# Instacart Market Basket Analysis

## Predicting product reorders

Cole Metzger, Hrithvik Ranka
Luddy School of Informatics, Computing, and Engineering, Indiana University, USA
Email: {cojmetz, hranka}@iu.edu

**Abstract – The Instacart dataset takes customers previous orders and asks challenge participants to answer key questions about what products will be reordered. We used a combination of different classification methods to determine reordering. We also showed association rules and a large portion of Exploratory Data Analysis in order to help Instacart see trends in the data.**

Keywords

Feature Engineering
Random Forest
Apriori
Naïve Bayes
Logistic Regression
Binary Classification

## I. Introduction

Instacart put out a large dataset in 2017 with their customer's orders over time. This dataset was added into a Kaggle competition which asked the primary question: "Which products will an Instacart consumer purchase again?" While the competition has been closed for over 5 years, we decided to see if we could go back and make predictions on the dataset using a multitude of different techniques. The goal for Instacart was to be able to predict which items that a customer has already purchased will also be reordered on subsequent transactions. Although that was one of our goals for analyzing this dataset, we also decided to gain general insights about Instacart users by performing Exploratory Data Analysis and using the Apriori algorithm to determine frequent item sets and association rules.

The Instacart data is comprised of relationally linked tables that when merged and combined, can be used to make predictions. There are two table structures that prove to be the most vital for understanding the dataset, order_products__*.csv and orders.csv. In order_products__*.csv (where the * is either prior or train), each row consists of one product, its associated order ID, whether or not it has been ordered in the past (labeled reordered), and also the number of the sequence in the order called add_to_cart_order (i.e., if an item was added to the cart 2nd, it will have a add_to_cart_order of 2). The goal for making predictions is to determine whether future products will have a 1 or a 0 in the reordered column, indicating whether or not a product will be reordered. Now, for the orders.csv table, we have more information regarding each individual cart transaction such as day of the week, hour of the day, and days since most recent prior order. Each cart transaction, or order, in orders.csv has 1 or more

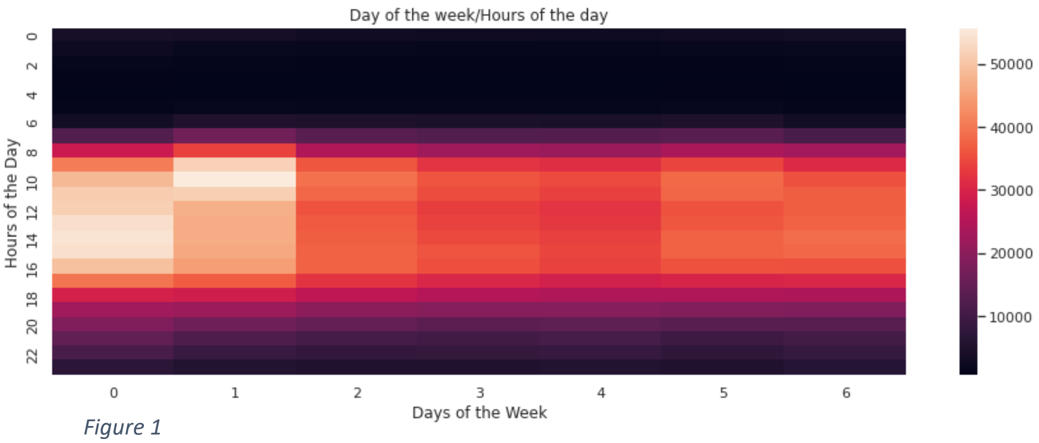
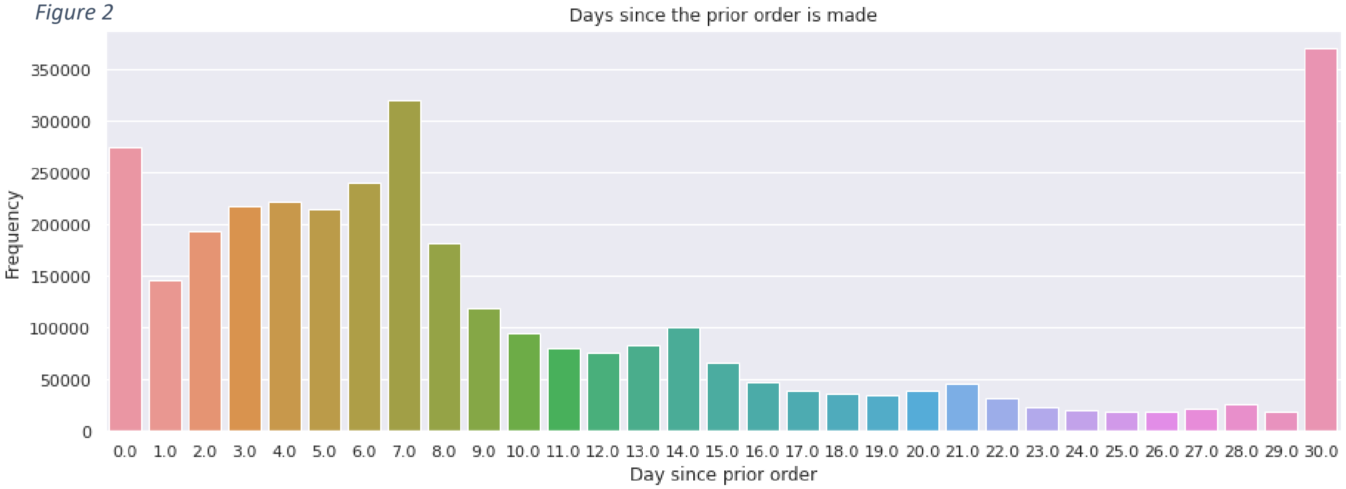


