Data Science Project

Project: Healthcare - Persistency of a Drug

Week 13 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 30, 2022

Submitted to: Data Glacier

Github Link: https://github.com/shonjeeyeon/DG_Week_13

Table of Contents

Problem Description

Medication persistence refers to completing the medication treatment using the duration set by the prescriber. (Cramer et al., 2008). Therefore, persistence is important in patients' positive outcomes as well as in pharmaceutical industries' profits.

In this project, a model was established and deployed to automate identifying persistency of a certain pharmaceutical product using a dataset provided by the client.

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 'Persistency_flag' is the target variable. The prediction used a classification process since the values of the target column has binary values.

Executive Summary

Deployed a Logistic Regression Model and an Heroku app to classify persistent vs. non-persistent patient using 7 predictors recommended by Recurrent Feature Elimination (RFE). The model has AUC of 0.8049 and accuracy of 0.8189.

Link to the Repository & App

- Repository: https://github.com/shonjeeyeon/DG_Week_13
- App: https://persistency.herokuapp.com/

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

Data Cleaning and Preprocessing

Basic Cleaning

Upper cases and special characters in the names of columns were removed.

Handling Missing or Minor Values

- In ‘Ntm_speciality’ column, values with <1% counts were integrated to ‘Others’ category.
- Also in ‘NTM_speciality’ column, missing values were integrated to ‘Unknown’ category.
- Other columns with missing values <40% were imputed using modes.
- Columns with missing values >40% were deleted.

Encoding of Categorical Features

The categorical values were dummy-encoded for ML processing.

Cleaning of Numeric Values

Skews and outliers in numeric features were addressed.

Addition of Extra Features

- ‘concom’ was added to calculate the total number of concomitant therapies of each patient.
- ‘risk’ was added to calculate the total number of risk factors of each patient.
- ‘comorb’ was added to calculate the total number of comorbidities of each patient.

Deletion of Redundant Features

- ‘dexa_during_rx’ and ‘count_of_risks’ were deleted because ‘dexa_frequency_during_rx’ and ‘risks’ had same information with more details.

Splitting to Target/Predictors, then to Train/Test sets

- The target of the dataset was 'persistency_flag', with other features being predictors.
- 25% of the total dataset was used for testing.

Handling Imbalances

SMOTE was used to oversample the training data as the dataset was unbalanced.

Findings from Exploratory Data Analysis (EDA)

Findings

- Certain regions and prescriber specialties were associated with higher persistency. (ANOVA, p values <0.05)
- Persistent group has higher mean number of comorbidities and concomitances compared to the non-persistent counterpart. (ANOVA, p values <0.05)
- All of the comorbidities and concomitant therapies in the dataset, as well as select risk factors, were associated with difference in persistency (Chi-square, p values<0.05).
- 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. (ANOVA, p values <0.05)

Model Recommendations

- This is a classification project, so using classifiers such as Logistic Regression Classifier, Naive Bayes, K-Neighbors, Random Forest Classifier, Support Vector Machines, or XGBoost is recommended.
- Starting from Logistic Regression, Decision Tree, or Naïve Bayes is recommended to save computational effort and time.

- The dataset has a substantial number of features, so using PCA or RFE to choose most important features is recommended.

Important Feature Selection

The dataset had more than 60 features, so in order to develop an application, choosing the most important features was necessary. Recurrent Feature Elimination (RFE) with Random Forest model was used to select 7 most important features, and below is the result:

- 'dexa_freq_during_rx'
- 'comorb_encounter_for_screening_for_malignant_neoplasms'
- 'comorb_encounter_for_immunization'
- 'comorb_long_term_current_drug_therapy'
- 'comorb'
- 'concom'
- 'risk'

These features all have p values <0.05.

Model Selection

Four models were tested for AUC, accuracy, and recall. GridSearchCV was used for optimizing each model's parameters.

Model	AUC	Accuracy	Recall
Logistic Regression	0.8049	0.8189	0.7484
Random Forest	0.7667	0.7850	0.6925
XGBoost	0.7877	0.7944	0.7609
Multi-Layer Perceptron	0.7959	0.8061	0.7547

