Instacart Product Recommendations

Ada Jing, Ilya Korzhenevich, Jessica Bao, Katy Kiefer, Kaylie Chen

Overview

Our object is to analyze over 3 million Instacart items in orders to draw insight and develop a recommendation system to help Instacart increase its sales and further develop its business strategies.

01

Exploratory Data Analysis

Examining popularity of products, lengths between orders, etc

02

Segmentation Clustering + PCA

Analyze customer buying patterns by clustering based on products bought

03

Market Basket Analysis

Looking at buying patterns between products and advise business strategies

04

Recommendation Systems

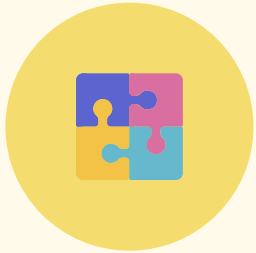
Recommend customers products to add based on previously added items

05

Conclusion

Wrap-up of our analysis and recommendations

EXPLORATORY ANALYSIS



Data Profile

Six csv files

Dating with the data

Exploring data via charts and numbers

Insight

Leverage understanding to spearhead mining

Orders.csv

3.4M+ rows

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15.0
2	473747	1	prior	3	3	12	21.0
3	2254736	1	prior	4	4	7	29.0
4	431534	1	prior	5	4	15	28.0

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
3421078	2266710	206209	prior	10	5	18	29.0
3421079	1854736	206209	prior	11	4	10	30.0
3421080	626363	206209	prior	12	1	12	18.0
3421081	2977660	206209	prior	13	1	12	7.0
3421082	272231	206209	train	14	6	14	30.0

Departments.csv

21 departments

	department_id	department
0	1	frozen
1	2	other
2	3	bakery
3	4	produce
4	5	alcohol

Aisles.csv

134 aisles

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars
3	4	instant foods
4	5	marinades meat preparation

Products.csv

49k+ products

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7
3	4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
4	5	Green Chile Anytime Sauce	5	13

Order_products__prior.csv
32M+ rows

	order_id	product_id	add_to_cart_order	reordered
0	2	33120	1	1
1	2	28985	2	1
2	2	9327	3	0
3	2	45918	4	1
4	2	30035	5	0

	order_id	product_id	add_to_cart_order	reordered
32434484	3421083	39678	6	1
32434485	3421083	11352	7	0
32434486	3421083	4600	8	0
32434487	3421083	24852	9	1
32434488	3421083	5020	10	1

order_products__train.csv
1.4M rows

	order_id	product_id	add_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0
3	1	49683	4	0
4	1	43633	5	1

	order_id	product_id	add_to_cart_order	reordered
1384612	3421063	14233	3	1
1384613	3421063	35548	4	1
1384614	3421070	35951	1	1
1384615	3421070	16953	2	1
1384616	3421070	4724	3	1

Prior/Train size split is 95.9%/4.1%

Customer Segmentation

Market and customer segmentation are some of the most important tasks in any company. The segmentation done will influence marketing, sales decisions, and potentially the company's survival.

