Performance Assessment D212 – Data Mining II  
Task III

Doug Haunsperger

College of IT, Western Governors University

April 5, 2024

# Part I. Rese**arch Questi**on

## A1. Question Proposal

I propose to research the question, “Is it possible to identify associations in our prescription data using market basket analysis?”

## A2. Goal

The goal of this analysis would be to identify prescriptions that are associated together, both for stocking purposes and to improve patient care. For instance, if a patient has been prescribed Medicine A that shows an association with Medicine B in our database, we can ensure we have adequate stock of Medicine B, and also propose to the medical care team that a prescription for Medicine B may be appropriate.

# Part II. Technique Justification

## B1. Explanation of Market Basket Analysis

Market basket analysis is “a kind of knowledge discovery in data” and “can be applied in various fields of work” (Maitra, 2019). It is applied by “implementing Association Rule Mining … a rule-based machine learning method that helps to uncover meaningful correlations between different [items] according to their co-occurrence in a data set” (Lim, 2022). Lim also states that the Apriori Algorithm is “one of the most popular algorithms used in association rule learning” (2022). This algorithm identifies items in a data set and groups them into larger and larger itemsets, but only if the sets are frequent, that is, “the probability of the itemset is beyond a certain predetermined threshold” (2022). Finally, itemsets are grouped together through association rules, in the form of ‘if itemset A, then itemset B’. These rules are scored based on metrics including support, confidence, lift, leverage, conviction, and Zhang’s metric.

I expect that running a Market Basket Analysis on the given patient prescription database will generate itemsets (containing one or more prescription drugs) and a set of association rules showing which itemsets are correlated with one another in our data set.

## B2. Transaction Example

Our data set includes a list of approximately 7500 patient prescription records with columns for up to 20 individual prescription drugs. An example record is given in Figure 1. Note that trailing fields are blank.

Figure 1  
*Record from given prescription dataset*



## B3. Market Basket Analysis Assumption

One of the underlying assumptions that Market Basket Analysis uses to make analysis more efficient is the Apriori principle, that is that “all subsets of a frequent itemset must also be frequent” (Garg, 2018). For example, if an itemset {a, b, c} has a certain level of support *s*, the support level of the subset {a, b} will be greater than or equal to *s.* This is known as the anti-monotone property of support (Garg, 2018).

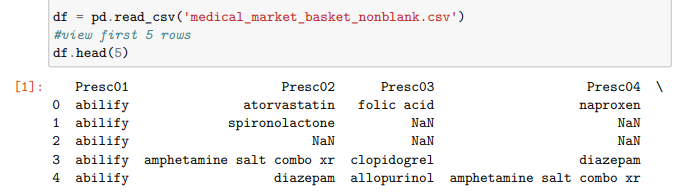
This principle is what allows the Apriori Algorithm to prune all supersets of an itemset which falls below a minimum level of support. From the anti-monotone property, it is easy to see that any such superset must have a support level less than or equal to the itemset being considered.

# Part III. Data Preparation & Analysis

## C1. Data Transformation

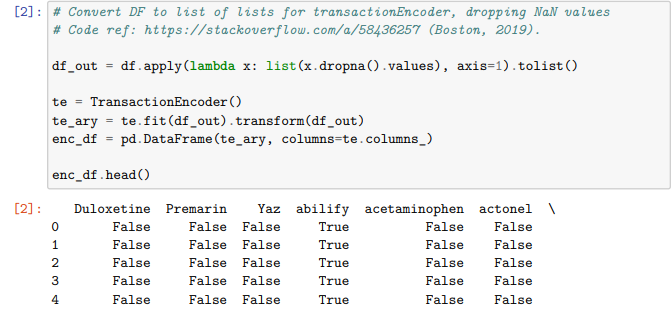
The given dataset has an empty row every other line. As a first step, I opened the CSV file in Excel, sorted by ‘Presc01’ and re-exported to CSV without the blank lines. I then read the CSV into a Pandas dataframe:

Figure 2



Following this step, I used the TransactionEncoder function from the mlxtend library to transform the original list of patient records to a one-hot encoded list with each possible drug given a True or False value. The input to the TransactionEncoder needs to be a list of lists with NaN values removed. I used this code:

Figure 3



I then exported the cleaned & encoded DataFrame back to a CSV file, ‘enc\_medical\_markbask.csv’, which is attached to my submission.

## C2. Code Execution

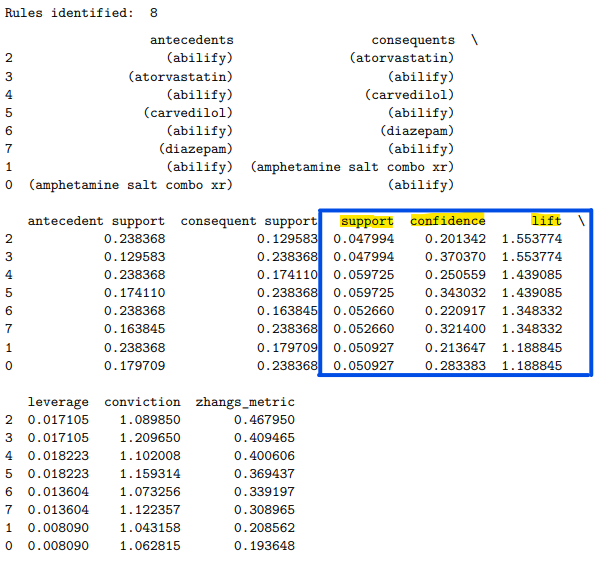
See attached Jupyter notebook, ‘D212\_PA3.ipynb’.

