How Digital Technologies are Transforming the Grocery Retailers? A Case Study of Instacart with Big Data

Luying Jiang
University of Chicago
May 26, 2020

Abstract

The surge of grocery delivery has changed the grocery retailing industry. Especially due to the current severe situation of COVID-19, people cannot do regular grocery shopping. More and more consumers are turning to the companies in the sharing economy for cost-effective access to goods and services including groceries. In this paper, I use Instacart, an online food-delivery company that does same-day delivery, as an example. I mainly discuss the impact of Instacart on the grocery retailers and labor market. Specifically, I will develop machine learning to build a model to show how to understand the consumer behavior better. In addition, I show some possible solutions with the aid of big data for grocery retailers to problems like balancing demands and supplies.

1. Introduction

2. Literature Review

3. Data

3.1 Data Collection

In 2017, Instacart announced its first public dataset release. It contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, it provides between 4 and 100 of their orders, with the sequence of products purchased in each order. It also provides the week and hour of day the order was placed, and a relative measure of time between orders.

3.2 Data Description

The whole dataset contains six individual csv files. There are two files that relates with the product information, namely *order_products_train* which contains the prior transaction records from all customers and *order_products_prior* which contains the latest transaction record from some customers only. There are 1,384,617 products in the *order_products_train* and 32,434,489 products in the *order_products_prior*. Both files have 4 feature columns: order_id (the ID of the order), product_id (the ID of the product), add_to_cart_order (the ordering of that product in the order), and reordered (whether that product was reordered). Overall, there are 3,346,083 unique orders for 49,685 unique products.

The orders.csv file has 3,421,083 orders and 7 feature columns: order_id (the ID of the order), user_id (the ID of the customer), eval_set (which evaluation datasets that the order is in — prior, train, or test), order_number (the number of the order), order_dow (the day of the week when that order occurred), order_hour_of_day (the hour of the day when that order occurred)
The number of days since the previous order (days_since_prior_order).

The departments and aisles contain information about the distinct department and aisle information of the application.

4. Method

4.1 Data Preprocessing

Here is some basic knowledge about the datasets. Firstly, each row of the order.csv file represents an order by a user who is identified by user_id. Orders are identified by their order_id. Each order of a user is characterized by an order_number which specifies when it has been made with respect to the other orders of the same user. Secondly, each order consists of a set of products. The add_to_cart_order feature indicates the sequence in which the products have been added to the cart. I will merge the order_product_prior and products on their product_id. Then merge with orders on the order_id. Lastly, I will merge with aisles on aisle_id.

4.2 Exploratory Analysis

In particular, the top 5 most ordered products are Banana (491,291), Bag of Organic Banana (394,930), Organic Strawberries (275,577), Organic Baby Spinach (251,705), and Organic Hass Avocado (220,877). The top 5 departments are Personal Care (6,563), Snacks (6,264), Pantry (5,371), Beverages (4,365), and Frozen (4,007). The top 5 aisles are Candy Chocolate (1,258), Ice Cream (1,091), Vitamins Supplements (1,038), Yogurt (1,026), and Chips Pretzels (989).

Figure 1 shows that the most common aisles are more likely to be reordered from. As we can see from the percentage of customers reorder behavior, milk is the most frequently reordered product, while fresh fruits have the largest number of purchases for each aisle. They were reordered more frequently than fresh vegetables. Baking ingredients are least likely to be reordered maybe because they are less frequently used.

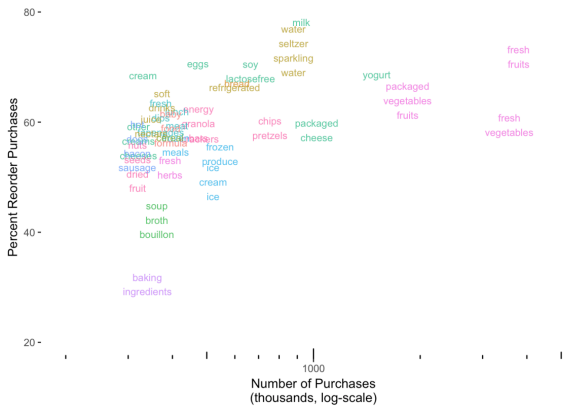

Figure 1: The percent of reorder purchases versus number of purchases for each aisle (colored by department).

We can also see the time of day that users purchase specific products.

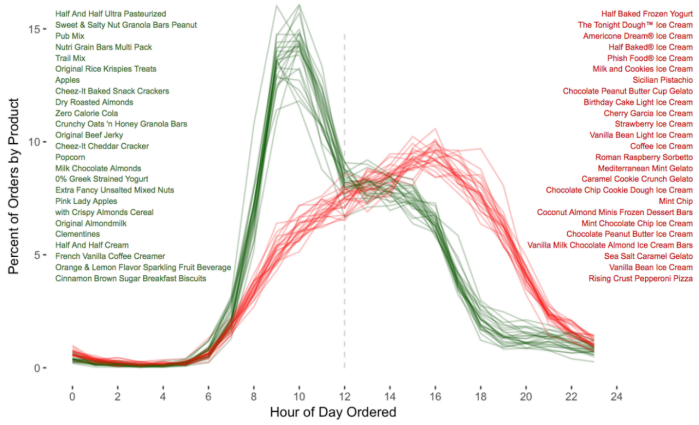

Figure 2: Popular products purchased earliest in the day (green) and latest in the day (red).

