Week8 deliverables

Group Name: Group Better

Name: Yuyang Xie

Email: yxie222@wisc.edu

Country: United States

College/Company: University of Wisconsin - Madison

Specialization: Data Science

1. **Problem description**

Problem Statement:

One of the challenges 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.

ML Problem:

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

2. **Data understanding**

The data provides individual information of 3424 patients, including their patient ID, their being persistent situations, demographics, provider attributes, clinic factors and disease/treatment factors.

3. **What type of data you have got for analysis**

After basic processing, we now have 66 categorical variables (including target variable), and 2 numerical variables.

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

The data does not include missing values. As for numerical data, if we define the observations with numerical data more than upper fence as outliers, we totally have 468 outliers. And the numerical data are overall left-skewed, close to zero.

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

Refer to this analysis, I decided not to drop any outliers because above-found outliers are based on only two dependent variables to remain the authenticity of data. I have also tried modeling both with these outliers and without these outliers, and the results were similar.

6. Github Repo link

https://github.com/YuyangXie1998/Healthcare-Persistency-of-Drug