D212 Task 3 Association Rules and Lift Analysis

Aaron Balke

S# 011005116

December 30th, 2023

Part I: Research Question

A1. Question

For this analysis, my research question is "Are any medication purchases associated with Duloxetine."

A2. Goals

As a business, an association between medication purchases would help determine resource and financial allocation to pharmaceutical products, and as a healthcare provider, an association could assist in correlating health problems and symptoms. For example, if Duloxetine, an antidepressant and anti-anxiety medication, is purchased often with Premarin, a menopause symptom relief drug, more research could be done to see if there is a correlation between those symptoms. (Duloxetine: MedlinePlus Drug Information)

Part II: Market Basket Justification

B1. Explanation

Market Basket Analysis is a data mining technique used to better understand customer purchasing patterns (WhatIs). This is completed through the creation and filtering of Frequency (support) of Itemsets using the apriori algorithm, which will only keep the discernible association rules, and remove the noisy rules. After this, association rules are evaluated, and those related to our chosen item that score high enough are kept.

We hope to get discernible association rules with Duloxetine as the antecedent or consequent. This would suggest a purchasing pattern between Duloxetine and another drug.

B2. Transaction Example

In [15]:
```
# Standard Imports
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

#MLxtend Imports
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Import Data
df = pd.read_csv('../medical_market_basket.csv')
```

In [16]:
```
# Transaction Example
df.loc[1]
```

Out[16]:
```
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
Presc16                  spironolactone
Presc17                   albuterol HFA
Presc18                    levofloxacin
Presc19                    promethazine
Presc20                       glipizide
Name: 1, dtype: object
```

B3. Assumption

The main assumption in Market Basket Analysis is that, with a large enough dataset, relationships form between products/options because those relationships have meaning. An example would be the previously mentioned hypothetical Duloxetine and Premarin. It is important to note, that with smaller datasets noise can play a factor and make relationships harder to see.

Part III: Data Preparation and Analysis

C1. Data Preparation

The following steps have to be completed before analysis:

1. Remove Even Rows: Every even row is completely blank. We have to remove these.
2. Convert Input Data: mlxtend Transaction Encoding cannot use a Dataframe as an input datatype. We have to convert the data frame into a native 2D Array (List of Lists).
3. Transaction Encoding: Fit and Transform data using the TransactionEncoder object
4. Convert the returned array back to a Dataframe
5. Drop NaN Columns: Column/Feature named 'nan' will have to be removed.
6. Export Cleaned Dataframe

In [17]:
```python
# Original Shape
print("Original Record Count: ", df.shape[0])

# Remove Completely Blank Rows
df.dropna(axis=0, how='all', inplace=True)
df.reset_index(drop=True, inplace=True)
print("After Blank Removal Count: ", df.shape[0])
```

```
Original Record Count:  15002
After Blank Removal Count:  7501
```

In [18]:
```python
# Convert Dataframe into List of Lists to fit/transform in Transaction Encoder
trans = []

for i in range(df.shape[0]):

    tran = []

    for j in range(df.shape[1]):
        tran.append(str(df.values[i,j]))

    trans.append(tran)
```

In [19]:
```python
# Fit and Transform data using TransactionEncoder
encoder = TransactionEncoder()
array = encoder.fit_transform(trans)
```

In [20]:
```python
# Convert Returned array back to Dataframe
cleaned_df = pd.DataFrame(array, columns=encoder.columns_)
```

In [21]:
```python
# Remove NaN columns from the Dataframe
cleaned_df.drop(['nan'], axis=1, inplace=True)
```

In [22]:
```
# Export Cleaned Dataframe as CSV
cleaned_df.to_csv('D212_task3_cleaned.csv')
```

C2. Apriori & Association Rules

1. Apriori Algorithm: Find the frequent itemsets in the dataset.

Frequent Itemsets are the Itemsets with support > the minimum support.

Minimum support of ~1.2% is chosen because we only want easily discernible rules, and we do not want to introduce noise from small sample sizes. Duloxetine only is in 1.2% of the transactions, so a max of 1.2% is required to have it included for association rule creation.

The Apriori Algorithm simplifies and removes the need to compute every possible rule since if a more general rule is not met, any more specific subset of that rule will not be met either. In our case, if {Duloxetine}{Cialis} is not a frequent itemset, then {Duloxetine, Gabapentin}{Cialis} will not be a frequent itemset. This makes it easier to compute since we will not have to compute every permutation/subset.

