Report

CUSTOMER BEHAVIOR MODELING

Mini-Project
Course Applied Statistics and Experimental Design 2023.2

Trieu Kinh Quoc - 20225524
Dao Van Tung - 20204932
Nguyen Tri Thanh - 20225457
Nguyen Xuan Thanh - 20225460

ACKNOWLEDGEMENT

We would like to express our deep gratitude to Assoc.Prof Nguyen Linh Giang, who gives us this golden opportunity to work on this project. Without his tremendous support and instruction, we cannot complete this project.

ABSTRACT

The primary objective of all sales departments is to understand client behavior. In this project, we'll uncover and extract data that can benefit the business. Our approach will follow two strategies. The first strategy involves using a classification model to predict whether a client will repurchase an item they have previously bought. The second strategy employs the Apriori algorithm, which uses statistical association rules to identify sets of items or purchase patterns (the purchasing habit rule effect) that frequently occur together. Both strategies will provide insights into consumer behavior and reveal trends in shopping carts. The report will include formulas, the implementation stage of the algorithms, our findings, and a conclusion.

Keywords. Customer behavior modelling, Apriori, SMOTE, Random Forest, XGBoost, LGBM.

Contents

1 Introduction

The application of technology in commercial operations has gained significant attention due to its potential to provide valuable insights into customer trends. These insights assist companies in formulating effective commercial strategies.

This project explores the field of Customer Reorder Behavior prediction using advanced machine learning methods, specifically random forests and gradient boosting algorithms, to predict the likelihood of product reorders. To achieve higher efficiency and accuracy, we will implement sophisticated feature engineering techniques and utilize association rules to analyze product-product and product-time relationships. Our evaluation metric is the F1 score, ensuring a balanced assessment of precision and recall.

By employing advanced machine learning as well as statistical analysis techniques, we aim to propose a suitable solution for making prediction regarding the current products that each customer bought.

2 Data description

The dataset that we used is from Instacart - an American company that operates a grocery delivery and pick-up service. This dataset contains a sample of over 3 million grocery orders from more than 200000 Instacart users . For each user, there are between 4 and 100 orders with the sequence of purchased products in each order. Moreover, the dataset also includes of the time the order was placed and a relative measure of time between orders. Our dataset contains 6 different *csv* files which are: *aisle.csv, departments.csv, order_products.csv, orders.csv, products.csv* . Each entity (customer, product, order, aisle, etc.) has an associated unique *id* so that we can more easily observe our data.

The *aisle.csv* contains the aisle categories of all products and their id. There are 134 different categories.

aisle_id	aisle
1	prepared soups salads
2	specialty cheeses
3	energy granola bars
4	instant foods
5	marinades meat preparation
6	other
7	packaged meat
8	bakery desserts
9	pasta sauce
10	kitchen supplies
11	cold flu allergy

Figure 1: The first 11 categories from *aisle.csv*

The department of all products are stored in *departments.csv*. There are total 21 departments with regard to the products.

departme	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol
6	international
7	beverages
8	pets
9	dry goods pasta
10	bulk
11	personal care
12	meat seafood
13	pantry
14	breakfast
15	canned goods
16	dairy eggs
17	household
18	babies
19	snacks
20	deli
21	missing

Figure 2: 21 departments in our dataset

The products.csv provides all the information of products in this market (id, name, aisle category, department id). Overal, there are 49688 products in this market.

product_id	product_name	aisle_id	departme
1	Chocolate Sandwich Cookies	61	19
2	All-Seasons Salt	104	13
3	Robust Golden Unsweetened Oolong Tea	94	7
4	Smart Ones Classic Favorites Mini Rigatoni With Vodka Cream Sauce	38	1
5	Green Chile Anytime Sauce	5	13
6	Dry Nose Oil	11	11
7	Pure Coconut Water With Orange	98	7
8	Cut Russet Potatoes Steam N' Mash	116	1
9	Light Strawberry Blueberry Yogurt	120	16
10	Sparkling Orange Juice & Prickly Pear Beverage	115	7
11	Peach Mango Juice	31	7
12	Chocolate Fudge Layer Cake	119	1
13	Saline Nasal Mist	11	11

Figure 3: The first 13 products

The orders_products.csv contains all the orders which includes: the id of the order, the products, the order that customer added to cart. Reorder indicates that the customer has a previous order that contains the products.

order_id	product_id	add_to_cart_	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1
1	13176	6	0
1	47209	7	0
1	22035	8	1
36	39612	1	0
36	19660	2	1
36	49235	3	0
36	43086	4	1

Figure 4: The detail of the order with id 1 and 36

The orders.csv represents the full information of each order: order_id, customer id, set of the order that it belongs to, the numbers of each order and the time(hour) of the order as well as days since prior order.

order_id	user_id	eval_set	order_number	order_dov	order_hour_of_da	days_since_prior_order
2539329	1	prior	1	2	8	
2398795	1	prior	2	3	7	15
473747	1	prior	3	3	12	21
2254736	1	prior	4	4	7	29
431534	1	prior	5	4	15	28
3367565	1	prior	6	2	7	19
550135	1	prior	7	1	9	20
3108588	1	prior	8	1	14	14
2295261	1	prior	9	1	16	0
2550362	1	prior	10	4	8	30

Figure 5: The orders in the market.

3 Exploratory data analysis

3.1 Ordering Time Analysis

First, we observe the distribution of time products are most likely to be ordered.

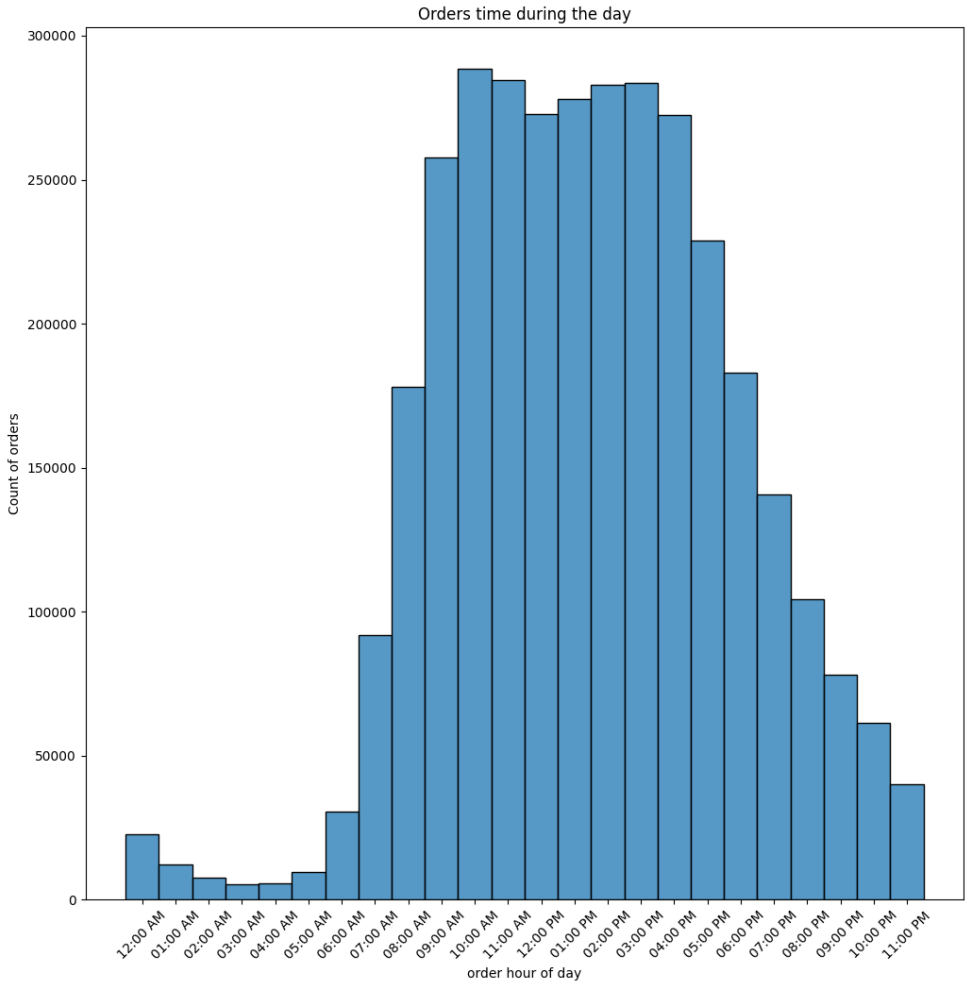

Figure 6

From Figure 6, we could see that the majority of products are ordered around 9:00 AM to 5:00 PM, which is the traditional working hour in the day. After 5:00 PM, the number of customers decreases and very few customers decide to place orders at midnight.

