**PA TASK 3: Association Rules & Lift Analysis**

**D212 – Data Mining II**

Michael Lawson

Western Governors University

D212 – Data Mining II

PA Task 3: Association Rules & Lift Analysis

Kesselly Kamara

09/17/2022

Michael C. Lawson | Student ID: #001479557 | Program: MS Data Analytics (5/1/21) |

Program Mentor: Emil Stoica | 859-265-1888 (EST) | [mlaws77@wgu.edu](mailto:mlaws77@wgu.edu)

**Part I:**

A.

1.  **One** question.

Do associations exist between prescriptions in the dataset that can be discovered using Market Basket Analysis (MBA).

2.  **One** Goal

The goal of using Market Basket Analysis on the ‘medical\_market\_basket.csv’ data set is to find links and associations between prescription drugs.

**Part II:**

B.

1.

Market Basket Analysis (MBA) defines the strength of the relationship between pairs of items bought together. In the data set used in this analysis, MBA will define the strength of relationships of prescription medications that are administered to patients. The analysis will seek to find association between prescribed medications and which additional medications are most likely to be prescribed to those patients. The expected outcome is finding the associations in the data that reveals if a patient receives a prescription medication, which additional medications is that patient likely to receive.

2.

The ‘medical\_\_market\_basket.csv’ data set contains a list of transactions administering prescription medications, and the prescriptions that patients had also taken either in the past or currently; for instance, a patient in the data set that took ‘abilify’ had also at some point taken ‘atorvastatin’, ‘folic acid’, ‘naproxen’, and ‘losartan’. Market based analysis will search for other rows containing each of those drugs and try to find similarities that exist that reveal how likely two medications are both going to be prescribed.

3.

One assumption of market basket analysis is that the association is happening naturally because of the associated items, and not because of other factors. According to the publication Towards Data Science, “… although market basket analysis may help you spot a trend, once you act on it, it’s difficult to assess the validity of the correlation” (Holst, 2021).

**Part III:**

C.

1.  Transform the data:

# Import libraries and packages

**import** pandas **as** pd

**import** numpy **as** np

**import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

**from** mlxtend.preprocessing **import** TransactionEncoder

**from** mlxtend.frequent\_patterns **import** apriori

**from** mlxtend.frequent\_patterns **import** association\_rules

**import** warnings

warnings**.**filterwarnings('ignore')

*# Import the dataset*

df **=** pd**.**read\_csv(r"C:\Users\mlaws\OneDrive - Western Governors University\Documents\WGU\D212\medical\_market\_basket.csv")

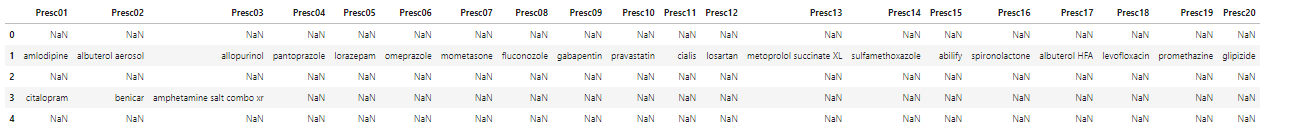
*# Get info on data*

df**.**info()

Text

Description automatically generated with low confidence

df**.**head()



*# Data dimensions*

df**.**shape



*# View column data types*

print(df**.**dtypes)

Table

Description automatically generated

*# Create list of lists for encoding*

trans**=**[]

**for** i **in** range (0,7501):

trans**.**append([str(df**.**values[i,j]) **for** j **in** range(0, 20)])

*# Prepare data for Apriori function*

TE **=** TransactionEncoder()

array **=** TE**.**fit(trans)**.**transform(trans)

*# Convert data set back to dataframe*

cleaned\_df **=** pd**.**DataFrame(array, columns **=** TE**.**columns\_)

cleaned\_df

A picture containing text, computer, indoor, white

Description automatically generated

**for** col **in** cleaned\_df**.**columns:

print(col)

# Drop empty cells

df\_cleaned **=** cleaned\_df**.**drop(['nan'], axis**=**1)

**for** col **in** df\_cleaned**.**columns:

print(col)

Text

Description automatically generated with low confidenceText

Description automatically generated

# Check the dimensions of the cleaned data

df\_cleaned**.**shape



*# View column data types*

print(df**.**dtypes)

Table, calendar

Description automatically generated

2.   Apriori algorithm:

*# Run Apriori to mine*

a\_rules **=** apriori(df\_cleaned, min\_support **=** 0.005, use\_colnames **=** **True**)

*# Review head*

a\_rules**.**head(3)

*Graphical user interface, text

Description automatically generated*

*# Set metric of associations rule table*

ass\_r **=** association\_rules(a\_rules, metric **=** 'lift', min\_threshold **=** 1)

ass\_r**.**head(3)

Graphical user interface, application

Description automatically generated

# print the # of transactions that include ‘abilify’

df\_cleaned['abilify']**.**sum()



*# Search with new metrics to find associations with the best metrics*

# Generate top 3 rules

ass\_r[ (ass\_r['lift'] >= 6.3) & ass\_r['confidence'] > 0.2][:3] Graphical user interface, application

Description automatically generated

# print the # of transactions that include ‘atorvastatin’, ‘metformin’ & #‘abilify’.

print('Atorvastatin:', df\_cleaned['atorvastatin']**.**sum())

print('Metformin:', df\_cleaned['metformin']**.**sum())

print('Abilify:', df\_cleaned['abilify']**.**sum())



3.  Provide values for the support, lift, and confidence of the association rules table.

The minimum value for support was set at 0.005. The top three rules generated for the table all have a support value of 0.005466.

