Briana Churchill

Student ID: 011009463

Dr. Kesselly Kamara

2024 April 13

ASSOCIATION RULES AND LIFT ANALYSIS

Part I: Research Question

A1. Propose one question relevant to a real-world organizational situation that you will answer using market basket analysis.

Although I am not a medical doctor or pharmacist, I believe that I can provide data that can tell a story that may assist medical professionals in informing their care practices. After working as a medical assistant in specialty medicine for 8 years, I recognize the value of data in medicine. In my experience, I noticed that many patients have numerous medical providers who do not always have access to view, or acquire reliable information on, the patient's medical history. Sometimes the patient is not aware of the drug names or dose in which they are taking, or they confuse names, dosages, etc. Other times patients see providers in different medical record systems, which means prior medical records are limited.

For example, a patient is sent to gastroenterologist specialists, a stand-alone clinic unassociated with the medical systems used in larger hospitals (Ie: Epic or EClinicalWorks) within the area. The doctor may focus on a referral from urgent care that has no prior medical history, or medications noted. The doctor asks the patient questions to learn this information, but the patient does not remember the name of the drugs they take, or what they are for. Therefore, the doctor focuses on the gastric issue without knowing that the patient is being seen for cardiac and endocrinological issues at the local hospital where they are prescribed medications that have contraindicated side effects if taken with the gastric medicine the doctor plans to prescribe.

In other instances, medical professionals may be unfamiliar with drugs prescribed by a different specialist and any counteractive effects they may have in combination with drugs prescribed at the time of the appointment. New drugs are always being introduced to the market and with so many specialties and patient loads, some prescribers may stick to being informed of new drugs within their scope of practice only.

As such, the question for this analysis is: can drug combinations with adverse reactions be found amongst the strong relationships identified within the market basket analysis? As I said before, I am not a medical doctor or pharmacist so I will use WebMD's Drug Interaction Checker to determine any potential combinations that are advised against. Providing medical providers with this information could allow them to reassess any patient files and limit the prescribing of contradictory medications in hopes of assisting in lowering the rates of readmission.

A2. Define one goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

The goal of the data analysis is to find strengthful relationships between at least three combinations of prescriptions. Having a minimum of three combinations will allow me to research multiple drugs, which may inform many providers. After assessing the information provided by WebMD regarding adverse reactions, I will be able to summarize the findings and inform the hospital system of the results and their indications.

Part II: Market Basket Justification

B1. Explain how market basket analyzes the selected data set. Include expected outcomes.

From Dr. Kamara's informational videos in the course content, I learned that market basket analysis (MBA) works by using the data obtained during single transactions to determine relationships of items purchased by a single buyer. In this case, the MBA will list the medications prescribed together for single patients to determine relationships between the prescriptions. We can think of these lists as items in a grocery store basket. MBA seeks to find relationships that dictate "if this, then that." "If this" is the antecedent, or otherwise the first item in the basket. "Then that" is the consequent, which is the item added to the cart after the first item was added. For example: when looking at physical shopping carts, you may notice spaghetti noodles being added to someone's cart. When seeing the same customer at the check-out lane later, and looking at the cart again, you are likely to see a spaghetti sauce or the ingredients to make some. If a business is able to find these relationships, the data can be used to inform business decisions related to incentives, product placement, and bundling.

Since many medical issues require multiple prescriptions, I expect that there will be at least one instance of a meaningful prescription combination with contraindications. Additionally, I expect at least one instance of a drug combination that has no counter interactions. In fact, there are likely to be more beneficial combinations than negative ones because many medications go hand and hand with one another. An example of this is the prescriptions written for individuals with type 1 diabetes mellitus. Not only do they need short-acting insulin (meal-time insulin) but they also need a basal insulin (background insulin). As such, it would come as no surprise to medical providers and pharmacists to see Novolog (short-acting) and Tresiba (basal) prescribed together. Hopefully there are patterns similar to the insulin example and this analysis results in no strong relationships between contradicting drugs, but this profession is known as "practicing" medicine for a reason. If we are able to find "baskets" of medications that should not be taken alongside one another, then the information will be valuable in updating care plans to avoid adverse effects.

B2. Provide one example of transactions in the data set

Seen below is the output for transaction 17. The first prescription is metoprolol, the second is carvedilol and the last medication is losartan.

In [5]:
```
df.iloc[17]
```

Out[5]:
```
Presc01    metoprolol
Presc02    carvedilol
Presc03      losartan
Presc04           NaN
Presc05           NaN
Presc06           NaN
Presc07           NaN
Presc08           NaN
Presc09           NaN
Presc10           NaN
Presc11           NaN
Presc12           NaN
Presc13           NaN
Presc14           NaN
Presc15           NaN
Presc16           NaN
Presc17           NaN
Presc18           NaN
Presc19           NaN
Presc20           NaN
Name: 17, dtype: object
```

B3. Summarize one assumption of market basket analysis

One assumption of market basket analysis is that the dataset is a representative sample of overall prescription transactions for the patients in the dataset. Having a representative patient sample is necessary to accurately find meaningful relationships within the data. However, it is important to note that the analysis may not accurately depict prescription trends if some patients see medical providers outside of the hospital system thus potentially resulting in a limited representative sample. (2023, Deniran)

Part III: Data Preparation and Analysis

C1. Transform the data set to make it suitable for market basket analysis. Include a copy of the cleaned data set.

