Data Science::Healthcare - Persistency of a drug

Project Performer detail (Individual):

Name	Amna Naeem
Email	amnanaeem675@gmail.com
Country	Pakistan
Specialization	Data Science

Problem Description:

One of the greatest problems faced by Pharmaceutical companies these days is of persistency related to drugs. Persistency is related to the behavior of patient towards the medication advised by doctor. This term reflects the duration of time a drug is taken by a patient from the start of treatment to the end of treatment as maintained by doctor. The persistency is important because it can save not only the health of patient but also the money endured on health care system on illness due to discontinuity of drugs.

Data Description:

Data contained information related to demographics, doctor specialty, clinical fact ors and disease/treatment factors mainly related to non-tuberculous mycobacteri a (NTM). The main features are further sub-categorized hence in total there are 3 424 observations related to 69 features including the Patient's id in data. However, 67 variables are non-numerical in nature, consist information regarding different categories and only two variables namely; Dexa_Freq_During_Rx, Count_Of_Risks are of numerical nature. The demographic variables contain information like age, race, region, gender and IDN indicator is also provided. The target variable is Persistency flag consists of two categories persistency and non-p ersistency. The data consist of information related to 3230 females and 194 males where most of the people are of age greater than 65 where less than 55 are 166 a nd greater than 75 are 1439. The people of different ethnicities mainly categorized as Hispanic, non-Hispanic and other are included from all sides of regi on.

Problems and Problem solution within dataset:

The basis data analysis is performed to look for possible problems like missing val ues, outliers and skewness. No missing value is found related to any feature. The bar and box-plot approach is deployed to visualize the presence of possible anom alies and both numerical variables found to have extreme values whereas NTM specialty variable has rare category issue. It consists of 40 different categories to differentiate the specialist's level but there is a large difference in numbers has be en found. Some have value less than 100 while other have values in thousands. In order to solve this problem the categories with minimum numbers will be merged into one category. The outliers from numerical variables are excluded using the interquartile method. This approach is used as it is easy to implement and to com

pare it with other categories. Further, the correlation check is performed using the encoding technique for all variables, the highest correlation is found to exist between three variables. This problem is solved by eradicating the variables in ord er to have features not influenced by other independent variables. Correlation be tween independent variables can make results to fluctuate significantly, unstable model and vary a lot given the little change in the value of data.

Github Repo Link:

https://github.com/AmnaNaaem/DataScience-Project