ESCOLA POLITÈCNICA SUPERIOR
UNIVERSITAT DE LLEIDA

Master's Degree in Informatics Engineering

Data Mining

Universitat de Lleida

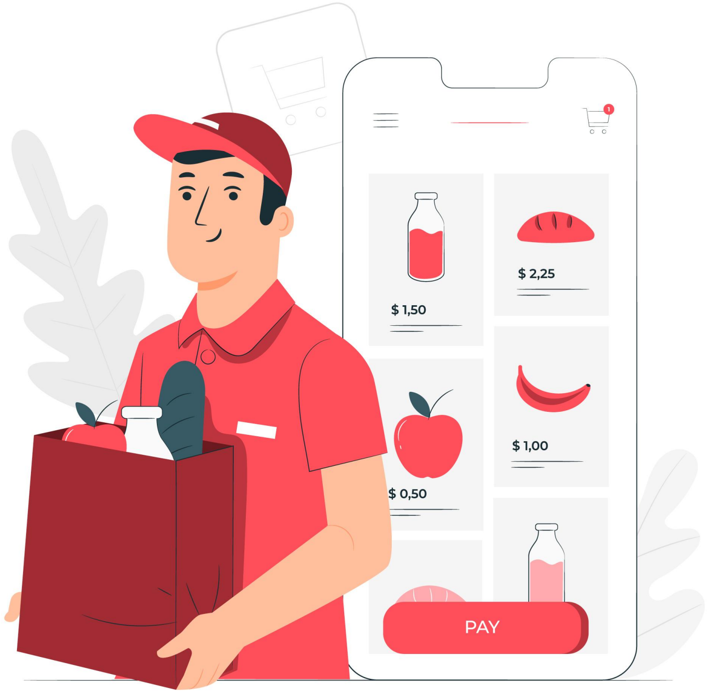
Recommending forgotten products

Grocery stores

Jeongyun Lee
Sergi Trujillo Agramunt

INDEX

1. Introduction



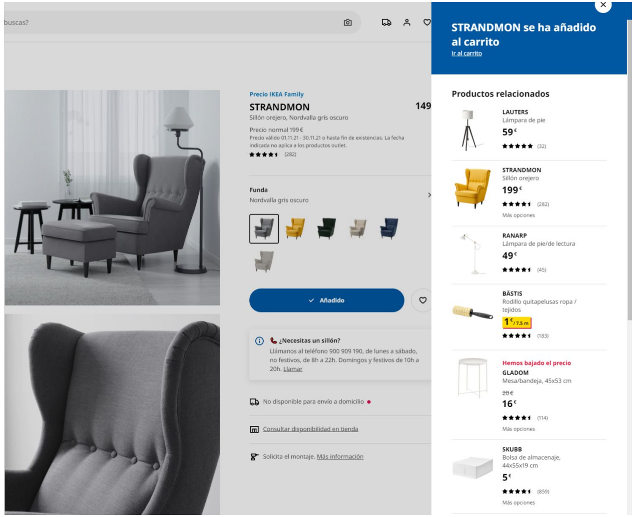
Introduction

1. Some ingredient can be forgotten for their recipes. How can we help them?

2. When they pay on the counter, can we recommend products that they need buy?

Tech & Business Goal

1. Analyze the sales history and predict associated product.
2. Present a recommendation solution considering Client environment.



1. Present user-oriented recommending system to increase the customer loyalty of the store.
2. Increase the number of potential customers and long-term sales performance

Steps to Follow

1. Data Exploration and Cleaning
2. Choose the recommendation systems
3. Determine methodology
4. Data Modeling
5. Training and testing
6. Analyze the results

2. DATA EXPLORATION

Data Exploration

Families

```
FAMILIA ;DESCRIPCIO.SECTOR ; DESCRIPCIO.SECCIO  ;DESC.FAMILIA ;;;;
01*01*01 ;Alimentacion y Bebidas  ;ALIMENTACIÓN SECA ;Aceites ;;;;
01*01*02 ;Alimentacion y Bebidas  ;ALIMENTACIÓN SECA ;Cafés y sucedáneos ;;;;
...
01*02*01 ;Alimentacion y Bebidas  ;CONSERVAS ;Conservas de pescado y marisco ;;;;
01*02*02 ;Alimentacion y Bebidas  ;CONSERVAS ;Conservas vegetales ;;;;
...
01*03*01 ;Alimentacion y Bebidas  ;Làcteos y derivados ;Leche ;;;;
01*03*02 ;Alimentacion y Bebidas  ;Làcteos y derivados ;Leches no liquidas ;;;;
...
01*04*01 ;Alimentacion y Bebidas  ;BEBIDAS ;Aguas ;;;;
01*04*02 ;Alimentacion y Bebidas  ;BEBIDAS ;Bebidas refrescantes ;;;;
```

- Family Description
- Family
- Section
- Sector

Family, Section, Sector } famiy_id

```
total families:  816
```

	family_id	family_desc
0	00*00*00	Desconeguda
1	01*01*01	Aceites
2	01*01*02	Cafés y sucedáneos
3	01*01*03	Infusiones
4	01*01*04	Chocolates

