TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Analýza nákupního košíku v jazyce Python

Semestrální práce

Václav Kesler
Liberec 2022

Obsah

Zadání

Obchodní řetězec potřebuje analyzovat nákupní zvyklosti zákazníků, aby bylo možné provádět cílené nabídky určitého typu zboží. Úloha má dva cíle 1. vytvořit akci na prodej konkrétního zboží (nabídnout nebo nenabídnout) a 2. vytvořit doporučení jaké zboží zákazníkovi ještě nabídnout. Jde o úlohy tzv. křížového a následného prodeje které se vyskytují v prostředí všech společností, které nabízí zákazníkům své zboží nebo služby.

Data

Po průchodu zákazníka pokladnou vznikají tzv. transakční data která jsou uložená v souboru “Shopping_items.sav”. Jednotlivé řádky datové matice obsahují identifikátor košíku a položku, která byla zakoupena. Řádky se stejným identifikátorem tvoří košíky, tedy samostatný nákup určitého zákazníka. Datová matice obsahuje celkem 2394 položek a ukázku počátečních a koncových řádků můžeme vidět na obrázku číslo jedna.

	ID	ITEM
0	1.0	READMADE
1	1.0	SNACKS
2	2.0	READMADE
3	2.0	TOILETRY
4	3.0	READMADE
...	...	...
2390	784.0	ALCOHOL
2391	785.0	FROZEN
2392	785.0	ALCOHOL
2393	785.0	SNACKS
2394	786.0	FROZEN

Obrázek 1 - Datová matice

Příprava prostředí

Byl stažen python z oficiálních stránek a nainstalován. Byla ověřena instalace pomocí příkazu "python --version" v příkazové řádce, jako je možné vidět na obrázku číslo dva.

```
C:\WINDOWS\system32>python --version
Python 3.9.5
```

Obrázek 2 - Ověření instalace

Dalším krokem bylo vytvoření virtuálního prostředí. V příkazové řádce pomocí příkazu "python -m venv env" bylo prostředí vytvořeno. Následně bylo vytvořené prostředí spuštěno skriptem ve složce “env\Scripts\activate.bat”. Projekt obsahuje soubor requirements.txt, ve kterém jsou zapsané všechny potřebné knihovny ke spuštění projektu. Příkazem “pip install -r .\requirements.txt” správce balíčků pip nainstaluje všechny potřebné závislosti.

Příkazem “jupyter notebook” se spouští server který na adrese http://localhost:8888/ zpřístupní prostředí. Samostatná práce se pak nachází v souboru “Semestral_Kesler.ipynb”

Import a příprava dat

Data v *.sav* formátu byla importována pomocí knihovny pyredstat. Naimportovaná data jsou uložené do takzvaného data framu, což je datová struktura z knihovny *pandas*, která je vlastně tabulka s popsanými řádky a sloupci. Naimportovaná data jsou vidět na obrázku číslo jedna.

Naimportovaná data jsou ve formátu záznamů nákupů a každá položka nákupu má vlastní řádek. Jednotlivé nákupy byly pro další zpracování potřebné agregovat. Z toho vyplývá, že místo každé položky na řádek bude mít každý řádek seznam položek. To znamená, že každý řádek je jeden nákupní košík. Ukázka prvních tří košíků je na obrázku číslo tři. Data byla převedena z data framu do pole polí.

```
['READMADE', 'SNACKS']
['READMADE', 'TOILETRY']
['READMADE', 'TOILETRY', 'SNACKS']
```

Obrázek 3 - Košíky

Pomocí knihovny *mlxtend* konkrétně *TransactionEncoderu* se z vytvořeného pole polí byla vytvořena matice příznaků. Každému druhu zboží byl přiřazen vlastní sloupec a hodnota nabývá pouze true/false. Hodnoty true nabývá, pokud byl druh zboží v nákupním košíku. Ukázku ze začátku a konce matice příznaků je možné vidět na obrázku číslo čtyři.