Initial Exploration

49677

Products

134

Aisles

200k+

Users

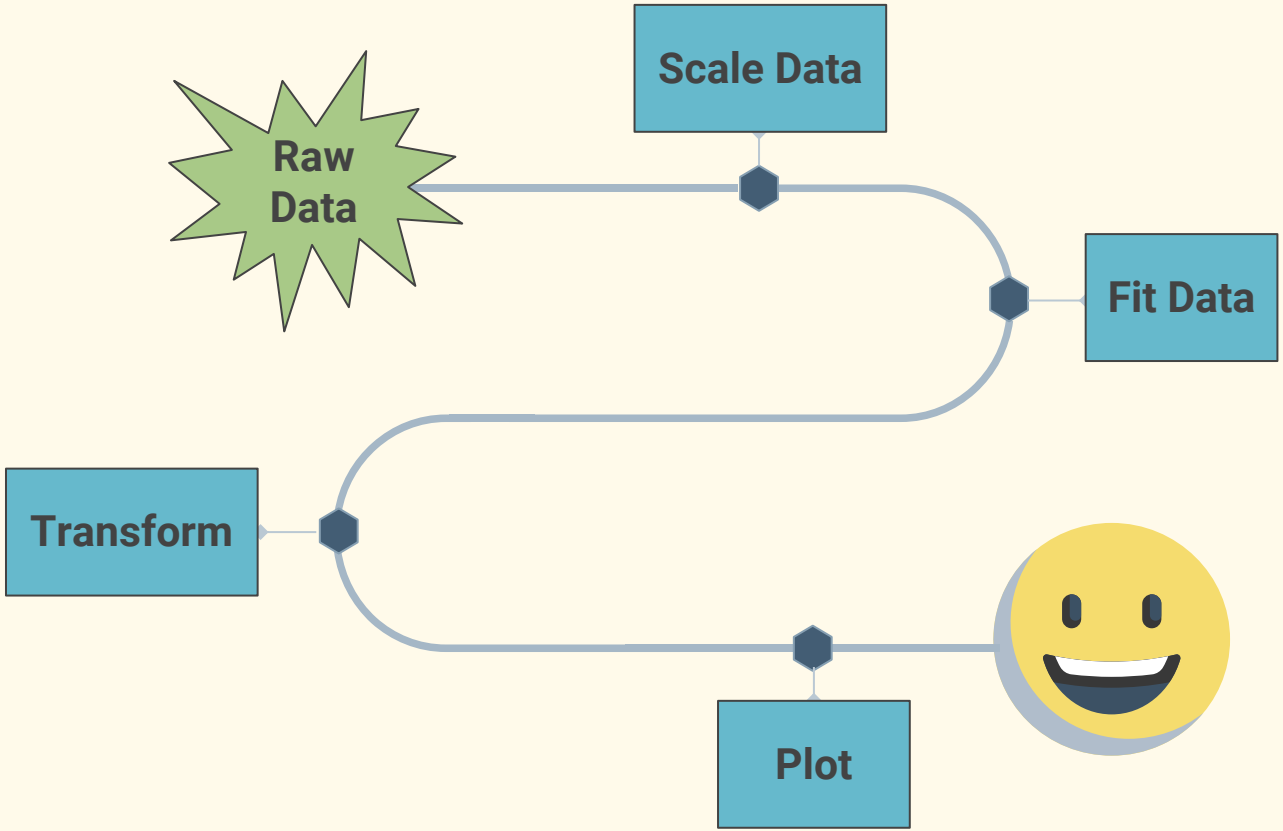
Dimension Reduction

PCA

Rotates data samples to be aligned with axes

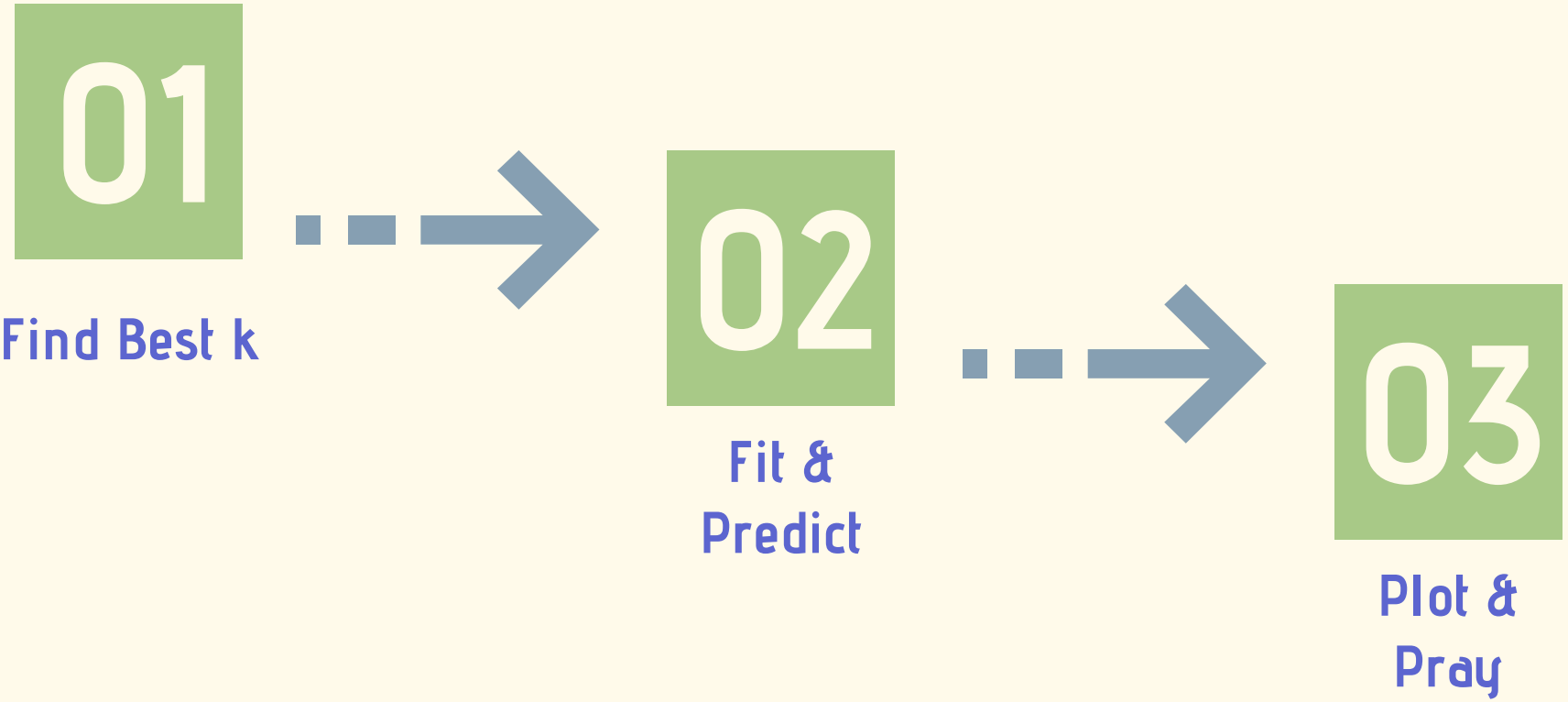
Shifts data samples so they have mean 0

No information is lost

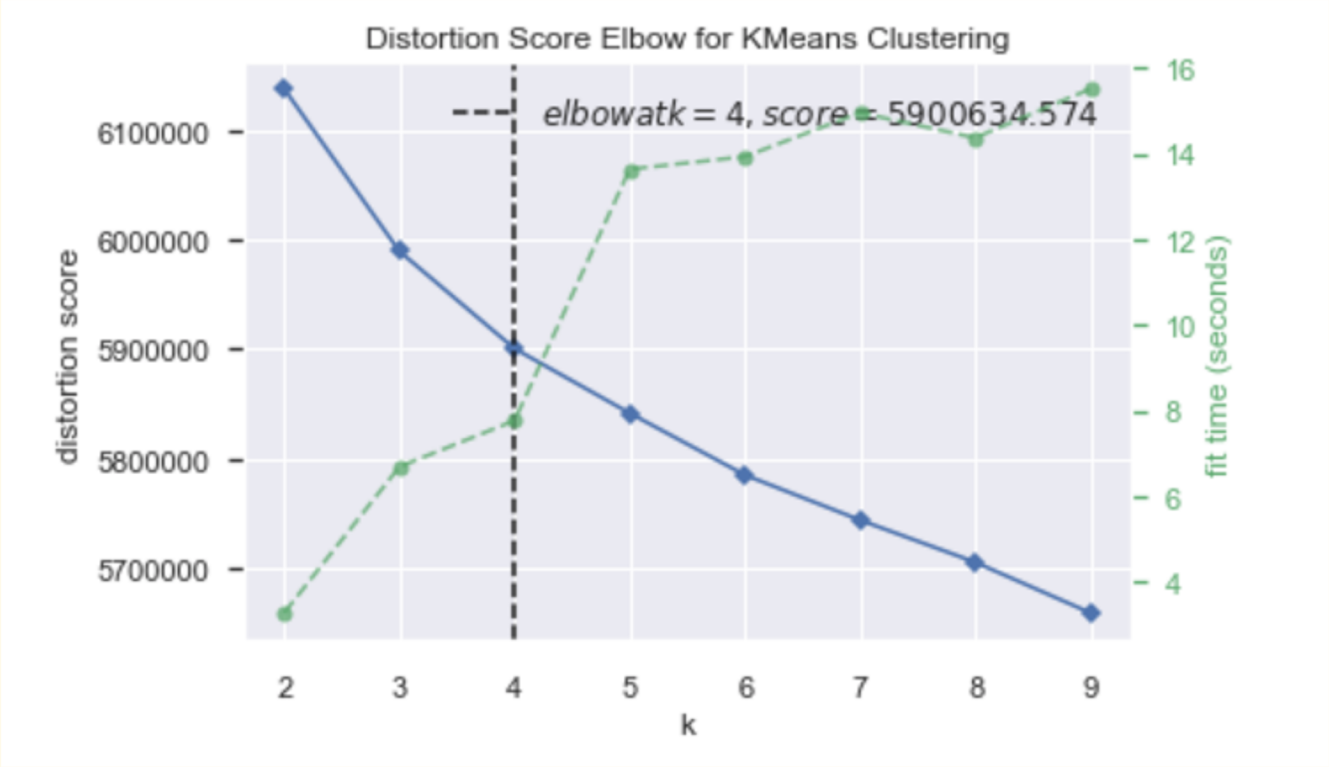


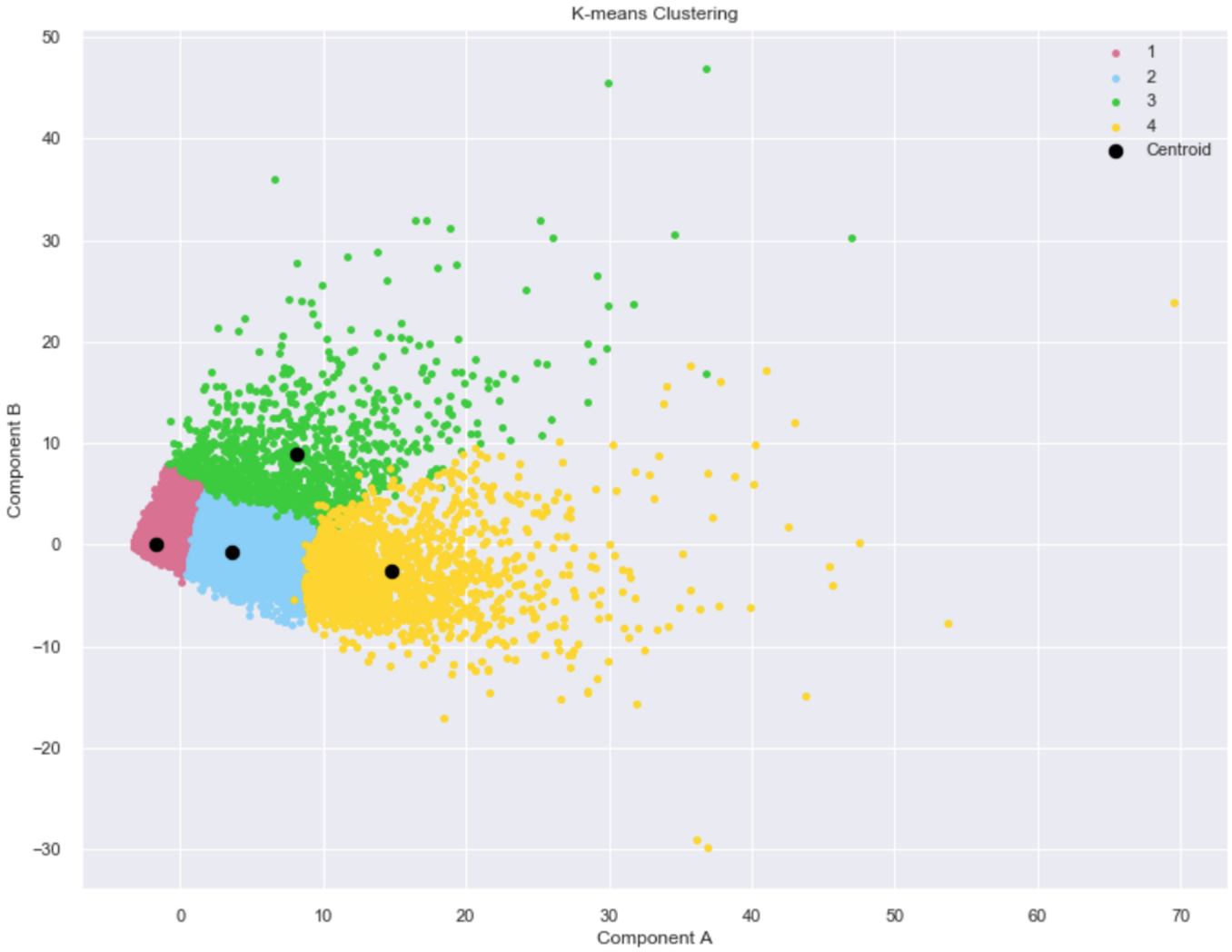
Explained Variance Plot

K-means Clustering



Best K


Distortion Score Elbow for KMeans Clustering
elbowatk = 4, score = 5900634.574
distortion score
6100000
6000000
5900000
5800000
5700000
fit time (seconds)
16
14
12
10
8
6
4
2
3
4
5
6
7
8
9
k


K-means Clustering
1
2
3
4
Centroid
Component B
Component A
50
40
30
20
10
0
−10
−20
−30
0
10
20
30
40
50
60
70

Other Clustering

DBSCAN

Does not require the number of clusters as a parameter

Infers the number of clusters based on the data

t-SNE

Technique for dimensionality reduction

Well suited for visualization of high-dimensional datasets

Cluster Characteristics

1

Aisle	Mean
Fresh Fruits	8.313
Fresh Veg	7.595
Packaged Veg	3.982
Yogurt	3.071
Seltzer	2.243
Cheese	2.064
Milk	1.917
Pretzels	1.620
Soy Milk	1.444
Refrigerated	1.309