## C3. Association Rules Table

I chose a minimum support level of five patients of the 7,501 total patients to identify itemsets using the apriori function. I then ran the association\_rules function, with a threshold lift value of 1.0. I identified 34 itemsets with three drugs or fewer and the minimum level of support. All itemsets found had either one or two drugs in them.

From these 34 itemsets, 8 rules were generated with a lift of at least 1.0. The full association rules table is shown in Figure 4 with the support, lift, and confidence columns highlighted.

Figure 4



## C4. Top Three Rules

Sorted by Zhang’s metric, which “is a measure designed to assess the strength of association (positive or negative) between two items” (Yasik, 2023), the top three rules can be expressed as , , and . These are seen in Figure 4, with Zhang’s metric values of 0.468, 0.409, and 0.401 respectively. Identifying these as the “top three” is influenced by the choice of scoring metric to use. If confidence or conviction was seen as more important, a different set of three rules would be at the top. I chose Zhang’s metric since it takes account of “both [itemsets’] co-occurrence and their non-co-occurrence” (Yasik, 2023).

# Part IV. Data Summary & Implications

## D1. Significance of Support, Lift, and Confidence

Support is simply measured by the percentage of transactions, or in our case, patient records, where an itemset is found. Looking at the rule , we see antecedent (*abilify*) support at 0.238, indicating that drug appears in 23.8% of patient records; consequent (*atorvastatin*) support at 0.130 – 13% of patient records; the combined rule support value is 0.048, indicating that the combination of these two drugs appear in 4.8% of patient records. Indeed, looking at the list of identified itemsets, we see that this combination had enough to support to count as a candidate itemset in its own right:

Figure 5  
*Snippet of the generated itemset list from the apriori function*



Lift is the “observed to expected ratio”, or how likely two items appear together “while controlling for how popular both items are” in the dataset (Lim, 2022).

Lim states that a lift of 1 indicates independence or no association (2022). Yasik states that lift greater than 1 indicates that “the presence of A has a positive effect on the likelihood of B” appearing in the basket, “suggesting a stronger association between A and B than would be expected by chance” (2023).

Neither support nor lift show a direction or causation of association. The top two rules this analysis identified have identical support and lift values since they are the complementary rules and . The lift value of 1.55 of these two rules shows a relatively strong positive association.

Confidence, on the other hand, is not symmetrical. It is a measure of the probability that if itemset A is present, that itemset B will also be present.

Note that the #2 rule, actually has a higher confidence (0.370) than its complementary #1 rule (0.201). This indicates a stronger likelihood that if atorvastatin is prescribed, then abilify will also be, versus the converse likelihood of if abilify is prescribed, then atorvastatin will also be. I have ranked the rules by Zhang’s metric, however, where the rule wins out.

## D2/D3. Practical Significance and Recommended Course of Action

Abilify is a very frequently prescribed drug in the hospital’s patient base – nearly one quarter of the patients in this dataset are receiving it. With these association rules, the hospital can identify these potential companion drugs to recommend to the care teams and/or to ensure sufficient stock: atorvastatin, carvedilol, diazepam, and “amphetamine salt combo XR”.

# Part V. Attachments

## E. Demonstration Video

## A video describing my methods and code can be found at: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c58d9ab9-9e64-431e-ba1e-b14901190ba9>

## F. Third-party Code Sources

Boston, Scott. (October 17, 2019). StackOverflow. *Answer to: Applying Transaction Encoder on Dataset.* <https://stackoverflow.com/questions/58435648/applying-transaction-encoder-on-dataset/58436257#58436257>

## G. References

Garg, Anisha. (September 17, 2018). Medium. *Complete guide to Association Rules (2/2).* Towards Data Science. <https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84>

Lim, Yenwee. (April 8, 2022). Medium. *Data Mining: Market Basket Analysis with Apriori Algorithm.* Towards Data Science. <https://towardsdatascience.com/data-mining-market-basket-analysis-with-apriori-algorithm-970ff256a92c>

Maitra, Sarit. (October 15, 2019). Medium. *Market Basket Analysis.* Towards Data Science. <https://towardsdatascience.com/market-basket-analysis-knowledge-discovery-in-database-simplistic-approach-dc41659e1558>

Yasik, IU. (October 4, 2023). Medium. *Market Basket Analysis* *& Apriori Algorithm using Zhang’s Metric.* <https://medium.com/@iuyasik/market-basket-analysis-apriori-algorithm-using-zhangs-metric-708406fc5dfc>