	ALCOHOL	BAKERY	FROZEN	MEAT	MILK	READMADE	SNACKS	TINNED	TOILETRY	VEG
0	False	False	False	False	False	True	True	False	False	False
1	False	False	False	False	False	True	False	False	True	False
2	False	False	False	False	False	True	True	False	True	False
3	False	True	False	False	True	True	False	False	False	False
4	False	False	False	False	False	True	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
780	True	False	True	False	False	False	False	True	True	False
781	True	False	False	False	False	False	False	False	False	False
782	False	True	True	False	False	True	True	False	False	False
783	True	False	False	False	False	False	False	False	False	False
784	True	False	True	False	False	False	True	False	False	False

Obrázek 4 - Matice příznaků

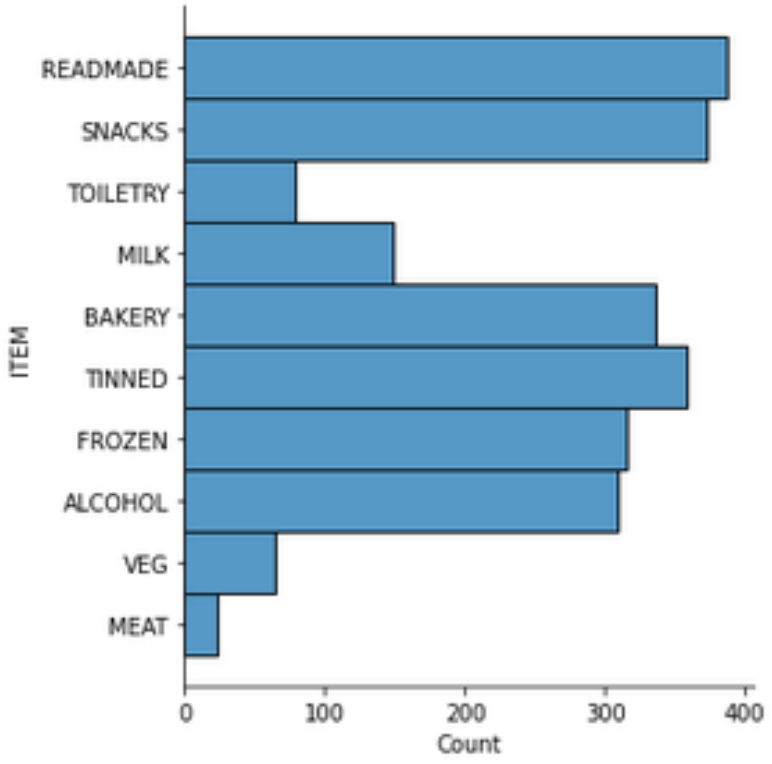
Porozumění datům

Z připravených dat bylo možné vypočítat procentuální zastoupení jednotlivých druhů zboží v košíkách. Na obrázku číslo pět vidíme, že maso se vyskytovalo pouze v 2.9% košících. Z celkových 784 košíků bylo maso pouze ve 23 z nich.

ITEM	Počet	Procenta
ALCOHOL	310	39.4
BAKERY	337	42.9
FROZEN	316	40.2
MEAT	23	2.9
MILK	148	18.8
READMADE	387	49.2
SNACKS	373	47.5
TINNED	358	45.5
TOILETRY	78	9.9
VEG	65	8.3

