Andreas Sebastian Dunn

Western Governor's University, D212: Data Mining II

Feburary 15, 2022

Task 3: Market Basket Analysis

A description of the scenario this paper is focused on:

"In the medical industry, readmission of patients is such a problem that an external organization penalizes hospitals for excessive readmissions (Centers for Medicare and Medicaid Services or CMS). When it comes to readmission penalties, studies show that many hospitals are overconfident and underprepared. The percentage of hospitals penalized for readmissions has increased each year since CMS began imposing penalties, and according to the CMS reporting, as much as 78 percent of hospitals were fined in fiscal year 2015. However, three-quarters of hospitals feel confident in their ability to reduce readmissions, and only 55 percent of them anticipate receiving a penalty this year. Given the historical trend and the addition of COPD and Hip and Knee replacement to the list of medical conditions measured, the percentage of hospitals penalized will likely be much higher than 55 percent. Additionally, although hospitals are applying various reduction strategies, fewer than 1 in 5 utilize technology that is specific to reducing their readmissions, so they may not be doing all that they can.

You are an analyst on a team of analysts for a popular medical hospital chain with patients in almost every state in the United States. Executives are interested in learning more about the typical prescriptions that are associated with patients. You have been asked to perform a market basket analysis on historical prescription data of patients for this chain of hospitals."

Part I

In our analysis we will be looking to answer the following business research question: which item sets (baskets of prescriptions for each patient) are associated with higher readmission rates to the hospital? Using metrics of association and pruning of the item sets, can we show some baskets of items that are correlated with readmission to the hospital?

The goal is to discover item sets that have some distinguishable level of association with readmission to the hospital. For instance, maybe certain prescriptions, or groups of prescriptions together, have a higher association with being readmitted than others. It would be important for the hospital chain to understand which prescriptions, or basket of prescriptions, could lead to higher readmission so that they can understand their patients risk level, improve service during initial admissions, or change financial incentives to lower readmission.

Part II

Market Basket works by finding groups of items associated with an observation. For our analysis, this means groups of prescriptions associated with a patient. Each observation has items associated with it and each item can be categorized as either an antecedent item or a consequent item. Below is an example of one transaction:

```
[15]: #an example of 1 transaction, the first row in the df
      df.iloc[0]
```

```
[15]: 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
```

One assumption of the market basket technique is that the relationship between transaction items are not casual relationships but they are correlated relationships. For example, if a grocery store finds during Thanksgiving that people tend to buy turkey and potatoes together, one does not necessarily cause the other but both do occur together. Market basket analysis has its aim towards finding those groupings.

Part III

C1

```
[4]: #drop all rows with missing data
     df.dropna(how='all', inplace=True)
     df.head()
```

[4]:

	Presc01	Presc02	Presc03	Presc04	Presc05	Presc06	Presc07	Presc08	Presc09	Presc10	Presc11	Presc12
1	amlodipine	albuterol aerosol	allopurinol	pantoprazole	lorazepam	omeprazole	mometasone	fluconozole	gabapentin	pravastatin	cialis	losartan
3	citalopram	benicar	amphetamine salt combo xr	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	enalapril	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	paroxetine	allopurinol	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	abilify	atorvastatin	folic acid	naproxen	losartan	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
[5]: df.shape
```

```
[5]: (7501, 20)
```

```
[8]: #convert df into list for apriori to work

     df_into_list = []
     for i in range(0, 7501):
         df_into_list.append([str(df.values[i, j]) for j in range(0, 20)])
     df_new = pd.DataFrame(df_into_list)

     df_new.head()
```

[8]:

	0	1	2	3	4	5	6
0	amlodipine	albuterol aerosol	allopurinol	pantoprazole	lorazepam	omeprazole	mometasone	flucono
1	citalopram	benicar	amphetamine salt combo xr	nan	nan	nan	nan	
2	enalapril	nan	nan	nan	nan	nan	nan	
3	paroxetine	allopurinol	nan	nan	nan	nan	nan	
4	abilify	atorvastatin	folic acid	naproxen	losartan	nan	nan	

C2

```
[9]: df_new.to_csv(r'C:\Users\dre2\Desktop\WGU\D212\Task3\task3_exported.csv')
```

[12]:
```
from apyori import apriori

assoc_results = apriori(df_into_list, min_support = 0.0035, min_confidence = 0.25, min_lift = 3, min_length = 2)
```

[13]:
```
assoc_results = list(assoc_results)
print(assoc_results[0])
```

