Week 9

Healthcare - Persistency of a drug: Group Project

Name: Judith Chepngetich

Email: chepngetich.judith@gmail.com

Group Name: N/A-(Self)

Country: Kenya

Company: Kenya Revenue Authority

Specialization: Data Science

GitHub Repo Link: https://github.com/JudieChep/Assignments.git

Problem Description

ABC pharma Company is in the pharmaceutical business and has a challenge in determining the persistence of a drug. The company would like to automate the process of identification of persistency since there are several factors that need to be considered to determine this. An insight into which factors affect persistency will be useful in automating this process.

Data Cleaning and Transformation

Data Cleaning

There were missing values that were not initally identified by the pandas profiling report tool, because they had been labelled as “Unknown” and “Other/Unkown”. These values were replaced as NaN values so that they could be imputed/ filled accordingly.

After replacement with NaN,the following columns were found to have missing values:

- Race
- Ethnicity
- Region
- Ntm_Speciality
- Risk_Segment_During_Rx
- Tscore_Bucket_During_Rx
- Change_T_Score
- Change_Risk_Segment

The above columns are all categorical, therefore mode was chosen as the most suitable method of filling in the missing values.

Data transformation

After closely analyzing the data, there is need to transform the skewed numerical column, *Dexa_Freq_During_Rx,* perform label encoding for the categorical columns and thereafter do column transformation. Based on the high cardinality nature of the categorical columns in the healthcare dataset, EDA will be performed first to get proper insight and visualization of the data. Once EDA is done, column transformation and transformation of skewed numerical column *Dexa_Freq_During_Rx* will be done.