AECOC
La Asociación de Fabricantes y Distribuidores

Families

Data Exploration

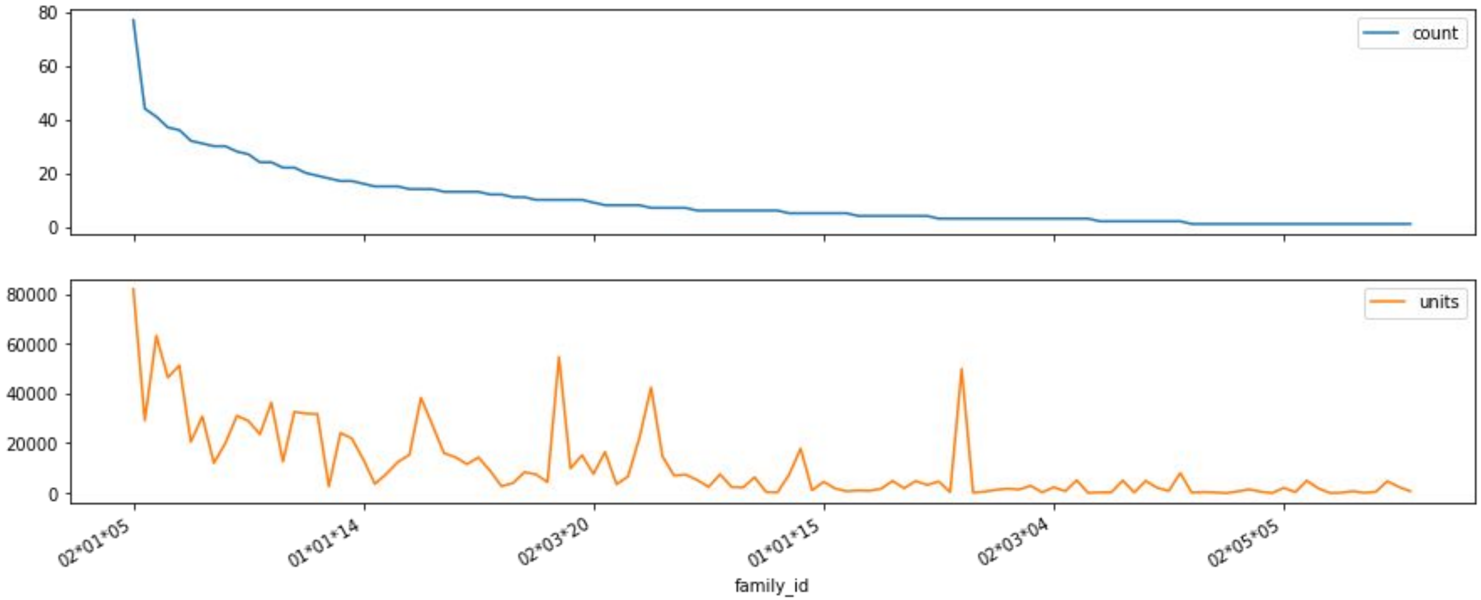
family_id	product_id	units	count	percent_units	family_desc
02*01*05	{3599, 5164, 5165, 5169, 5174, 5175, 5177, 517...	81842.0	77	6.567467	Aves
02*01*02	{4481, 4354, 4483, 4355, 4486, 4487, 4488, 448...	63167.0	41	5.068879	Porcino
01*03*01	{7552, 7558, 7870, 8492, 8493, 7796, 7797, 789...	54684.0	10	4.388155	Leche
02*01*10	{5890, 5895, 4493, 4496, 6164, 6165, 6166, 681...	51250.0	36	4.112591	Elaborados Frescos
03*01*07	{6969, 9117, 8055}	49840.0	3	3.999445	Utiles de limpieza
01*01*20	{2437, 2442, 2443, 2445, 2447, 2448, 2449, 245...	46409.0	37	3.724122	Frutos secos y fruta seca
02*01*07	{6215, 6251, 6252, 6253, 6255, 3248, 6330}	42370.0	7	3.400009	Huevos
02*05*01	{3072, 8229, 8202, 1642, 1647, 1648, 3091, 306...	38234.0	14	3.068113	Postres
02*03*06	{6920, 6170, 6940, 7069, 6182, 7079, 7081, 708...	36321.0	24	2.914603	Fuet
02*09*01	{8348, 8095, 8485, 8121, 1602, 1603, 1610, 161...	32615.0	22	2.617213	Curado, semi y tierno
02*06*02	{2688, 2562, 2564, 2567, 2707, 2716, 2880, 288...	31907.0	20	2.560399	Verduras y hortalizas
01*02*01	{8225, 7587, 7588, 8232, 7602, 7603, 7605, 760...	31711.0	19	2.544671	Conservas de pescado y marisco
02*06*01	{2176, 2309, 2053, 2056, 2060, 2189, 2210, 208...	31029.0	28	2.489943	Frutas
02*02*02	{3585, 6149, 6150, 6151, 6152, 3474, 5919, 592...	30778.0	31	2.469802	Base pasta y arroz
01*01*13	{7683, 8456, 7690, 8337, 8468, 8596, 8598, 911...	29170.0	44	2.340766	Pastelería y bollería industrial

Work with Families is a good idea?

