D212: Data Mining II

Task 3

Lucas Benish

Western Governors University

# Part I: Research Question

## Describe the purpose of this data mining report by doing the following:

### Propose **one** question relevant to a real-world organizational situation that you will answer using market basket analysis.

One organizational question I would like to answer based on the medical dataset is the following: what medications are patients often prescribed together over the course of 2 years for patients seen by our organization?

### Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

One goal of this analysis is to determine which 3 to 5 medications have commonly been prescribed together to patients within a 2-year period.

# Part II: Market Basket Justification

## Explain the reasons for using market basket analysis by doing the following:

### Explain how market basket analyzes the selected dataset. Include expected outcomes.

Market basket analysis takes in a transactional dataset and looks for items that are frequently purchased together. In the case of medications, such as this dataset, a transaction would be a medication history. Market basket analysis looks for relationships between items to determine “rules.” Rules are a way of determining “if x happens, we can expect y follow” (Rizka, 2019). The expected outcome therefore is a set of rules that state based on a specified number of items purchased, or in this case prescribed, what other medications would likely also be prescribed?

### Provide **one** example of transactions within the dataset.

From the original dataframe provided, we can see that line 2 is our first true entry without a full row of NAs. There are a total of 15 prescriptions, or transactions, within this row. Presc16 through Presc20 are null. The prescriptions within this row are as follows:

Presc01: amlodipine

Presc02: albuterol aerosol

Presc03: allopurinol

Presc04: pantoprazole

Presc05: lorazepam

Presc06: omeprazole

Presc07: mometasone

Presc08: fluconozole

Presc09: gabapentin

Presc10: pravastatin

Presc11: cialis

Presc12: losartan

Presc13: metoprolol succinate XL

Presc14: sulfamethoxazole

Presc15: abilify

### Summarize **one** assumption of market basket analysis.

Per Indeed.com, market basket analysis operates on the assumption that when individuals purchase one item, they are therefore likely to purchase another specific item or group of items (Indeed Editorial Team, 2022).

# Part III: Data Preparation and Analysis

## Prepare and perform market basket analysis by doing the following:

### Transform the dataset to make it suitable for market basket analysis. Include a copy of the cleaned dataset.

The data utilized within this analysis was transformed from a standard dataset into a transactional dataset. This transformation of data was completed utilizing the code provided within Dr. Kamara’s video, “How to Perform Market Basket Analysis in R” (2022). The code snippets will be noted below for citation, and the cleaned dataset will be attached to the final submission.

The first line of the following code, after the comment, was adopted from SparkBy{Examples} (Naveen, 2022) to remove rows in which all entries were NA. The second line was a modified version of the first line, meant to remove all columns in which all entries were NA.

#removing rows where they are all NAs, followed by removing columns that are all NAs

medical\_mba <- medical\_mba[rowSums(is.na(medical\_mba)) != ncol(medical\_mba), ]

medical\_mba <- medical\_mba[, colSums(is.na(medical\_mba)) != nrow(medical\_mba)]

The following lines of code, except for the comments, were provided within Dr. Kamara’s video as noted above (Kamara, 2022). The names of the dataframes have been changed according to my naming convention. This code was utilized to transform the data into a transactional dataset rather than a dataframe.

#next is from Dr. Kamara's code; adding transaction ID and factorizing dataframe

medical\_mba$TransactionID <- factor(seq.int(nrow(medical\_mba)))

medical\_mba <- as.data.frame(unclass(medical\_mba), stringsAsFactors = TRUE)

#pivoting into two columned dataframe

medical\_mba\_pre\_transactions <- pivot\_longer(medical\_mba, cols = 1:15, names\_to = "PrescriptionNBR", values\_to = "Medication")

head(medical\_mba\_pre\_transactions)

medical\_mba\_pre\_transactions <- medical\_mba\_pre\_transactions[, c(1, 3)]

medical\_mba\_pre\_transactions <- medical\_mba\_pre\_transactions[!(medical\_mba\_pre\_transactions$Medication == ""), ]

list\_medical\_mba <- as.data.frame(medical\_mba\_pre\_transactions)

list\_medical\_mba <- split(list\_medical\_mba$Medication, list\_medical\_mba$TransactionID)

str(list\_medical\_mba)

#creating "basket"

medical\_mba\_trans <- as(list\_medical\_mba, "transactions")

medical\_mba\_basket <- as(medical\_mba, "matrix")

### Execute the code used to generate association rules with the Apriori algorithm. Provide screenshots that demonstrate the error-free functionality of the code.

The code utilized to perform the aprior() algorithm was also provided within Dr. Kamara’s video, “How to Perform Market Basket Analysis in R” (2022). The code will be noted below, as well as the outputs following the code. Similarly to the pre-processing in section C1, the code utilized will match up with Dr. Kamara’s except for utilizing my naming conventions.

#running apriori

medical\_mba\_arules <- apriori(medical\_mba\_basket, control = list(verbose = F), parameter = list(supp = 0.008, conf=0.4, minlen=2))

#removing redundancy

medical\_redundant <- is.redundant(medical\_mba\_arules)

non\_redundant\_medical\_mba <- medical\_mba\_arules[!medical\_redundant]

inspect(head(sort(non\_redundant\_medical\_mba, by = "lift", decreasing = T), 5))

summary(non\_redundant\_medical\_mba)

The following screenshot demonstrates the output in the console from running the above lines of code in R Studio.