Obrázek 5 - Procentuální zastoupení skupin zboží v celkovém počtu košíků

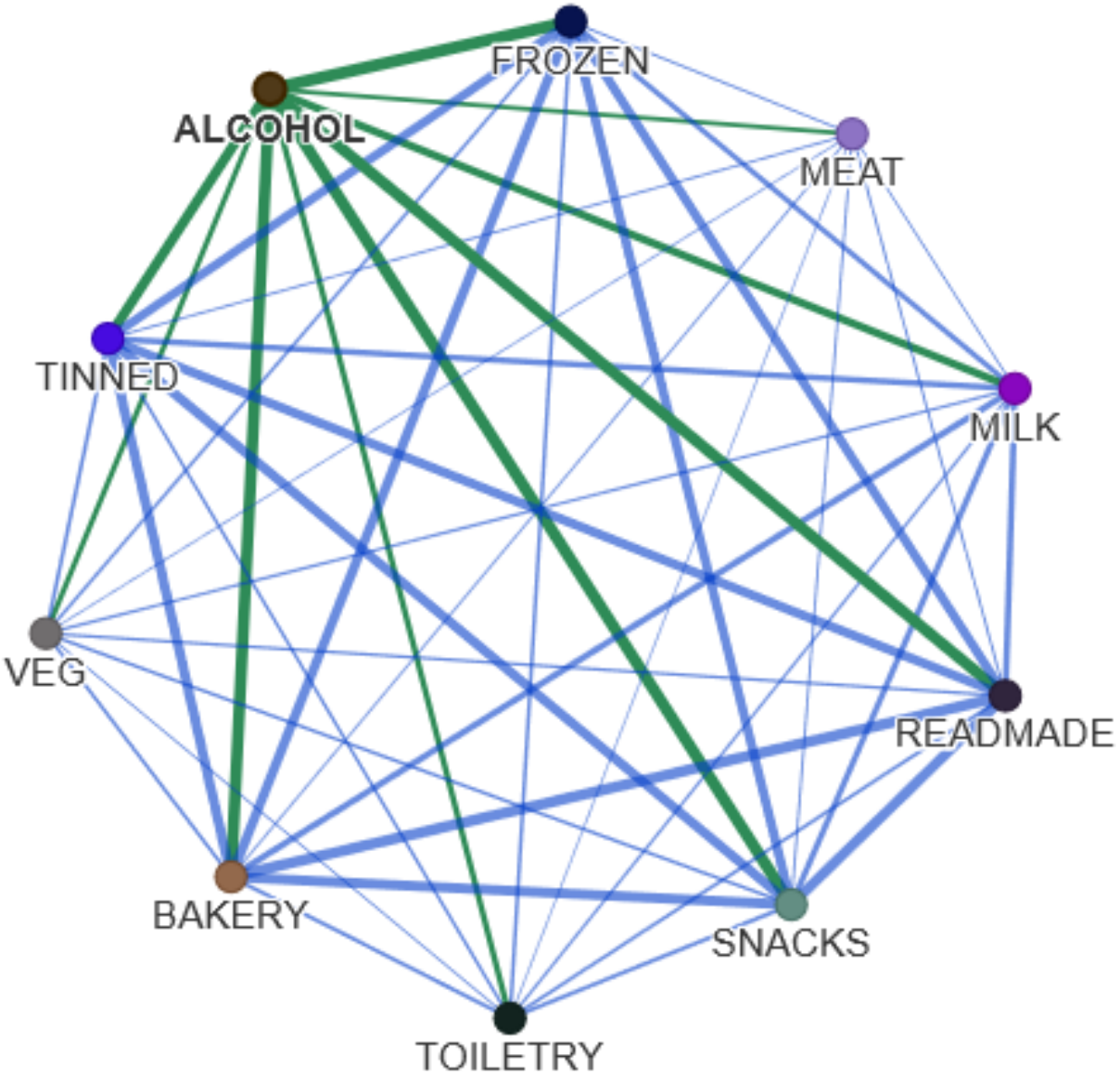
Pro lepší vizuální představu nad rozložením dat byl vytvořena distribuce. Tento graf byl vytvořen pomocí knihovny *seaborn,* která byla zavolána nad původním data framem. Graf je vidět na obrázku číslo šest. Mléko, zelenina, maso a drogerie má oproti ostatním položkám malé zastoupení v nákupních košících.