*Figure 1*

*Figure 2*



products associated with it in the order_products__*.csv table, indicating a one-to-many relationship between them. Aside from these two tables, we also have a couple of tables that help describe each product. These are departments.csv, aisles.csv, and products.csv. Each product is named and connected with department ID and aisle ID inside of the products.csv file. In department.csv and aisles.csv we simply get a list of all the possible named departments and aisles respectively.

## Exploratory Data Analysis

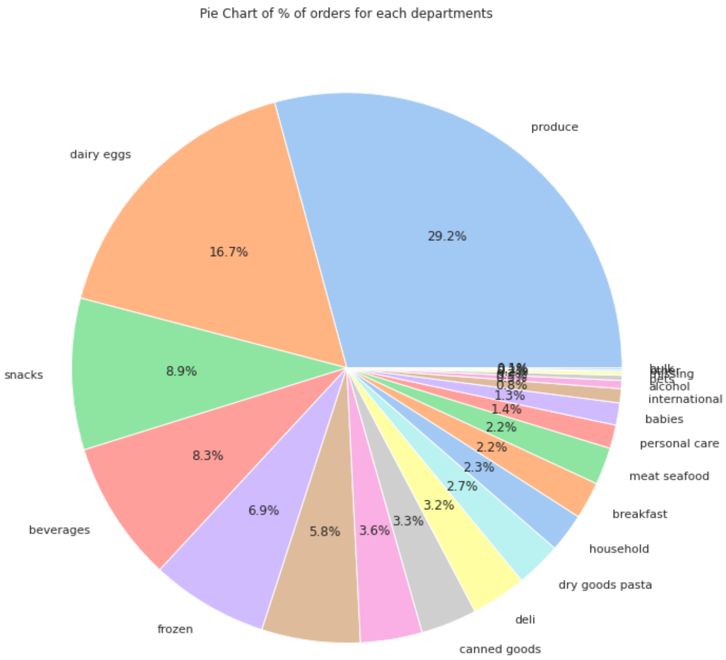


*Figure 3*

When exploring the dataset, we found a few trends that would help us to generate features later. We took the approach of doing EDA on all of the separate files, as well as on merged data. Some of the most important insights can be seen from merged data, but there are also helpful insights from the unmerged datasets. In figure 1, we can see for example that most people are ordering early in the week (Monday and Tuesday) between the hours of 8 a.m. and 4 p.m. (index 0 indicates the first day of the week, Monday).

Figure 1 can help Instacart determine when the most orders are placed, and how to plan accordingly. In figure 2, we charted the number of days since the prior order was made, with the frequency on the y-axis showing the number of total orders. From figure 2, there are clear spikes at day 7 and day 30, which indicates that Instacart users are ordering either weekly, or monthly.

Now, when we move on to analyze the merged data frames, where we merge together the orders and each product associated with that order, as well as department and aisle data, we get even more telling results about what users are ordering. In figure 3 we can see that departments produce, dairy/eggs, snacks, and beverages are the most commonly ordered from.

Figure 4, which shows the most commonly ordered form aisles, also shows similar trends for which kinds of products are being ordered. The top 2 aisles are fresh fruits, and fresh vegetables. From both of these figures we can see that spoilable food is the most common type of food that is reordered the most.