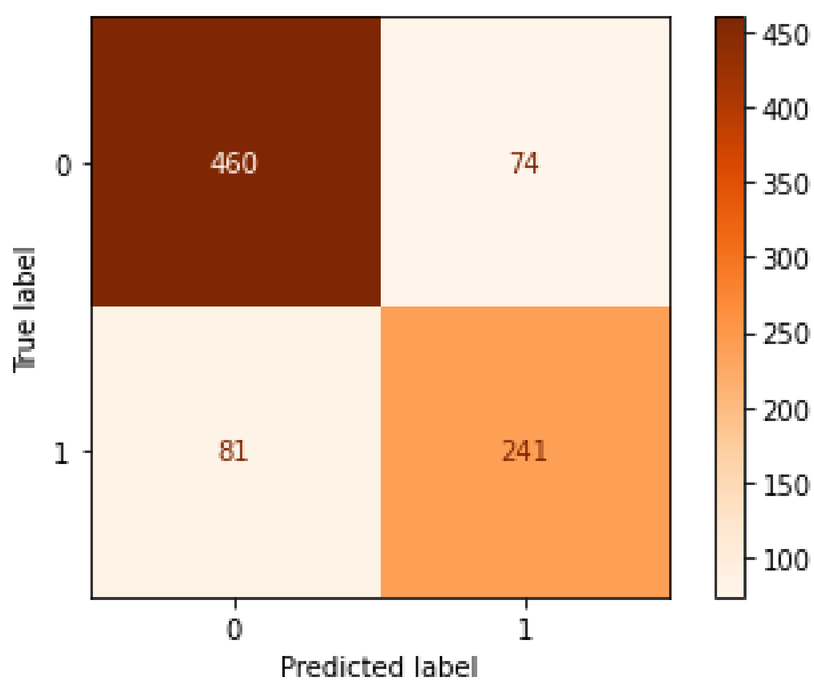
Based on AUC and accuracy scores, Logistic Regression was selected for model and app development.

Model Evaluation

AUC and Accuracy Scores

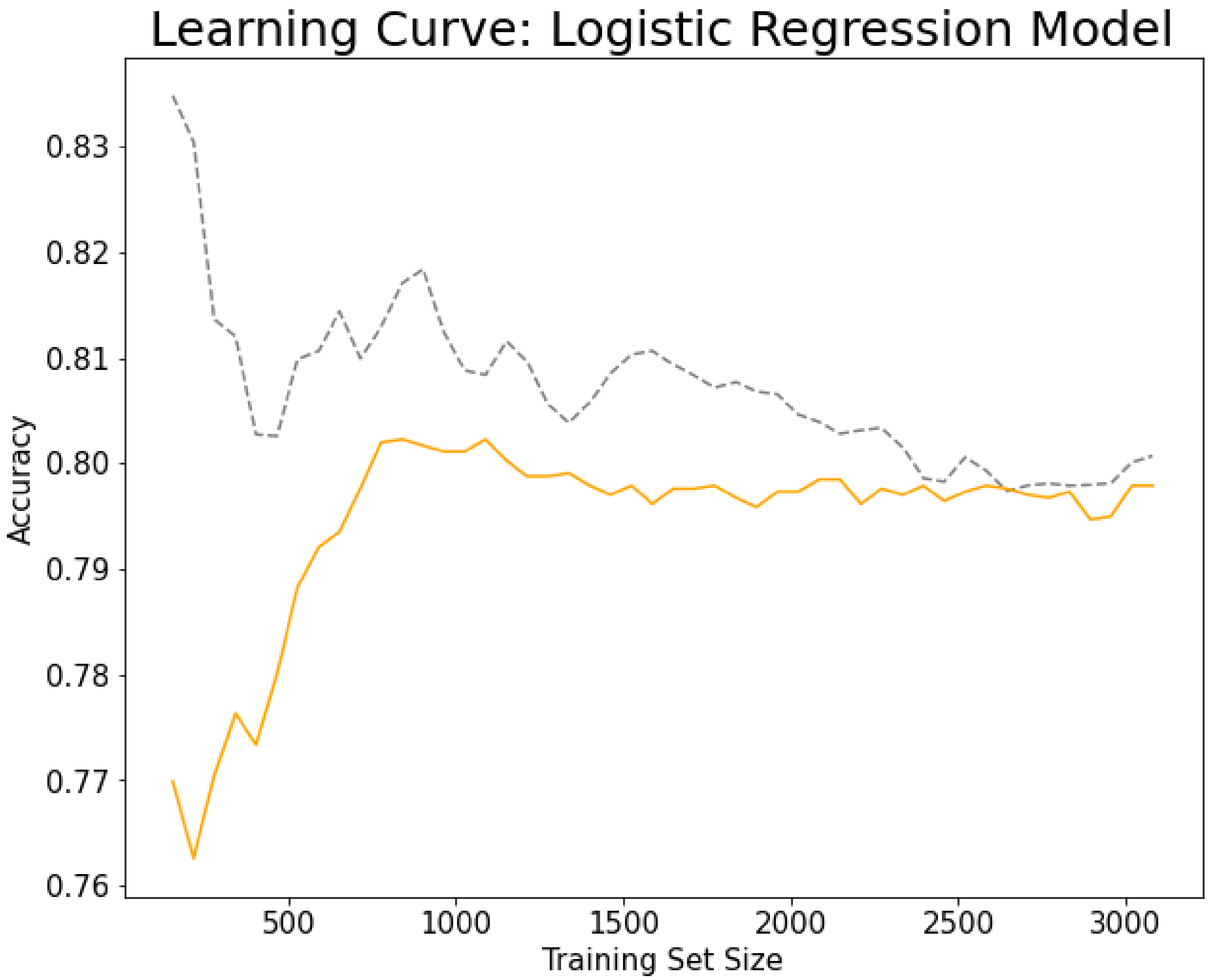
The Logistic Regression model had AUC of 0.8049 and accuracy of 0.8189 on the test set.

