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

Luying Jiang

MA in Computational Social Science
University of Chicago
luyingj@uchicago.edu

Introduction

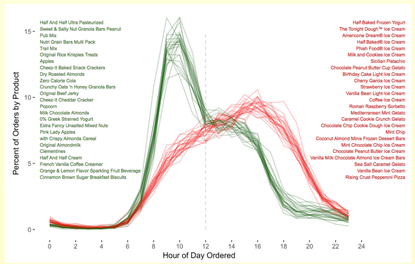
The surge of grocery delivery has changed the grocery retailing industry. Especially due to the current severe situation of COVID-19, people cannot do regular grocery shopping. More and more consumers are turning to the companies in the sharing economy for cost-effective access to goods and services 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. Specifically, I use machine learning to model the consumer repurchasing behavior. In addition, I will discuss some possible solutions with the aid of big data for grocery retailers to problems like balancing demands and supplies.

Data

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. It also provides the week and hour of day the order was placed, and a relative measure of time between orders.

The whole dataset contains six individual csv files. There are two files that relates with the product information, order_products_train and order_products_prior 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 The departments and aisles contain information about the distinct department and aisle information of the application.



An example from the exploratory analysis

Methods

Data 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?

⇩

Feature Engineering

Based on the information from step 1, I create different types of features. For example, how do we take time since a user last purchased an item into account? The dependent variable is whether a user reorder the product.

⇩

Model Training

I choose **Logistic Regression**, **Gradient Boosting** and **XGBoost** for my classification models. I expect XGBoost to be the most accurate, since it computes second-order gradients and has advanced regularization (L1 & L2).

⇩

Model Selection

I select the model based on their accuracy classification score.

References

Eric T. Bradlow, Manish Gangwar, Praveen Kopalle and Sudhir Voleti, "The Role of Big Data and Predictive Analytics in Retailing", *Journal of Retailing*

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

Sandra C Matz and Oded Netzer, Using Big Data as a window into consumers' psychology, *ScienceDirect*

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

Results

Feature Engineering:

Based on the data exploratory analysis, I create two kinds of features: user features and item features.

User features:
- Total amount of orders user made
- How far is last purchase
- Total amount of items a user has purchased
- User reorder ratio
- Average item number per order
- …

Item features:
- Number of times user purchased the item
- Product reorder ratio
- Average position in the cart
- Order rate
- Order since last order
- …

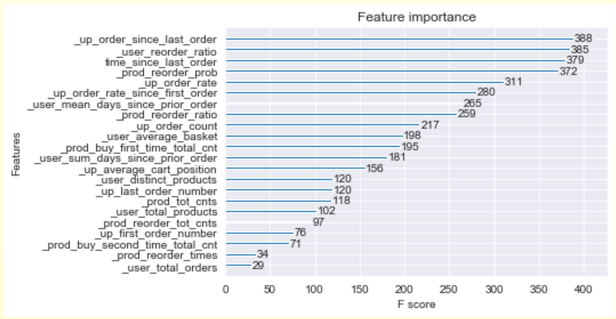
I discuss how I select and create these features more specifically in my paper.

Model Training and Selection:

Accuracy Score:
Logistic Regression: 0.1337124591674921
Gradient Boosting: 0.9021615416938993
XGBoost: 0.9097000928912923

XGBoost has the highest accuracy score among others. We can see that the difference between accuracy scores of Gradient Boosting and XGBoost is not very large. But one thing to notice is that the computation time of Gradient Boosting is significantly larger compared with XGBoost.

Feature importance:



Conclusion

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

The top 5 important features in my model are:
- User's order since last order
- User reorder ratio
- Product order date
- Product's order rate since first order
- Mean days since prior order

XGBoost has the highest accuracy score among the three models that I choose. Also, XGBoost training is very fast and can be parallelized across clusters.

Limitations

The dataset is very large and hence requiring a huge amount of computing power. Due to the limit of my computer, I was not able to do cross validation on the dataset. Features are manually created and hence may be incomplete.

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 access the accuracy.

Acknowledgements

I would like to say a special thanks to our amazing advisor Dr. Richard W. Evans, for his invaluable advice and incredible support. I would like to also thank my family and all the classmates from MACS 30250. They have provided me with generous help during this hard time.