Team member's details: Group Name: Healthcare Team

Emily Yao, yaowemily@gmail.com, United States of America, University of Florida, Data Science

Problem description:

Persistence of drugs, otherwise defined as the duration between initial drug use and its discontinuation, is a big challenge for pharmaceutical companies. However, with the use of machine learning, we can automate its identification by gathering insights on the factors impacting drug persistence.

Data understanding:

What type of data you have got for analysis:
Our data is mostly qualitative with only a few quantitative observations (like Age_Bucket for example). Furthermore, all the quantitative observations appear to be continuous data. There are also instances of nominal data (gender, race, ethnicity). There seems to be no instances of ordinal data as there are no columns displaying any ranking of the desired represented variable.
What are the problems in the data (number of NA values, outliers , skewed etc):
Skewness- The data is skewed left for quantitative variables 'Count_of_Risks' and "Dexa_Freq_During_Rx". "Dexa_Freq_During_Rx" is skewed more by ~8.5x.

Outliers- There exists outliers in "Dexa_Freq_During_Rx" and 'Count_of_Risks' however there exists more outliers in "Dexa_Freq_During_Rx"

Missing Information- There are no NA values to consider in the data but there is missing information to consider from the dataset labeled as "unknown". This is specifically in "Risk_Segment_During_Rx", "Tscore_Bucket_During_Rx", "Chane_T_Score", "Tscore_Bucket_During_Rx"
Mixed Range- Specifically, in "Age_Bucket" there is no proper index (there are ranges instead) which is in improper form for the quantitative variable analysis.

Incomplete Information- There is limited amount of testing that can be conducted on the quantitative variables present because of limited, undetailed information that is not of the appropriate index for quantitative variable analysis. This limitation should be noted when interpreting the data.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

Outliers- If it is an insignificant portion of the data that are outliers, we can safely drop them and presume them to be due to human error. This will be performed on "Count_of_Risks" as it only has 8 outliers out of 3424 possible data entries. If there is a significant number of outliers there are several techniques to handle them such as

replacing them with either the mean or the median value. This will be performed on “Dexa_Freq_During_Rx” as it has 460 outliers.

Skewness- If manipulation of outliers does not solve skewness, then we can transform the quantitative data (log transformations, normalization, square root, cube root, reciprocal).

Mixed Ranges- It is best to divide the “Age_Bucket” into further categories and perform analysis from there (lower, middle, and upper age range).

Missing Information- Similar to outliers, there are two different methods: we can completely delete them from the data set as in remove the entire row however this might not be an appropriate approach if too much data is removed. For categorical data I would replace it with the mode. For quantitative data, we can “forward fill” it or fill it with the previous data input.