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

Luying Jiang
University of Chicago
June 10, 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.

Keywords: digital economy, online shopping, consumer behavior, grocery retailing.

1. Introduction

With the surging of electronic commerce, consumers' shopping behavior have been largely transformed from offline to online. By using online platforms, consumers tend to have better pictures of pricing of a product than before. In the meanwhile, companies are able to use those data to model the consumer behavior. Walmart collects around 2.5 petabytes (1 petabyte = 1,000,000 gigabytes) of information every hour about transactions, customer behavior, location and device (McAfee et al., 2012). With those data, companies can better assess consumer information, reduce operating cost and improve service efficiency. In this paper, I try to understand how the digital technologies are transforming the grocery retailers. Specifically, how we can use data to model consumer's repurchase behavior? What models can we use? What factors are the most important ones to the model?

Due to the current severe situation of COVID-19, people cannot go outside and do grocery shopping in a regular manner. More and more consumers are turning to the companies in the sharing economy for cost-effective access to goods and services including groceries. There are many different types of online retailing and each of them consists unique characteristics. This paper focuses on the grocery online retailing stores. I choose grocery because the grocery retailing is a simpler model to start with. For example, compared with over 100 different brands and styles of dresses, we do not have that many different bananas in the grocery stores. Many products are not highly related with consumer loyalty to certain brands and the probability of purchase based on others' review is much less compared with other retailing industry. Hence it requires less features in the dataset to build a model. In this particular study, I use Instacart, the largest third-party online grocery delivery service in the United States, and its data as an example to model the consumer repurchase behavior.

With a better understanding of the consumer repurchase behavior, grocery retailers can predict resource requirements, personalize digital engagement and improve return on marketing campaigns. Therefore, an understanding of how consumers leverage the features of the internet to make purchasing decisions in the e-commerce environment would help managers devise suitable marketing strategies (Wu and Lin, 2006). They can predict the distribution of expected outcomes in the future based on the model. They also want to update that forecast as new information arises.

I firstly merge and clean the dataset and perform exploratory analysis based on different existing features. Then, I use some important findings to create features used for modeling. I choose Logistic Regression, Gradient Boosting and XGBoost for my classification models and assess the performance through confusion matrixes and accuracy scores. In the end, I conclude that XGBoost has the highest accuracy score and fast computing time to train the model. The top 5 important features in my model are: product reorder ratio, user reorder ratio, product ordered by user since last order, time since last order, and product order rate by the user.

2. Literature Review

Many literatures try to examine the role of big data and predictive analytics in retailing. For example, Bradlow et al. (2014) discuss about opportunities in data pertaining to customers, products, time, location and channel and highlight the importance of theory in solving retailing problems. The role of big data is not just increasing the data volume but improving the data quality. Martz and Netzer (2017) highlights some useful analytical techniques as a window to customers' psychology and hence used for marketing. By using the digital records of customers, we are able to infer psychological traits and states and see how those can affect customer behavior.

There is an increasing number of literatures discussing the digital economy environment and its impact on consumer purchasing behavior. For example, Baye and Morgan (2001) presents a model examining the market equilibrium between price information and the product. Baye and Morgan mainly discusses about when consumer and sellers participate in the online market and in what kinds of circumstances an information gatekeeper can maximize its profit. However, in this model, the products are assumed to be homogeneous which usually will not be the case in a real-world situation. Dinerstein, Einav, Levin, and Sundaresan (2017) estimate the consumer demand and retail margins using detailed browsing data on search results, consumer purchasing decisions, and product prices. The model is applied to quantitatively analyze a large-scale redesign of the search process on eBay in 2011. It specifically discusses about the role of search design in reducing consumer search frictions and determining optimal market outcomes.

Most of the papers focus on the theoretical analysis of models due to the difficulty of collecting commercial dataset. More recent studies tend to realize the importance of empirical analysis and apply various data into the model. There are some studies about applications of different online purchasing models with data supported. For example, De los Santos, Hortaçsu, and Wildenbeest (2012) tests consumer purchasing theories using online retailing data from Amazon and other dominant book sellers. Honka and Chintagunta (2013) uses data in Auto insurance industry in the U.S. to compare two different product search models. The paper shows that the large insurance companies are better off when consumers use a sequential search method, while smaller companies are better off when consumers search with a fixed sample size.

