Data Science Project

Project: Healthcare - Persistency of a Drug

Week 10 Deliverables

Team Name: Team Healthy Bones

Member: Jeeyeon Shon

E-mail: shon.jeeyeon@gmail.com

College/Company: Data Glacier

Batch Code: LISUM 10

Country: United States of America

Specialization: Data Science

Submission Date: August 9, 2022

Submitted to: Data Glacier

Github Link: https://github.com/shonjeeyeon/DG_Week_10

Table of Contents

Problem Description 1
Data Understanding 1
Link to the Repository 1
Exploratory Data Analysis (EDA) 2
Data Intake Report 4
Summary of Columns and Data Types 5
Problems and Suggested Actions 9

Problem Description

A model will be established and deployed to automate identifying persistency of a certain pharmaceutical product.

Data of patients who take the medication will be used for analysis, and correlation between medication persistency and other factors such as patient demographics, provider attributes, clinical factors, and disease/treatment factors will be investigated. Finally, an optimal model to predict persistency based on above features will be developed.

Data Understanding

The dataset includes 3,424 records of patients on a certain medication. 69 features pertaining the demographics of the patient, attributes of the prescriber, and clinical/disease/treatment factors of the disease progression are present.

The client requested to build a model to predict a patient's drug persistency, so the column 'Persistent' will be the target variable. The prediction will use a classification process since the values of the target column are binary ('Persistent' vs. 'Non-Persistent')

Link to the Repository

https://github.com/shonjeeyeon/DG_Week_10

Exploratory Data Analysis (EDA)

Summary of Actions

- Prior to EDA, ‘risk’, ‘concom’, and ‘comorb’ columns were added.
    - These columns refer to each patient’s sum of risk factors, concomitant treatments, and comorbidities. (‘Y’s and ‘N’s were converted to 1 and 0 prior to the calculation)
    - There are a substantial number of risks, concomitances, and comorbidities listed in the dataset. Moreover, relatively small portion of patients has each of the conditions.
    - Therefore, incorporating the features in larger groups might help reduce computational efforts and improve predictability.
    - The original features used in the calculations will be maintained because some of the conditions may contribute more to the persistency compared to the rest of the condition.

- Count plots, violin plots and box plots were plotted with hue=persistency_flag to compare count, mean, and range of values between persistent and non-persistent populations.

Findings

- Persistent group has higher mean number of comorbidities and concomitances compared to the non-persistent counterpart.
- Patient groups who have had concomitant encounters for below reasons have higher persistency levels:
    - General exam without complaints
    - Immunizations
- Patients who have cancer have higher persistency level.
- Persistent group has higher mean number of DEXA scans compared to the non-persistent group. Persistent group also has higher proportion of patient who has had at least one DEXA scan compared to non-persistent group.
- Prescribers with certain specializations (e.g. Oncology and Endocrinology) have higher proportion of persistent patients compared to the others.

Model Recommendations

- This is a classification project, so using classifiers such as Linear Regression Classifier, Naive Bayes, K-Neighbors, Random Forest Classifier, Support Vector Machines, or XGBoost is recommended.
- To save computational efforts, starting from simple Linear Regression then trying kernel or ensemble models is recommended.
- The dataset has a substantial number of features, so using PCA or RFE to choose most important features is recommended.

Data Intake Report

Name: Healthcare – Persistency of a Drug
Report date: August 9, 2022
Internship Batch: LISUM 10
Version:<1.0>
Data intake by: Jeeyeon Shon
Data intake reviewer:
Data storage location:
https://github.com/shonjeeyeon/DG_Week_8/blob/main/Healthcare_dataset.csv
(Original xslx file:
https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view)

Tabular data details:
Healthcare_dataset.csv

Total number of observations	3,424
Total number of files	1
Total number of features	69
Base format of the file	.csv
Size of the data	892 KB

Proposed Approach:

- Ptid can be used to identify and remove duplicate observations
- The dataset has been deidentified already
- No missing values, however there are practical missing values such as ‘unknown’. These values should be imputed appropriately
- Most of the features are categorical; will need encoding to enable ML

Summary of Columns and Data Types

Bucket	Variable	index #	Dtype	Notes
Target	Persistency	1	Object	Non-Persistent: 62.35% Persistent: 37.65% (→ Imbalanced data)
Unique Row ID	Patient ID	0		[redacted]
Demographics	Gender	2		
	Race	3		NaN='Other/Unknown' (2.85%) Mode='Caucasian' (91.94%)
	Ethnicity	4		NaN='Unknown' (2.66%) Mode='Non-Hispanic' (94.48%)
	Region	5		NaN='Other/Unknown' (1.75%) Mode='Midwest' (40.39%)
	Age_Bucket	6		[redacted]
Prescriber Attributes	Ntm_Speciality	7		NaN='Unknown' (9.05%) Mode='General Practitioner' (44.83%)
	Ntm_Specialist_Flag	8		[redacted]