Data Exploration

Families

Relation of number of products in families and units sold



Data Exploration

Products

total products: 72200

	product_id	family_id	product_desc
0	10000	03*01*06	LIMPIACRISTALES CON
1	10001	03*01*06	LIMPIACRISTALES RECA
2	10002	03*01*06	LIMPIACRISTALES BONA
3	10003	03*01*06	LIMPIACRISTALES BONA
4	10004	03*01*06	FREGASUELOS CAG 1 L

```
ARTICLE;DESCRIPCIO;SECTOR;SECCIO;FAMILIA;DESC
10002;LIMPIACRISTALES BONA;3;1;6;03*01*06
10003;LIMPIACRISTALES BONA;3;1;6;03*01*06
10004;FREGASUELOS CAG 1 L;3;1;6;03*01*06
10006;VAJILLAS CONCENTRADO;3;1;4;03*01*04
10007;VAJILLAS VERDE BONAC;3;1;4;03*01*04
...
```

- Family Code
- Product Description
- Product Code

Families — Products — Sales

Products

Data Exploration

```
products.nunique()
```

```
product_desc    34217
family_id         450
dtype: int64
```

⚠️ Products file: 72,200 rows
⚠️ Families file: 720 rows

```
products[products['product_desc'].isnull()]
```

product_id	product_desc	family_id
307355	NaN	60*04*04
307356	NaN	60*04*04
307357	NaN	60*04*04
307358	NaN	60*04*04
307365	NaN	60*04*03
...	...	...
313380	NaN	40*01*01
313383	NaN	40*01*03
313384	NaN	50*01*08
313386	NaN	40*03*03
313388	NaN	60*04*01

555 rows × 2 columns

Data Exploration

Sales

One store 2019 ~= 2020

```
2027-T0101C01-100089;3055;1000000;24;01;19343;35707
2027-T0101C01-100089;3055;1000000;24;01;19343;35707
2027-T0101C01-100089;6989;1000000;0;01;19343;35707
2027-T0101C01-100188;8939;1000000;226;01;19343;44143
2027-T0101C01-100188;8939;1000000;228;01;19343;44143
...
```

- Hour
- Date
- Checkout
- Amount (x10^2)
- Units (x10^6)
- Product Code
- Invoice number
- Store number
- Year

Data Exploration

Sales

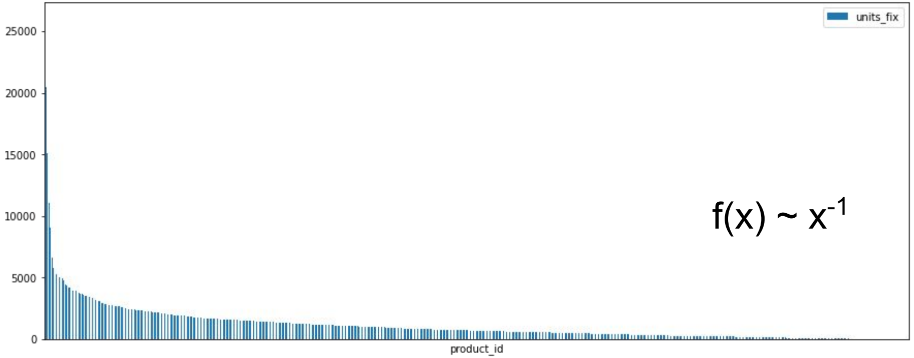
	invoice_id	product_id	quantity	amount	checkout	datetime
0	2027-T0105C01-100089	5379	1.0	1.71	1	2020-10-15 17:53:16
1	2027-T0105C01-100089	5379	1.0	1.60	1	2020-10-15 17:53:16
2	2027-T0105C01-100089	3482	1.0	0.63	1	2020-10-15 17:53:16
3	2027-T0105C01-100089	3059	1.0	0.45	1	2020-10-15 17:53:16
4	2027-T0105C01-100089	3059	1.0	0.45	1	2020-10-15 17:53:16

Number of rows: 8,096,494
Different products sold: 1,000
Without nulls and nan

Data Exploration

Sales

Top Sales



product_id	units_fix	product_desc
9117	26066.0	BOLSAS CAMISETA GALG
8055	23754.0	BOLSAS CAMISETA CON
7550	20488.0	LECHE ENTERA BONAREA
7551	15087.0	LECHE SEMIDESNATADA
7665	14994.0	AGUA BONï¿½REA 1,5 L.
6253	14086.0	HUEVOS M RUBIO BONAR
6252	11105.0	HUEVOS L RUBIO BONAR

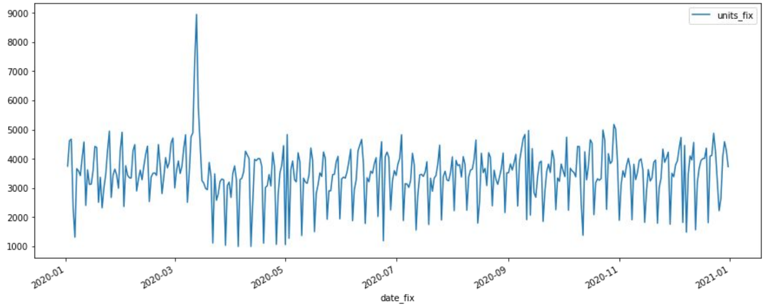
Remove from dataset:

product_id	units_fix	product_desc
9117	26066.0	BOLSAS CAMISETA GALG
8055	23754.0	BOLSAS CAMISETA CON
8419	439.0	BOLSAS RAFIA COLOR N