2

Aisle	Mean
Fresh Fruits	39.824
Fresh Veg	39.130
Packaged Veg	19.382
Yogurt	15.410
Cheese	10.565
Milk	9.450
Seltzer	7.982
Pretzels	7.502
Soy Milk	7.362
Bread	6.24

3

Aisle	Mean
Fresh Fruits	24.904
Fresh Veg	18.820
Yogurt	14.650
Seltzer	14.308
Packaged Veg	13.635
Paper Goods	12.628
Cheese	12.084
Pretzels	11.806
Soft Drinks	11.619
Milk	10.699

4

Aisle	Mean
Fresh Fruits	106.501
Fresh Veg	96.122
Packaged Veg	49.863
Yogurt	44.727
Cheese	31.537
Milk	25.959
Pretzels	19.852
Baby Formula	18.123
Soy Milk	17.855
Bread	17.661

Summary

Business Uses

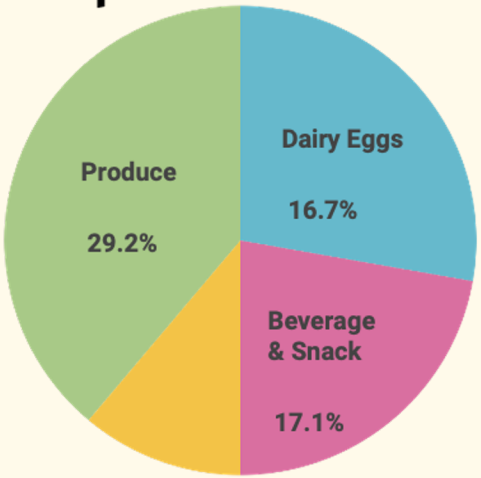
- Determine appropriate pricing
- Develop marketing campaigns
- Design optimal distribution strategy
- Offer promotion to loyal customers
- Incentives for new shoppers

Future Scope

- RFM (recency, frequency, monetary) analysis
- Product2Vec
- Hierarchical clustering
- Add demographics
- Customer segments as labels for predictive classification algorithm

MARKET BASKET ANALYSIS

Dept Distribution



	product_id	purchase quantity	product_name
0	24852	472565	Banana
1	13176	379450	Bag of Organic Bananas
2	21137	264683	Organic Strawberries
3	21903	241921	Organic Baby Spinach
4	47209	213584	Organic Hass Avocado
5	47766	176815	Organic Avocado
6	47626	152657	Large Lemon
7	16797	142951	Strawberries
8	26209	140627	Limes
9	27845	137905	Organic Whole Milk
10	27966	137057	Organic Raspberries

- The order and product are essential data used for rule of association
- The list ranking shows most frequent purchased items are fruits, veggie and diary product
- Customer purchase trend

MARKET BASKET ANALYSIS

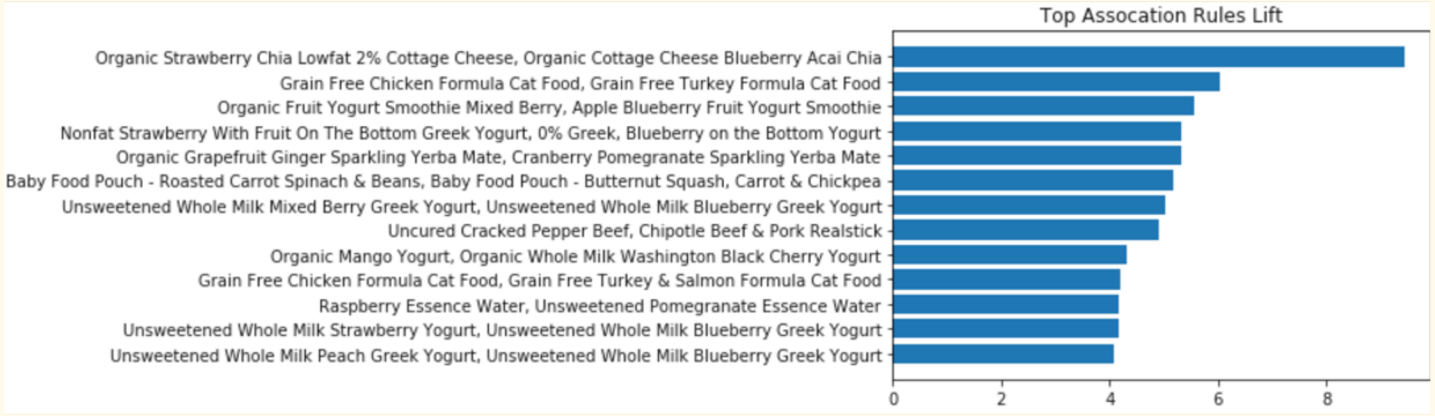
antecedents	consequents	antecedent support	consequent support	support	confidence	lift
((product_id, Organic Hass Avocado))	((product_id, Bag of Organic Bananas))	0.066596	0.119702	0.018776	0.281935	2.355307
((product_id, Bag of Organic Bananas))	((product_id, Organic Hass Avocado))	0.119702	0.066596	0.018776	0.156854	2.355307
((product_id, Organic Raspberries))	((product_id, Bag of Organic Bananas))	0.043139	0.119702	0.012937	0.299883	2.505249
((product_id, Cucumber Kirby))	((product_id, Banana))	0.030706	0.149502	0.010772	0.350820	2.346594
((product_id, Organic Avocado))	((product_id, Banana))	0.055472	0.149502	0.016863	0.303993	2.033374
((product_id, Organic Fuji Apple))	((product_id, Banana))	0.027585	0.149502	0.010520	0.381387	2.551054
((product_id, Organic Strawberries))	((product_id, Organic Baby Spinach))	0.078375	0.073895	0.012131	0.154785	2.094656
((product_id, Organic Baby Spinach))	((product_id, Organic Strawberries))	0.073895	0.078375	0.012131	0.164169	2.094656

