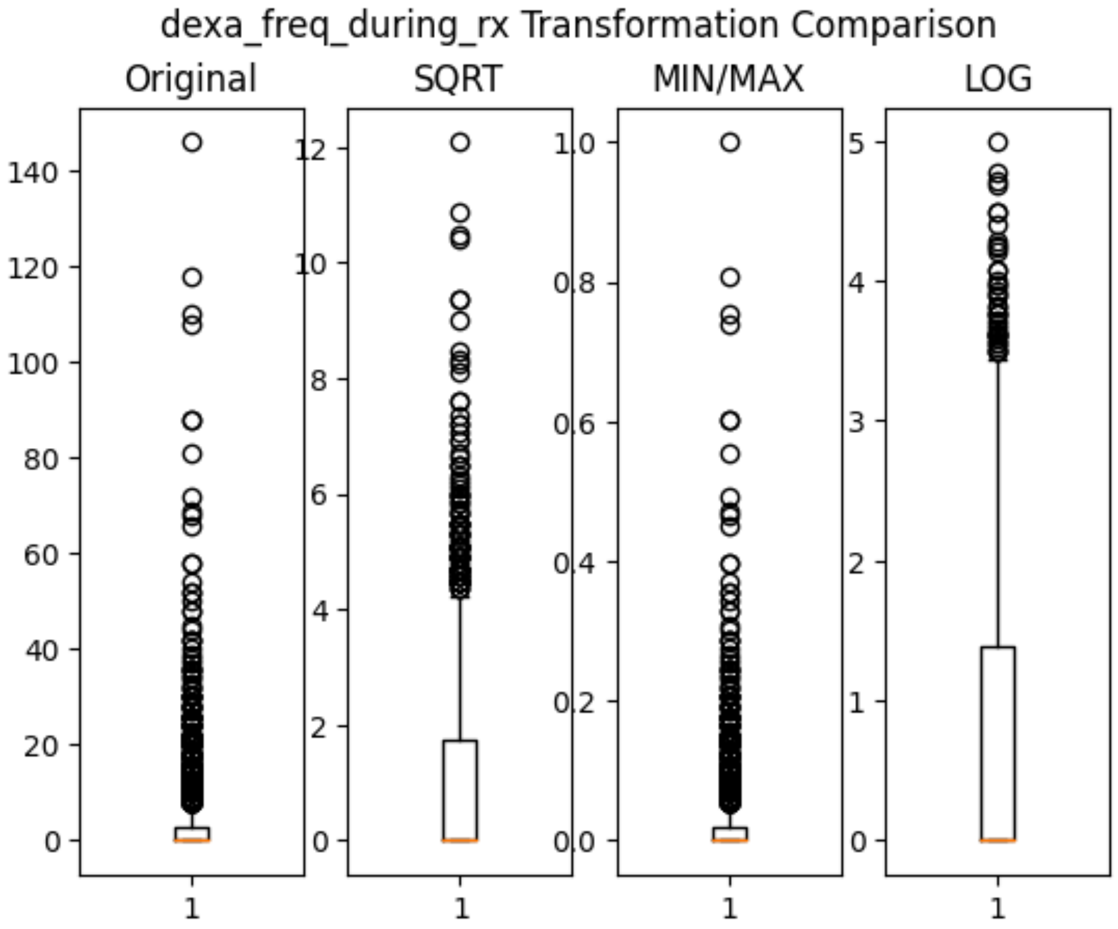
Missing Values: There is lots of missing data in the dataset which required a tiered approach:

- Dropped rows with NaNs if fields that totaled less than 5% of the overall data.
- Dropped fields with missing values that also had high correlation to other features, such as ntm_specialist_flag and ntm_specialist_bucket. Since we already have an ntm_speciality field, we do not need these other aggregated features.
- For the missing ntm_speciality features, I imputed data to match the proportion of specialties in the original dataset to avoid over weighting the mode, general practitioner.
- Finally, there is roughly a third of the data set that is missing values for the same rows in the following features: change_t_score, risk_segment_during_rx, and tscore_bucket_during_rx. This represents too much of the data set to execute an arbitrary imputation. I conducted a two sample T-test to confirm that NaNs for these fields are not equally distributed in both persistent and not persistent groups.

Approach

Multi-collinearity: With nearly 70 features, there are numerous redundant columns in the dataset. Indicators that have both a flag (binary) and measurement field will need to be consolidated to avoid multi-correlation.

Skewness: Although most fields are binary, the several numeric fields both suffer from skewness. Log transformations reduced the effect of outliers. Square root and min-max method did not produce adequate results.



EDA

After EDA and transformation, the dataset now contains 107 features to be used for modeling.

Thank You

Data Glacier

Your Deep Learning Partner