A Project Report On

Instacart Market Basket Analysis

By
Garima Singh

TABLE OF CONTENT

LIST OF GRAPHS

LIST OF FIGURES

LIST OF TABLES

1. Problem Statement

Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you. The aim is to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order. We will use this data to test models for predicting products that a user will buy again or add to cart next during a session.

2. Data Used

The dataset is a relational set of files describing customers' orders over time. The goal is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, we are provided between 4 and 100 of their orders, with the sequence of products purchased in each order. We are also provided with the week and hour of day the order was placed, and a relative measure of time between orders.

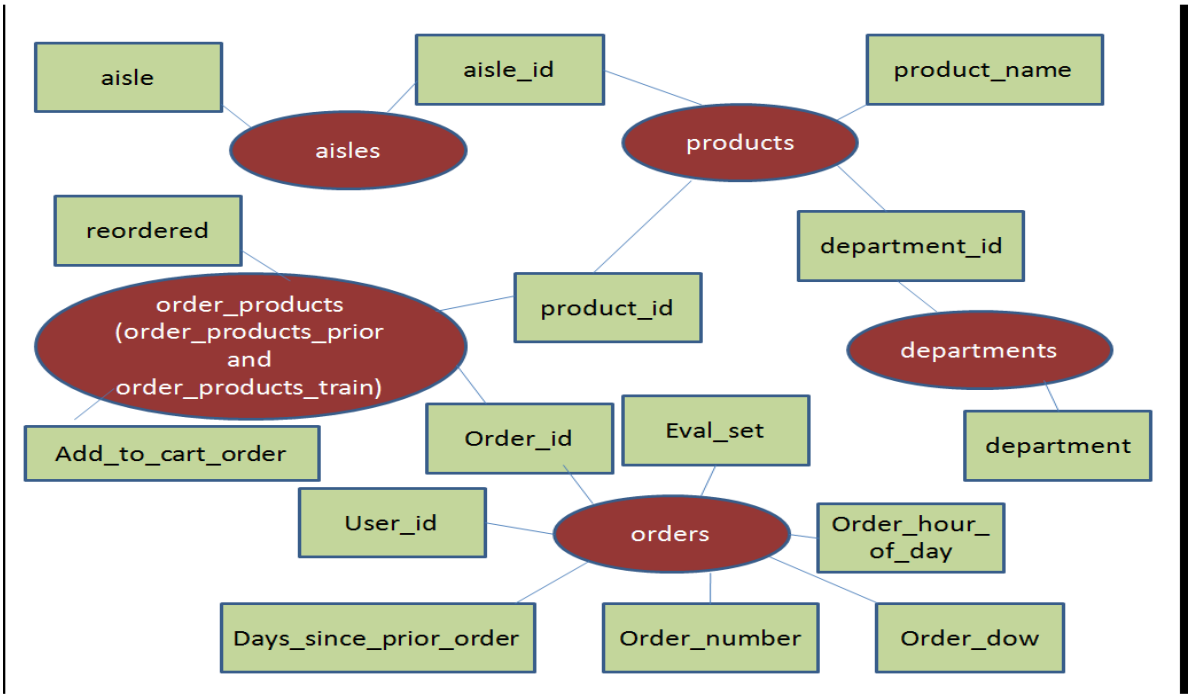


Figure (a) : Entity Relationship Diagram

3. Exploration of Numerical Variable:

We will discuss each dataset seperately and then will perform some operations over them like merging, grouping, mathematical calculations, summarising, etc.