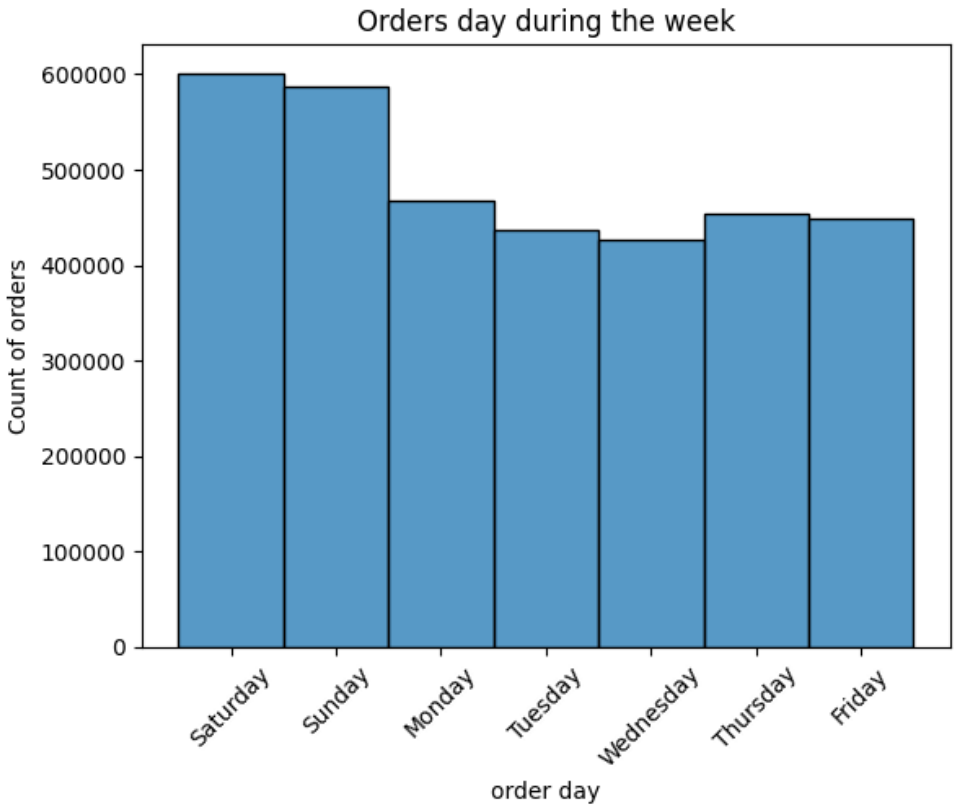

Figure 7

Surprisingly, the number of orders remained quite stable during the whole week, with just a slight jump in the weekends.

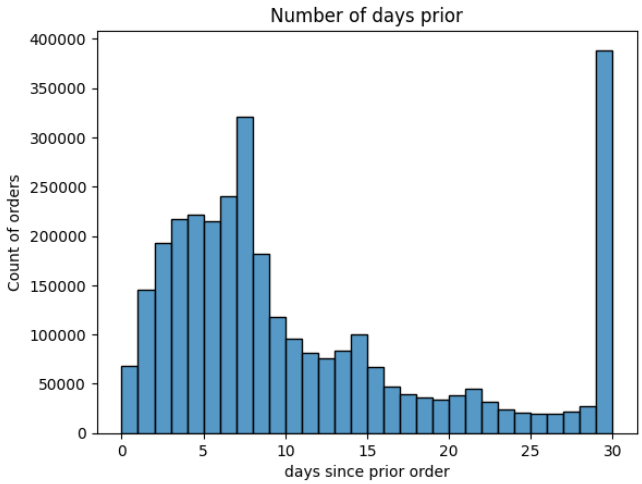

Figure 8

Most customers place their order around 3 to 8 days after their previous one. Customers seem to reorder after one week as we see peaks of order at 7, 14, 21 days. From Figure 8, we also notice that there are

many orders that are 30 days since prior order, this is because all the values greater than 30 are coerced to 30.

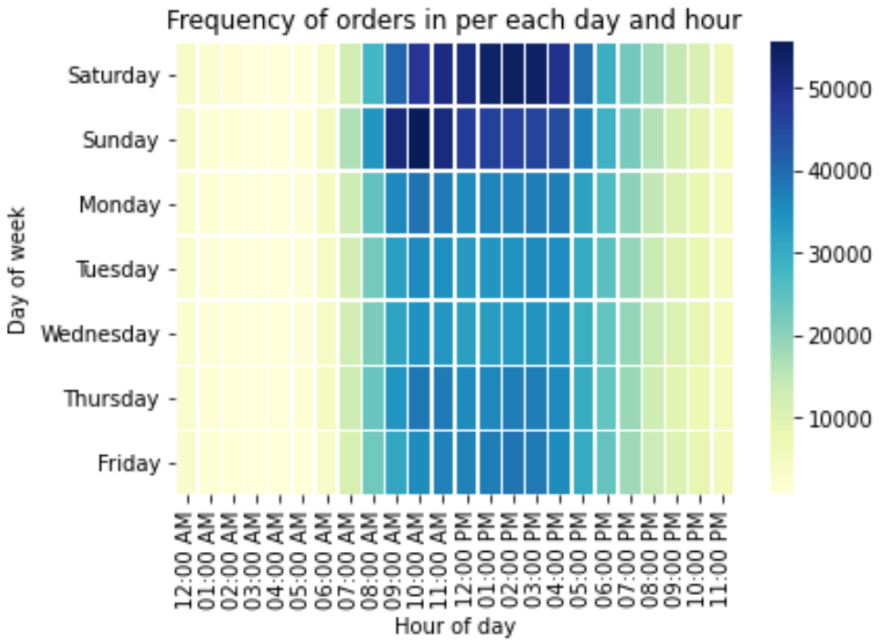

Figure 9

We create the heatmap of the count of numbers with each hour of each day, as shown in Figure 9. Saturday afternoon and Sunday morning seem to be the prime time for ordering.

3.2 Product Analysis

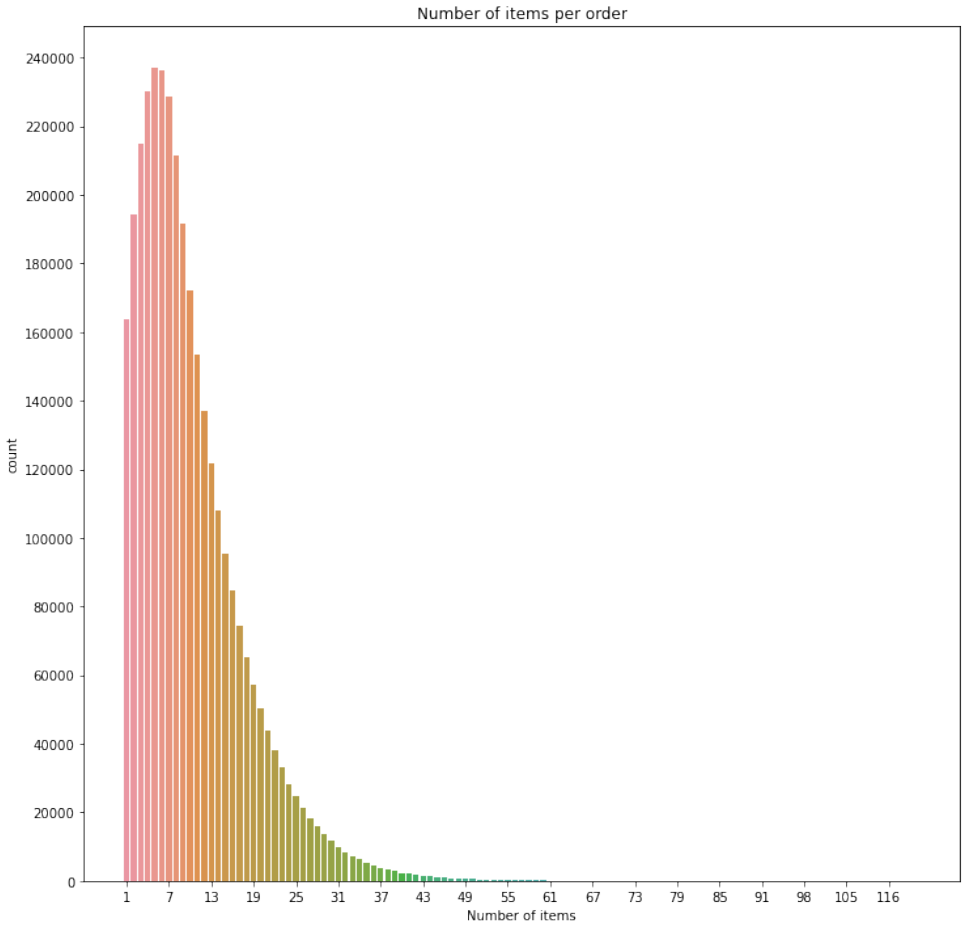

Figure 10

Figure 10 illustrates the number of items per order in all customers. Most orders contain between 3 and 13 products.