Support Threshold of 0.01

Confidence - probability of seeing the consequent in a transaction given that it also contains the antecedent

Lift Score - A->C occur together than we would expect if they were statistically independent

MARKET BASKET ANALYSIS



Conclusion

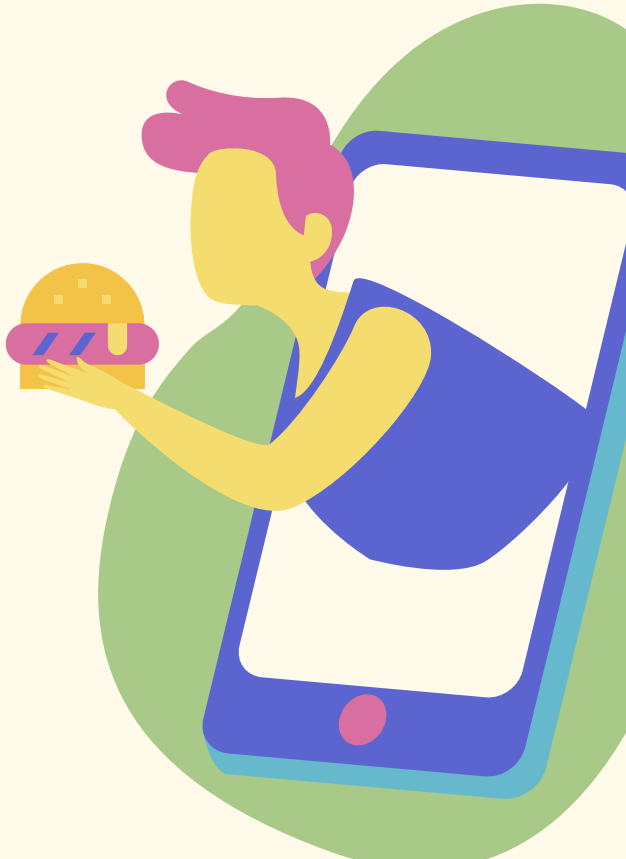
Customers are likely to purchase products from the same department.

The analysis can be used for marketing campaigns can be used for product promotions, arranging store layout & catalog design according to purchase trend

Apriori Algorithm is good at identifying product with frequent purchase but rather computationally expensive.

