Market Basket Analysis

410885010 劉嘉蕎

目錄

一、 介紹

沃爾瑪曾經利用資料探勘技術分析會員資料，在這個過程中偶然發現週五晚上男性的結帳清單中啤酒和尿布具有高度相關性。經過深入解析，發現這個現象是因為年輕爸爸抓住到超市採購小朋友尿布的機會，並順手拿了幾罐啤酒好迎接將到來的週末。因此，沃爾瑪於是調整商品陳列[1]，將啤酒與尿布擺放在鄰近的位置。這樣的調整有助於提升商品的銷售率，使得啤酒和尿布的銷售率提高了 30%。

基於這個經驗，本專題預計使用關聯規則學習（association rule learning）分析顧客交易資料，找出熱門的購物組合。這些資訊可以幫助零售業者了解顧客的購物習慣與市場趨勢，做出明智的決策，例如改變商品擺放位置，或順勢推出促消組合以刺激營收。透過這些措施，零售業者可以更有效地經營店鋪，提高營收。

二、 資料集

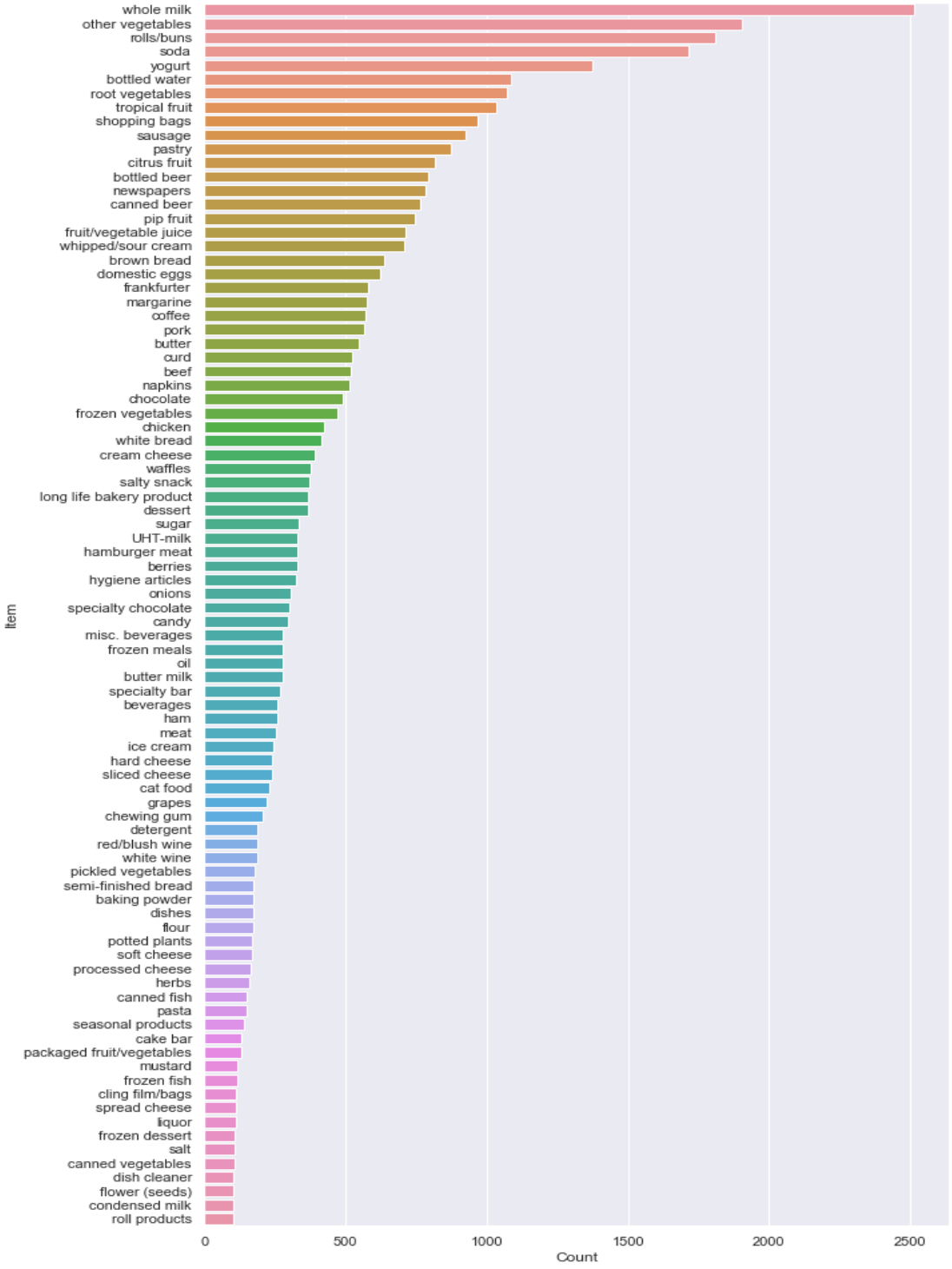
使用 kaggle 的 The Groceries Market Basket Dataset[2]，共包含 9835 筆交易資訊與 169 種商品。第一欄為此筆交易的購買總數，剩餘欄位為購買商品清單。