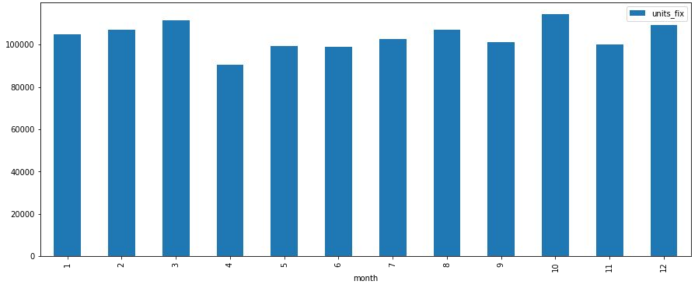
Data Exploration

Sells

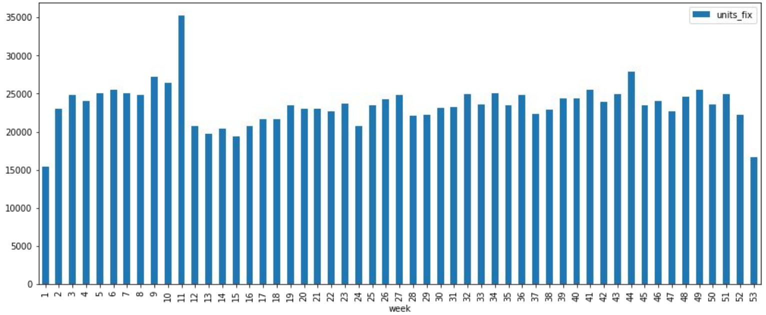
Sales by day



Sales by month



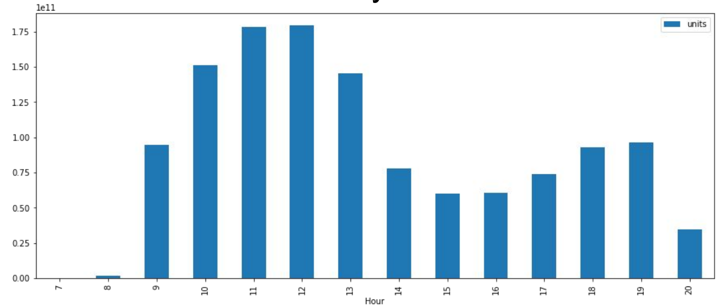
Sales by week



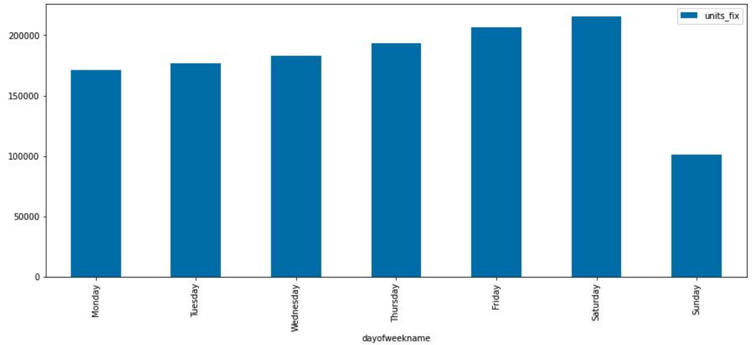
Data Exploration

Sales

Sales by hour



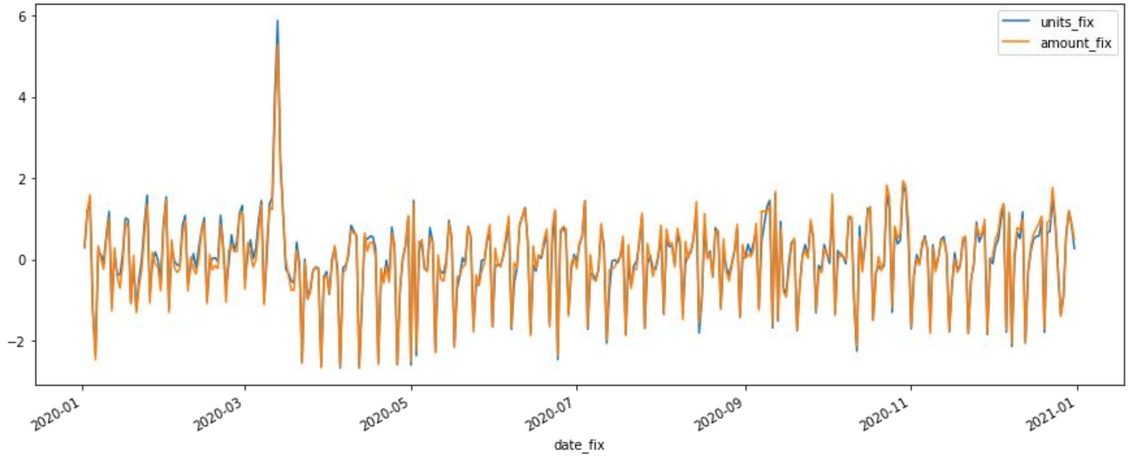
Sales by weekday



Data Exploration

Sales

Reduce dimensionality of units and amount



Pearsons

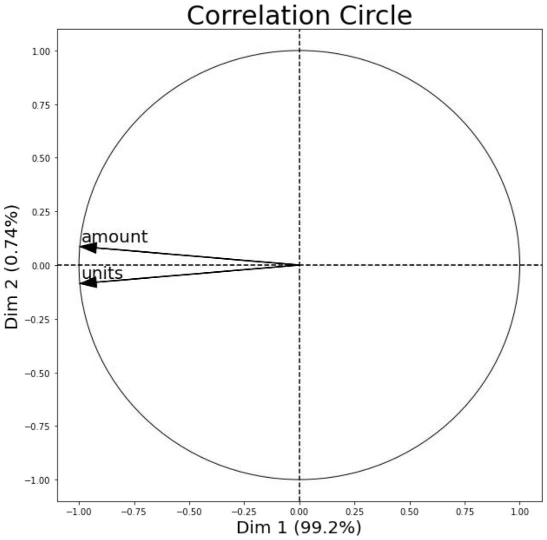
	units_fix	amount_fix
units_fix	1.000000	0.985044
amount_fix	0.985044	1.000000

