PERSISTENCY OF A DRUG: GATHERING INSIGHTS FOR A PHARMA COMPANY

Name: Shruthi Madgi
Email:Shruthi.madgi05@gmail.com
Country: USA
University: Visvesvaraya Technological University
Specialization: Data Science
Batch no:LISUM17
Reviewer: Data Glacier

Problem Description:

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem, ABC pharma company has approached an analytics company to automate the process of identification.

Objective is to build a classification model for drug persistency identification.

The dataset contains 69 medical variables. The data might have NaN values, format errors, outliers, incorrect datatypes, so on and so forth. It is essential to wrangle the dataset before using it.

For performing data wrangling, I will be making use of the dplyr package and tidyr. Tidyr is a package that is used for tidying the data. Data is considered to be tidy when each variable represents a column and each row represents an observation.

For imputing the missing values and for outliers, Pandas/Scikit-Learn will be used.

Understanding the data:

The dataset is in .xlsx format and has 3424 points and feature vectors have 69 variables. It has 2 discrete value features/columns, 1 continuous value feature and 66 categorical value columns/features including the target which is the Persistency flag.

The goal of using this dataset is to predict whether the patient will be persistent in completing his/her treatment.

Apart from individual identificators and the target variables, there are four other buckets:

- Demographics
- Provider attributes
- Clinical factors
- Disease and treatment factors

Problems with the data:

By a preliminary analysis, the dataset seems to have:

1.Number of missing values: There are many missing values specifically in the features-'Race','Ethnicity','Region','Ntm_Speciality','Risk_Segment_During_Rx',Tscore_Bucket_During_Rx','Change_T_Score','Change_Risk_Segment’. Rest of the data did not return any missing values.

2.Number of duplicated values- 0

3.Outliers – There are outliers in the numerical columns: Dex_Freq_During_Rx and Count_Of_Risk.

4.Skewed data – The numerical columns have positively skewed data.

Overcoming the Data Errors:

For missing values in Race, Region and Ethnicity, since the data is categorical in nature, ***mode*** values are used to impute. Since Mode imputation assumes that the missing data looks like the rest of the data, it is easier to implement and manipulate.

The downside of the Mode imputation is the distortion of the data that can occur. However, in the given Healthcare dataset:

1.Only 2.83% in race (97 instances out of 3424) are “Other/Unknown”. The mode imputer can be used to fill the NaN values since it works perfectly for data that is MAR and is not more than 6% of the dataset.

2. The Mode alone accounts for 91.94% of the Race data. This might ensure that original variation distortion will be minimal.

3. 100% of “Other/Unknown” values in the Region variable, the instance Ethnicity falls under “Not Hispanic”.

4. Mode accounts for 94.48% (3235 instances out of 3424) of the values for Ethnicity. There are only 2.66% of missing values in Ethnicity. Mode is used since the missing values are not high.

5. NTM_Speciality variable accounts for less than 10% of the data. Mode is used again since it is categorical in nature and will not have higher missing values.

For NTM - Injectable Experience, Risk Factors, Comorbidity and Concomitancy (group of variables) in handling categorical data,“Y” will be replaced with 1 and “N” with 0.

Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score and Change_Risk_Segment have more than 40% values missing. They will have to be eliminated and cannot be used.

Another approach to impute the missing values with minimal data distortion is using KNNImputer from Scikit-Learn. However, this imputer only works for numerical values, and data that is categorical has to be changed for the imputer to work. KNN, when implemented correctly, will fill the missing values that fit the model properly with minimal data leakage. To use KNN with categorical variables:

1.Subset the object's data types(all) into another container.
2.Change np.NaN into an object data type, say None. Now, the container is made up of only objects data types.
3. Change the entire container into categorical datasets.
4. Encode the data set.
5. Change back the value of encoded None into np.NaN.
6. Use KNN (from fancyimpute) to impute the missing values.
7. Re-map the encoded dataset to its initial names.

Downside of this approach is higher computing power and time required to execute this. Hence, Mode imputer is the perfect fit for the given dataset.

Outliers and Avoiding Skewed data:

Outliers in the given data are present in numerical columns: *Dex_Freq_During_Rx and Count_Of_Risk.*

By using graphical methods such as Boxplot or ScatterPlot or statistical methods as Percentile capping, it can be determined on how outliers affect the data.
In the given data, the values are positively skewed. To balance it out, there are few methods:
1. The values in the outlier are minimal enough so as to not affect the data on the whole. However, it is cumbersome to manually delete the values and replace them; it is time consuming and taxing.
2. To impute the values. By using a regression model to predict the missing value or by using mean or median or mode imputation based on data distribution.
3. Using a Tree based modelling like Random Forest which are less impacted by outliers.
4.Square root or Log transformations work well when the outlier is a dependent variable and can reduce the impact of a single point if the outlier is an independent variable.

Using a Mode imputation on the columns works since the outliers are replaced with value that is most frequent value of the entire column. This approach fits perfectly since Mode imputation was already used for missing values as well.

Link: https://github.com/shru0405/DataGlacierInternship/tree/main/Week_8