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 Understanding

- There are 69 features to the dataset that range from the Patient Id to Glucose Records(broken into different columns).
- Target variable is Persistency Flag column
- There are 3424 entries that have to be dealt with.
- Data types are currently that of generic objects, ints and floats.
- Majority of columns are categorical
- The data is also grouped together into 4 different buckets. The buckets being:
    - Demographics
    - Provider Attributes
    - Clinical Factors
    - Disease/Treatment Factor.

List of features with missing value.

- Race
- Region
- Ethnicity
- Ntm_Specialty(Provider Attributes)
- Ntm_Specialty(Clinical Factors)
- Risk_Segment_During_Rx
- Tscore_Bucket_During_Rx
- Change_T_Score
- Change_Risk_Segment

In these cases the missing values aren't that of NaN of typical numeric sort but rather they are inputted as Unknown.

Data Approach

In such cases such as Race and Ethnicity they can be forward filled using a groupby function if one is known and the other isn't and vice versa.
If neither are known they can be filled in using the mode since the mode accounts for roughly 94% of the total dataset.

For Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score and Change_Risk_Segment they all roughly have 40% of the data missing and since these are more concrete measurements they shouldn't be filled in with external computations such as the mean, median, or mode of the data since so much is missing. They will most likely be removed from the dataset.

GitHub Repo Link:

https://github.com/RogerPineda13/Healthcare_Persistency_of_a_drug