# Data Science Project

**Project:** Healthcare - Persistency of a Drug

**Week 8** 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:** July 25, 2022

**Submitted to:** Data Glacier

**Github Link:** https://github.com/shonjeeyeon/DG_Week_8

# Table of Contents

# 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 selected and developed.

# Data Understanding

The dataset includes 3,424 records of patient who is on a certain medication. 69 features pertaining the demographics of the patient, attributes of the prescriber, and indicators/risk 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 variables. The prediction will use a classification process since the values of the target column are binary ('Persistent' vs 'Non-Persistent')

## Summary of Columns and Data Types

| Bucket | Variable | index # | Dtype | Notes |
|---|---|---|---|---|
| Target | Persistency | 1 | Object | Non-Persistent: 62.35%<br>Persistent: 37.65%<br>(→ Imbalanced data) |
| Unique Row ID | Patient ID | 0 | | . |
| Demographics | Gender | 2 | | . |
| | Race | 3 | | NaN='Other/Unknown' (2.85%)<br>Mode='Caucasian' (91.94%) |
| | Ethnicity | 4 | | NaN='Unknown' (2.66%)<br>Mode='Non-Hispanic' (94.48%) |
| | Region | 5 | | NaN='Other/Unknown' (1.75%)<br>Mode='Midwest' (40.39%) |
| | Age_Bucket | 6 | | . |
| | Ntm_Speciality | 7 | | NaN='Unknown' (9.05%) |

<table>
<tr><td rowspan="3">Prescriber Attributes</td><td></td><td></td><td rowspan="5"></td><td>Mode=’General Practitioner’ (44.83%)</td></tr>
<tr><td>Ntm_Specialist_Flag</td><td>8</td><td>.</td></tr>
<tr><td>Ntm_Speciality_Bucket</td><td>9</td><td>.</td></tr>
<tr><td rowspan="7">Clinical Factors</td><td>Gluco_Record_Prior_Ntm</td><td>10</td><td>.</td></tr>
<tr><td>Gluco_Record_During_Rx</td><td>11</td><td>.</td></tr>
<tr><td>Dexa_Freq_During_Rx</td><td>12</td><td>int64</td><td>• Outlier issues<br>• The data is skewed (6.81)<br><br><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="5">Object</td><td>.</td></tr>
<tr><td>Frag_Frac_Prior_Ntm</td><td>14</td><td>.</td></tr>
<tr><td>Frag_Frac_During_Rx</td><td>15</td><td>.</td></tr>
<tr><td>Risk_Segment_Prior_Ntm</td><td>16</td><td>.</td></tr>
<tr><td>Tscore_Bucket_Prior_Ntm</td><td>17</td><td>NaN='Unknown' (43.72%)</td></tr>
</table>

| | | | | |
|---|---|---|---|---|
| | | | | **The other two categories have very few differences in percentages**<br>HR_VHR 28.18%<br>VLR_LR 28.10% |
| | Risk_Segment_During_Rx | 18 | | NaN='Unknown' **(43.72%)**<br>**The other two categories have very few differences in percentages**<br><=-2.5 29.70%<br>>-2.5 26.56% |
| | Tscore_Bucket_During_Rx | 19 | | NaN='Unknown' **(43.72%)**<br>Mode=’No Change’ (48.48%) |
| | Change_T_Score | 20 | | NaN='Unknown' **(65.01%)**<br>Mode=’No Change’ (30.72%) |
| | Change_Risk_Segment | 21 | | . |
| Disease/ Treatment Factors | Adherent_Flag | 22 | | . |
| | Idn_Indicator | 23 | | . |
| | Injectable_Experience_During_Rx | 24 | | . |
| | Comorbidities columns (Column names start with ‘Comorb_’) | 25-38 | | . |

<table>
<tr><td rowspan="3"></td><td>Concomitant drugs use columns (Column names start with ‘Concom_’)</td><td>39-48</td><td rowspan="2"></td><td>.</td></tr>
<tr><td>Risk factors columns</td><td>49-67</td><td>.</td></tr>
<tr><td>Count_of_Risks</td><td>68</td><td>Dtype: int64</td><td>
<ul>
<li>Outlier issues</li>
<li>The data is skewed (0.88)</li>
</ul>
<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 Type | Columns | Index | Details | Categorical / Quantifiable | Suggested Actions |
|---|---|---|---|---|---|
| Missing Data | Race | 3 | NaN='Other/Unknown' (2.85%) Mode='Caucasian' (91.94%) | Categorical | Impute with mode |
| | Ethnicity | 4 | NaN='Unknown' (2.66%) Mode='Non-Hispanic' (94.48%) | | |
| | Ntm_Speciality | 7 | NaN='Unknown' (9.05%) Mode='General Practitioner' (44.83%) | | |
| >40% Missing Data | Tscore_Bucket_Prior_Ntm | 17 | NaN='Unknown' **(43.72%)** HR_VHR 28.18% VLR_LR 28.10% | Categorical | Delete the columns because the proportions |

<table>
<tr><td></td><td>Risk_Segment_During_Rx</td><td>18</td><td>NaN='Unknown'<br>(43.72%)<br>The other two categories have very few differences in percentages<table><tr><td><=-2.5</td><td>29.70%</td></tr><tr><td>>-2.5</td><td>26.56%</td></tr></table></td><td></td><td>of missing data are too large to impute without contributing to potential biases</td></tr>
<tr><td></td><td>Tscore_Bucket_During_Rx</td><td>19</td><td>NaN='Unknown'<br>(43.72%)<br>Mode=’No Change’ (48.48%)</td><td></td><td></td></tr>
<tr><td></td><td>Change_T_Score</td><td>20</td><td>NaN='Unknown'<br>(65.01%)<br>Mode=’No Change’ (30.72%)</td><td></td><td></td></tr>
<tr><td>Outliers/ Skews</td><td>Dexa_Freq_During_Rx</td><td>12</td><td><ul><li>Outlier issues</li><li>The data is skewed (6.81)</li></ul><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></table></td><td>Quantifiable</td><td>Use Tukey’s rule to remove outliers<br><br>Also implement other</td></tr>
</table>

<table>
<tr><td></td><td></td><td></td><td><table><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><td></td><td>methods to remove skews, such as log transformation or box-cox method</td></tr>
<tr><td></td><td>Count_of_Risks</td><td>68</td><td><ul><li>Outlier issues</li><li>The data is skewed (0.88)</li></ul><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><td></td><td></td></tr>
<tr><td>Imbalanced Target Data</td><td>Persistency</td><td>1</td><td>Non-Persistent: 62.35%<br>Persistent: 37.65%</td><td>Categorical</td><td>Consider SMOTE</td></tr>
<tr><td>Encoding</td><td>Applies to every categorical column</td><td></td><td>Categorical values are written in alphabet, which ML cannot process</td><td>Categorical</td><td>Label, dummy, or one-hot encoding</td></tr>
</table>

| | | | | | |
|---|---|---|---|---|---|
| Basic Cleaning | All columns | | Will need to remove upper cases, special characters, or spaces | Categorical/ Quantifiable | Switch the col names and values to all lower cases<br>Remove special characters and spaces |

## Link to the Repository

https://github.com/shonjeeyeon/DG_Week_8