The value for lift was set to find any rules with a lift above above 6.3. For the top three rules, the values for lift are 7.729491, 7.729491, & 6.341052.

The metric for confidence was set at 0.2 for searching the association rules table for the top three rules. The top confidence level for the top three rules is 0.422680, 0.209184, & 0.201970.

4.  Top **three** rules.

# Generate top 3 rules

ass\_r[ (ass\_r['lift'] >= 6.3) & ass\_r['confidence'] > 0.2][:3]

Graphical user interface

Description automatically generated with low confidence

# Print how many times each prescription appears in the table

print('Atorvastatin:', df\_cleaned['atorvastatin'].sum())

print('Metformin:', df\_cleaned['metformin'].sum())

print('Abilify:', df\_cleaned['abilify'].sum())

Text

Description automatically generated

**Part IV:**

D.

1.  Summarize the significance of support, lift, and confidence from the results of the analysis.

Support is how popular an item (antecedent) is in a dataset alongside another item (consequent). For the table, the minimum value for support was set at 0.005. The top three rules generated for the table all have a support value of 0.005466. Support is calculated by dividing the number of instances a prescription appeared in the data by the total number of transactions. In the provided data set, the strongest rules generated appear in transactions together 0.55% out of all transactions.

Imagine no relationship exists between two prescriptions, the lift would be the measurement of how much the two prescriptions appearing together exceeded the expectation that they would not appear together. Lift greater than 1 indicates that when the antecedent is prescribed, the chance of the consequent being prescribed increases. If the lift is below 1, the opposite is true, and the chance of the consequent also being prescribed reduces because of the antecedent. The top three rules were found by setting the lift metrics to find any rules with a lift value above 6.3. For the top three rules, the values for lift are 7.729491, 7.729491, & 6.341052.

“While the support emphasizes how popular an itemset is, confidence denotes the likelihood of certain items are purchased together” (Saygin, 2021). Confidence is measured by dividing the proportion of transactions that include the antecedent and the consequence, by all transactions that contain the antecedent. The metric for confidence was set at 0.2 for searching the association rules table for the top three rules. The top confidence level for the top three rules is 0.422680, 0.209184, & 0.201970.

2.  Discuss the practical significance of the findings from the analysis.

According to the strongest association rules discovered using the Apriori algorithm, patients are highly likely to be prescribed abilify, metformin and atorvastatin at the same time. All three of these drugs appear in the top three association rules discovered.

3.  Recommend a course of action for the real-world organizational situation from part A1 based on your results from part D1.

Informational pamphlets that include abilify might be a good place to also give information about metformin and atorvastatin. Advertisements for all three prescriptions could be placed in the same general area. The hospital could perform a study to find more data that might explain why patients are likely to take all three prescriptions and whether other factors than a medical need for all three prescriptions exist, such as doctors being trained to prescribe the medications together. The pharmacies in the hospitals might place the medications closer together on the shelves so that they can fill prescriptions faster because all three are likely going to be filled together.

**Part V:**

E.  Panopto:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=02672873-0632-4e5c-a7d6-af130106b4a1>

**References**

Data Mining II <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=dbe89ddb-e92f-4d40-a87a-af030178abf1>

Sivek Ph. D., Susan Currie. November 16, 2020. *Market Basket Analysis 101: Key Concepts.* <https://towardsdatascience.com/market-basket-analysis-101-key-concepts-1ddc6876cd00>

Holst, Jared. February 16, 2021. *Analyze This: What Retailers Can and Can’t Do with Market Basket Analysis.* [https://cb4.com/blog/market-basket-analysis/#:~:text=What%20are%20the%20Limitations%20of,still%20leave%20room%20for%20improvement.&text=Averages%20tend%20to%20lie.,ll%20hit%20some%20speed%20bumps](https://cb4.com/blog/market-basket-analysis/%23:~:text=What%20are%20the%20Limitations%20of,still%20leave%20room%20for%20improvement.&text=Averages%20tend%20to%20lie.,ll%20hit%20some%20speed%20bumps)

Saygin, Eser. February 20,2021. *Affinity Analysis (Market Based Analysis).* <https://towardsdatascience.com/affinity-analysis-market-basket-analysis-c8e7fcc61a21>