a) Merging the data and taking out insights

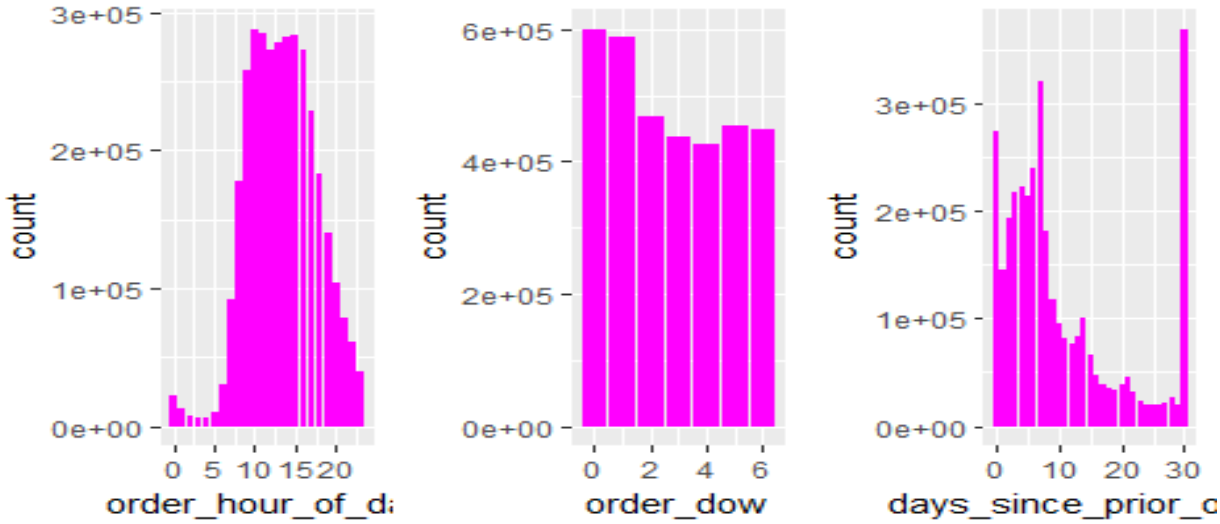
- The orders table contains variable eval_set consisting of 3 distinct values test(75,000), train(131,209) and prior (3,214,874)
- Test is the actual test set which needs to be predicted

- Train is the recent order by the user.
- Prior are the previous orders ordered by the user
- The factor variable reordered tells which items have been reordered (1) and not (0).
- After getting the data altogether it has dimension of 7 variables and 3421083 observations. Out of 7 variables 7 were numerical variable and 5 were categorical variable, out of which one is dependent variable.

b) Checking the distribution of the variables:

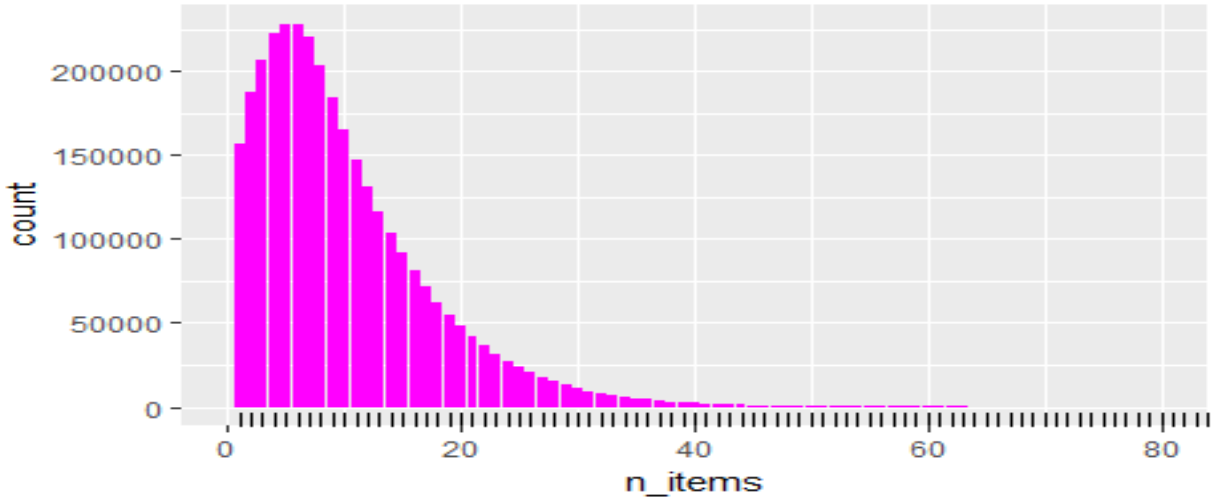
Histogram was plotted to check the distribution of the variables. What we get to know is:

- Most orders are between 7.00 am-19.00pm
- Maximum orders are placed on day 0 and 1
- Most of the orders are either new or placed after a week or after a month



Graph (a) : Histogram plot for order_hour_of_day, order_dow and days_since_prior_order

- Now lets see how many products are being ordered at a go and how many are reordered product.



Graph (b) : Number of items ordered at once

From the graph we can see people usually order 3-8 products and approximately 58-60% products are reordered.

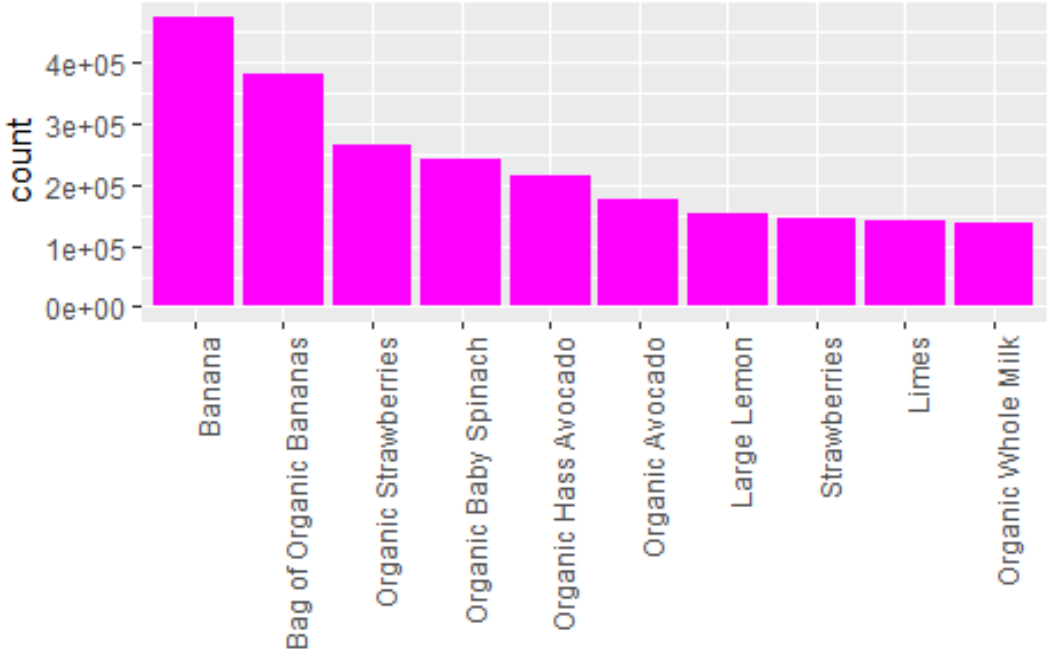
Reordered	Total Products
0	13863746
1	19955360

Table (a) : Number of products reordered

Our job is basically to create a new train set which contains user_id order_number order_dow order_hour_of_day days_since_prior_order, probablity of reorder and product_id and we will train our model based on this data and predict the product_id's for the test data.
So we merge the order_products and orders dataset on order_id.
Upon summarising the tables, we find that there are no missing data except for orders table where we find that 206209 orders are first time orders and so the values in days_since_prior_order is NA. We will impute it with 0.