1. Association Rules: Based on the inputted itemsets, we will return evaluation metrics to correlate purchases. A minimum lift threshold of 1 is chosen to only get a positive correlation.

In [23]:
```
# See Minimum Support required for Duloxetine (Support = Freq / Total)
true_count = cleaned_df['Duloxetine'].value_counts()[1]
false_false = cleaned_df['Duloxetine'].value_counts()[0]

min_support = true_count / (false_false + true_count)
print(min_support)
```

```
0.011998400213304892
```

In [24]:
```
# Generate Frequent Itemsets through Apriori Algorithm, min_support = 1%
itemsets = apriori(
    cleaned_df,
    min_support= min_support,
    use_colnames=True
)
itemsets.head()
```

Out[24]:

	support	itemsets
0	0.011998	(Duloxetine)
1	0.046794	(Premarin)
2	0.238368	(abilify)
3	0.015731	(acetaminophen)
4	0.011998	(actonel)

In [25]:
```
# Generate Association Rules from freq. itemsets
# Lift Metric > 1 is used to filter rules
rules = association_rules(
    itemsets,
    metric='lift',
    min_threshold=1
)
```

C3. Association Rules

In [26]:
```
rules
```

Out[26]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(amlodipine)	(abilify)	0.071457	0.238368	0.023597	0.330224	1.385352	0.006564	1.137144	0.299568
1	(abilify)	(amlodipine)	0.238368	0.071457	0.023597	0.098993	1.385352	0.006564	1.030562	0.365218
2	(amphetamine salt combo)	(abilify)	0.068391	0.238368	0.024397	0.356725	1.496530	0.008095	1.183991	0.356144
3	(abilify)	(amphetamine salt combo)	0.238368	0.068391	0.024397	0.102349	1.496530	0.008095	1.037830	0.435627
4	(amphetamine salt combo xr)	(abilify)	0.179709	0.238368	0.050927	0.283383	1.188845	0.008090	1.062815	0.193648
...	...	...	...	...	...	...	...	...	...	...
277	(carvedilol, metoprolol)	(abilify)	0.027863	0.238368	0.011998	0.430622	1.806541	0.005357	1.337656	0.459252
278	(abilify, metoprolol)	(carvedilol)	0.035729	0.174110	0.011998	0.335821	1.928784	0.005778	1.243475	0.499381
279	(carvedilol)	(abilify, metoprolol)	0.174110	0.035729	0.011998	0.068913	1.928784	0.005778	1.035640	0.583054
280	(abilify)	(carvedilol, metoprolol)	0.238368	0.027863	0.011998	0.050336	1.806541	0.005357	1.023664	0.586184
281	(metoprolol)	(carvedilol, abilify)	0.095321	0.059725	0.011998	0.125874	2.107549	0.006305	1.075674	0.580885

282 rows × 10 columns

C4. Top Rules

The following 3 rules are the most relevant according to the lift metric. It is important to note the difference between metrics.

Support: A measure of itemset frequency

Confidence: The likelihood of the consequent given the antecedent

Lift: The rise in the probability of having the consequent given the antecedent, over just having the consequent. If lift has a value of 1, the antecedent and consequent are completely independent. (Garg)

We want associations that show the consequent is strongly included if given the antecedent. The lift metric is also the same metric we to create our association rules, leading to consistency. They are sorted from highest lift metric to lowest, however both the first and second are mirrors of each other, and the lift is identical.

antecedents	consequents
methylprednisone	lisinopril
lisinopril	methylprednisone
carvedilol, abilify	lisinopril

In [27]:
```
# the top 3 rules using lift metric, sorted with highest at the top
rules.sort_values('lift', ascending=False).head(3)
```

Out[27]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
228	(methylprednisone)	(lisinopril)	0.049460	0.098254	0.015998	0.323450	3.291994	0.011138	1.332860	0.732460
229	(lisinopril)	(methylprednisone)	0.098254	0.049460	0.015998	0.162822	3.291994	0.011138	1.135410	0.772094
272	(carvedilol, abilify)	(lisinopril)	0.059725	0.098254	0.017064	0.285714	2.907928	0.011196	1.262445	0.697788