# Item(s)	Item 1	Item 2	Item 3	Item 4	Item 5
4	citrus fruit	semi-finished bread	margarine	ready soups	
3	tropical fruit	yogurt	coffee		
1	whole milk				
4	pip fruit	yogurt	cream cheese	meat spreads	
4	other vegetables	whole milk	condensed milk	long life bakery product	
5	whole milk	butter	yogurt	rice	abrasive cleaner
1	rolls/buns				
5	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)
1	potted plants				

三、 方法

A. 資料前處理

因為在總共 9835 筆資料中，有部分商品很少出現，若將較冷門的商品也納入關聯規則學習，會拉高 support 的分母，不利於找出關聯性高的熱門商品，所以先進行資料前處理，統計出購買次數≧100 的商品，如下圖所示。



接著將所有交易中，有出現購買次數<100 的商品交易紀錄全部刪除，只保留長條圖中商品所組成的交易紀錄。這樣總共剩下 7311 筆交易紀錄和 88 種商品，降為原始資料集的約 74%。

把剩餘的交易紀錄轉化為 one-hot vector 以進行關聯規則學習。

	UHT-milk	baking powder	beef	berries	beverages	bottled beer	bottled water	brown bread	butter	butter milk	...
0	False	False	False	False	False	False	False	False	False	False	...
1	False	False	False	False	False	False	False	False	False	False	...
2	False	False	False	False	False	False	False	False	False	False	...
3	False	False	False	False	False	False	False	False	False	False	...
4	False	False	False	False	False	False	False	False	False	False	...
...	...	...	...	...	...	...	...	...	...	...	...
7306	False	False	False	False	False	False	False	False	False	False	...
7307	False	False	False	False	False	False	False	False	False	False	...
7308	False	False	False	False	False	False	False	False	True	False	...
7309	False	False	True	False	False	False	False	False	True	False	...
7310	False	False	False	False	False	True	True	False	False	False	...

B. 關聯規則學習

關聯規則學習是一種常見的資料探勘技術，用於發現在大型資料集中隱藏的 Frequency Itemset 和關聯規則。因此，本專題使用 python 知名的資料科學套件 mlxtend [3]，來實現 Apriori Algorithm、Association Rules 等等常見的算法。

資料前處理僅保留購買次數≧100 的商品所組成的交易資料，所以計算兩個商品同時購買的機率，同樣只保留同時購買次數≧100 的 itemset。經實驗發現將閥值再提高一倍的效果較好，所以設定最小 support = (100/7311) ^ 2 * 2 ≒ 0.0037。

```
frequent_itemsets = apriori(df, min_support=0.0037, use_colnames=True)
```