This paper focuses on the online retailing industry. Data is a very important factor when evaluating consumer behaviors and data-driven methods are adopted by most companies. The biggest contribution of my paper is to applies empirical data to validify in the real-world scenarios. This paper also limits the category to only the grocery retailers to construct a better model. Gunawan, Saleha and Muchardie (2018) used linear regression and cluster analysis to understand the consumer online purchase behavior of groceries. They found out that brand preference, shopping convenience and consumer adoption level significantly influence purchase behavior.

In the section 3, I will discuss the data used for empirical analysis and its key variables. In section 4, I will talk about the appropriate advanced computational methods and models I used for understanding the consumer purchase behavior. In section 5, I will explain the result I get from following the steps in section 4. In section 6, I will discuss some potential limitations of my study and some ways to improve and get more accurate result.

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)and 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

Through exploratory analysis, I would like to find out some important relationships between the existing features in the dataset and the probability of reordering. For example, do users buy different kinds of items at different times of the day? Does the category of the product relate with the probability of repurchase?

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.

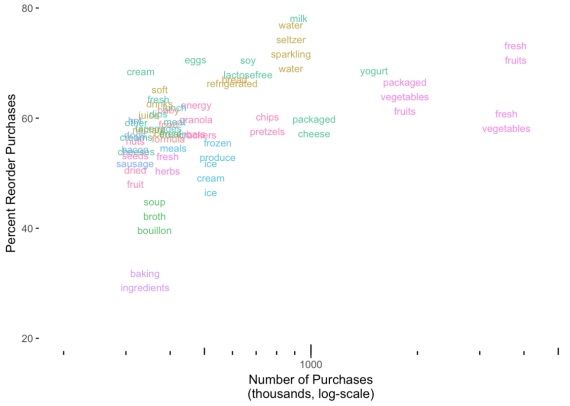

Figure 1: The percent of reorder purchases versus number of purchases for each aisle (colored by department).

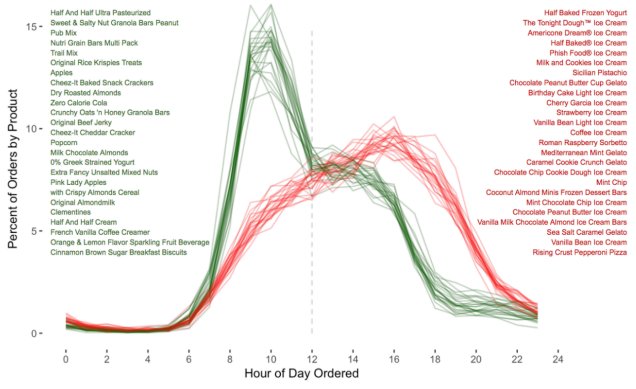

Figure 2: Popular products purchased earliest in the day (green) and latest in the day (red).

From figure 2, it shows that customers mostly made their orders between 8 AM and 7 PM. Also, we can see that snacks that are healthier tend to be purchased earlier in the day, whereas

there are far more ice creams ordered in the evening than the daytime. There might be a relationship between the time of order and the category of the product.

reordered	Total no. products	Ratio
0	13863746	0.409938
1	19955360	0.590062

Table 1: Reorder Frequency

In all the products purchased, there are around 59% of the products are previously purchased by customers.

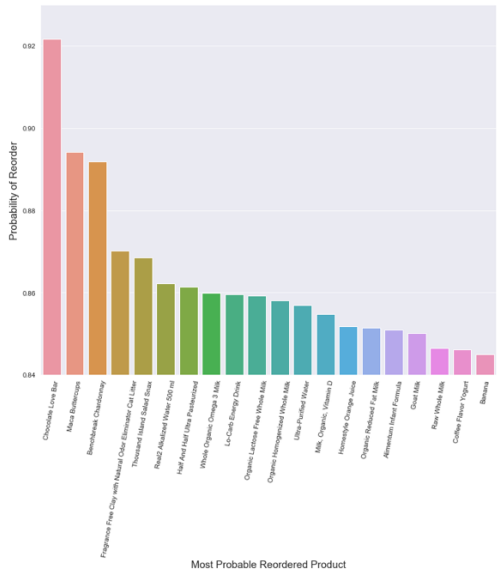

Figure 3: Reordered Products Distribution

Among all of the products with reorder number over 100 as we can see from Figure 3, the chocolate love bar is the product with the highest reorder ratio and follows by Maca Buttercups. Benchbreak Chardonnay, Fragrance Free Clay and so on.

4.3 Feature Engineering

Based on the data exploratory analysis, I create three kinds of features: user features, item features and user x item feature. User features are the ones relates with a particular user and his/her particular shopping behavior. Product features are the ones associated with the product and its particular characteristics. User x item feature is an interactive feature which are the ones that

evaluate the specific purchase behavior a user has towards a particular product. I came up with as many features as possible in the first stage after the exploratory analysis and then removed those duplicates which will make the computation time for the model longer and also unnecessary.