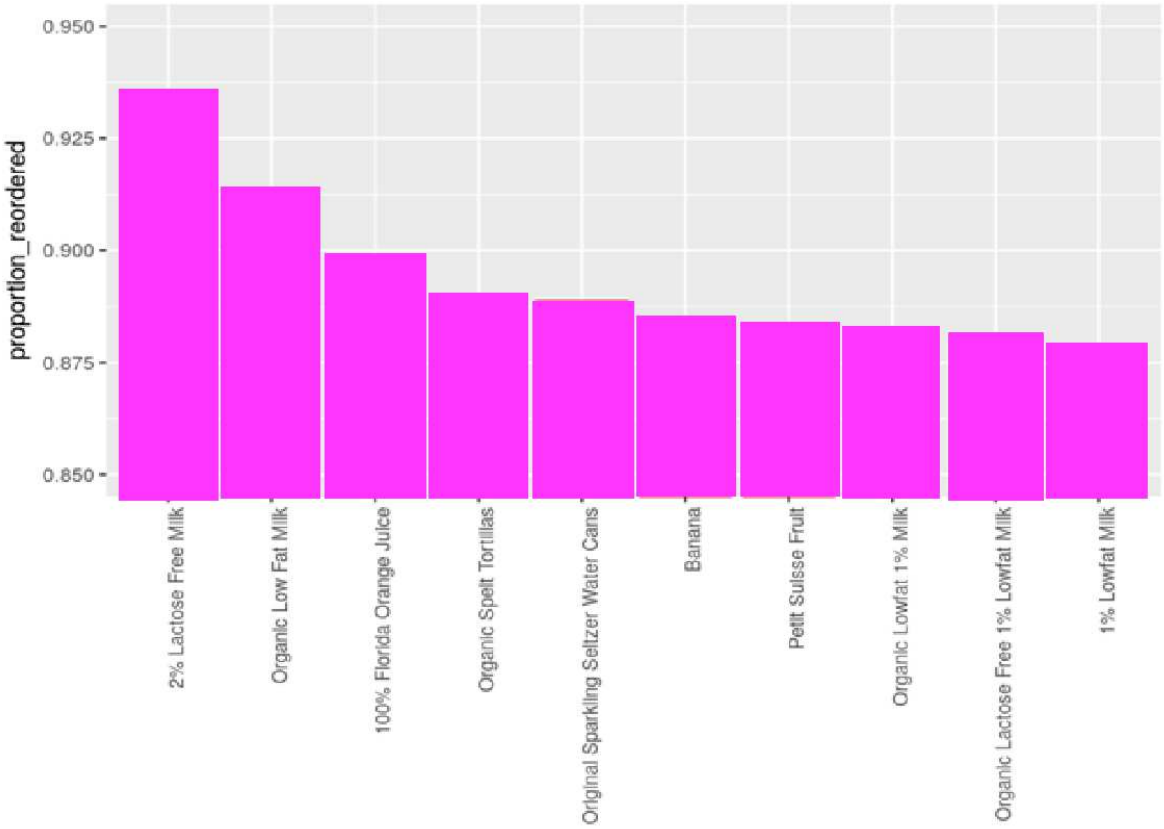
- Now we will connect various datasets
  - products table with the aisles on aisle_id
  - products table with orders on product_id to get the most ordered product.

  From the graph we may conclude that fruits and milk are the most commonly ordered products with bananas having the highest frequency.



Graph (c) : Most commonly ordered items

- Now lets check which are the most reordered items.
- Next we calculate the probablity of reordering the products as that would decide how much people liked the product.



Graph (d) : Most reordered items

Based on the data gathered and feature extracted, we try to check our training data by dividing it in 80:20 ratio for train and test. We applied XGBoost Model and C5.0 model to check whether the product is reordered next time.

(c) Applying XGBoost Algorithm

Upon applying this algorithm, Confusion Matrix and Accuracy is checked and the model is 60.2% accurate.

Predicted/observed	**0**	**1**
0	1111	113088
1	200	170218

Table (b) : Confusion Marix for XGBoost Algorithm

(d) Applying C5.0 Model

Upon applying this algorithm, Confusion Matrix and Accuracy is checked and the model is 67.01% accurate.