關聯規則學習的另一個常見指標為 confidence，代表購買某商品 A 的前提下購買另一商品 B 的機率。設定 confidence(A, B) > 0.2。

```
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.2)
```

僅用 confidence 挖掘商品關聯性的最大問題是，當 A、B 兩項商品本身就很熱門，那麼 confidence(A, B)就會很高，無法反映這兩項商品同時出現的獨立性。所以新增：

- conviction(A, B) < 1.2，代表購買 A 卻沒有購買 B 與同時購買 A、B 的比例須 < 1.2。
- 由於 whole milk、rolls/buns、soda、yogurt 與 bottled water 本來就是熱門商品，故先將這五種商品與剩餘商品的 itemset 移除。
- other vegetables 根據統計也是熱門商品，但是此類別無法反映顧客購買傾向所以也先移除。

```
rules[  (rules['antecedent_len'] == 1) &
        (rules['conviction'] < 1.2) &
        (rules['antecedents'] ≠ {'whole milk'}) &
        (rules['consequents'] ≠ {'whole milk'}) &
        (rules['antecedents'] ≠ {'other vegetables'}) &
        (rules['consequents'] ≠ {'other vegetables'}) &
        (rules['antecedents'] ≠ {'rolls/buns'}) &
        (rules['consequents'] ≠ {'rolls/buns'}) &
        (rules['antecedents'] ≠ {'soda'}) &
        (rules['consequents'] ≠ {'soda'}) &
        (rules['antecedents'] ≠ {'yogurt'}) &
        (rules['consequents'] ≠ {'yogurt'}) &
        (rules['antecedents'] ≠ {'bottled water'}) &
        (rules['consequents'] ≠ {'bottled water'}) ]
```

最後共剩餘七項 itemset。

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
66	(flour)	(citrus fruit)	0.012173	0.071673	0.002462	0.202247	2.821811	0.001590	1.163678
68	(citrus fruit)	(tropical fruit)	0.071673	0.091780	0.017371	0.242366	2.640746	0.010793	1.198760
97	(flour)	(root vegetables)	0.012173	0.089181	0.002872	0.235955	2.645809	0.001787	1.192102
138	(hard cheese)	(sausage)	0.018465	0.085488	0.003967	0.214815	2.512818	0.002388	1.164709
171	(oil)	(root vegetables)	0.020654	0.089181	0.004651	0.225166	2.524824	0.002809	1.175502
211	(pip fruit)	(tropical fruit)	0.063603	0.091780	0.015456	0.243011	2.647767	0.009619	1.199780
238	(salt)	(root vegetables)	0.008617	0.089181	0.002052	0.238095	2.669807	0.001283	1.195450

四、 討論與結論

由於進行在資料前處理時已知全脂牛奶(whole milk)、麵包類(rolls/buns)、汽水(soda)、優格(yogurt)與瓶裝水(bottled water)這五項商品的熱門度遠高於其他商品(count > 1000)，所以建議零售商家可直接將這五項商品放在顯眼位置，甚至可以在商場的多個位置同時擺放，刺激消費數量。

另外通過關聯規則學習能找出七個商品之間的關聯性，可以總結成四條商品擺放位置或促銷組合的建議：

- No. 68, 211 的 itemset：將柑橘類水果(citrus fruit)、梨類水果(pip fruit)與熱帶水果(tropical fruit)擺放在一起。
- No. 97, 171, 238 的 itemset：在麵粉(flour)、油(oil)與鹽(salt)旁擺放根菜(root vegetable)，或是在根菜旁擺放這三項商品。
- No. 66 的 itemset：麵粉(flour)和柑橘類水果(citrus fruit)可以組合促銷。
- No.138 的 itemset：硬起司(hard cheese)和香腸(sausage)可以組合促銷。

五、 參考資料

[1] https://dataholic.wordpress.com/2016/11/03/一個門外漢的資料科學學習之旅/

[2] https://www.kaggle.com/datasets/irfanasrullah/groceries

[3] http://rasbt.github.io/mlxtend/