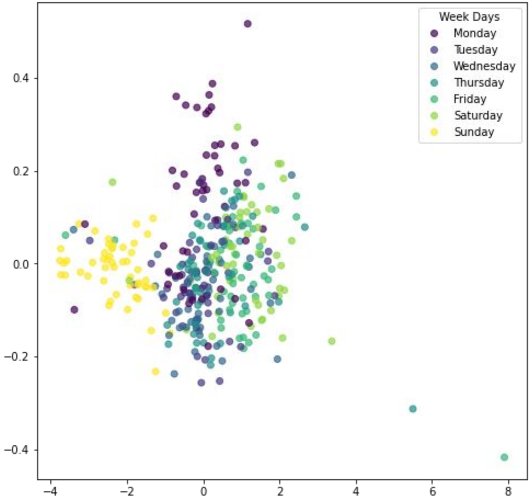
It's no necessary use amount and units to get the same meaning.
One of them is enough.

Data Exploration

Sales

Reduce dimensionality of units and amount





3. Methodology

Methodology

Try 3 recommendation systems

- **Data modeling**
- **Train** the model
- **Test** dataset : remove 1 product for each buyer
- **Apply** the model
- **Compare** the recommendation with the product removed
- → **Precision**

Methodology

- Benchmark algorithm : recommend a top 10 best-selling product

Precision : 0.8%

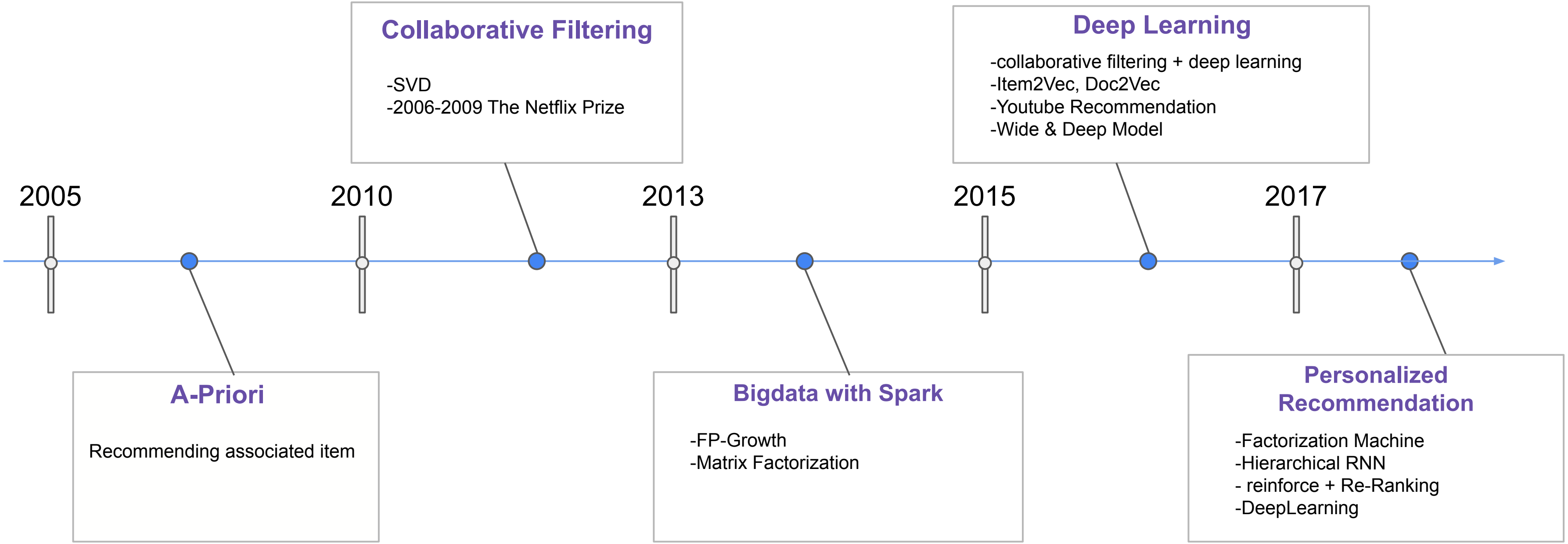
product_id	sum(quantity)	family_id	product_desc
7550	20488.0	01*03*01	LECHE ENTERA BONAREA
7551	15087.0	01*03*01	LECHE SEMIDESNATADA
7665	14994.0	01*04*01	AGUA BONï¿½REA 1,...
6253	14086.0	02*01*07	HUEVOS M RUBIO BONAR
6252	11105.0	02*01*07	HUEVOS L RUBIO BONAR
7552	9068.0	01*03*01	LECHE DESNATADA BONA
2111	8574.0	02*06*01	PLATANOS CANARIAS BO
6255	6727.0	02*01*07	HUEVOS DE CORRAL RUB
8127	6642.0	01*04*04	CERVEZA BONï¿½REA...
3060	5812.0	02*05*01	YOGUR NATURAL BONARE

