Week 8 Submission

Group: Single Member Group

Name: Hassan Faheem

Batch: LISUM06

Email: Hassan_hsn9@hotmail.com

Country: UAE

University: Heriot-Watt University

Specialization: Data Science

Problem Description

The problem given here is that the Pharmaceutical Company, ABC is in need to understand the persistency of drug as per the physician prescription. The company ABC has thus approached a company that specializes in Analytics, to get this process of identification to be automated. The company has assigned the case to the relevant member to figure out the solution for the automation of persistency of drug for the company ABC.

Data Understanding

The Healthcare Dataset provided has 69 columns and 3424 number of observations. The target variable is Persistency_Flag. This variable is of Boolean data type with values that are either True or False. After understanding and analyzing the data, its been found that there are few columns that are of numerical data type. Most of the columns are of either Boolean data type or String data type. The column of “Ptid” which refers to Patient ID has no value in terms of model training and thus will be removed the dataset.

Exploratory Data Analysis (EDA)

After performing Exploratory Data analysis on the dataset, the results show that most of the columns are of the Boolean data type and have the values of “Y” and “N”. These values will contribute to the model training in their current type and hence were mapped to the values of 1 and 0. Further analysis shows that no Null values were found in the dataset and so did not require any sort of data handling. The analysis show that a certain feature has some outliers and needed to be handled. To fix this, log transformation was performed on this feature to handle the outliers.