In order to prepare the dataset for this analysis, we needed to ensure that there were no null values. Using the guidance from Dr. Kamara, the dataset was cleaned of its null values first.

Once the nulls were taken care of, a list of lists was created, and this list was fed into the transaction encoder. Once the list was fed into the transaction encoder, each prescription drug became its own column. Therefore, 119 columns (medications) and 7501 rows (transactions) are available. Additionally, each row now has "True" or "False" in each column indicating whether the patient purchased that prescription or not. A copy of the cleaned data set was saved as a CSV file and attached to this assessment.

C2. Execute the code used to generate association rules with the Apriori algorithm.

Again, using code information learned from Dr.Kamara's instructional videos, the Apriori object was created with a minimum support value of 0.02. The association rules were then generated, creating an association table with lift values of at least 1 and support values of 0.02 or more. The code used for these steps has been attached below.

In [15]:
```
# Create apriori object
rules = apriori(clean_df, min_support = 0.02, use_colnames = True)
rules.head(5)
```

Out[15]:

	support	itemsets
0	0.046794	(Premarin)
1	0.238368	(abilify)
2	0.020397	(albuterol aerosol)
3	0.033329	(allopurinol)
4	0.079323	(alprazolam)

In [16]:
```
# Create rules table
rules_table = association_rules(rules, metric = 'lift', min_threshold = 1)
rules_table.head(20)
```

Out[16]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(abilify)	(amlodipine)	0.238368	0.071457	0.023597	0.098993	1.385352	0.006564	1.030562	0.365218
1	(amlodipine)	(abilify)	0.071457	0.238368	0.023597	0.330224	1.385352	0.006564	1.137144	0.299568
2	(abilify)	(amphetamine salt combo)	0.238368	0.068391	0.024397	0.102349	1.496530	0.008095	1.037830	0.435627
3	(amphetamine salt combo)	(abilify)	0.068391	0.238368	0.024397	0.356725	1.496530	0.008095	1.183991	0.356144
4	(abilify)	(amphetamine salt combo xr)	0.238368	0.179709	0.050927	0.213647	1.188845	0.008090	1.043158	0.208562
5	(amphetamine salt combo xr)	(abilify)	0.179709	0.238368	0.050927	0.283383	1.188845	0.008090	1.062815	0.193648
6	(abilify)	(atorvastatin)	0.238368	0.129583	0.047994	0.201342	1.553774	0.017105	1.089850	0.467950
7	(atorvastatin)	(abilify)	0.129583	0.238368	0.047994	0.370370	1.553774	0.017105	1.209650	0.409465
8	(abilify)	(carvedilol)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008	0.400606
9	(carvedilol)	(abilify)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314	0.369437
10	(abilify)	(cialis)	0.238368	0.076523	0.023997	0.100671	1.315565	0.005756	1.026851	0.314943
11	(cialis)	(abilify)	0.076523	0.238368	0.023997	0.313589	1.315565	0.005756	1.109585	0.259747
12	(abilify)	(citalopram)	0.238368	0.087188	0.024397	0.102349	1.173883	0.003614	1.016889	0.194486
13	(citalopram)	(abilify)	0.087188	0.238368	0.024397	0.279817	1.173883	0.003614	1.057552	0.162275
14	(abilify)	(clopidogrel)	0.238368	0.059992	0.022797	0.095638	1.594172	0.008497	1.039415	0.489364
15	(clopidogrel)	(abilify)	0.059992	0.238368	0.022797	0.380000	1.594172	0.008497	1.228438	0.396502
16	(abilify)	(dextroamphetamine XR)	0.238368	0.081056	0.027463	0.115213	1.421397	0.008142	1.038604	0.389252
17	(dextroamphetamine XR)	(abilify)	0.081056	0.238368	0.027463	0.338816	1.421397	0.008142	1.151921	0.322617
18	(abilify)	(diazepam)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256	0.339197
19	(diazepam)	(abilify)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357	0.308965

C3. Provide values for the support, lift, and confidence of the association rules table.

As seen in the script outputs above, the association table provides values for the support >= 0.02, lift of >=1, and confidence of the associations within the table.

C4. Explain the top three relevant rules generated by the Apriori algorithm. Include a screenshot of the top three relevant rules.

The top three relevant rules generated by the Apriori algorithm and sorted by lift values of > 0.08 in descending order. The association table can be seen in the table below. Rule 75 shows that there is a strong likelihood that lisinopril will be prescribed if carvedilol has already been prescribed. Rule 74 indicates a strong likelihood of the reverse, which shows that carvedilol and lisinopril have a strong association. Lastly, rule 72 indicates that there is a high likelihood that carvedilol will be prescribed if glipizide is prescribed.