Obrázek 6 - Distribuce

Pro pochopení závislostí mezi jednotlivými skupinami zboží byl sestrojen pomocí knihovny *pyvis* pavučinový graf. Graf je možné vidět na obrázku číslo sedm. *Pyvis* je postaven na JavaScript knihovně *VisJS* a proto výsledný graf je webová stránka. V projektu byla pojmenována web.html.

Každá skupina zboží má vlastní bod ze kterého vede k ostatním bodům hrana jejíž velikost se odvíjí podle počtu košíků ve kterých byly obě položky. Graf je interaktivní a je možné klikat na jednotlivé body. Po kliknutí na bod jsou obarveny hrany vedoucí z tohoto bodu zelenou barvou. Na obrázku číslo sedm je možné vidět, že alkohol byl ve větším počtu košíku s pečivem než třeba s mlékem.



Obrázek 7 - Pavučinový graf

Modelování

Knihovna *mlxtend* obsahuje modul s názvem *apriori*, který z matice příznaků generuje frekventované množiny. Generování frekventovaných množin bylo omezeno požadavkem na support větší než deset procent.

Celkem bylo vygenerováno 52 množin a obrázku číslo osm se nachází prvních deset množin a jejich support, tedy procentuální zastoupení v celkovém počtu košíků. Při porovnání s obrázkem číslo pět můžeme vidět, že se do frekventovaných množin nedostaly toaletní potřeby, maso a zelenina, jelikož jejich support byl menší než deset procent.

	support	itemsets
0	0.394904	(ALCOHOL)
1	0.429299	(BAKERY)
2	0.401274	(FROZEN)
3	0.188535	(MILK)
4	0.492994	(READMADE)
5	0.475159	(SNACKS)
6	0.456051	(TINNED)
7	0.215287	(ALCOHOL, BAKERY)
8	0.230573	(ALCOHOL, FROZEN)
9	0.114650	(ALCOHOL, MILK)
10	0.212739	(ALCOHOL, READMADE)

Obrázek 8 - Frekventované množiny

Z frekventovaných množin byla vytvořena asociační pravidla pomocí knihovny *mlxtend* konkrétně modulu *association_rules.* Ukázku prvních a posledních asociačních pravidel můžeme vidět na obrázku číslo deset. Tabulka je seřazena od největší confidence po nejmenší. Při vytváření pravidel byl požadavek na confidenci větší než padesát procent.

Confidence znamená v kolika košících kde byl splněn předpoklad (antecedent), byl také přítomen závěr (consequent). Support je rozdělen do tří sloupců. Support pro předpoklad, support pro závěr a support pro předpoklad i závěr. Dále tabulka obsahuje sloupec lift, který udává poměr confidence a consequent support. Leverage celkový support minus vynásobený support předpokladu a závěru. Conviction odpovídá sloupci deployability v modeleru a odpovídá vzorci na obrázku číslo devět.

$$\text{conviction}(A \rightarrow C) = \frac{1 - \text{support}(C)}{1 - \text{confidence}(A \rightarrow C)}, \quad \text{range: } [0, \infty]$$