Part IV: Data Summary and Implications

D1. Results

In [28]:
```
ant = rules[rules.antecedents == 'Duloxetine']
con = rules[rules.consequents == 'Duloxetine']

print(f"Rules with Duloxetine Antecendent: {ant.shape[0]}")
print(f"Rules with Duloxetine Consequents: {con.shape[0]}")
```

```
Rules with Duloxetine Antecendent: 0
Rules with Duloxetine Consequents: 0
```

Unfortunately, none of our final association rules have Duloxetine included. While Duloxetine itemsets made it through the apriori algorithm, at the time we created association rules with lift values > 1, all Duloxetine itemsets were removed. This means Duloxetine has no positive item association with any other drugs in our dataset.

In our analysis, the support metric was the most significant and what I believe became an immediate problem. Support is a measure of the item's frequency in the transactions. For us, this was 1.2% or around 90 records. With so few records, I do not believe it is possible to move beyond the noise and accurately see the association rules associated with Duloxetine.

Lift was the measure of the rise in probability of having Duloxetine given the other drug, over just having Duloxetine. Using this metric to filter our rules removed Duloxetine from the analysis since in our data, Duloxetine purchases, are completely independent of the other drugs.

Finally, confidence is the likelihood of buying Duloxetine at the same time as the other drug. While this metric was not used to filter rules, if we did have Duloxetine rules that mirrored each other, Duloxetine as the antecedent and Duloxetine as the consequent for the same drug, the confidence value would help determine which drug was the true antecedent and consequent.

D2. Practical Significance

While our analysis provides no significant associations, I do think we should outline why this is the case from understanding the dataset. To begin, Duloxetine is one of many antidepressants - if it had a monopoly on this relief type, it may have been easier to see patterns. Furthermore, if all antidepressants were grouped for association rule creation, we may be able to see results from a generalized perspective. Another point to take into account is Duloxetine takes 2-4 weeks for relief, while other antidepressants usually take 1-2 weeks. Perhaps it is recommended less because it provides less value than other antidepressants. These are possible causes for the low support and should be taken into consideration while choosing the next steps.

D3: Recommended Action

Since no association rules were found regarding Duloxetine, I believe no financial or medical action should be taken. However, I do believe further data gathering is required to have a dataset that not only is sufficient for the more common drugs but also the less used ones such as Duloxetine. With a larger dataset, further analysis can be done to create well-supported statements on association with Duloxetine.

E1. Presentation

https://youtu.be/ITkcfkTbGA4

F. Web Sources

GeeksforGeeks. (2018, September 4). Apriori Algorithm - GeeksforGeeks. GeeksforGeeks. https://www.geeksforgeeks.org/apriori-algorithm/

Isaiah, H. (n.d.). Market Basket Analysis in Python. Datacamp. from https://app.datacamp.com/learn/courses/market-basket-analysis-in-python

Li, S. (2017, September 25). A Gentle Introduction on Market Basket Analysis — Association Rules. Medium; Towards Data Science. https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

Mlxtend Docs:

TransactionEncoder - mlxtend. (n.d.). Rasbt.github.io. https://rasbt.github.io/mlxtend/user_guide/preprocessing/TransactionEncoder/

Raschka, S. (n.d.). Association rules - mlxtend. Rasbt.github.io. https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

Mlxtend.frequent patterns - mlxtend. (n.d.). Rasbt.github.io. https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/

G. Other Sources

Common questions about duloxetine. (2022, February 17). Nhs.uk. https://www.nhs.uk/medicines/duloxetine/common-questions-about-duloxetine/

Duloxetine: MedlinePlus Drug Information. (2020, March). Medlineplus.gov. https://medlineplus.gov/druginfo/meds/a604030.html

Garg, A. (2018, September 3). Complete guide to Association Rules (1/2). Towards Data Science; Towards Data Science. https://towardsdatascience.com/association-rules-2-aa9a77241654

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

Gbemudu, A. (2018). The Comprehensive List of Antidepressants. RxList. https://www.rxlist.com/the_comprehensive_list_of_antidepressants/drugs-condition.htm

WhatIs.com. What is market basket analysis? Definition from WhatIs.com. (n.d.). SearchCustomerExperience. https://www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis