Trends and Patterns in Instacart Data

José Gutiérrez
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
jogu0215@colorado.edu

Anna Malawista
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
anma7757@colorado.edu

Michael Whitlock
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
miwh5549@colorado.edu

Keaton Hoeger
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
keho3676@colorado.edu

Aidan O'Connor
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
aioc7481@colorado.edu

1 Problem Statement/Motivation

Instacart is an online grocery delivery service. Our group's motivation is to analyze patterns from Instacart orders. We plan to determine if there are trends in the types or rate of products ordered depending on the time of day, week, and frequency of repurchases. Potential questions include: "Which items are most likely to be added order, based on the items already in the cart?"; "Based on purchase history, can we predict which new items a customer is most likely to purchase?"; and "Can we predict the length of time between orders?"

Ideally, our research could be used to help companies determine how best to target and advertise to customers, what products to stock in the highest quantity and when, and possibly area where the market appears to be untapped. One very interesting area for possible future research would be to compare these finding to similar measures for standard grocery stores (purchased from in person) to determine what, if any, differences are seen when grocery purchases are provided through a digital, delivery service. The potential application for this research are quite broad, and trends that are found could have implications or not only Instacart but also any other grocery and food delivery service, as well as any companies selling products through these services.

2 Literature Survey

Instacart posted the dataset on their website and a brief description on Medium [1, 2]. We have found that most of the prior work on this dataset is exploratory in nature. Instacart ran a competition on Kaggle in early 2016. Various analyses of this dataset were conducted as a result of this competition. One of the primary areas of research done for this competition looked at the trends related to repurchasing [3, 4].

In Instacart's Medium post two short examples of patterns derived from the dataset are provided, one on repurchasing and item popularity, and the other on difference in "healthiness" of orders depending on the time of day [2]. This data (and other similar datasets) have also been used in a number of studies looking at buyer recommendation and customer targeting strategies [5, 6, 7].

3 Proposed Work

Since the dataset appears to be relatively clean and does not contain many missing values, we will begin by using exploratory data analysis techniques to examine the data. This analysis will use charts, graphs, and other visualization methods to give us an initial idea of the relationship between attributes. The Instacart dataset is comprised of 6 tables that are linked by ID attributes for orders, departments, etc. In order to complete our project, we will need to join the tables by these attributes, so we can fully analyze all the order information available. We have 3.2 million data points in total for analysis. There are pre-designated subsets of data for training and testing machine learning models.

Ideally, our group will implement some simple machine learning and data mining techniques to predict future order activity.

4 Data Set

Our dataset is from Instacart, a company which created an app that provides a grocery ordering and delivery service. This dataset is made up of 6 different CVS files [1]. The dataset is very clean and simple to follow; there are no null or missing values. According to Instacart, the data is anonymized and is a sample of over 3 million orders from more than 200,000 unique Instacart users [1]. There are prelabeled training and test sets. The data set provides information on what was ordered (including department and aisle of item), the day of week and time of day that the order was placed, as well as the amount of time between orders. However, no information is provided on what time of year items were purchased, so no information can be drawn on the fluctuations in ordering across seasons. A “data dictionary” was provided by Instacart and is available on GitHub [8].

There are a few issues we have noted about this dataset. The purchase day of week is provided as an integer from 0-6; however, thus far we have been unable to find any information on what day of the week each integer corresponds to. We will continue to research this question, but others have documented this same issue, and most have concluded that based on the rate of ordering 0 represents the beginning of the weekend (saturday) and that the days of the week are in order from that point [4,8]. One other issue we will face is that all data in the “days_since_prior” order attribute is capped at 30 and the number of orders for any user is capped at 99. Rather than truncating the data at these points we have concluded that all data greater than these limits were simply assigned that max value (censored). This conclusion is based on a large spike in the data at 30 for “days_since_prior” and 99 for number of orders per user. In our most of our data visualization, we will be able to simply label this part of the data as >=30 or >=99. However, we are still assessing if we should try to account for the possibility of this censored data skewing our results through preprocessing, and if so, what the best method would be. Over all, this data set is large, clean, and fairly comprehensive; and it should provide ample opportunity for analysis and prediction.