Methodology

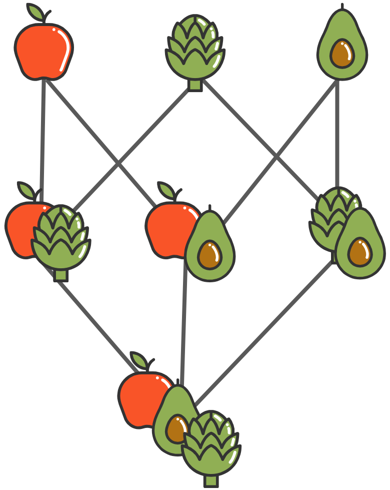
invoice_id	datetime	product_id	quantities	amounts	product_expected	product_without_expected
206	2020-01-02 13:16:45	[7551, 3046, 7696...	[6.0, 1.0, 1.0, 1...	[3.35, 1.29, 0.9,...	7551	[3046, 7696, 6473...
1159	2020-01-04 18:27:19	[2111, 4490, 4355...	[1.0, 1.0, 1.0, 1...	[1.72, 1.88, 2.05...	2111	[4490, 4355, 8510...
1202	2020-01-04 19:13:29	[2111, 6647, 6647...	[1.0, 1.0, 1.0, 1...	[2.32, 0.9, 0.9, ...	2111	[6647, 6647, 7596...
1245	2020-01-04 19:57:46	[6253, 6253, 6253...	[1.0, 1.0, 1.0, 1...	[0.99, 0.99, 0.99...	6253	[6253, 6253, 8345...
1614	2020-01-06 13:11:51	[6253, 7036, 5954]	[1.0, 1.0, 1.0]	[0.99, 2.65, 4.31]	6253	[7036, 5954]
1671	2020-01-07 10:48:35	[6255, 5588, 8202...	[1.0, 1.0, 1.0, 1...	[0.74, 1.35, 0.68...	6255	[5588, 8202, 5480...
1888	2020-01-07 18:00:24	[7552, 5572, 7075...	[1.0, 1.0, 1.0, 1...	[0.52, 1.53, 1.62...	7552	[5572, 7075, 7075...
2173	2020-01-08 13:27:23	[7551, 338, 5867,...	[6.0, 1.0, 1.0, 1...	[3.35, 0.96, 2.35...	7551	[338, 5867, 4475,...
2625	2020-01-09 15:01:08	[6252, 6252, 2454...	[1.0, 1.0, 1.0, 4...	[1.05, 1.05, 2.63...	6252	[6252, 2454, 3248...
3050	2020-01-10 14:16:00	[7550, 8092, 5852...	[1.0, 1.0, 1.0, 1.0]	[0.58, 1.71, 2.26...	7550	[8092, 5852, 6102]
3214	2020-01-11 10:28:09	[6252, 7620, 6893...	[1.0, 1.0, 1.0, 1...	[1.05, 0.71, 1.03...	6252	[7620, 6893, 7660...
3491	2020-01-11 19:03:47	[7551, 5378, 5378...	[2.0, 1.0, 1.0, 1...	[1.12, 1.7, 1.95,...	7551	[5378, 5378, 8110...
3797	2020-01-13 10:40:57	[2111, 5233, 5361]	[1.0, 1.0, 1.0]	[2.18, 3.06, 2.71]	2111	[5233, 5361]
4029	2020-01-13 16:37:01	[7550, 8595, 7582...	[6.0, 1.0, 1.0, 1...	[3.45, 1.7, 0.89,...	7550	[8595, 7582, 2562...
4765	2020-01-15 16:39:45	[6253, 7757, 207,...	[1.0, 1.0, 1.0, 1...	[0.95, 0.42, 4.34...	6253	[7757, 207, 7629,...
4919	2020-01-16 10:17:41	[2111, 8202, 8202...	[1.0, 1.0, 1.0, 1.0]	[2.4, 0.68, 0.68,...	2111	[8202, 8202, 324]
5310	2020-01-17 11:03:58	[6252, 5545, 5907...	[1.0, 1.0, 1.0, 1...	[1.05, 1.14, 0.75...	6252	[5545, 5907, 5907...
5818	2020-01-18 12:43:25	[7550, 7550, 350,...	[1.0, 1.0, 1.0, 1...	[0.58, 0.58, 1.94...	7550	[7550, 350, 231, ...
5839	2020-01-18 13:23:48	[6252, 4488, 4490...	[1 Ajouter une cellule de texte	[1.04, 0.53, 1.83...	6252	[4488, 4490, 2880...
5880	2020-01-18 14:04:41	[6255, 364, 4181,...	[1.0, 1.0, 1.0, 1...	[0.74, 0.97, 1.9,...	6255	[364, 4181, 3638,...