<table>
<tr><td></td><td>Ntm_Speciality_Bucket</td><td>9</td><td rowspan="3"></td><td rowspan="3">[redacted]</td></tr>
<tr><td rowspan="9">Clinical Factors</td><td>Gluco_Record_Prior_Ntm</td><td>10</td></tr>
<tr><td>Gluco_Record_During_Rx</td><td>11</td></tr>
<tr><td>Dexa_Freq_During_Rx</td><td>12</td><td>int64</td><td>Outlier issues
The data is skewed (6.81)
<table><tr><td>Count</td><td>3,424</td></tr><tr><td>Mean</td><td>3.02</td></tr><tr><td>Std</td><td>8.14</td></tr><tr><td>Min</td><td>0.00</td></tr><tr><td>25%</td><td>0.00</td></tr><tr><td>50%</td><td>0.00</td></tr><tr><td>75%</td><td>3.00</td></tr><tr><td>Max</td><td>146.00</td></tr></table></td></tr>
<tr><td>Dexa_During_Rx</td><td>13</td><td rowspan="6">Object</td><td rowspan="5">[redacted]</td></tr>
<tr><td>Frag_Frac_Prior_Ntm</td><td>14</td></tr>
<tr><td>Frag_Frac_During_Rx</td><td>15</td></tr>
<tr><td>Risk_Segment_Prior_Ntm</td><td>16</td></tr>
<tr><td>Tscore_Bucket_Prior_Ntm</td><td>17</td></tr>
<tr><td>Risk_Segment_During_Rx</td><td>18</td><td>NaN='Unknown' (43.72%)
The other two categories have very few differences in percentages</td></tr>
</table>

				HR_VHR 28.18% VLR_LR 28.10%
	Tscore_Bucket_During_Rx	19		NaN='Unknown' **(43.72%)** **The other two categories have very few differences in percentages** <=-2.5 29.70% >-2.5 26.56%
	Change_T_Score	20		NaN='Unknown' **(43.72%)** Mode='No Change' (48.48%)
	Change_Risk_Segment	21		NaN='Unknown' **(65.01%)** Mode='No Change' (30.72%)
Disease/ Treatment Factors	Adherent_Flag	22		[redacted]
	Idn_Indicator	23		
	Injectable_Experience_During_Rx	24		
	Comorbidities columns (Column names start with 'Comorb_')	25-38		
	Concomitant drugs use columns (Column names start with 'Concom_')	39-48		
	Risk factors columns	49-67		

<table>
<tr><td></td><td>Count_of_Risks</td><td>68</td><td>Dtype: int64</td><td>Outlier issues
The data is skewed (0.88)
<table><tr><td>Count</td><td>3,424</td></tr><tr><td>Mean</td><td>1.24</td></tr><tr><td>Std</td><td>1.09</td></tr><tr><td>Min</td><td>0.00</td></tr><tr><td>25%</td><td>0.00</td></tr><tr><td>50%</td><td>1.00</td></tr><tr><td>75%</td><td>2.00</td></tr><tr><td>Max</td><td>7.00</td></tr></table></td></tr>
</table>

Problems and Suggested Actions

Problem	Column	Details	Actions Taken	Rationale
Missing Data	Race	NaN='Other/Unknown' (2.85%) Mode='Caucasian' (91.94%)	Impute mode	The Modes are high in proportion while the NaNs are relatively small in proportion.
	Ethnicity	NaN='Unknown' (2.66%) Mode='Non-Hispanic' (94.48%)		
	Region	NaN='Other/Unknown' (1.75%) Mode='Midwest' (40.39%)		
	Ntm_Speciality	NaN='Unknown' (9.05%) Mode='General Practitioner' (44.83%)	Keep 'Unknown' as a separate value	The value 'Unknown' is relatively high in proportion (9.05%), while the category has many values with smaller proportions (as small as <1%). Therefore, it will be

				prudent to leave the unknown as it is.
>40% Missing Data	Risk_Segment_During_Rx	NaN='Unknown' **(43.72%)**	Delete columns	The columns have very high proportion of 'Unknown'. Imputation may cause serious distortion of the data.
	Tscore_Bucket_During_Rx	NaN='Unknown' **(43.72%)**		
	Change_T_Score	NaN='Unknown' **(43.72%)**		
	Change_Risk_Segment	NaN='Unknown' **(65.01%)**		
Outliers/ Skews	Dexa_Freq_During_Rx	• Outlier issues • The data is **skewed** (6.81)	Remove outliers, and try skewness reduction strategies as needed	Outliers were removed using quantiles and it reduced the skewness. Then, square root was used to additionally reduce skew. Skews after above steps are: 1.28 for Dexa_Freq_During_Rx, and 0.38 for Count_of_Risks
	Count_of_Risks	• Outlier issues • The data is **skewed** (0.88)		

Basic Cleaning	All columns	Will need to remove upper cases, special characters, or spaces	Use df.replace() to clean the column names	df.replace() used to remove upper cases, special characters, and spaces
Typo in Value	Ntm_Speciality	‘OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY’	Use df.replace() to correct the value	Replaced with 'OBSTETRICS AND GYNECOLOGY'
Imbalanced Target Data	Persistency	Non-Persistent: 62.35% Persistent: 37.65%	Use SMOTE	SMOTE will be implemented during the process of model development
Encoding	Applies to every categorical column	Categorical values are written in alphabet, which ML cannot process	Label or one hot encoding	Values will be encoded after EDA