In [20]:
```
# Retreive sorted rules with lift > 0.08
sorted_rules = rules_table[(rules_table['lift'] > 0.08)].sort_values(by=['lift'], ascending= False)
sorted_rules.head(3)
```

Out[20]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
75	(carvedilol)	(lisinopril)	0.174110	0.098254	0.039195	0.225115	2.291162	0.022088	1.163716	0.682343
74	(lisinopril)	(carvedilol)	0.098254	0.174110	0.039195	0.398915	2.291162	0.022088	1.373997	0.624943
72	(glipizide)	(carvedilol)	0.065858	0.174110	0.022930	0.348178	1.999758	0.011464	1.267048	0.535186

Part IV: Data Summary and Implications

D1. Summarize the significance of support, lift, and confidence from the results of the analysis.

Support measures how many times an item appears in a dataset and is used to decide if it is worth pursuing an association rule or not. (2022, Kamara) If a particular item appears only twice in a dataset with 1,000 transactions, it would not be worth pursuing an association rule as that item would have a low support value. However if an item appears in 400 of 1,000 transactions, then an association rule would be worth pursuing as the support value for that item would be high. For the top three rules included above, the support values are much higher than the minimum 0.02 value I implemented which indicates that the prescriptions occurred often within the dataset and are reliable. As we can see from rule 75 to rule 72 the support values are 0.039, 0.039, and 0.229. Alone, support is not enough to analyze the data, we will need confidence and lift values as well.

Confidence measures the likeness that an item will be purchased if another item is already in the basket. For example, we would expect a high confidence that customers would add hot dog buns after adding hotdogs to the cart if 90 out of 100 customers in a dataset added both items to their carts. Using the results above again, we can see that the confidence is higher in rule 74 with a value of 0.398 and rule 72 with a value of 0.348. Now that we have the support and confidence values, we can better understand the lift value.

Lift measures how popular items are within a dataset. Lift is a ratio between confidence and expected confidence, which can be used to determine the likelihood of item "b" being added when item "a" is present. (2022, Dobilas) With lift values of > 1 for all top three rules, we can expect that there is a likelihood that the consequent will be prescribed after the antecedent is prescribed. Without a high lift parameter we may believe there is a strong relationship between milk and fireworks during the week before independence day; when in reality, one of the items is a "staple" grocery item and the other has a seasonal uptick in sales and thus both items together have weak association overall.

D2. Discuss the practical significance of your findings from the analysis.

As I described in C4, there are indications that carvedilol and lisinopril have a strong relationship and are likely to be picked up together in a prescription transaction. Fortunately for these patients, WedMD's drug interaction checker found no adverse effects. Additionally, the results for the combination of carvedilol and glipizide found no adverse effects either. As such, the findings indicate higher chances of safe drug prescription practices within this hospital system.

D3. Recommend a course of action for the real-world organizational situation from part A1 based on the results from part D1

Now that the hospital system is aware of strong relationships between safe drug combinations, they are a step closer to understanding why their hospital system is experiencing higher readmissions rates. Ruling out prescribing practices allows the hospital system to focus on other potential factors. Therefore, a recommended course of action would be to review other available data such as labs ordered during time of admittance, patient satisfaction surveys, etcetera. By assessing different stages of care such as lab ordering or length of hospital stay can help the hospital narrow down potential factors in readmissions rates. Additionally, reviewing patient surveys for further insight can assist in allowing the healthcare system in seeing the bigger picture.

Part V: Attachments

E1. Include the presenter and a vocalized demonstration describing the programs used to complete this task in the Panopto video recording.

A Panopto video was created and attached to the submission for this assessment.

F. Record all web sources you used to acquire data or segments of third-party code

Kamara, K.(2022)“Market Basket Analysis in Python.” Western Governors University.

Kamara, K.(2022)“How To Install mlxtend in Anaconda Environment.” Western Governors University.

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

Kamara, K.(2022) “Market Basket Analysis in Theory.” Western Governors University.

Dobilas, S. (May, 2022) "One Minute Overview of the Apriori Association Rule Learning." LinkedIn. Retrieved from: https://www.linkedin.com/pulse/one-minute-overview-apriori-association-rule-learning-saulius-dobilas#:~:text=Support(A)%20%3D%20Frequency(,B)%20%2F%20Support(B)

Garg, A. (Sep, 2018) "Complete Guide to Association Rules." Towards Data Science. Retrieved from: https://towardsdatascience.com/association-rules-2-aa9a77241654

Deniran, O. (Nov, 2023) "Boosting Sales With Data: The Power of Market Basket Analysis in Retail." Medium. Retrieved from: https://medium.com/@chemistry8526/boosting-sales-with-data-the-power-of-market-basket-analysis-in-retail-c79cc10a14df#:~:text=Transaction%20Independence%3A%20It%20is%20assumed,item%20associations%20within%20individual%20transacti

N.A. (n.d.) "Drug Interaction Checker." WebMD. Retrieved from: https://www.webmd.com/interaction-checker/default.htm