A screenshot of a computer

Description automatically generated with medium confidence

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

The values for median and average support, lift, and confidence can be seen in B2. They will be provided once again here as well. Average and median confidence were 0.45 and 0.44, respectively. Average and median support were 0.013 and 0.011, respectively. Average and median lift were 2.02 and 1.91, respectively. These were provided within the output of the summary(non\_redundant\_medical\_mba) call.

A screenshot of a computer

Description automatically generated with medium confidence

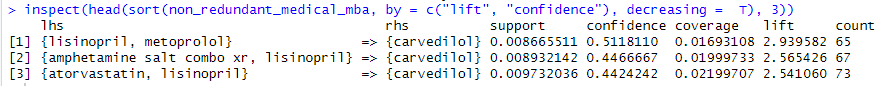
The entire association rules table can be seen using inspect() on the ruleset. The output of inspect(non\_redundant\_medical\_mba) is noted below.

A screenshot of a computer

Description automatically generated with medium confidence

### Identify the top **three** rules generated by the Apriori algorithm. Include a screenshot of the top rules along with their summaries.

The top three rules of the association rules table, based on lift, are determined once again from code provided by Dr. Kamara. In his video, “How to Perform Market Basket Analysis in R” (2022), he demonstrates how to view the first 5 rows based on lift. I decided to inspect only the top 3 rules, based on lift then confidence, for the purposes of the top 3 rules. Confidence would be utilized as a tiebreaker in the event of a tied lift-value. A screenshot is noted here.



# Part IV: Data Summary and Implications

## Summarize your data analysis by doing the following:

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

Prior to discussing the support, lift, and confidence results, it is important to understand what these entail. Support represents how often an antecedent, or left-hand side of a rule, appears within a transactional dataset (Zhang, 2021). Confidence represents the probability of transactions that contain the antecedent also containing the consequent, or the right-hand side of the rule (Zhang, 2021). Finally, lift represents a ratio between the confidence and support, or what is the likelihood of the consequent happening because of the antecedent (Zhang, 2021).

Taking my top rule as an example, we can see the following. The support is 0.008, or the probability of my antecedent, {lisinopril and metoprolol} occurring together within my transactional dataset is 0.8%. The confidence is 0.512, meaning that the probability of carvedilol being present within the same “transaction” as both lisinopril and metoprolol is 51.2%. Finally, the lift of the top rule is 2.9. With the lift being greater than one, we can conclude that the presence of lisinopril and metoprolol prescriptions increases the likelihood of a patient also being prescribed carvedilol.

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

Within the top 3 rules, it is noted that carvedilol is the consequent of all 3 rules. It can also be noted that lisinopril is present within all 3 antecedents. Lisinopril is used to treat hypertension (high blood pressure) within adults (NIH, 2021). Carvedilol is a beta-blocker that is used to treat both hypertension and heart failure (NIH, 2017). Based on these results, it is highly likely that if a patient has been prescribed lisinopril within the last 2 years, they have also been prescribed carvedilol. It cannot be assumed that they have a diagnosis of hypertension, as no patient data outside of prescriptions is present within this data.

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

My question from part A1 was as follows: what medications are patients often prescribed together over the course of the last 2 years for patients seen by our organization?

Based on the results of this analysis, we can see that the lisinopril is often prescribed, and when it is it is likely that the patient is also prescribed carvedilol. One actionable insight from this is that it may be worth performing more pre-emptive blood pressure screenings during annual physicals to address blood pressure concerns prior to the need for medication arising. It is also worth looking at side effects and interactions of medications that are often prescribed together.

An appropriate next step may be performing another apriori analysis, this time with a single medication as an antecedent to determine what other medications are often prescribed with each other and to explore other antecedents.

# Part V: Attachments

## Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

## Record *all* sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

1. Kamara, K. (2022, August 28). How to Perform Market Basket Analysis in R [Video Presentation]. Western Governors University. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5674b196-a9f1-4e85-a322-af0000021f3f>
2. Naveen (2022, July 19). How to Remove Rows with NA in R. SparkBy{Examples}. Retrieved June 23, 2023, from <https://sparkbyexamples.com/r-programming/remove-rows-with-na-in-r/>

## Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

1. Indeed Editorial Team (2022, October 12). FAQ: What Is Market Basket Analysis? (Types Plus Examples). Retrieved June 23, 2023, from <https://sg.indeed.com/career-advice/career-development/market-basket-analysis#:~:text=The%20approach%20relies%20on%20the,rules%20or%20if%2Dthen%20statements>
2. NIH - National Library of Medicine (2017, February 15). Carvedilol. MedlinePlus. Retrieved June 23, 2023, from https://medlineplus.gov/druginfo/meds/a697042.html
3. NIH - National Library of Medicine (2021, February 15). Lisinopril. MedlinePlus. Retrieved June 23, 2023, from https://medlineplus.gov/druginfo/meds/a692051.html
4. Rizka, Y. (2019, July 9). Market Basket Analysis with R. Medium. Retrieved June 23, 2023, from <https://medium.com/@yolandawiyono98/market-basket-analysis-with-r-8001417a8e29>
5. Zhang, L. (2021, September 21). Understanding Support, Confidence, Lift for Market Basket (Affinity) Analysis. Retrieved June 23, 2023, from <https://www.thedataschool.co.uk/liu-zhang/understanding-lift-for-market-basket-analysis/>

## Demonstrate professional communication in the content and presentation of your submission.