Team of One

Roger Pineda
Email: roger.pinedaquijada@gmail.com
Country of Origin: USA
Recent School: Flatiron School
Specialization: Data Science

Problem Description:

Solving the issue of how patients will be with the persistence of the drug given. The dataset given may have outliers, null or unknown values or mis labeled data types.

Data Cleansing and Transformation

- Patient ID column
  - Had every patients Id number and therefore wasn't necessary for any future EDA or modeling so it was removed from the dataset
- Risk_Segement_During_RX
  - Had 43.72% of the data unknown and couldn't be filled in and therefore it was dropped from the dataset
- Change_T_Score
  - Had 43.72% of the data unknown and couldn't be filled in and therefore it was dropped from the dataset
- Tscore_Bucket_During_Rx
  - Had 43.72% of the data unknown and couldn't be filled in and therefore it was dropped from the dataset
- Change_Risk_Segment
  - Had 65.1% of the data unknown and couldn't be filled in and therefore it was dropped from the dataset
- Race
  - Changed all unknown values into the mode value of "Caucasian" since it amounted for 91.94% of the data
- Ethnicity
  - Changed all unknown values into the mode of "Not Hispanic" since it amounted for 94.48% of the data
- Region
  - The mode for the most unknown ethnic group of Not Hispanic found to be for an unknown region therefore it was safe to use the most common region to fill in the few missing unknown regions
- Changed all known column values with string values of 'Y' for "Yes" into a 1 and "N" for "No" into a zero throughout the entire dataset
- Used One Hot Encoder to create a series of dummy columns that either had the value of 0 or 1 with 0 meaning that row did not have said value while 1 saying that it did. This tactic was used on the columns that were deemed to have categorical values.

- Conceited all newly made dummy columns into data frame while removing original columns from dataset
- Then Saw that the column “Dexa_Freq_During_Rx” had numerical values whose range was too wide to be categorical and saw that the distribution of values did not follow a bell curve and therefore use a Min Max Scaler to transform the values and updated the column
- Then changed the values in our target column of “Persistency_Flag” into 0’s and 1’ with 0 meaning “Non-Persistent’ and 1 meaning Persistent

GitHub Repo Link:

https://github.com/RogerPineda13/Healthcare_Persistency_of_a_drug/tree/main/Week9