4.Algorithms

History of Recommendation System



FP-Growth



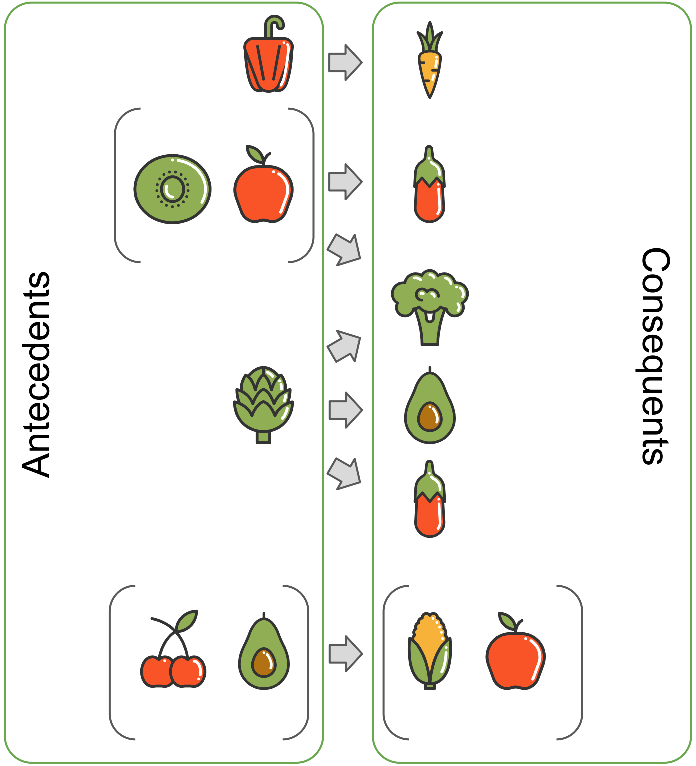
Improvement of Apriori algorithm.

Not requires candidates

Antecedent: Product/s origin of the relation
Consequent: Product/s derived from Antecedent
Support: Ratio of number of repeats of correlation with the total
Confidence: Is the ratio of one relation fit in its contexts. It's a great unit to measure the quality of the relation.

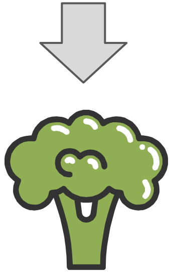
FP-Growth → Association rules → Recommendation

FP-Growth



Recommender:

argmax(∑ (consequent, confidence))



Precision : 2,4%

FP-Growth

observations

FP-Growth

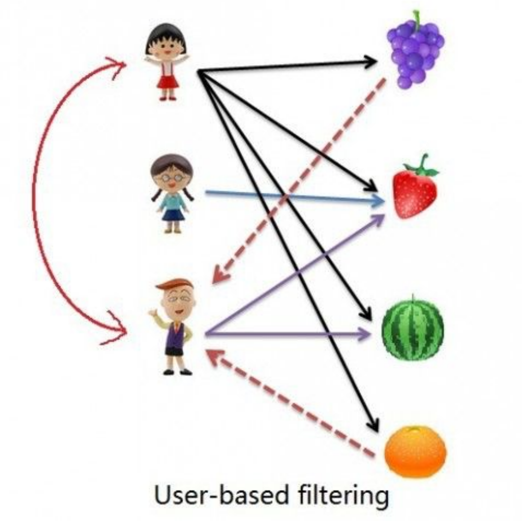
Top-selling products greatly affect results

100% of confidence with drinks with sugar and special tax. This tax should be removed from dataset

product_id		
9771	IMPOST SOBRE BEGUDES	00*00*00

As rules aren't hidden, these are a good tool to study the customer habits and they are easy to modify

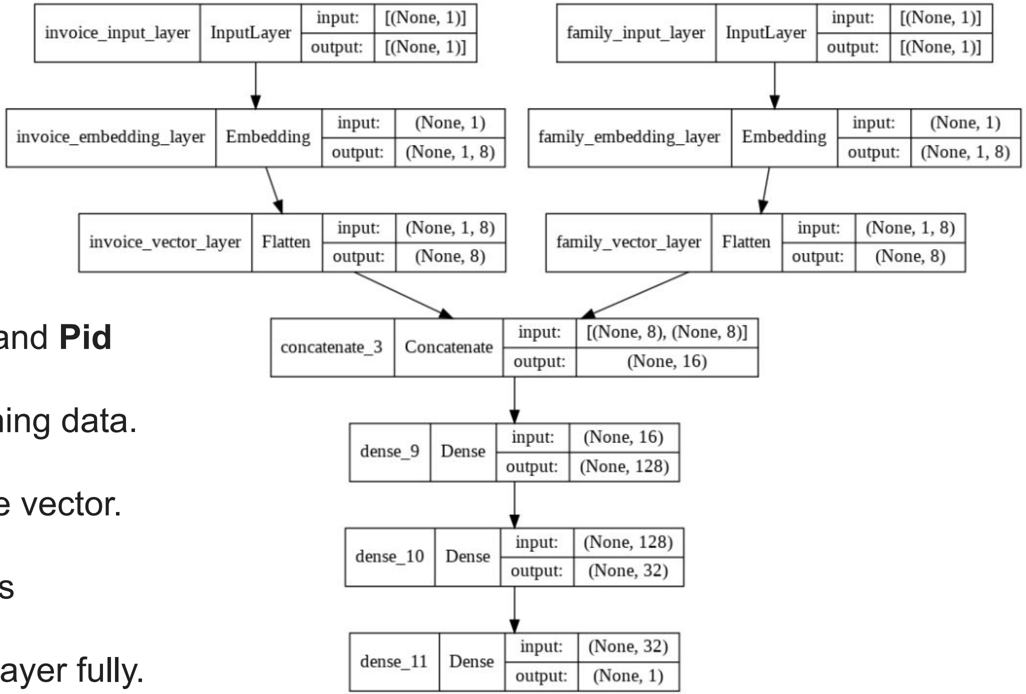
Collaborative Filtering



- ALS algorithm
- Fill in the missing entries of a user-item association matrix
- Similarity between items and users
- Works with a rating column → artificial

Precision : 3,2%

Neural Network



1. **Input layer :** get the inputs **'invoice ID'** and **Pid**
2. **Embedding layer :** give weights for training data.
3. **Flatten layer :** reduce 1 dimension of the vector.
4. **Concatenate layer :** merge the branches
5. **Dense layer :** connect input and output layer fully.