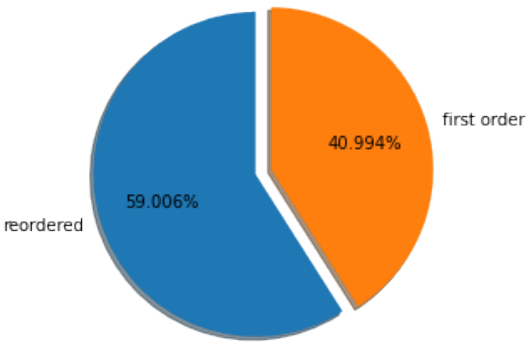

Figure 11

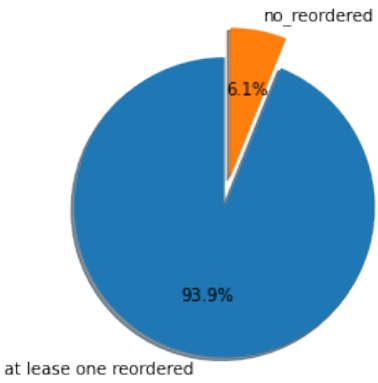

Figure 12

The number of reordered products accounts for 59%, while only 41% of products are the first time being ordered. Moreover, 93,9% of orders contain at least one reordered product, which infer that customers have a high tendency to rebuy products that they have already ordered and used before.

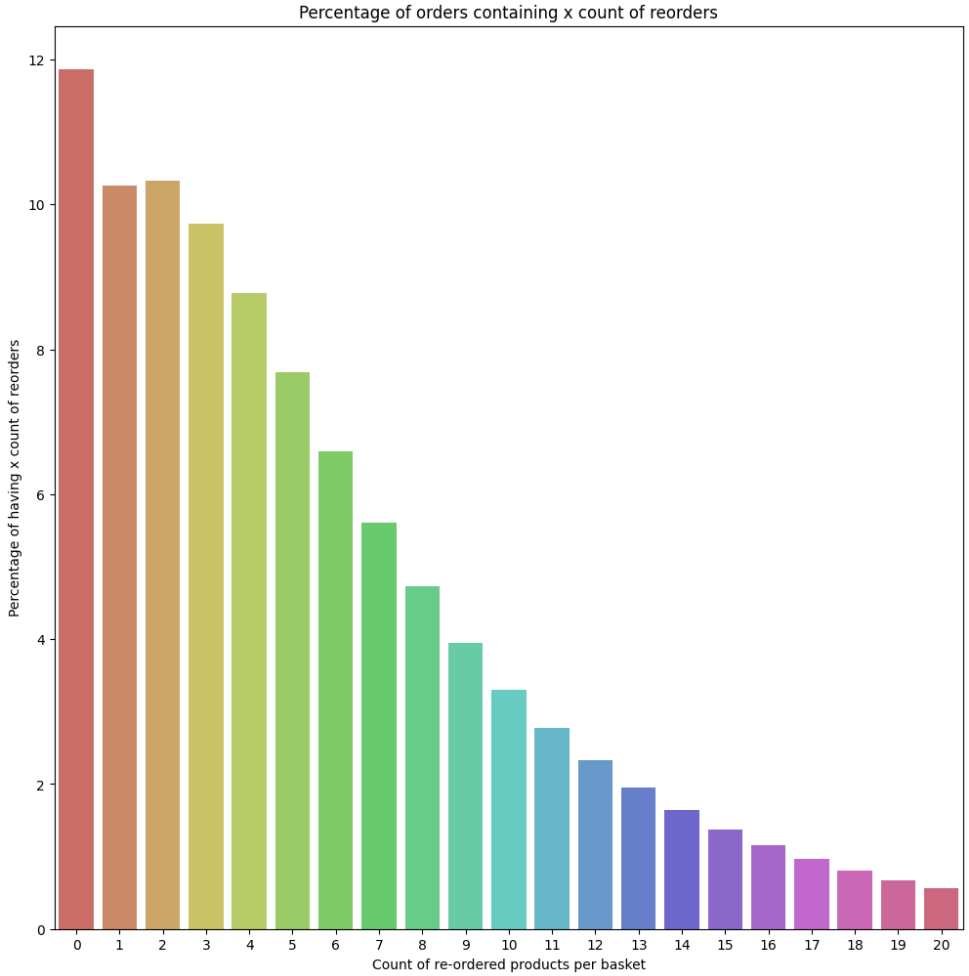

Figure 13

As we can see, most baskets contain from 0 to 6 reordered products.

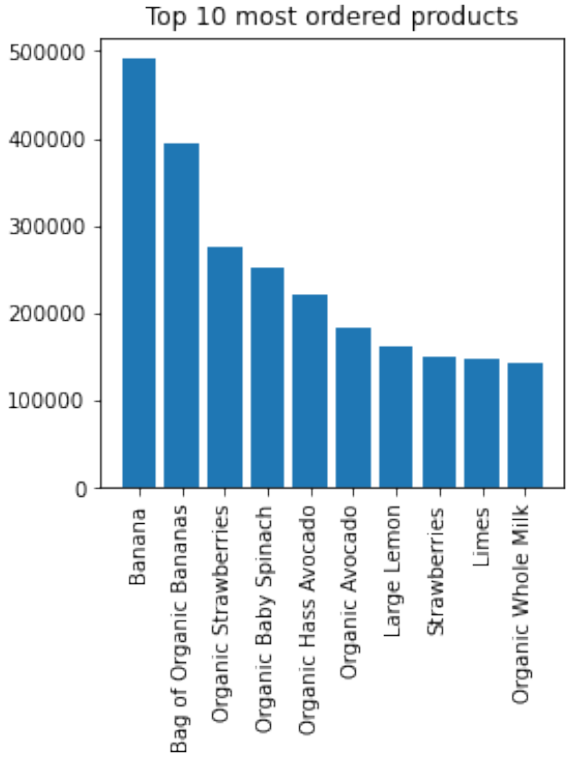

Figure 14

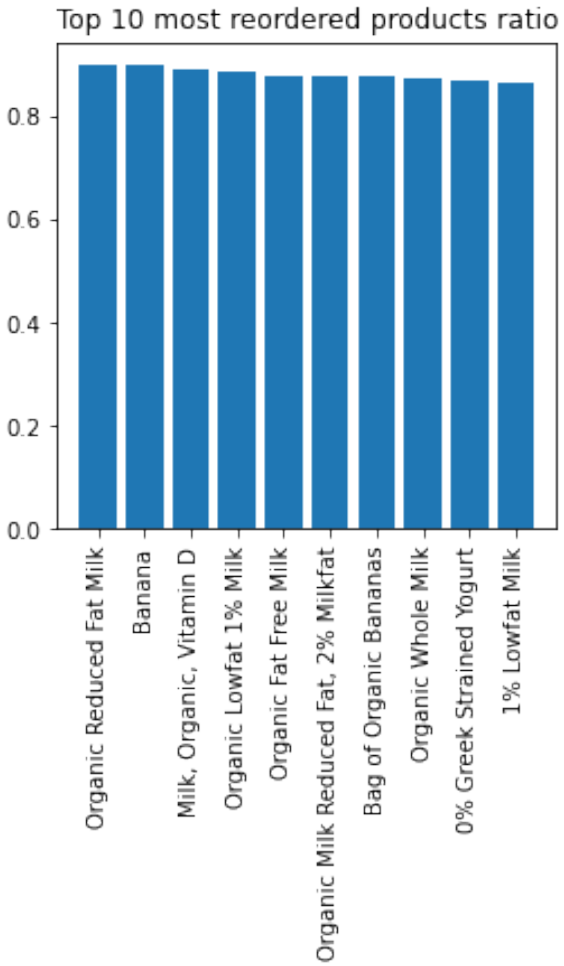

Figure 15

Banana is the bestseller, followed by Organic strawberries and organic baby spinach. Meanwhile, Organic reduced fat milk is the product with the highest probability of being reordered.