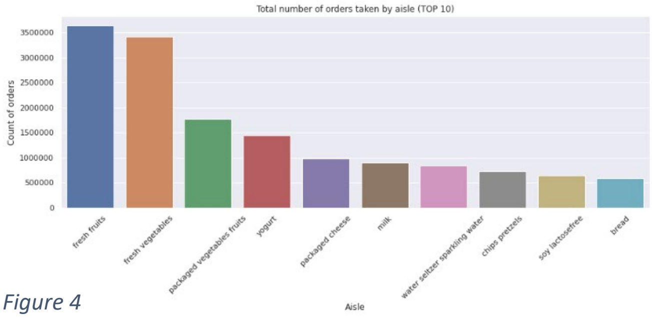


*Figure 4*

This will help us when we start to engineer features for classification.

## II. Methods

For binary classification, the 3 methods we used were Random Forest, Naïve Bayes, and Logistic Regression.

Random forest classifiers are a popular choice for market basket analysis because they are able to handle large datasets and can identify complex relationships between items. Additionally, random forests are resistant to overfitting, which can be a common problem when working with market basket data. This means that a random forest classifier can provide accurate predictions on new data, even when dealing with many variables and large amounts of data. Overall, the use of a random forest classifier for market basket analysis can provide valuable insights into customer purchasing behavior and can help businesses make informed decisions about their product offerings. Naive Bayes classifiers are another type of algorithm that can be used for market basket analysis. This method is based on the assumption that each feature (i.e., item in the market basket) is independent of the others. This allows the algorithm to make predictions based on the probabilities of each individual feature, rather than trying to consider the interactions between all of the features.

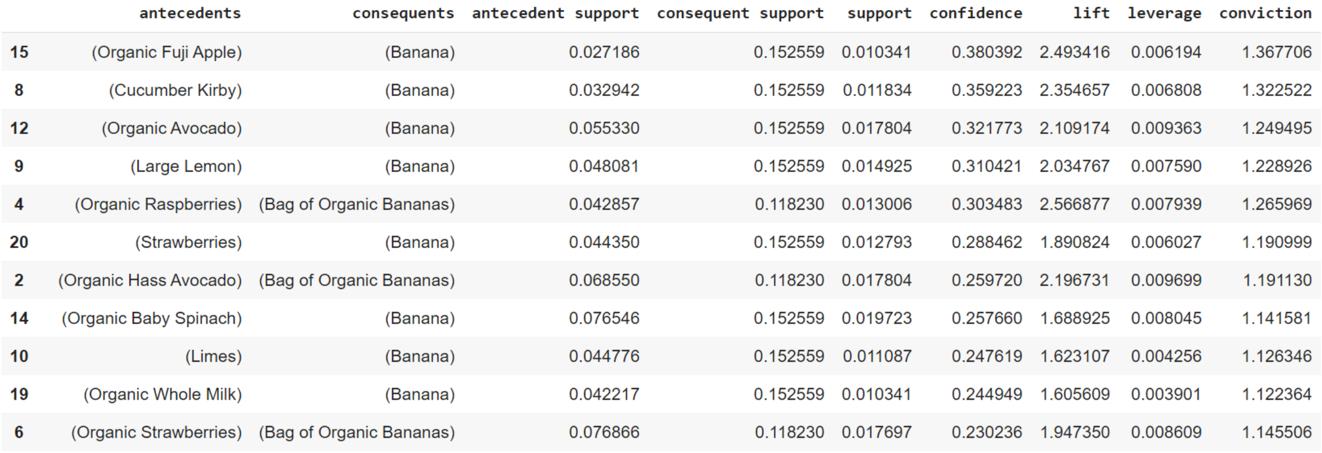
Logistic regression classifiers are a type of regression model that is often used for binary classification tasks. In the context of market basket analysis, this method could be used to predict whether a given customer will purchase a particular item. Logistic regression models are able to handle large datasets and can account for complex relationships between features.

Overall, each of these three classifiers has its own strengths and limitations, and the choice of which one to use will depend on the specific requirements and characteristics of the market basket data.

We also did feature engineering based on the EDA that we performed to try to get a better accuracy. For our features, we added 4 department features to check if a given product was in one of the top 4 departments. Part of the idea for department-based feature engineering came from Sagar [1]. Also, we decided to add a feature called "*isSpoilable*" in order to better make predictions. This feature was labeled as a 1 if the department was one of the following: {bakery, produce, meat seafood, dairy eggs, deli}. We also made the feature labeled as 1 if the aisle was in one of the following: {fresh fruits, fresh vegetables, packaged vegetables fruits, yogurt, packaged cheese, milk, bread}.