Obrázek 9 - Rovnice výpočtu conviction

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALCOHOL, READMADE, TINNED)	(BAKERY)	0.121019	0.429299	0.100637	0.831579	1.937061	0.048684	3.388535
1	(ALCOHOL, BAKERY, TINNED)	(READMADE)	0.123567	0.492994	0.100637	0.814433	1.652015	0.039719	2.732201
2	(READMADE, MILK)	(BAKERY)	0.133758	0.429299	0.105732	0.790476	1.841317	0.048310	2.723798
3	(TINNED, MILK)	(BAKERY)	0.127389	0.429299	0.100637	0.790000	1.840208	0.045949	2.717622
4	(ALCOHOL, READMADE, SNACKS)	(BAKERY)	0.135032	0.429299	0.101911	0.754717	1.758020	0.043942	2.326703
...	...	...	...	...	...	...	...	...	...
97	(BAKERY)	(FROZEN)	0.429299	0.401274	0.221656	0.516320	1.286703	0.049389	1.237857
98	(SNACKS)	(READMADE)	0.475159	0.492994	0.244586	0.514745	1.044122	0.010336	1.044825
99	(TINNED, SNACKS)	(FROZEN)	0.224204	0.401274	0.114650	0.511364	1.274351	0.024683	1.225300
100	(BAKERY)	(ALCOHOL)	0.429299	0.394904	0.215287	0.501484	1.269886	0.045754	1.213793
101	(TINNED)	(BAKERY)	0.456051	0.429299	0.228025	0.500000	1.164688	0.032243	1.141401

Obrázek 10 - Asociační pravidla

Nasazení

Nasazení je už řešení jednotlivých úloh ze zadání. Byl vytvořen košík, který simuluje nákupní košík zákazníka na kterém je prováděna anlýza. První úloha požadovala zda nabídnout nebo nenabídnout zákazníkovi alkohol. Vytvořená asociační pravidla byla vyfiltrována podle řádků jejichž závěr (consequent) byl alkohol. Poté byla tato pravidla for cyklem procházena a pokud je pravidlo podmnožinou košíku a není v košíku alkohol je zákazníkovi nabízen alkohol. Tato skutečnost je reprezentována vytisknutím “Nechcete ještě alkohol?” jak je možné vidět na obrázku číslo jedenáct. V košíku měl zákazník mléko, plechovky a noviny. Pokud by nebylo nalezeno žádné pravidlo, které by splňovalo podmínku výše, nic se nevypíše.

```
Nechcete ještě alkohol?
```

Obrázek 11 - Nabídka alkoholu

Druhá úloha požadovala doporučení konkrétního zboží zákazníkovi. For cyklem byla procházena všechna vytvořená asociační pravidla. Pokud je předpoklad (antecedent) podmnožinou košíku a závěr (consequent) není ještě doporučený a ani v košíku. Vypíše se doporučení. Po třech doporučeních je výpis ukončen. Může nastat situace kdy se nic nevypíše, například pokud zákazník má již všechny druhy zboží v košíku. Příklad výstupu pro košík ve kterém bylo mléko, plechovky a noviny je možné vidět na obrázku číslo dvanáct.

```
1 . Nechcete také koupit: BAKERY | Confidence: 0.7904761904761906
2 . Nechcete také koupit: SNACKS | Confidence: 0.6621621621621621
3 . Nechcete také koupit: ALCOHOL | Confidence: 0.6081081081081081
```

Obrázek 12 - Doporučení skupin zboží

Závěr

Byl vytvořen model, který řeší obě zadané úlohy, tedy úlohu rozhodování zda doporučit produkt a úlohu, které produkty mají být zákazníkovi doporučeny. Toto řešení by mohlo pomoci například internetovému obchodu ke zvýšení prodejů. Třeba pokud by na základě tohoto modelu bylo na pokladně zobrazeno další zboží z doporučené kategorie.

Při řešení úloh jsem se nesetkal s většími problémy. Největší překážkou bylo vytvoření pavučinového grafu, jelikož bylo potřeba vytvořit speciální nastavení aby byl graf dobře čitelný. Řešení odpovídá úloze řešené v modeleru na hodině. Asociační pravidla se shodují s těmi vytvořeními v modeleru pomocí bloku CARMA. Pro získání stejných pravidel jaké byly vytvořené v modeleru pro algoritmus apriori je nutné nastavit nižší úroveň supportu, při vytváření frekventovaných množin. Knihovna mlxtend nastavuje hodnotu support, která neodpovídá supportu v modeleru. Support v modeleru odpovídá sloupci antecedent support v mlxtend, který se nedá nastavit při vytváření frekventovaných množin.