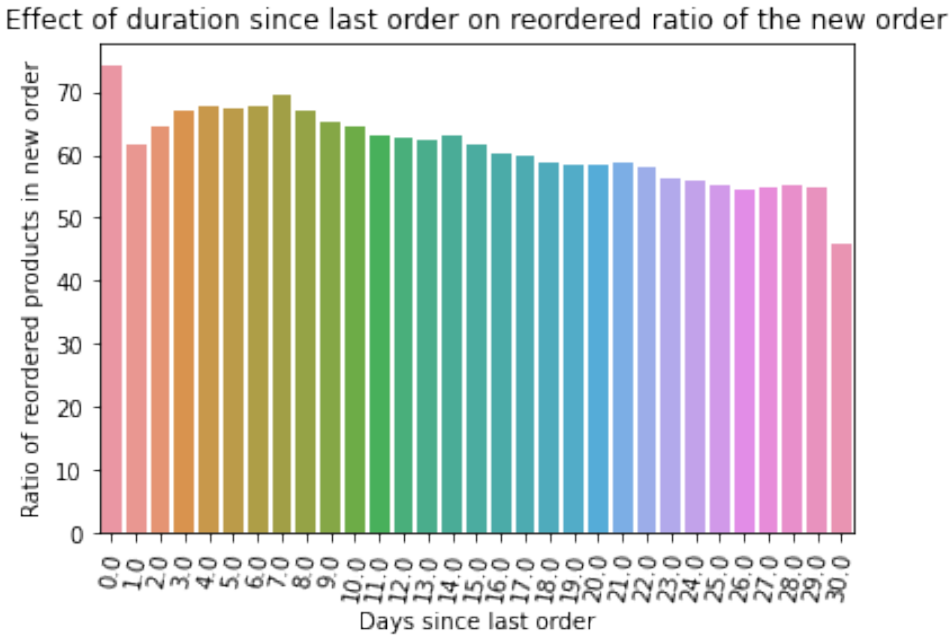

Figure 16

74% of products bought at the same day of previous order are reordered and 69% of products bought after 1 week of previous order, are reordered.

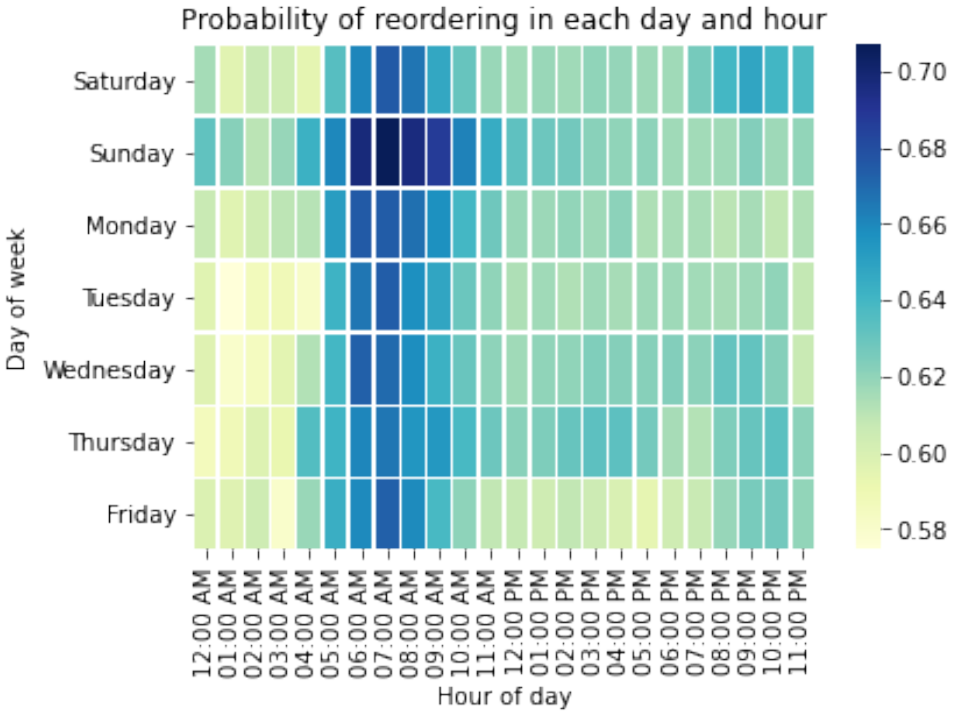

Figure 17

Customers tend to reorder products around 6AM to 8AM (especially on Sunday), these are the times

we should recommend customers with products that they have bought before.

3.3 Departments and Aisles analysis

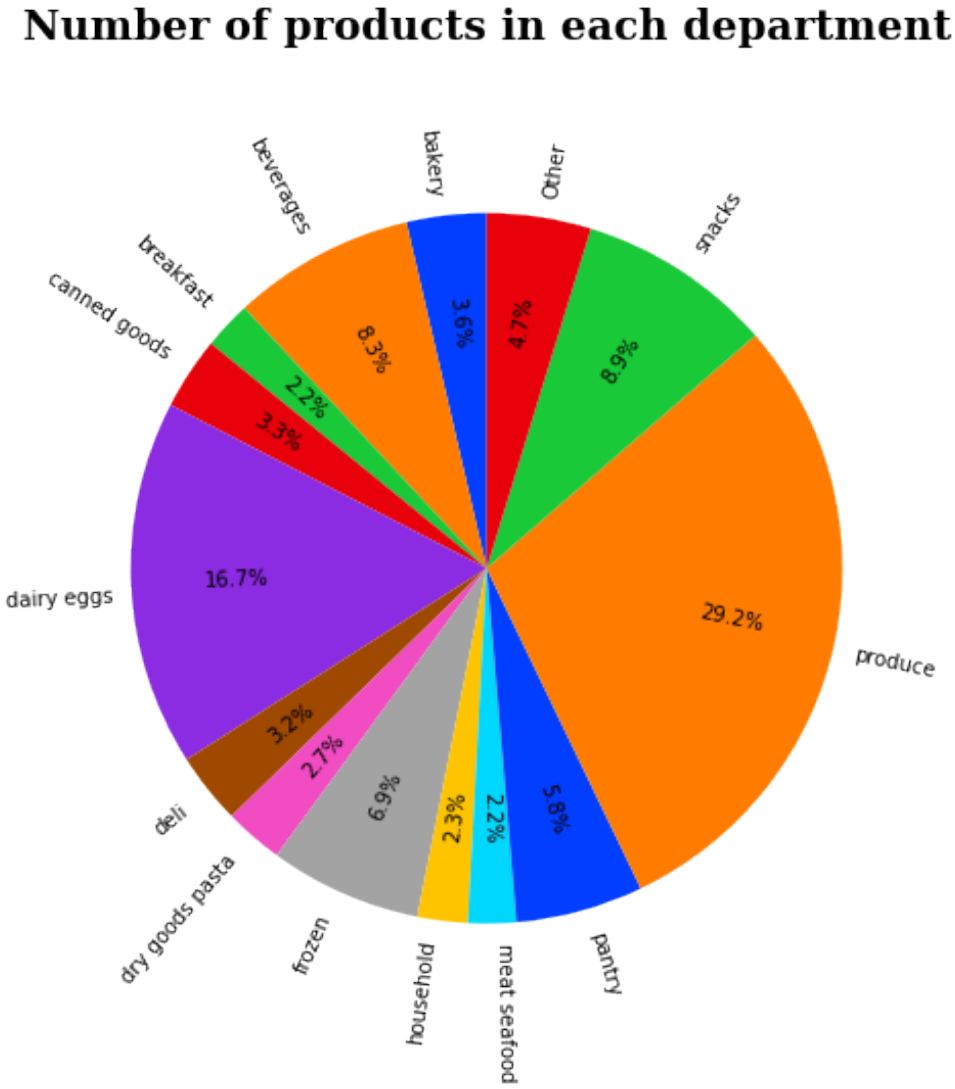

Figure 18

The produce department is the dominant sector with 29.2% of products, which is almost double the number of products in department dairy eggs in second place. Beverage and frozen products are also two important departments, as they account for 8.3% and 6.9% of products, respectively.

