PERSISTENCY OF A DRUG: GATHERING INSIGHTS FOR A PHARMA COMPANY

Name: Shruthi Madgi
Email:Shruthi.madgi05@gmail.com
Country: USA
University: Visvesvaraya Technological University
Specialization: Data Science
Batch no:LISUM17
Reviewer: Data Glacier

Problem Description:

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription, i.e whether a patient will be persistent in completing his/her dose. To solve this problem, ABC pharma company has approached an analytics company to automate the process of identification.

Objective is to build a classification model for drug persistency identification.

The dataset contains 69 medical variables. The data has missing values, outliers and skewed data.

For imputing the missing values and dealing with the outliers, Pandas/Scikit-Learn will be used.

Problems with the data:

By a preliminary analysis, the dataset seems to have:

1.Number of missing values: There are many missing values specifically in the features- 'Race','Ethnicity','Region','Ntm_Speciality','Risk_Segment_During_Rx',Tscore_Bucket_During_Rx','Change_T_Score','Change_Risk_Segment’. Rest of the data did not return any missing values.

2.Number of duplicated values- 0

3.Outliers – There are outliers in the numerical columns: Dex_Freq_During_Rx and Count_Of_Risk.

4.Skewed data – The numerical columns have positively skewed data.

Data Cleaning and Transformation:

1.Only 2.83% in race (97 instances out of 3424) are “Other/Unknown”. The *mode* imputer can be used to fill the NaN values since it works perfectly for data that is MAR and is not more than 6% of the dataset.

2. Mode accounts for 94.48% (3235 instances out of 3424) of the values for Ethnicity. There are only 2.66% of missing values in Ethnicity. Mode is used since the missing values are not high.

3.NTM - Injectable Experience, Risk Factors, Comorbidity, Concomitancy & Frag_Frac_During_Rx (categorical data):Y is replaced with 1 and N with 0.

4. Group rare values in 'Ntm_Speciality' column are marked as 'OTHER' if the percentage of that category is less than 1% .

5.Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score and Change_Risk_Segment missing values: These are the variables that have more than 40% of the missing values. They are eliminated.

6.Replaced‘Unknown/Other’ values with NaN.

7.Tscore_Bucket_Prior_Ntm(categorical data): >-2.5 is replaced with 1 and <=-2.5 with 0.

8.Risk_Segment_Prior_Ntm(categorical data) :VLR_LR is replaced with 1 and HR_VHR with 0.

9.Replaced all the missing values in each column with the mode of that column.

10.Converted the 'Age_Bucket' column to numeric values based on the provided mapping age_bucket.

11.Detected and removed the outliers using IQR.

12.Exported cleaned data to a CSV file.