Predicted/observed	0	1
0	49712	64487
1	29398	141020
Accuracy : 67.01%		

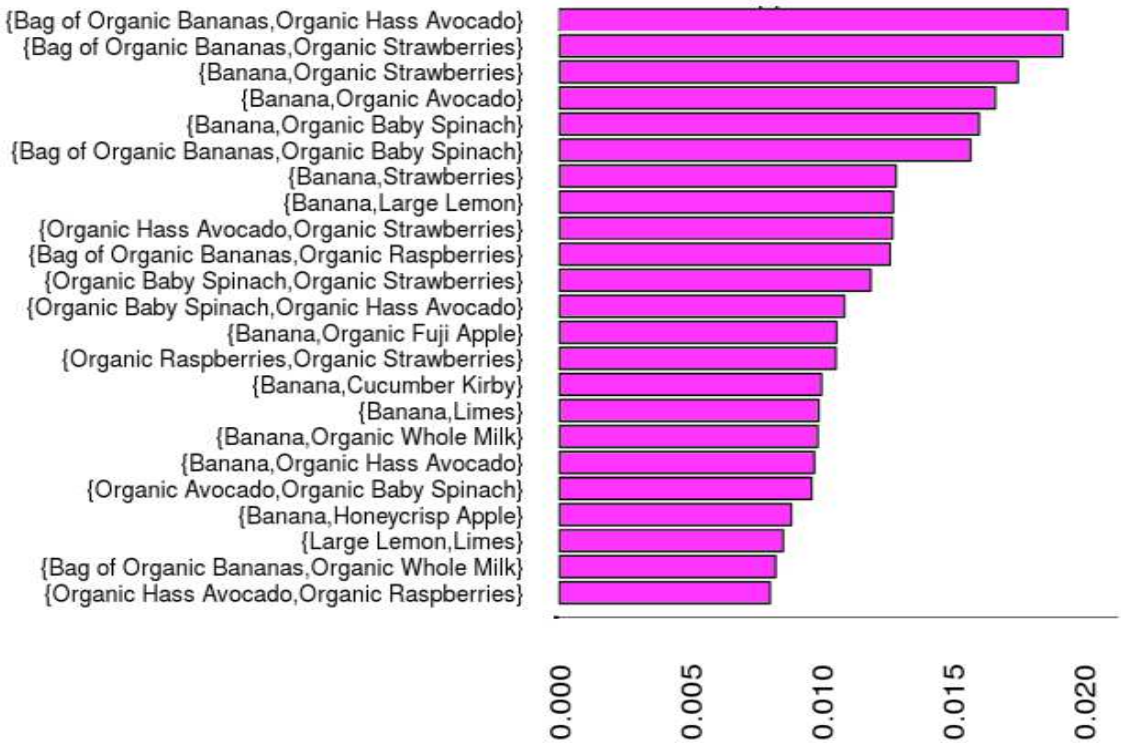
Table (c) : Confusion Matix for C5.0 Model

Upon running both the models we found that the C50 model produced better results with better accuracy. But this way are not able to predict the cart items for next order.