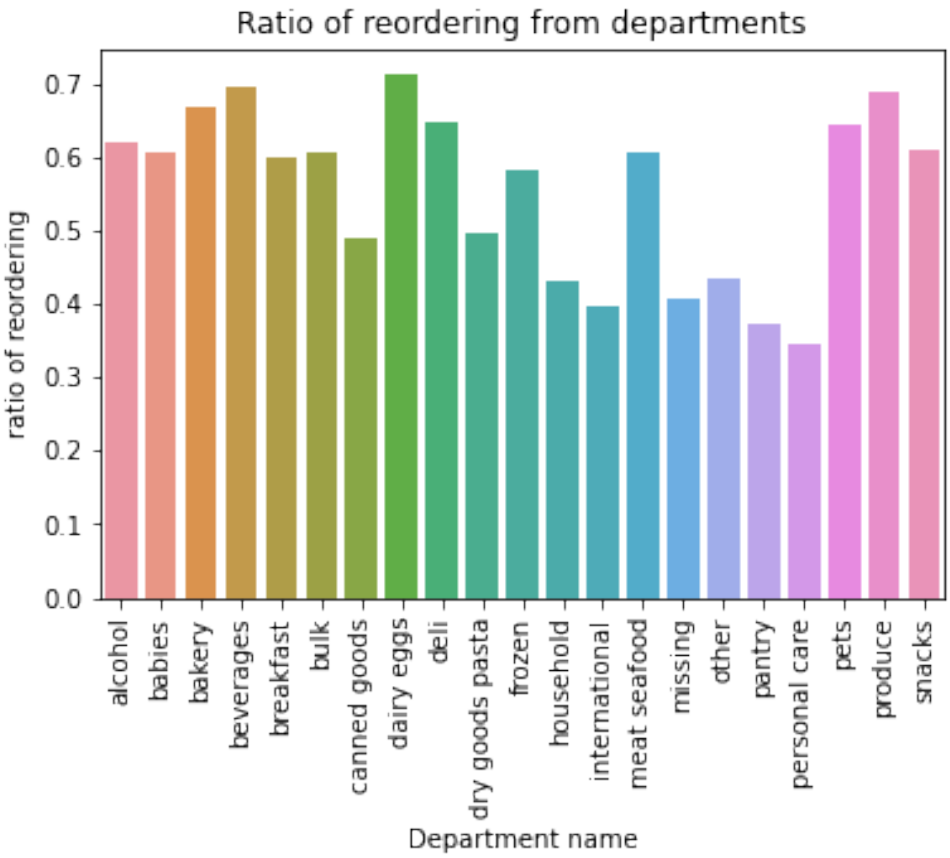

Figure 19

Dairy eggs, produce, pets, snacks, beverages, bakery, deli, meat seafood, bulk, alcohol are ten departments with the most reordered products. These aisles are necessities of life, showing a higher chance of reordering compared to others.

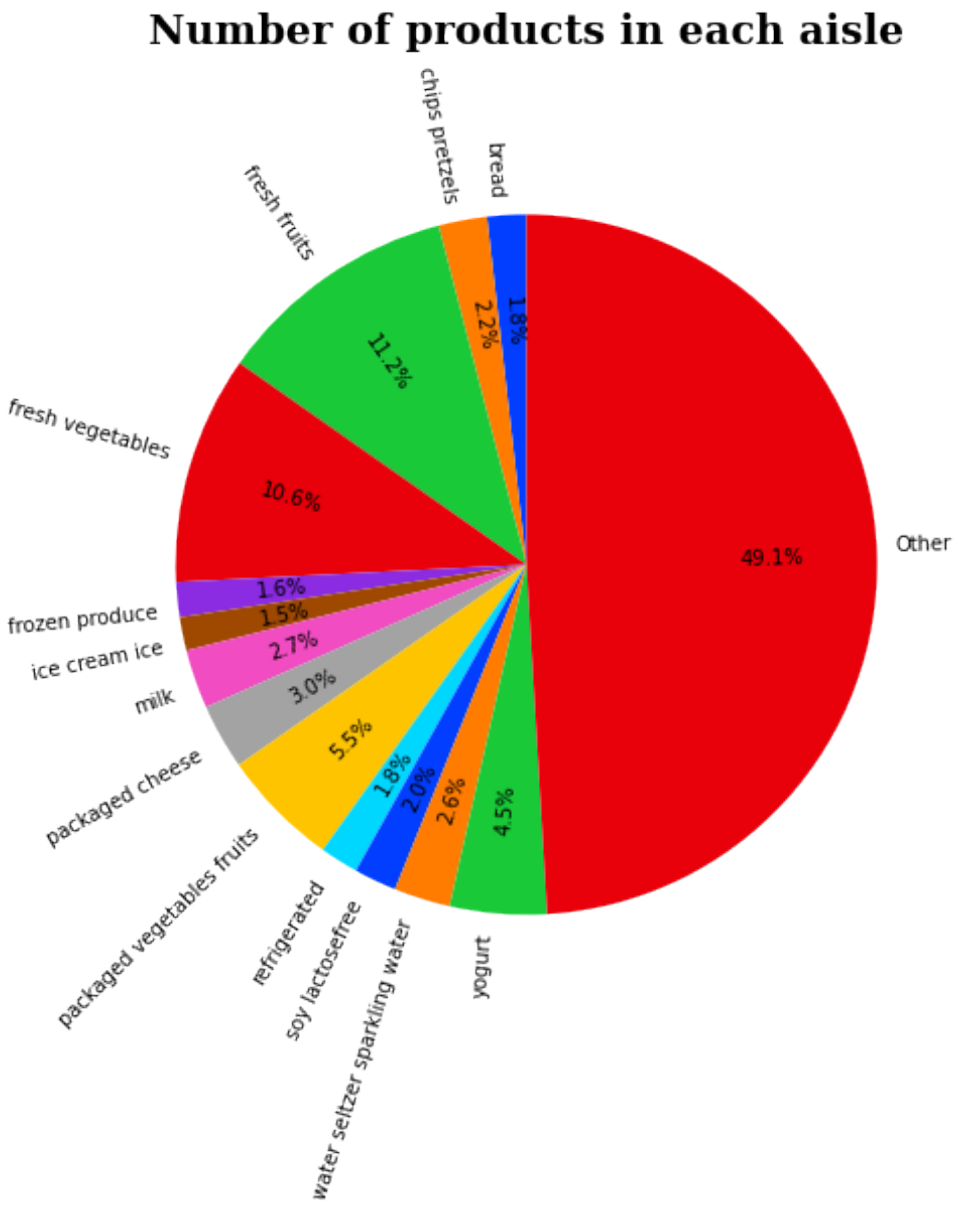

Figure 20

There are 121 aisles in total with 12 aisles accounting for 51% of products in the whole market. Fresh fruits and fresh vegetables are the two aisles with the most products, making up of 11.2% and 10.6% of products, respectively