From figure 2, it looks like customers mostly made their orders between 8 AM and 7 PM. Also, we can see that healthier snacks tend to be purchased earlier in the day, whereas ice cream are far more popular when customers are ordering in the evening.

4.3 Customer Segmentation with PCA and K-Means model

The first step is to create a data frame that contains the information of individual customers about their purchases. This data frame has 206,209 rows which represents distinct customers. Then, I

will break customers into segmentations. The intuition here is that I segment the customers into subgroups based on their characteristics. There are 49,685 unique products, so I will group them into different categories. Luckily, we can use aisle to group the product. There are 134 types, which is also too many features for us to use. I use Principal Component Analysis to reduce the dimensions. Then, I apply the K-Means model for the clustering to assign clusters to different customers. There are four clusters of customers and assign each of them to a specific number.

4.4 Association Rule Mining Model

I combine the Association Mining with Apriori Algorithm to construct a recommendation list. Specifically, I calculate the frequency, support and confidence of a set of products. However, due to the computational power required for this method, I use the customer data only for each cluster. This might lead to an incomplete recommendation list, but it greatly saves time and computational cost.

Association Mining searches for frequent items in the dataset. In frequent mining usually the interesting associations and correlations between item sets in transactional and relational databases are found. Apriori Algorithm uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent item sets are used to find k+1 item set. I follow the exact steps of Apriori Algorithm to discover all the frequent item-sets. Then, we need to calculate confidence of each rule.

Support is one of the measures of interestingness. This tells about usefulness and certainty of rules. 5% Support means total 5% of transactions in database follow the rule.

$$\text{Support}(A \rightarrow B) = \text{Support\_count}(A \cup B)$$

A confidence of 60% means that 60% of the customers, who purchased milk and bread also bought butter.

$$\text{Confidence}(A \rightarrow B) = \text{Support\_count}(A \cup B) / \text{Support\_count}(A)$$

5. Result

aisle	
fresh fruits	12.997293
fresh vegetables	11.264617
packaged vegetables fruits	6.532016
yogurt	4.838682
packaged cheese	3.754675
milk	3.303355
water seltzer sparkling water	3.168569
chips pretzels	2.782964
soy lactosefree	2.349505
bread	2.279440

aisle	
fresh fruits	84.445473
yogurt	62.984685
packaged vegetables fruits	28.129081
water seltzer sparkling water	25.795860
fresh vegetables	22.891787
milk	22.726523
chips pretzels	19.449680
packaged cheese	19.042915
energy granola bars	19.022383
refrigerated	16.012959

aisle	
fresh vegetables	96.941836
fresh fruits	51.419980
packaged vegetables fruits	27.925411
fresh herbs	11.318104
packaged cheese	10.646082
yogurt	9.926398
soy lactosefree	8.805224
milk	8.353379
frozen produce	7.815187
water seltzer sparkling water	6.770039

aisle	
baby food formula	90.031453
fresh fruits	72.334056
fresh vegetables	50.059111
packaged vegetables fruits	34.557484
yogurt	33.242950
packaged cheese	24.305315
milk	23.996746
bread	12.200651
chips pretzels	11.457701
crackers	11.247831

Table 1: top 10 product goods in cluster 1 - 4

Some characteristics are displayed in each cluster. For example, “Baby Food Formula” product in cluster 3 is significantly different from other clusters. By looking at products in those four clusters, fresh fruits, fresh vegetables, packaged vegetable fruits, yogurt, packaged cheese, milk, water seltzer sparkling water, and chips pretzels are products which are generically bought by most of the customers.

	fresh fruits	fresh vegetables	packaged vegetables fruits	yogurt	packaged cheese	milk	water seltzer sparkling water	chips pretzels
0	26.720216	23.158130	13.428710	9.947504	7.718970	6.791135	6.514038	5.721298
1	29.581621	8.019094	9.853741	22.063813	6.670817	7.961201	9.036403	6.813309
2	23.611072	44.513837	12.822815	4.558012	4.888477	3.835712	3.108672	2.661403
3	27.769415	19.217949	13.266795	12.762139	9.330935	9.212474	4.041622	4.398670

Table 2: Percentage of Generally Consumed Goods in each cluster

Customers in cluster 1 consume fresh fruits in a much higher percentage than other products. Customers in cluster 2 consume the fresh vegetables much more than customers in other clusters.

Those are some crucial information that we can use to do recommendations based on different clusters.

Here are some related items that display high confidence.

itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB
Organic Strawberry Chia Lowfat 2% Cottage Cheese	Organic Cottage Cheese Blueberry Acai Chia	306	0.010155	1163	0.038595	839	0.027843	0.263113
Grain Free Chicken Formula Cat Food	Grain Free Turkey Formula Cat Food	318	0.010553	1809	0.060033	879	0.029170	0.175788
Organic Fruit Yogurt Smoothie Mixed Berry	Apple Blueberry Fruit Yogurt Smoothie	349	0.011582	1518	0.050376	1249	0.041449	0.229908

Table 3: Associations between paired products

From the above table, we can clearly see that the different flavors of a product are highly connected together. This kind of recommendation pattern seems to be highly efficient. Also, once we find the patterns from the customer clustering, we can use the same recommendation strategies to save cost.

6. Discussion

7. Conclusion

Reference:

3 Million Instacart Orders, Open Sourced, Jeremy Stanley, https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

Apriori Algorithm, https://www.geeksforgeeks.org/apriori-algorithm/

The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017