```
RelationRecord(items=frozenset({'alprazolam', 'acetaminophen'}), support=0.005865884548726837, ordered_statistics=[OrderedStatistic(items_base=frozenset(
```

[14]:
```
print(len(assoc_results))
```

```
78
```

C3

[16]:
```
for item in assoc_results:

    item_sets  = item[0]
    items = [x for x in item_sets]

    value0 = str(items[0])
    value1 = str(items[1])
    value2 = str(item[1])
    value3 = str(item[2][0][2])
    value4 = str(item[2][0][3])

    row = (value0, value1, value2, value3, value4)
    results.append(row)

labels = ['1', '2', 'Support', 'Confidence', 'Lift']
```

[17]:
```
rules = pd.DataFrame.from_records(results, columns = labels)
print(rules)
```

```
                1                   2                Support  \
0      alprazolam        acetaminophen   0.005865884548726837
1   acetaminophen           amlodipine   0.005065991201173177
2     hydrocodone           alprazolam   0.005732568990801226
3     clopidogrel   salmeterol inhaler   0.004532728969470737
4       glipizide  hydrochlorothiazide   0.007998933475536596
..            ...                  ...                    ...
73        abilify                  nan   0.003999466737768298
74        abilify                  nan   0.004532728969470737
75        abilify                  nan   0.003999466737768298
76        abilify                  nan  0.0037328356219170776
77        abilify                  nan   0.004399413411545127

            Confidence              Lift
0    0.3728813559322034   4.700811850163794
1    0.3220338983050847   4.506672147735896
2    0.3006993006993007   3.790832696715049
3   0.29059829059829057    4.84395061728395
4    0.2714932126696833   4.122410097642296
..                  ...                 ...
73  0.29702970297029707   3.023093354111531
74  0.28813559322033894  3.0228043143297376
75   0.4109589041095891  3.1714019956029094
76   0.3373493975903614   3.539101861993428
77  0.25984251968503935   4.350622187851519

[78 rows x 5 columns]
```

C4

```
[18]: #top 1 rule
      rules.iloc[0]
```

```
[18]: 1                      alprazolam
      2                    acetaminophen
      Support       0.005865884548726837
      Confidence      0.3728813559322034
      Lift             4.700811850163794
      Name: 0, dtype: object
```

```
[19]: #top 2 rule
      rules.iloc[1]
```

```
[19]: 1                    acetaminophen
      2                       amlodipine
      Support       0.005065991201173177
      Confidence      0.3220338983050847
      Lift             4.506672147735896
      Name: 1, dtype: object
```

```
[20]: #top 3 rule
      rules.iloc[2]
```

```
[20]: 1                      hydrocodone
      2                       alprazolam
      Support       0.005732568990801226
      Confidence      0.3006993006993007
      Lift             3.790832696715049
      Name: 2, dtype: object
```

Part IV

There are certain metrics to measure the strength of the results as shown above in the association rules table. Support is a number between 0 and 1 that shows us the frequency of a certain item set in the total number of transactions. Low support basically means that the rule associated with an item-set doesn't show up often, which is what we have here for our top three results. Confidence is the likelihood that a consequent will happen given an antecedent, it is calculated by taking number of transactions that have both antecedent and consequent and dividing by the number of transactions with only the antecedent. Our confidence is a bit better for our top three results, indicating a stronger relationship (occurring roughly 30% of the time).

Finally, the Lift metric measures the likelihood that we get an antecedent given the consequent and is calculated by diving the Confidence of (if A then B) divided by the Support for A.

Looking at the top performing rule (if alprazolam, then acetaminophen), we have a low level of support, indicating it doesn't appear very often, a decent level for confidence, indicating both drugs having a 1/3 chance of appearing together, and a Lift of 4.7, indicating that there is roughly a 4.7 times higher chance that these drugs will be prescribed together then if just alprazolam was prescribed.

Given the ambiguous nature of Market Basket Analysis (MBA), it is difficult to determine how this specific data mining technique can be associated with higher rates of readmission. This type of analysis is usually better suited to find buying habits so that organizations can update and refine their marketing strategies. Since the hospital is not selling these drugs and is just recording their patients drug information, our task is to find which rules are stronger and therefore which baskets are associated with higher rates of readmission. Given the fact that these baskets do not tell use anything about readmission rates, my suggestion would be for the hospital to consider alternative data science methods, like regression, to predict readmission rates and to focus their data mining efforts using alternate algorithms like k-means or support vector machines.

Web Sources:

https://medium.com/@fabio.italiano/the-apriori-algorithm-in-python-expanding-thors-fan-base-501950d55be9

https://www.section.io/engineering-education/apriori-algorithm-in-python/

All other sources from course lectures.