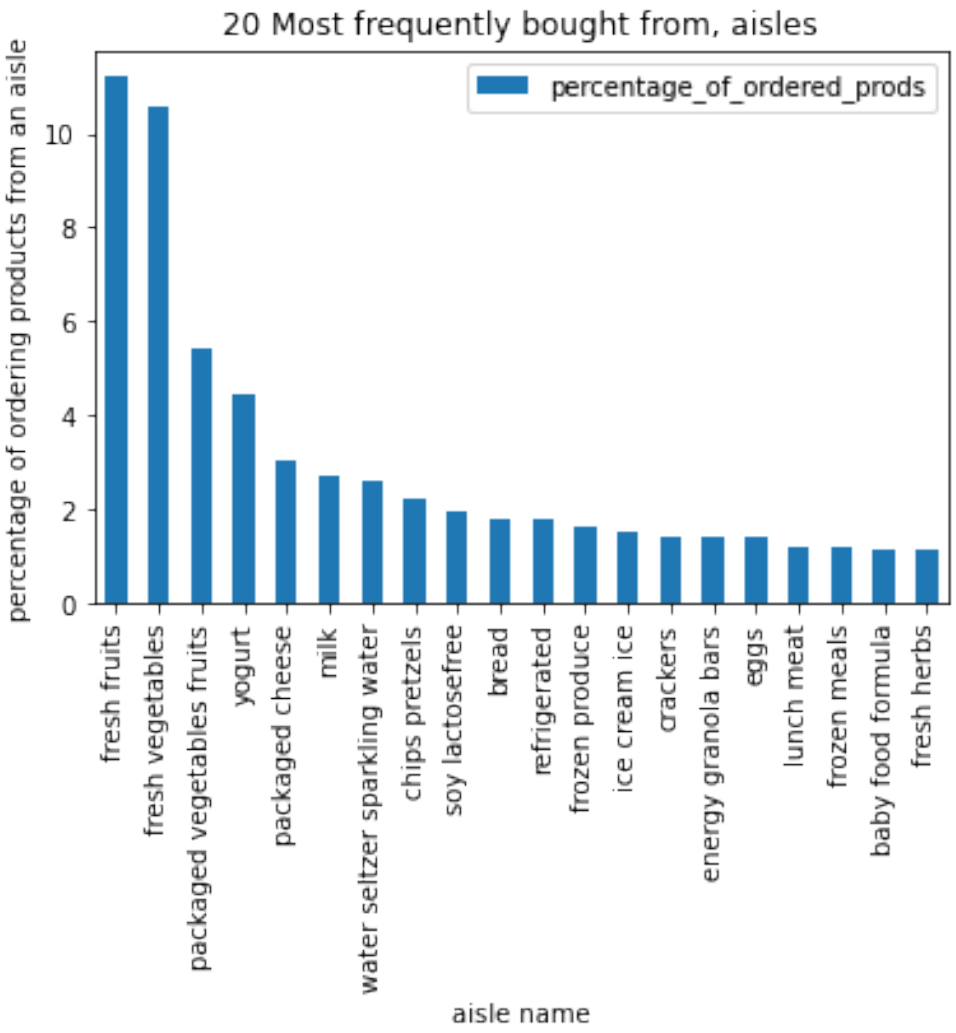

Figure 21

Fresh fruits and fresh vegetables are the two most frequently bought from aisles, with around 14% and 12%, respectively. In the 20 most frequently bought from aisles list, we could also see some familiar names, including milk, yogurt, cheese, bread, eggs and water.

4 Association rule using Apriori

4.1 About Apriori

What is Apriori? Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects. It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule learning that analyzes that people who bought product A also bought product B.

Components in Apriori: The given three components comprise the apriori algorithm: Support, Confidence, Lift.

- **Support:** Refers to the default popularity of any product.

$$\text{support}(A) = \frac{\text{number of transactions containing } A}{\text{total number of transactions}} \tag{1}$$

- **Confidence:** Refers to the possibility that the customers bought both products A and B together.

$$\text{confidence}(A \rightarrow B) = \frac{\text{number of transactions containing } A \text{ and } B}{\text{number of transactions containing } A} \tag{2}$$

- **Lift:** Refers to the increase in the ratio of the sale of A when you sell B. On the other hand, if the lift value is below one it requires that the people are unlikely to buy both the items together. Larger the value, the better the combination.

$$\text{lift}(A \rightarrow B) = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)} \tag{3}$$

4.2 Problem approaches

In this problem, we were about to divide into two separate approaches:

1. **Approach 1**: We applied this algorithm to the products data to figure out the customer buying habits.
2. **Approach 2**: We used date_time data with apriori algorithm to find out the specific type of timezone customers and more.

4.3 Products - products Apriori

Data selection: From the orders table, we chose the train set only; then we selected the products that were consumed at least 1500 and at most 3500 because we wanted to eliminate bias caused by high consuming products. Moreover, the orders must have had at least three items inside as we want to survey the association between many items.

Problem solving: In this problem, we did in these main steps:

1. First, we started with itemsets containing just a single item (Individual items) then we determined the support for these itemsets.
2. Next, we keep the itemsets that meet the minimum support threshold and remove itemsets that do not support minimum support (In this problem we use 0.003).
3. After that, we used the itemsets that are kept from Step 1 and generated all the possible itemset combinations, computed the lift of all possible combinations among times and items with a minimum threshold of 1.2.
4. Finally, we repeated steps 1 until there are no more new itemsets.

Here are the top 5 support items:

support	itemsets
0.079135	(Sparkling Water Grapefruit)
0.062153	(Raspberries)
0.081682	(Organic Fuji Apple)
0.090060	(Small Hass Avocado)
0.090966	(Broccoli Crown)

Here are top 9 rules:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(Soda)	(Hass Avocados, Raspberries)	0.003227	0.004528	0.000340	0.105263	23.244737	0.000325	1.112586
(Hass Avocados, Raspberries)	(Soda)	0.004528	0.003227	0.000340	0.075000	23.244737	0.000325	1.077593
(Small Hass Avocado, Raspberries)	(Blueberries, Organic Blackberries)	0.004472	0.007076	0.000623	0.139241	19.678582	0.000591	1.153544
(Blueberries, Organic Blackberries)	(Small Hass Avocado, Raspberries)	0.007076	0.004472	0.000623	0.088000	19.678582	0.000591	1.091588
(Organic Granny Smith Apple, Jalapeno Peppers)	(Organic Kiwi, Organic Small Bunch Celery)	0.003000	0.005830	0.000340	0.113208	19.416743	0.000322	1.121085
(Organic Kiwi, Organic Small Bunch Celery)	(Organic Granny Smith Apple, Jalapeno Peppers)	0.005830	0.003000	0.000340	0.058252	19.416743	0.000322	1.058670
(Lime Sparkling Water, Organic Whole String Ch...	(Organic Half & Half, Sparkling Water Grapefruit)	0.003340	0.005830	0.000340	0.101695	17.442159	0.000320	1.106717
(Organic Half & Half, Sparkling Water Grapefruit)	(Lime Sparkling Water, Organic Whole String Ch...	0.005830	0.003340	0.000340	0.058252	17.442159	0.000320	1.058309
(Red Vine Tomato, Green Bell Pepper)	(Organic Red Bell Pepper, Boneless Skinless Ch...	0.005661	0.003736	0.000340	0.060000	16.060000	0.000318	1.059855

In this observation, we can see that Organic food tends to go with diet foods like jalapeno Peppers, Roma tomatoes, or Lime Sparkling Water. Thus these products can be recommended for diet users when they buy Organic food.

4.4 Products - time Apriori

Making the data ready for further exploration: From the orders table, we filter out the train set only; then merge it with the order_products_train set and the product set to get a final table like this.

	order_id	product_id	product_name	order_dow	order_hour_of_day	days_since_prior_order
0	1187899	196	Soda	4	8	14.0
1	2757217	196	Soda	0	11	5.0
2	632715	196	Soda	0	13	26.0
3	1167274	196	Soda	4	10	8.0
4	3347074	196	Soda	3	21	5.0
...	...	...	...	...	...	...
1384612	3351563	22165	Chewy Reduced Sugar Granola Bars Variety Pack	3	13	7.0
1384613	2629221	31540	Plain Flavor Probiotic Acidophilus	6	13	30.0
1384614	2721635	44507	100% Juice, Rio Red Grapefruit	6	10	30.0
1384615	2078948	47814	Puppy Complete Nutrition Chicken & Beef Dinner Wet Dog Food	3	11	15.0
1384616	243575	49653	Organic Aromatherapeutic Moroccan Argan Oil Set	0	22	11.0

We can take all products to see relationships for an in-depth assessment but because of the computational limit we take product range from 600 to 650; this range helps us to eliminate bias and it is a great combination of healthy organic food and unhealthy instant food. Therefore we can observe how the purchases relate to the time of purchasing.

```
Whole Milk Ricotta Cheese                    347
Zero Calorie Cola                            347
Gluten Free 7 Grain Bread                    347
Grade A Large Brown Eggs                     347
Organic Hot House Tomato                     346
Organic Extra Large Brown Eggs               346
Large Greenhouse Tomato                      345
Unrefined Virgin Coconut Oil                 345
Organic Sea Salt Roasted Seaweed Snacks      345
Mineral Water                                344
Sweet Onions                                 344
Plain Mini Bagels                            344
Dark Chocolate Pretzels with Sea Salt        343
Hot Dog Buns                                 343
Flaky Biscuits                               343
Drinking Water                               342
Avocado                                      342
Seven Grain Crispy Tenders                   341
Celery                                       340
Root Beer                                    340
Organic Baby Bella Mushrooms                 339
Frozen Organic Blueberries                   338
Chicken Breast Tenders Breaded               337
White Corn Tortillas                         337
Organic Golden Delicious Apple               337
Mediterranean Mint Gelato                    337
Shredded Hash Browns                         336
Margherita Pizza                             336
Organic Red Chard Greens                     336
Cane Sugar                                   336
Natural Classic Pork Breakfast Sausage       335
Veggie Chips                                 334
Birthday Cake Light Ice Cream                333
Stringless Sugar Snap Peas                   333
```

Last but not least, we encode our data into categories and time frames so we can dummy them

later.

