Week 8

Group Name: Ignotus

Team Member Details:

- Corey Hamren, hamhams86@gmail.com, USA, Self-teaching, Data Science
- Motamen MohammedAhmed, Motamen.salih@hotmail.com, UAE, Zayed University, Data Science
- Francis Kim, fkim39@gmail.com, Fkim39@gmail.com, USA, University of Maryland, Data Science
- Inna Soltsman-Groysman, innasol90@gmail.com, USA, Data Science

Problem description:

There are different diseases that affect people around the world. Pharmaceutical companies started to manufacture cures to defeat them. One of the challenges for these companies is to understand the persistency of drug as per the physician prescription and the related factors. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

Data understanding:

There are different diseases that affect people around the world. . Pharmaceutical companies started to manufacture cures to defeat them. One of the challenges for these companies is to understand the persistency of drug as per the physician prescription and the related factors. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

What type of data you have got for analysis:

Healthcare_dataset which includes int, float, binary, and categorical data.

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

- There are no NA values, but there are some labeled as unknown in the ntm_specialty, risk_segment_prior_ntm, tscore_bucket_during_rx, change_tscore, and change_risk_segment columns

- change_t_score, change_risk_segment, tscore_bucket_during_rx, risk_segment_during_rx, have ‘unknown’ as the highest value count while change_t_score has it as the second highest value.

- both the count_of_risks and dexa_freq_during_rx are skewed right, especially the dexa_freq_during_rx column

- the ntm_specialist_flag column has many values with less than 20 counts, but others with hundreds

- most of the binary columns have more than twice as man ‘no’ values than ‘yes’

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

Approaches for Data in terms of cleaning it is:

#Dropping duplicated

data = data.drop_duplicates()

#Dropping Null

data = data.dropna()

#Changing the column names to lower case

data.columns = [x.lower().strip() for x in data.columns]

data.columns

GitHub Repo link - https://github.com/TheHamhams/dg_group_project