Recommendation System

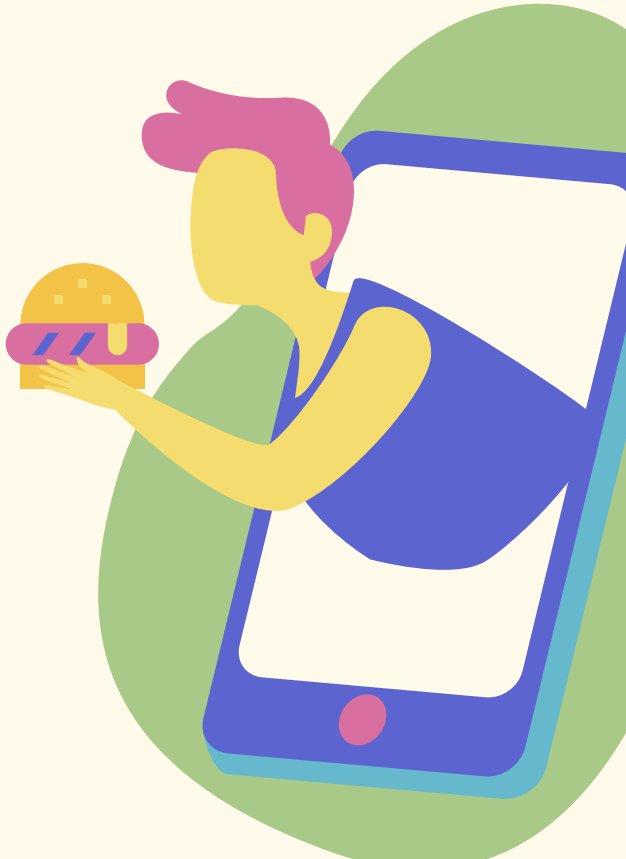
Can we give people what they want,
before they know they want it?



Recommendation System

Can we give people what they want,
before they know they want it?

...maybe



Recommendation System

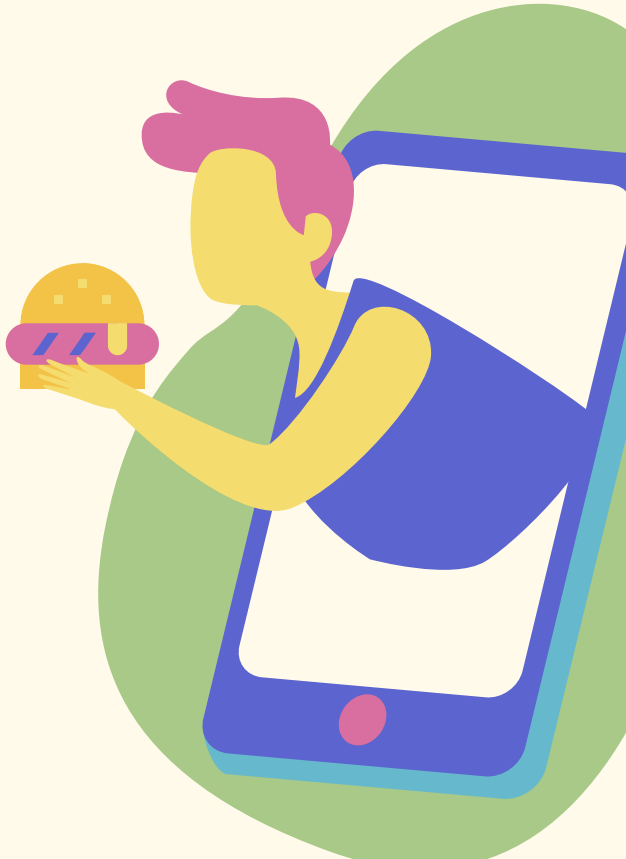
Can we give people what they want, before they know they want it?

Product2vec

Turning all products into a big matrix

Collaborative filtering

Find what other people also bought



Word2Vec Overview

Model

Skip-gram, to fit our objective of providing recommendations based on products

Validation

Compared our recommended items to the rest of the order

Methodology

- Model trained based on product_id, where sentences were all the IDs per order
- Use an average vector of all products in the order to make recommendations
- Assumed each item was of equal importance (no weighting for items added first)

Similar Products?

What is apple juice similar to?

Similar Products?

What is apple juice similar to?

	Item	Similarity Score	Mean of difference of vectors
1.	Cranberry Juice Cocktail,	.809	n/a
2.	Original Orange Juice,	.776	.141
3.	Lemonade,	.772	.156
4.	Pulp Free Orange Juice,	.727	,155
5.	Raspberry Lemonade,	.712	.173

Recommendations

Results

Let's look at order_id 232454. There are originally 17 items in the order

- All Natural Marinara Sauce
- European Cucumber
- Vine Ripe Tomatoes
- Red Onion
- Honeycrisp Apple
- Coconut Flavored Sparkling Water
- Fresh Ginger Root
- Low Fat Split Pea Soup
- Select-A-Size White Paper Towels
- Heavy Duty Scrub Sponges
- Tall Kitchen Bags, Drawstring, Lavender, 13 Gal, Mega Pack
- Yellow Onions
- 100% Natural Beef Broth
- Super Spinach! Baby Spinach, Baby Bok Choy, Sweet Baby Kale
- Coffee, Coffee BuzzBuzzBuzz! Ice Cream
- Traditional Rope Hung Smoked Scottish Salmon
- Green Tea with Ginseng and Honey

The ten predicted items are:

- Ice Cream Cake Celebration
- Cookies
- Unsweetened Lemon Flavor Real Brewed Tea
- Variety Pack Grab & Snack
- Organic Fruit Snacks Bunch O' Berries
- Natural Premium Coconut Water & Pineapple Juice from Concentrate
- Strawberry Banana on the Bottom Greek Yogurt
- Chips Deluxe Mini Rainbow Cookies
- Clear Strips
- Frosted St. Patrick's Day Cookies

Business Uses and Considerations

Is average vector the way to go?