Confusion Matrix



The model has True Positive Rate = 0.7484 and True Positive Rate = 0.8337 on the Test Set.

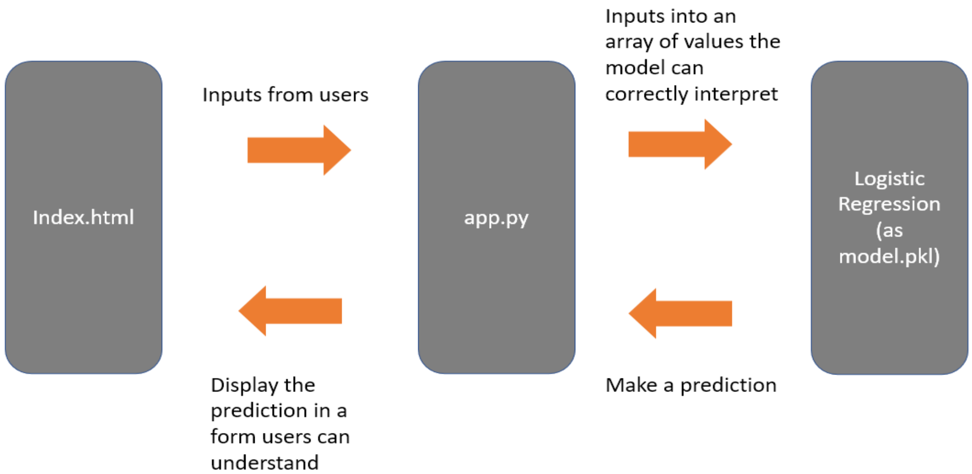
Learning Curves



The differences in accuracy scores decrease as size of the training set increases, then minimizes when the training set size exceeds 2,250. The training set used for the model development had 2,568 records (75% of the original dataset).

Application Development

Overview of the Application



The app displays index.html in the /templates folder when users access.

Via the form on the index.html, the page receives inputs from the users.

(Memudualimatou (2020)'s GitHub was consulted to establish categorical forms; full hyperlink available in References section)

The inputs are processed with app.py so it can be interpreted and further processed with the picked model.

- The categorical values are converted to 0 and 1.
- Frequency of DEXA and number of comorbidities will be square rooted as those in the training set had been.

The result will be displayed on the index.html page.

Links to the Application

- The application can be accessed by clicking the link: https://persistency.herokuapp.com/
- The application is also accessible on 'Environment' menu on the right side of the GitHub repo: https://github.com/shonjeeyeon/DG_Week_13

Conclusion

Logistic Regression model was able to predict persistence of a drug with approximate AUC of 80% and approximate accuracy of 82%. Seven select features were used for prediction.

References

Cramer, J.A., Roy, A., Burrell, A., Fairchild, C. J., Fuldeore, M.J., Ollendorf, D.A., Wong, P.K. (2008). Medication compliance and persistence: terminology and definitions. Value Health. 11(1), 44-47

Memudualimatou (2020), index.html. GitHub. Retrieved from https://github.com/memudualimatou/INSURANCE-CHARGES-WEB-APPLICATION/blob/main/templates/index.html on Aug 20, 2022.

[Appendix 1] 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		
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		
Prescriber Attributes	Ntm_Speciality	7		NaN='Unknown' (9.05%) Mode=’General Practitioner’ (44.83%)
	Ntm_Specialist_Flag	8		

<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>

[Appendix 2] Problems and Actions Taken Prior to EDA

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 was 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 were dummy encoded using EDA