I list the features with its detailed explanation in the following table to make it clearer:

Feature type	**Variable Name**	**Explanation**
User Feature	user_total_orders	User's total order number
	user_sum_days_since_prior_order	Days since prior order (sum)
	user_mean_days_since_prior_order	Days since prior order (mean)
	user_reorder_ratio	User's reorder ratio
	user_total_products	Total number of products bought
	user_distinct_products	Total number of distinct products bought
	user_average_basket	Average products purchased per order
Product Feature	prod_tot_cnts	Total time of purchase
	prod_reorder_tot_cnts	Total time of reorder
	prod_buy_first_time_total_cnt	Total time of product bought only once
	prod_reorder_ratio	Product reorder ratio
	prod_reorder_times	Product reorder times
	times_user_buy	The n^{th} time bought the product
User x Product Feature	up_order_count	Number of times the user bought the product
	up_first_order_number	Order no. of user bought product for the first time
	up_last_order_number	Order no. of user bought product most recently
	up_average_cart_position	Average cart position of the product
	up_order_rate	Product order rate by the user
	up_order_since_last_order	Product ordered by user since last order
	up_order_rate_since_first_order	Product order rate since last order

Table 2: Features used for the models

All the above features need to be calculated based on the original data set. In total, I made 20 features related with user and product which will be used in the next step.

4.4 Classification Models Training

4.4.1 Logistic Regression

I firstly use the logistic regression for the basic classification model. I assume a linear relationship between the predictor variables and the log-odds of the event that Y=1 in which Y stands for binary variable "reorder". This linear relationship can be written in the following mathematical form:

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where ℓ is the log-odds, b is the base of the logarithm, β_i are parameters of the model, and x_i are the features we get from step 3.

4.4.2 Gradient Boosting

I think the tree-based method are very suitable for this kind of multi-variable classification problems. Gradient Boosting is an ensemble learning method where multiple models (aka "weak learners") are trained and combined to get better results. Observations have an unequal probability of appearing in subsequent models and the ones with the highest errors appear most.

We first initialize the model with a constant value: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$ where L stands for the loss function, y_i refers to the observed values, and γ refers to log(odds) value. $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ Then for $m = 1\ to\ M$, we compute for $i = 1, \dots, n$.

We fit a regression tree to the r_{im} values and create terminal regions R_{im} for $j = 1, \dots, J_m$. For $j = 1, \dots, J_m$ we compute $\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$. And finally, we update the model $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.

4.4.3 XGBoost

At last, I will implement the XGBoost model for comparison. XGBoost model has become very popular in the industry for its characteristics and improvement of over Gradient Boosting. XGBoost stands for Extreme Gradient Boosting. It uses more accurate approximations to find the best tree model. It computes second-order gradients and has advanced regularization (L1 & L2), which improves model generalization. The training time is greatly reduced for the large-scale data model as the Instacart one.

I use accuracy score to assess the performance of the models. It can represent the proportion of correctly classified observations. Confusion matrix, which is 2x2 table showing four parameters, including the number of true positives, true negatives, false negatives and false positives, is also used for reference. Besides, I also take computation time into consideration. The computation time is a relative idea which I do not have function to count but my perceivable value.

5. Result

5.1 Model training results

Logistic Regression's accuracy classification score is 0.9074262556549295.
Confusion Matrix:

	Predicted: 0	Predicted:1
Actual: 0	6807861 (TP)	73466 (FN)
Actual: 1	632612 (FP)	113256 (TN)

Gradient Boosting's accuracy classification score is 0.9091826025163904.
Confusion Matrix:

	Predicted: 0	Predicted:1
Actual: 0	6805451 (TP)	75876 (FN)
Actual: 1	616806 (FP)	129062 (TN)

XGBoost's accuracy classification score is 0.9097613211672181.

Confusion Matrix:

	Predicted: 0	Predicted:1
Actual: 0	6797556 (TP)	83771 (FN)
Actual: 1	604497 (FP)	141371 (TN)

As a result, XGBoost has the highest accuracy classification score.

5.2 Feature importance

By using the data of Instacart, I model the consumer repurchase behavior using machine learning models.