Along with our classification methods, we also performed the Apriori algorithm on the orders. The

*Figure 5*

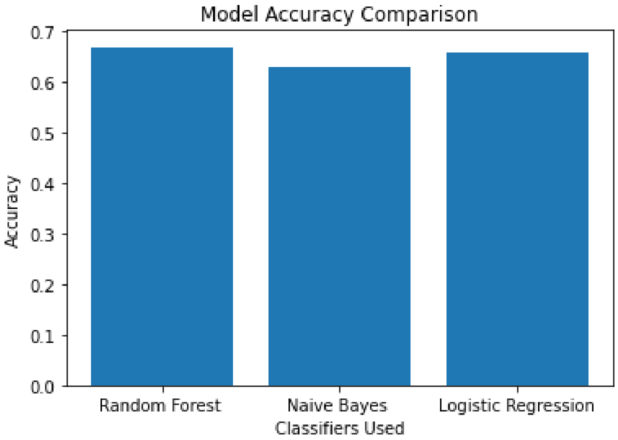


| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| **15** | (Organic Fuji Apple) | (Banana) | 0.027186 | 0.152559 | 0.010341 | 0.380392 | 2.493416 | 0.006194 | 1.367706 |
| **8** | (Cucumber Kirby) | (Banana) | 0.032942 | 0.152559 | 0.011834 | 0.359223 | 2.354657 | 0.006808 | 1.322522 |
| **12** | (Organic Avocado) | (Banana) | 0.055330 | 0.152559 | 0.017804 | 0.321773 | 2.109174 | 0.009363 | 1.249495 |
| **9** | (Large Lemon) | (Banana) | 0.048081 | 0.152559 | 0.014925 | 0.310421 | 2.034767 | 0.007590 | 1.228926 |
| **4** | (Organic Raspberries) | (Bag of Organic Bananas) | 0.042857 | 0.118230 | 0.013006 | 0.303483 | 2.566877 | 0.007939 | 1.265969 |
| **20** | (Strawberries) | (Banana) | 0.044350 | 0.152559 | 0.012793 | 0.288462 | 1.890824 | 0.006027 | 1.190999 |
| **2** | (Organic Hass Avocado) | (Bag of Organic Bananas) | 0.068550 | 0.118230 | 0.017804 | 0.259720 | 2.196731 | 0.009699 | 1.191130 |
| **14** | (Organic Baby Spinach) | (Banana) | 0.076546 | 0.152559 | 0.019723 | 0.257660 | 1.688925 | 0.008045 | 1.141581 |
| **10** | (Limes) | (Banana) | 0.044776 | 0.152559 | 0.011087 | 0.247619 | 1.623107 | 0.004256 | 1.126346 |
| **19** | (Organic Whole Milk) | (Banana) | 0.042217 | 0.152559 | 0.010341 | 0.244949 | 1.605609 | 0.003901 | 1.122364 |
| **6** | (Organic Strawberries) | (Bag of Organic Bananas) | 0.076866 | 0.118230 | 0.017697 | 0.230236 | 1.947350 | 0.008609 | 1.145506 |

algorithm was used to determine which association rules could be found from a subset of the data. This would help in determining which items were bought in pairs or in tandem with others. Association rules

## III. Results

For Apriori, we can see the top 10 results for association rules in figure 5.For the classification algorithms, the results can be seen in figure 6.



## IV. Discussion

For the future, we would make changes to the project by adding in more features related to product and customer before running the algorithms. We got low scores on all the algorithms, which means changes need to be made to our sampling technique, feature engineering, or feature selection. Because all 3 of the classifiers got similarly low scores, this means that the problem isn't in the algorithms but in the input data.

As for the Apriori algorithm, we found an overabundance of rules that were generated containing a "banana". This is because we took a small portion of the data. Running Apriori on large amounts of data can be extremely costly and expensive even with GPU performance. In the future we would run the Apriori multiple times on different subsets of the data, and then combine all the results in the end.

A future possibility would be to somehow use the association rules to help make classifications. This would require us to go back to the ground up models, and not just use the basic scikit models for Random Forest and others. With association rules, we could help the classifier see if a product is frequently bought with another. If it is frequently bought with another, then we would increase the chance of that product being reordered.

## REFERENCES

[1] Sagar, Arun. "Instacart Market Basket Analysis&nbsp;: Part 2 ( Fe &amp; Modelling)." Medium, Medium, 5 Apr. 2021, https://asagar60.medium.com/instacart-market-basket-analysis-part-2-fe-modelling-1dc02c2b028b.

[2] Kaur, Manpreet, and Shivani Kang. "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining." CyberLeninka, Elsevier BV, 1 Jan. 1970, https://cyberleninka.org/article/n/1371972.

[3] "Instacart Market Basket Analysis." Kaggle, https://www.kaggle.com/c/instacart-market-basket-analysis.

[4] A Comparative Study on Market Basket Analysis and Apriori Association. https://ieeexplore.ieee.org/abstract/document/7231468.

[5] Team, Kaggle. "Instacart Market Basket Analysis." Medium, Kaggle Blog, 7 Jan. 2020, https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded.