Healthcare - Persistency of a drug
Data Science Final Project Report

Group Name: Data Science Warrior
Name: Ugur Selim Ozen
Email: ugur_ozen58@hotmail.com
Country: Turkey
Collage/Company : Yıldız Technical University
Specialization : Data Science
Submission date: 11.08.2021
Github Repo Link:
https://github.com/UGURSELIMOZEN/Data_Glacier_DS_Internship/tree/main/DataScience_Healthcare_Final_Project

1.Problem Description

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

2.Data Understanding

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	1.8 MB
Null/NA Values	0

Also, we find that **almost all features are object dtype** and with the int64 ones, all of them are **categorical features**.

2.a. What type of dataset is this?

Firstly, this is a **classification problem** and generally for the classification problems we can have imbalanced dataset in real-life, as seen from the dataset we can say that this **dataset is imbalanced,** so we need to apply **oversampling or under sampling methods** in model building step.

2.b. Null / NA Values Problem in Dataset

Secondly, by checking null values in all features we can see there is **no null values** but now we need to focus on much more to observe any missing or unknown value assigned for null values.

2.b. Null / Missing Values Handling Approach

Thirdly, we can see that some features have 'Unknown' value in the rows, this can be considered that **null or missing values are filled with 'Unknown' value**. However, there is **no need extra transformation on this Unknown value** because this is classification problem and **Unknown rows represent other categorical values**.