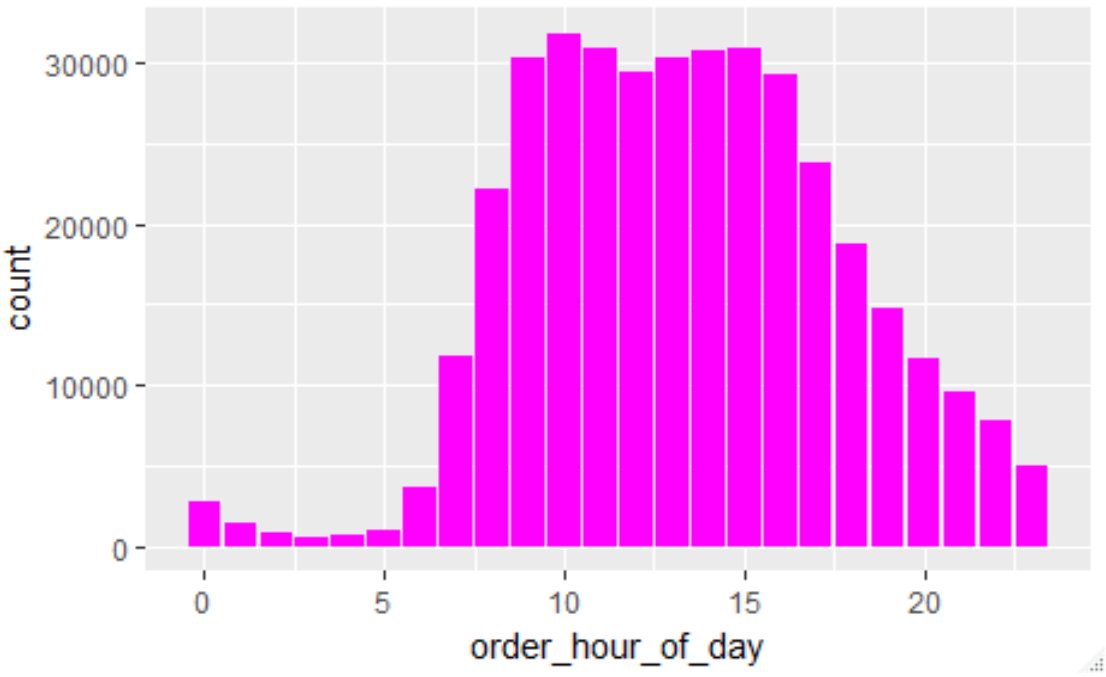
Thus we need to make Rules for predicting the basket of future orders.

4. Deriving Association Rules from the Dataset

- We will now check which itemsets are frequent with support of 0.8% and we found that people mostly ordered combination of Bag of Organic Bananas, Avacados and Strawberries together. Also the orders are maximum during the morning hours between 10am-5pm.



Graph (e) : Most common itemsets



Graph (f) : Time between which the itemsets in Graph (e) are ordered

Likewise, we may apply certain set of association rules to predict the actual basket items. Here we use Apriori Rule by translating the dataset into transaction type for applying arules algorithm. To make the most of the dataset, the support is kept high and the confidence level is kept low.
Next we will apply arules apriori algorithm with minimum support of 0.1% and confidence of 25%

support	confidence	lift
Min. :0.001004	Min. :0.2500	Min. : 1.763
1st Qu.:0.001165	1st Qu.:0.2718	1st Qu.: 2.415
Median :0.001386	Median :0.3012	Median : 3.369
Mean :0.001950	Mean :0.3221	Mean : 5.789
3rd Qu.:0.001808	3rd Qu.:0.3540	3rd Qu.: 4.592
Max. :0.023279	Max. :0.6250	Max. :86.850

Table (d) : Summary of Quality Measures

Next when we inspect the rules after running the algorithm and applying the mentioned support and confidence.

LHS	RHS	Support	Confidence	Lift
Baby Cucumbers	Bag of Organic Bananas	0.2515723	0.001205122	2.152274

Sweet Potato Yam	Banana	0.001174994	0.3861386	2.719988
Organic Honey Sweet Whole Wheat Bread	Bag of Organic Bananas	0.001064524	0.3642612	3.116359
---	---	---	---	---

Table (e) : Apriori Rules

Thus using the rules, we may predict the relative itemsets to be present in the next time order cart. 36% of people who are buying Bag of Organic Bananas also bought Organic Honey sweet whole wheat bread and 38% bought Sweet potato Jam along with Organic Bananas.
Though only 0.11% people bought from the former category, latter was bought by 0.10% people.
Again if we further classify the data according to time, day and other factors, one may predict more accurate results.