Figure 4: Feature Importance

The top 5 important features in my model are:

- Product reorder ratio
- User reorder ratio
- Product ordered by user since last order
- Time since last order
- Product order rate by the user

6. Discussion

The result of model performance makes much sense in terms of the tradeoff between the model simplicity and accuracy. Logistic Regression is the simplest among the three models I choose and use the shortest time to train the classification model. However, it has the lowest accuracy score. XGBoost takes shorter time than the Gradient Boost model but longer than Logistic Regression. It has the highest accuracy score. I expect that with an increase of the train set, the accuracy score will increase.

For the feature importance part, we can see that features related with reorder rate and time since last order are the ones more important than other features. If the product is more likely to be reordered such as a popular reordered item for most of the customers, then it will lead to a higher reorder probability for the user. Likewise, if the user has a high reorder rate then it may indicate the purchase style of the user who is more stick to purchased items. Time since last order also tends to play an importance role in a way that a user may follow a specific pattern of purchasing an item. Product order rate indicates a specific purchase behavior of a user towards a product. This also make a lot of sense in real life.

For future studies, if there are more information about the product, for example, the prices of products and brands and also about users personal information and IP address tracking, cookie tracking, loyalty card usage, spatial analysis, I will be able to add more features into the model and hence have a better trained model to understand consumer repurchase behavior.

There also some limitations of this model. The dataset is very large and hence requiring a huge amount of computing power. Due to the limit of my computer, I choose the test size of 0.9 to train the model. I was not able to do cross validation on the dataset. Features are manually created and hence may be incomplete. If I have access to higher computational power hardware, I am able to use more train set and perform cross validation to more accurately compare model performance.

Another thing to notice is that the dataset provides with a set of testing data but in a large amount and with no reorder label (y_test). To be more computational efficient and effective, I did

not use the testing data provided to assess the accuracy. I used the test set split with the test size of 0.9 as testing samples and compare them with the predicted results from the model.

7. Conclusion

I analyze the dataset that was open sourced by Instacart to analyze the consumer repurchase behavior and train a model to predict the reorders. Exploratory analysis helps me find some existing features to work on. Then, I use those important findings to create features used for modeling. In the feature engineering part, I create 20 features and fit them to the models. I choose Logistic Regression, Gradient Boosting and XGBoost for my classification models and assess the performance through confusion matrixes and accuracy scores. The Logistic Regression has the shortest computation time while the lowest accuracy score. The Gradient Boost model has the longest computation time. In the end, I conclude that XGBoost has the highest accuracy score and relative fast computing time to train the model. The top 5 important features in my model are: product reorder ratio, user reorder ratio, product ordered by user since last order, time since last order, and product order rate by the user. XGBoost will perform the best to model the consumer repurchase behavior.

Reference

BAYE, M. R., AND J. MORGAN (2001): "Information Gatekeepers on the Internet and the Competitiveness of Homogeneous Product Markets," *American Economic Review*, 91 (3), 454–474. [1262]

BRADLOW, ERIC T., M. GANGWAR, P. KOPALE, S. VOLETI (2014), "The Role of Big Data and Predictive Analytics in Retailing", *Journal of Retailing*

DE LOS SANTOS, B., A. HORTAÇSU, AND M. R. WILDENBEEST (2012): "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior," *American Economic Review*, 102 (6), 2955–2980. [1262]

DINERSTEIN, M., L. EINAV, J. LEVIN, AND N. SUNDARESAN (2017): "Consumer Price Search and Platform Design in Internet Commerce," *American Economic Review* [1262]

GUNAWAN, ANNETTA, RACHMAWATI ANGGUN SALEHA AND BRIAN GARDA MUCHARDIE (2018), "Online Groceries Segmentation of Brand, Shopping Convenience, and Adoption to Influence Consumer Purchase Intention", *Social Science & Humanities*, 26 (T): 21-32

HONKA, ELISABETH AND CHINTAGUNTA, PRADEEP K. (2013): "Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry", *Marketing Science*. 36(1), 21-42

MATZ, SANDRA AND ODED NETZER (2017), "Using Big Data as a window into consumers' psychology", *Behavioral Sciences*, 18: 7-12

MCAFEE, ANDREW, ERIK BRYNJOLFSSON, THOMAS H. DAVENPORT, D. J. PATIL, AND DOMINIC BARTON (2012) "Big data: The management revolution," *Harvard Bus Rev 90*, no. 10 (2012): 61-67.

STANLEY, JEREMY (2017), 3 Million Instacart Orders, Open Sourced, https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

WU, L.L. AND LIN, J.Y. (2006), "The quality of consumers' decision-making in the environment of e-commerce", *Psychology and Marketing*, Vol. 23 No. 4, pp. 297-311.

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