Missing data?

Modeling: why we chose LightFM

01

What is it

A hybrid latent representation recommender model that's suitable for high-dimension.

02

What does it do

LightFM produces scores for every item for a given user, high score indicating higher likelihood of purchase

03

How does it work

Maximises the rank of positive examples with WARP loss function.

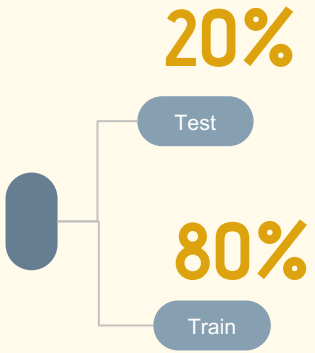
Data Processing

Original dataset

User	Order	…
ID	ID	…

Sparse.coo.matrix

	Product	…
User	Quantity	…
…	…	…



Model output

Customer already have:

- Cinnamon Swirl French Toast
- Rose Hips Tangerine Slippery Elm Lozenges
- Fresh Organic Carrots
- Lifewater 0 Cal Blood Orange Mango Vitamin Enriched Water
- Sweet Orange Marmalade
- Simply Clean & Fresh HE Liquid Laundry Detergent, Daybreak Fresh Scent, 89 Loads
- Steel Cut Oats
- Sourdough Hearty Sliced Bread

Top Recommendations:

- Garbanzos Classic Chickpeas
- Pure Demerara Cane Sugar
- Peppermint Patties
- Throat Shield Rapid Relief Herbal Tea
- Original Cheese
- Acti-Fresh Body Shape Regular To Go Pantiliners
- Snack Bites
- BelVita Golden Oat Breakfast Biscuit Packets
- White Whole Mushrooms
- Sweet Orange Marmalade

Modeling Evaluation - Collaborative Filtering

AUC

Likelihood of a random positive item recommended over a negative one

90%

Precision at k

Percentage of total actually purchased items that were on the top k list

5%

Potential Improvements

Challenges

- Our system is able to rate relevant items higher, when it comes to putting these items up front, it does a terrible job.
- We don't have information readily available on a use level.

Opportunity

- Can we incorporate user and product information to make our recommendation system smarter?

Inferring customer gender

We think that by utilizing the data we have, it is possible to infer a users' gender based on their purchase history. In fact, we have a specific item feature that indicates whether or not a product belongs to the feminine care aisle.

While we fully acknowledge that this assumption is by no means perfect for many reasons… here, we implemented this user feature out of pure academic curiosity.

Modeling Evaluation - Hybrid Model

AUC

Model with user/item or with only user features: 92%
Model with only item feature: 93%

93%

Precision at k for all three models

Percentage of total actually purchased items that were on the top k list

5%

Comparing our results

Toilet Paper is most similar to...

Similar products based on collaborative filtering

'Organic Cream Cheese Bar',
'Baby Food Blueberry, Parsnip & Buckwheat Stage 2',
'Women Over 55 Multivitamin',
'Quinoa Cereal Honey Almond',
'Balsamic Soy Ginger Sauce',

Similar products based on hybrid method with user/item features

'12 G. Protein Bar Coffee Chocolate',
'Eye Makeup Remover Pads With Kiwi Extract No Fragrance Added',
'PM Sleep Aid Plus Pain Relief',
'Soft White Halogen 72 Watt Double Life Light Bulb',
'Exfoliating Hydro Gloves 1 pair'

Similar products based on Word2Vec model

Proactive Health Sensitive Stomach Cat Food
Whisk Broom with Dust Pan
Gold Temptation Refreshing Shower Gel
Himalayan Pink Salt Liquid Hand Soap

Conclusion

- We've uncovered that there are definite patterns that emerge in the Instacart dataset, such as:
  - Orders peak on Saturday evenings and Sunday mornings- good time to restock or launch new products?
  - Marketing campaigns can be targeted to customers based on the cluster they are in to enhance success
- Relationships found in market basket analysis can help to further refine recommendation systems
  - Some relationships have already been incorporated naturally, such as 'Organic Fuji Apples' being recommended as a similar product to 'Bananas' in the Word2Vec model

Thanks!