5 Evaluation Method

We want to try using as many of the following analysis methods as can be reasonably applied: data visualizations (box plots, scatter plots and trend lines, histograms, etc.), data analysis (correlation coefficient, chi-square test, and other correlation statistics, association rules), classification and prediction with machine learning. We intend on using the approximately 3.2 million order information for the analysis methods discussed above. Then for the machine

learning models there are additional orders for both training and testing our machine learning models of size 131,000 and 75,000 respectively. We intend to test multiple models in order to find the one that performs best and then optimize that model through hyperparameter tuning. While accuracy is likely a useful measurement for this dataset other measurements such as precision, recall, ect. will be considered for the final model. We will continue to modify and add to our intended evaluation methods as we learn more throughout the course.

6 Tools

We will mainly be using Python, as the language provides a lot of libraries that are helpful in both data analysis and any machine learning we want to attempt on the datasets. NumPy is one library we plan to use, which grants access to useful functions and features. Pandas is the main Python library we plan to use for data analysis, it will give us the ability to load and interpret the CSV files which the dataset is made up of. Having them loaded, allows us to combine and perform the necessary data analysis on the instacart dataset. MySQL is another tool that could be valuable for the manipulation of the dataset, and could be useful in other exploratory analysis of the dataset. Finally Scikit-learn is a Python library that we can use for any machine learning we want to try and do on the data set.

Other tools include Google docs for creating our milestones, allowing for easy editing by our whole group. GroupMe and Zoom will serve as our communication platforms. And Github for communication and version control [9].

7 Milestones

To date, our group members have each downloaded and saved the dataset on their local machines. We've also completed an initial project description that outlines our goals and proposed work. We have also performed an initial survey of the data and previous research that has been performed with this dataset. Looking ahead, these are our groups planned milestones.

March 24th: Part 3 Progress Report

End of March:

- Data preprocessing
- Joining data tables
- Initial exploratory data analysis
- Classification and clustering

End of April:

- Narrow down machine learning models
- Optimize selected model's hyperparameters
- Visualizations for final model's performance
- Visualizations for data analysis

May 4th:

- Final paper

8 References

[1] Instacart. 2017. The Instacart Online Grocery Shopping Dataset 2017. Retrieved March 10, 2019 from https://www.instacart.com/datasets/grocery-shopping-2017

[2] Stanley, Jeremy. 2017. 3 Million Instacart Orders, Open Sourced. (May 2017) Retrieved March 10, 2019 from https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

[3] Annie George. 2017. Instacart Market Basket Analysis - Reorder Analysis. (Sept. 2017) Retrieved March 10, 2019 from https://nycdatascience.com/blog/student-works/capstone/instacart/

[4] Philipp Spachtholz. 2017. Exploratory Analysis - Instacart. Retrieved March 10,

2019 from https://www.kaggle.com/philippsp/exploratory-analysis-instacart

[5] Wan, Mengting, Di Wang, Jie Liu, Paul Bennett, and Julian McAuley. 2018. Representing and Recommending Shopping Baskets with Complementarity, Compatibility and Loyalty. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, (Oct. 2018), 1133-1142. DOI: https://doi.org/10.1145/3269206.3271786

[6] Bhade, Kalyani, Vedanti Gulalkari, Nidhi Harwani, and Sudhir N. Dhage. 2018. A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), (Jul. 2018) 1-6. DOI: https://doi.org/10.1109/ICCCNT.2018.8494019

[7] Asha, K. N., and R. Rajkumar. 2017. Pre-processing of user behaviour for e-commerce. In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), (Aug, 2017) 715-719. DOI: https://doi.org/10.1109/ICCCNT.2018.8494019

[8] Stanley, Jeremy. 2017. data_description. (May 2017) Retrieved March 10, 2019 from https://gist.github.com/jeremystan/c3b39d947d9b88b3ccff3147dbcf6c6b

[9] Our Group's GitHub repository: https://github.com/MichaelWhitlock/CSCI4502Group1.git