A glimpse on our final table:

	product_name	order_dow	order_hour_of_day	days_since_prior_order
0	Zero Calorie Cola	weekday	morning_hour ([7,9])	two_week
1	Zero Calorie Cola	weekday	peak_hours ([10,16])	one_week
2	Zero Calorie Cola	weekend	peak_hours ([10,16])	two_week
3	Zero Calorie Cola	weekday	morning_hour ([7,9])	more_than_a_month
4	Zero Calorie Cola	weekday	night_hour ([17,23])	one_week
...	...	...	...	...
16859	Large Greenhouse Tomato	weekday	night_hour ([17,23])	three_week
16860	Large Greenhouse Tomato	weekend	peak_hours ([10,16])	one_week
16861	Large Greenhouse Tomato	weekday	peak_hours ([10,16])	one_week
16862	Large Greenhouse Tomato	weekday	peak_hours ([10,16])	more_than_a_month
16863	Large Greenhouse Tomato	weekend	night_hour ([17,23])	more_than_a_month

Problem-solving:

- Computing the support of all possible combinations among times and items with a minimum threshold of 0.003.

	support	itemsets
0	0.627609	(weekday)
1	0.576079	(peak_hours ([10,16]))
2	0.372391	(weekend)
3	0.349917	(peak_hours ([10,16]), weekday)
4	0.255870	(night_hour ([17,23]))
...	...	...

- Computing the lift of all possible combinations among times and items with a minimum threshold of 1.0.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedents_length	consequents_len
(weekday, 100% Lactose Free Fat Free Milk)	(early_morning_hours (<7))	0.012808	0.026269	0.000949	0.074074	2.819831	6.123044e-04	1.051630	2	
arly_morning_hours (<7))	(weekday, 100% Lactose Free Fat Free Milk)	0.026269	0.012808	0.000949	0.036117	2.819831	6.123044e-04	1.024182	1	
(weekday, arly_morning_hours (<7))	(100% Lactose Free Fat Free Milk)	0.018797	0.019568	0.000949	0.050473	2.579333	5.809325e-04	1.032548	2	
100% Lactose Free Fat Free Milk)	(weekday, early_morning_hours (<7))	0.019568	0.018797	0.000949	0.048485	2.579333	5.809325e-04	1.031200	1	
(Shredded Hash Browns, weekday)	(early_morning_hours (<7))	0.013223	0.026269	0.000830	0.062780	2.389902	4.828047e-04	1.038957	2	
...	...	...	...	...	...	...	...	...	...	
Gluten Free 7 Grain Bread)	(weekday)	0.020576	0.627609	0.012927	0.628242	1.001009	1.302416e-05	1.001703	1	
(weekend, Organic Edamame)	(morning_hour ([7,9]))	0.008361	0.141781	0.001186	0.141844	1.000442	5.239201e-07	1.000073	2	
orning_hour ([7,9]))	(weekend, Organic Edamame)	0.141781	0.008361	0.001186	0.008365	1.000442	5.239201e-07	1.000004	1	
(weekday, Large reenhouse Tomato)	(peak_hours ([10,16]))	0.012453	0.576079	0.007175	0.576190	1.000193	1.385400e-06	1.000263	2	
(peak_hours ([10,16]))	(weekday, Large Greenhouse Tomato)	0.576079	0.012453	0.007175	0.012455	1.000193	1.385400e-06	1.000002	1	

- Printing the top 20 of the relation to observe:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedents_length	consequer
0	(weekday, 100% Lactose Free Fat Free Milk)	(early_morning_hours (<7))	0.012808	0.026269	0.000949	0.074074	2.819831	0.000612	1.051630	2	
1	(early_morning_hours (<7))	(weekday, 100% Lactose Free Fat Free Milk)	0.026269	0.012808	0.000949	0.036117	2.819831	0.000612	1.024182	1	
2	(weekday, early_morning_hours (<7))	(100% Lactose Free Fat Free Milk)	0.018797	0.019568	0.000949	0.050473	2.579333	0.000581	1.032548	2	
3	(100% Lactose Free Fat Free Milk)	(weekday, early_morning_hours (<7))	0.019568	0.018797	0.000949	0.048485	2.579333	0.000581	1.031200	1	
4	(Shredded Hash Browns, weekday)	(early_morning_hours (<7))	0.013223	0.026269	0.000830	0.062780	2.389902	0.000483	1.038957	2	
5	(early_morning_hours (<7))	(Shredded Hash Browns, weekday)	0.026269	0.013223	0.000830	0.031603	2.389902	0.000483	1.018979	1	
6	(weekday, Drinking Water)	(early_morning_hours (<7))	0.014231	0.026269	0.000889	0.062500	2.379233	0.000516	1.038646	2	
7	(early_morning_hours (<7))	(weekday, Drinking Water)	0.026269	0.014231	0.000889	0.033860	2.379233	0.000516	1.020316	1	
	(early_morning_hours										

When printing the top 20 rules, we can see that the early morning purchases are greatly associated with healthy food like mineral water, lactose-free fat milk, chicken breast, and hash browns. This makes a lot of sense as people with a healthy lifestyle will choose healthy food and wake up early. This is a great insight to the sales department where it can be suggested to early buyers with a group of healthy food.

Printing tuples that contain "pizza"

0	(Margherita Pizza, weekday)	(night_hour ([17,23]))	0.012334	0.255870	0.004507	0.365385	1.428006	0.001351	1.172568	2	1
1	(Margherita Pizza)	(night_hour ([17,23]), weekday)	0.019924	0.169177	0.004507	0.226190	1.337005	0.001136	1.073679	1	2
2	(Margherita Pizza)	(night_hour ([17,23]))	0.019924	0.255870	0.006523	0.327381	1.279479	0.001425	1.106316	1	1
3	(Margherita Pizza)	(weekend, night_hour ([17,23]))	0.019924	0.086694	0.002016	0.101190	1.167220	0.000289	1.016129	1	2
4	(Margherita Pizza, peak_hours ([10,16]))	(weekend)	0.010792	0.372391	0.004566	0.423077	1.136110	0.000547	1.087856	2	1
5	(Margherita Pizza, morning_hour ([7,9]))	(weekend)	0.002253	0.372391	0.000949	0.421053	1.130674	0.000110	1.084052	2	1
6	(night_hour ([17,23]), Margherita Pizza)	(weekday)	0.006523	0.627609	0.004507	0.690909	1.100859	0.000413	1.204794	2	1
7	(weekend, Margherita Pizza)	(peak_hours ([10,16]))	0.007590	0.576079	0.004566	0.601562	1.044236	0.000193	1.063958	2	1
8	(weekend, Margherita Pizza)	(night_hour ([17,23]))	0.007590	0.255870	0.002016	0.265625	1.038123	0.000074	1.013283	2	1
9	(Margherita Pizza)	(weekend)	0.019924	0.372391	0.007590	0.380952	1.022991	0.000171	1.013830	1	1
10	(Margherita Pizza)	(weekend, peak_hours ([10,16]))	0.019924	0.226162	0.004566	0.229167	1.013284	0.000060	1.003898	1	2

Pizza is considered an unhealthy kind of food, it is high in calories and fat. Here we can see that Pizza will be more likely to be ordered at later times of the day. This can be purchased to support parties, gamers' nights etc. Using this fact the sales department can suggest late customers with a group of unhealthy instant food like pizza, ramen, or coke.

5 Predicting user next product

5.1 Feature engineering

In addition to the existing features, we hand-crafted some more features to increase the accuracy and robustness of our model User related features:

- user_total_orders: Total number of orders by user
- user_total_items: Total number of items the user had ordered
- total_distinct_items: Total number of distinct items purchased by each user.
- user_average_days_between_orders: average days between orders
- user_average_basket: average number of item in each order

Order related features:

- order_hour_of_day: The hour the order is placed
- days_since_prior_order: Number of days since the last order
- days_since_ratio: days_since_prior_order / user_average_days_between_orders

Product related features:

- product_orders: total number of orders of a product
- organic: check if the product is organic or not
- product_reorders: total number of reorders of a product
- product_reorder_rate: product_reorders/product_orders

UserXProduct (UP) features:

- UP_orders: total number of a product ordered by a user
- UP_orders_ratio: UP_orders / user_total_orders
- UP_last_order_id: ID of the last order
- UP_average_pos_in_cart: average position of a product when the user purchased it
- UP_reorder_rate: UP_orders / user_total_orders
- UP_orders_since_last: user_total_orders - UP_last_order_id
- UP_delta_hour_vs_last: difference between the hour that the user makes the current order and that of the previous order.

