

Team Name: Data Minders**Team Members:**

Jacob Farrington; jtsfarrington@gmail.com , USA, University of North Texas, Data Science

Xiaoyan Zhang; xiaoyanhouston@gmail.com , USA, University of Texas at Dallas, Data Science

Marina Tsvetkova; marinatsv07@gmail.com , Poland, N/A , Data Science

Pamela S. D. Martey; pmlmartey@gmail.com, USA, Clark Atlanta University, Data Science

Problem description:

ABC is a pharmaceutical company keen on understanding persistency of a drug based on physician prescriptions for patients. To tackle this, ABC has engaged an analytics company to automate the identification of persistency. This analytics firm has tasked Team Data Minders with creating an automated solution to evaluate and enhance the persistency of a drug for ABC.

Data Understanding:

The healthcare dataset comprises 69 columns with data types including 'object' and 'int64'. The target variable, "Persistency_Flag," is binary.

Most of the columns contain binary data labeled "Y" and "N."

Upon analysis, it was determined that only a few columns are numerical, while the rest are string data stored as 'object' type.

Exploratory Data Analysis:

No missing values ("N/A") were found in the dataset. Outliers identified in certain columns were removed, decreasing the dataset size from 3,424 to 2,956 entries.

The dataset primarily consists of binary data, with entries marked as "Y" or "N."

A correlation analysis using the phik matrix exposed multicollinearity among some columns, which will require attention during subsequent model training stages.

What type of data you have got for analysis:

- Patient Information:
 - o Ptid: Patient ID
 - o Gender: Gender of the patient
 - o Race: Race of the patient
 - o Ethnicity: Ethnicity of the patient
 - o Region: Geographic region
- Medical Data:
 - o Persistency_Flag: Indicates whether the patient is persistent or non-persistent
 - o Age Bucket: Age range of the patient
 - o Ntm_Speciality: Specialty of the treating physician
 - o Ntm_Specialist_Flag: Indicates if the specialist is an NTM specialist
 - o Ntm_Speciality_Bucket: Bucketed specialty information
 - o Idn_Indicator: Indicator for IDN involvement
 - o Injectable_Experience_During_Rx: Experience with injectables during treatment

The dataset contains 69 columns, mainly with binary data labeled "Y" or "N," and data types include 'object' and 'int64'. The target variable, "Persistency_Flag," is binary.

What are the problems in the data (number of NA values, outliers , skewed etc):

Outliers:

- Summary statistics such as minimum, maximum, mean, and standard deviation show a potential presence of outliers in our data. Outliers will be further examined visually using box plots and histograms.
- Outliers were found and removed from the dataset, specifically in columns such as 'dexa_freq_during_rx' and 'count_of_risks', reducing the dataset size from 3,424 to 2,956 entries.

Skewness:

- Notable skewness in columns like Dexa_Freq_During_Rx and Count_Of_Risks.

NA Values: There are no NA values in the dataset.

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

We will approach skewness using log or square root transformations, and we will approach outliers via capping in order to maintain data integrity. Outliers in the columns 'dexa_freq_during_rx' and 'count_of_risks' were identified and removed to improve the reliability of the data for accurate analysis and modeling.

Data Handling: With no NA values present, there was no need for strategies related to missing data.

Github Link:

<https://github.com/S0n0f1saac/LISUM33-GroupProject>