Team Member Details

Group Name: Team Pedram

Name: Pedram Doroudchi

Email: pdoroudchi@gmail.com

Country: USA

College: University of California, San Diego

Specialization: Data Science

Problem Statement

Drug persistency may be defined as the extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen. One of the challenges for all pharmaceutical companies is to understand the persistency of a drug as per the physician's prescription. To solve this problem, ABC Pharma Company would like us to automate this process of identification. With an objective to gather insights on the factors that are impacting the persistency, we will build a classification for the given dataset which contains patient demographics, provider attributes, clinical factors, and disease/treatment factors as well as the target variable indicating whether the patient was persistent or not in their medication usage.

Data Understanding

In our pandas healthcare dataset, we have 3,424 observations of 69 features. Two variables are of the integer type while all other features are of the object type. The vast majority of these object type features are categorical in nature. The pandas duplicated() method was used to find zero duplicate rows in the dataset. The isnull() method was then used to detect zero missing values. Upon further analysis of the integer type features with the pandas describe() method, we see that there are a few patients with rather large amounts of DEXA scans to measure bone density, with the maximum value being 146 in a calendar year. However, this could be due to other clinical factors that warrant so many scans. Thus, there is no strong evidence pointing to outliers or skewness in the dataset. In general, this dataset behaves very nicely.