Neural Network

Embedding layer

Embedding (number_of_unique_product+1)

Range of Value

invoice id : 0 - number of unique invoice ID [0, 127422]
product id: 0 - number of unique product ID [0, 2840]

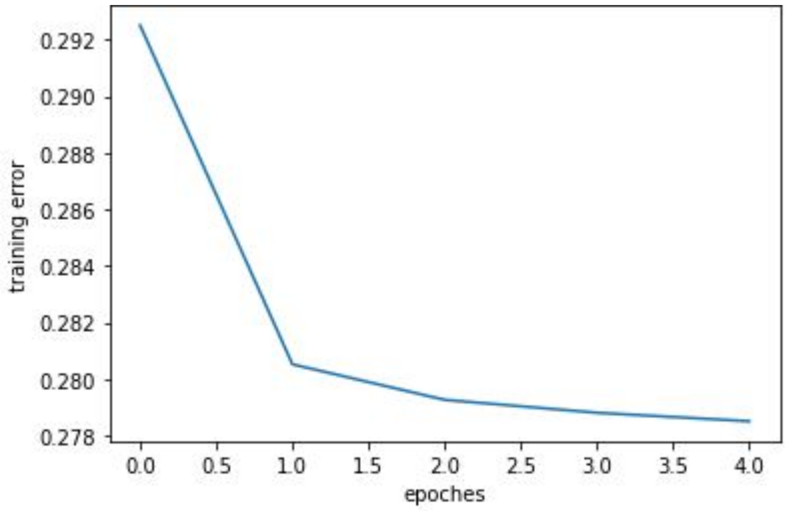
PID process

	product_id	pid
0	200	0
1	202	1
2	206	2
3	207	3
4	208	4
...	...	...
2835	91534	2835
2836	91541	2836
2837	91549	2837
2838	91554	2838
2839	91555	2839

TEST DATASET

	invoice_id	product_id	units	pid
686806	49813	24038	1	1725
1016710	96471	9117	1	1211
285495	96469	218	1	11
837392	112984	87962	1	2151
149277	117785	7895	1	966
...	...	...	...	...
541964	75542	7764	1	905
95409	57265	8055	1	971
453919	36876	5824	1	528
725647	32123	8507	1	1167
641490	91068	86663	1	1921

Neural Network



MSE : 0.2804

RSME : 0.5294829105987167

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Neural Network

Outputs

Predictions of units per items.

output

	test_units	predictions
;92429	1	1.317959
348598	2	1.353539
'00331	1	1.073415
;12349	1	1.067494
158184	1	1.125359
...	...	...
216584	1	1.214687
'48362	1	1.274643
'55685	2	1.137260
250863	2	1.461086
45233	1	1.109274

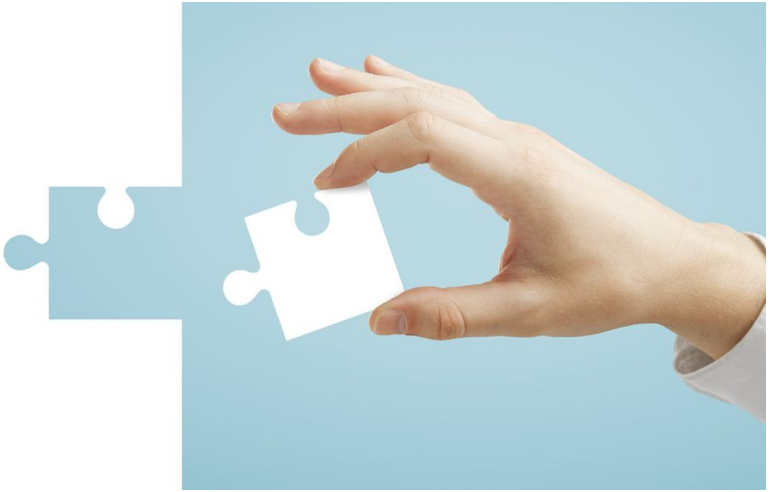
Top 5

product_id	product_desc
20762	BEBIDA ENERGETICA RE
90897	CERVEZA ESTRELLA GAL
91195	CERVEZA CORTES BOTEL
1504	VIENTRE DE VACUNO

Conclusions

Personalized Recommendation X
Retraining For Recommendation

RNN Model can be solution



5.Conclusions

Conclusions

Model Selection

→ FP-Growth: visibility and easy to manipulate.

→ Collaborative filtering : good balance between precision and complexity

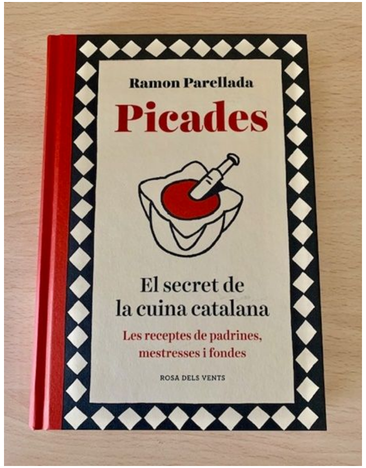
→ Neural Network for the future with better model

Conclusions

Do we have to change dataset of training?

- Can it be more interesting to use kitchen recipes?

- Can the company make a dataset with its interests? (Change from unsupervised to supervised learning)



Thanks