Group Name: Yusuf Shamsi

Name: Yusuf Shamsi

Email: Yushamsi@outlook.com

Country: United Kingdom

College/Company: None

Specialization: Data Science

Data Understanding of Healthcare_dataset.xlsx:

This excel file contains two sheets:

Sheet 1: Feature Description

- **Bucket**: Categorizes variables into groups such as 'Unique Row Id', 'Target Variable', and 'Demographics'.
- **Variable**: Lists the names of individual variables within each bucket, including 'Patient ID', 'Persistency_Flag', 'Age', 'Race', and 'Region'.
- **Variable Description**: Provides detailed descriptions for each variable, explaining what each variable represents.

Sheet 2: Dataset

- **Ptid**: Patient ID, a unique identifier for each patient.
- **Persistency_Flag**: Indicates whether a patient was persistent (continued their treatment as prescribed) or non-persistent (did not follow the treatment as prescribed).
- **Demographic Information**: Includes gender, race, ethnicity, region, and age bucket of patients.
- **Medical History and Treatment**: Details on the patient's medical history, treatment speciality (e.g., general practitioner), and specific conditions such as risk factors (e.g., family history of osteoporosis, vitamin D insufficiency, etc.).
- **Risk Factors**: Various risk factors associated with osteoporosis treatment outcomes, such as low calcium intake, vitamin D insufficiency, poor health frailty, excessive thinness, hysterectomy/oophorectomy, estrogen deficiency, immobilization, and recurring falls.
- **Count_Of_Risks**: A numerical value indicating the total number of risk factors identified for each patient.

Type of Data in this Dataset:
This data contains information about medical patient's who are have been prescribed a drug, this dataset includes patient demographics, medical history and treatments, risk factors that the patient's have, a numerical measurement of these fore-said risks, and the persistency in taking the drug.

It is unclear from the data, what specific drug this is, and which specific condition this drug treats. This would require more domain knowledge into this related field.

Potential Data Problems (number of NA values, outliers, skewed etc)

Categorical Data

Majority of the variable in this dataset is categorical including the available Age data which is in the form of an Age Bucket. Majority of the variables are also Boolean, However some of the other variables have more categories with 'unknown' data.

- Ntm_Speciality – contains 310 observations that are unknown
- Ntm_Speciality_Bucket – though values are in buckets, the structure of the buckets are unintuitive, with others and unknown mixed with OB/GYN/PCP, so information
- Ethnicity – have 91 unknown values, with the only ethnicity's explicitly stated were Hispanic and Non-Hispanic, data not very useful to identify if a specifically ethnicity correlated with Drug Persistency.
- Race – 97 Other/Unknown Values
- Region – 60 Other/Unknown values
- Risk_Segment_During_Rx – 1497 – Majority of the values are Unknown
- Tscore_Bucket_During_Rx - 1497 – Majority of the values are Unknown
- Change_T_Score – 1497 - Majority of the values are Unknown
- Change_Risk_Segment - 2229 - Majority of the values are Unknown

To address these firstly we will re-structure the Ntm_Speciality_Bucket, if a specialty observed is less than 5% of the total specialties observed in count, with 5% and less of the values observed for each specialty being placed in a synthesised bucket of 'Other'.

Unknown values in the other variables that in count are less than 5% shall be removed, all other values will be kept as removing them would mean a loss of information. Sensitivity Analysis will be done - it will be observed whether these variables will be of use when building a classification model by comparing perfomce with and without the outliers, if performance is better then these values will be dropped and the difference in performance will be observed.

Numerical Data

There are only two variables that are numerical, **Dexa_Freq_During_Rx & Count_Of_Risks.** "Dexa_Freq_During_Rx" is thought to refer to the frequency of DEXA (Dual-Energy X-ray Absorptiometry) scans conducted during the course of the prescription. There are 3424 observations of these (in line with the total number of observations seen in this dataset). Both these variables have a minimum value of 0 – distinguishing the possibility if these values are actual recordings of reality or if they are recorded as 0 if the data are missing – we will assume the former. When visualising the data, both variables were found to have outliers, Dexa_Freq_During_Rx was found to have more outliers with a Median of 0, a Mean of 3, a Standard Deviation of 8 and a max value of 146.

In terms of Count_Of_Risks 2 outliers were detected, with observations have a count of 6 and 7 risks. As for Dexa_Freq_During_Rx 460 were found to be outliers above the upper limit of the IQR.

As domain knowledge is required to understand whether these were errors in data recording or if these rarities are in fact valid, we will be keeping these observations with the basic understanding that 0 would be the majority when it comes to patient's and that of higher values would in fact be rarities but also may be key when it comes to correlation with persistency.