5.2 Dealing with imbalance classes

In practice, an imbalance class creates a bias where the machine learning model tends to predict the majority class, and ignore the minority class. Due to the fact that our dataset class has a skewed proportion, where the reordered class is around 10% of the whole dataset, we have investigated different resampling methods. In the end, a technique for oversampling called SMOTE (Synthetic Minority Oversampling Technique) [1] is chosen to overcome the imbalance problem. The main idea of SMOTE is to utilize a k-nearest neighbor algorithm to create synthetic data. The main idea of SMOTE is that it first selects a minority class instance at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.

5.3 Classification technique

The techniques that will be used to classify the two classes are: RandomForest [3], Extreme gradient boosting (XGBoost) [2] and Light Gradient Boosting Machine (LGBM) [4]. Random Forest is currently one of the most popular and accurate methods that can be implemented easily and efficiently whereas XGBoost and LGBM have become the de-facto algorithm for winning numerous competitions just because they are too powerful.

5.4 Experimental results

In the training phases of the classification methods, we split the training set into 2 parts: one used for training and one used for validation to tune the hyperparameters. After that, we use our models to

Method	F1-score	
	with Smote	without Smote
Random Forest	0.18258	0.31546
XGBoost	0.34518	
LGBM	0.23972	0.33209

Table 1: F1-score of all classification methods

make predictions and submit to the original kaggle page to get the results.

Based on Table 1, it can be concluded that among the three models that we considered and implemented, XGBoost gives the highest F1-score, followed by LGBM and RandomForest. Furthermore, the performance of Random Forest and LGBM improved drastically after oversampling with SMOTE, proving its effectiveness in imbalance class problems. However, XGBoost works well with an imbalanced dataset even when not applying SMOTE by making use of the parameter scale_pos_weight to set the suitable weight for each class. One note is that three classifiers perform really well in classifying the not reordered class, while underperforming in the reordered class.

5.5 Predicting customer behavior

In this work, we will use LGBM to predict customer behavior and identify importance feature.

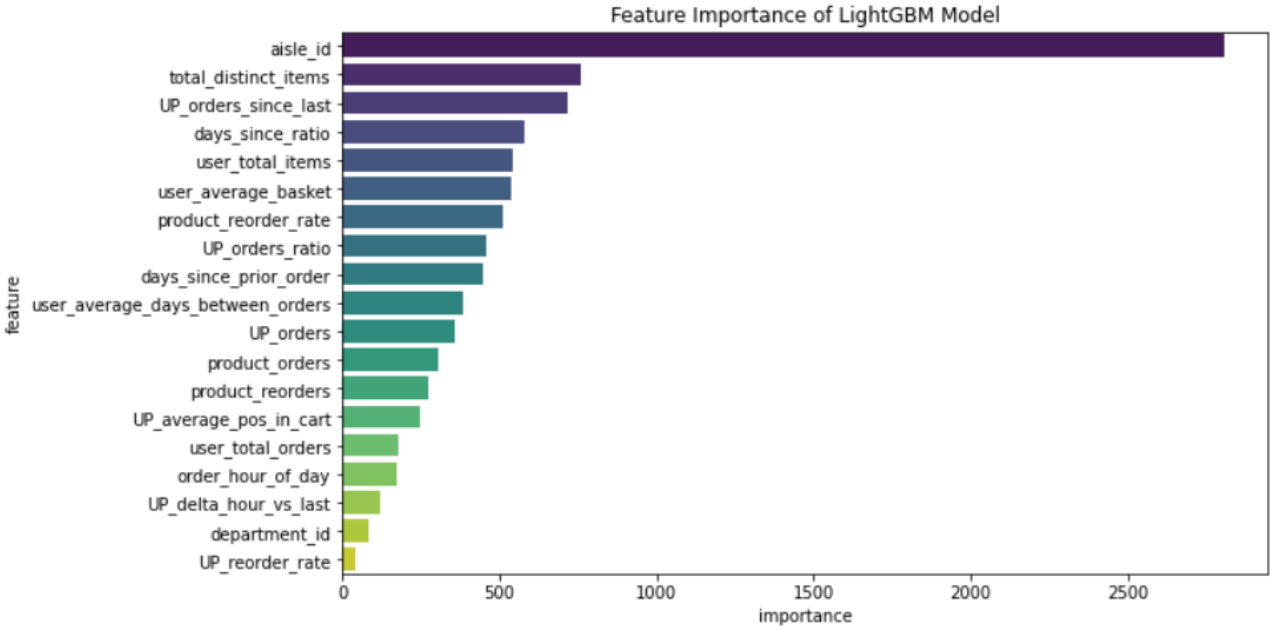

Figure 22 Feature importance of LGBM

Figure 22 visualizes the impact of various features on the LGBM model's predictions regarding

customer behavior, particularly focusing on their purchase history. The model ranks features by their importance, helping to identify which variables most significantly influence the prediction outcomes. The plot clearly shows that **aisle_id** is the most critical feature, significantly more influential than the rest. This prominence suggests that **aisle_id** contains implicit, multifaceted information that is crucial for understanding customer purchasing patterns. This feature inherently categorizes products into specific aisles, which typically represent broader product types and categories. This categorization helps capture customer preferences for particular types of products. For instance, a customer frequently purchasing from the "Dairy" aisle is likely to have different shopping behaviors compared to one who often buys from the "Snacks" aisle. Features related to the recency and frequency of purchases (like **total_distinct_items**, **UP_orders_since_last**, **days_since_ratio**, and **user_total_items**) also play crucial roles. These insights can guide marketing strategies, inventory management, and personalized recommendations to enhance customer engagement and satisfaction. Overall, the feature importance plot is a powerful tool for understanding which factors most influence customer purchasing behavior, enabling more informed business decisions.

order_id	product_id	user_total_	user_total_	total_distin	user_avera	user_avera	order_hou	days_since	days_since	aisle_id	departmer	product_o	product_re	product_re	UP_orders	UP_orders	UP_averag	UP_reorde	UP_orders	UP_delta_	pred
2774568	17668	13	88	33	12	6.769231	15	11	0.916667	91	16	2110	1220	0.578199	5	0.384615	3.6	0.384615	2	3	0.8520829284287534
2774568	44683	13	88	33	12	6.769231	15	11	0.916667	83	4	22275	11981	0.537868	2	0.153846	9.5	0.153846	7	1	0.3930252713320075
2774568	48523	13	88	33	12	6.769231	15	11	0.916667	37	1	5129	2376	0.463248	2	0.153846	6.5	0.153846	4	1	0.4843983084856877
2774568	21903	13	88	33	12	6.769231	15	11	0.916667	123	4	241921	186884	0.7725	8	0.615385	4.25	0.615385	1	0	0.9331176538491984
2774568	14992	13	88	33	12	6.769231	15	11	0.916667	83	4	29069	16942	0.58282	2	0.153846	7	0.153846	6	0	0.4691221742125461
2774568	21137	13	88	33	12	6.769231	15	11	0.916667	24	4	264683	205845	0.777704	1	0.076923	7	0.076923	11	4	0.39929527009730065
2774568	32402	13	88	33	12	6.769231	15	11	0.916667	78	19	2056	1328	0.645914	3	0.230769	8.333333	0.230769	3	1	0.7126103346624069
2774568	22035	13	88	33	12	6.769231	15	11	0.916667	21	16	59676	45639	0.76478	3	0.230769	3.666667	0.230769	5	2	0.5985032206085983
2774568	49683	13	88	33	12	6.769231	15	11	0.916667	83	4	97315	67313	0.691702	1	0.076923	4	0.076923	10	1	0.22757719106210747
2774568	39190	13	88	33	12	6.769231	15	11	0.916667	91	16	10972	6294	0.573642	10	0.769231	1.8	0.769231	1	0	0.9690166347696902
2774568	47766	13	88	33	12	6.769231	15	11	0.916667	24	4	176815	134044	0.758103	9	0.692308	3.777778	0.692308	1	0	0.9535089049019598
2774568	42265	13	88	33	12	6.769231	15	11	0.916667	123	4	76896	50472	0.656367	1	0.076923	9	0.076923	8	2	0.35067239881641327
2774568	1819	13	88	33	12	6.769231	15	11	0.916667	88	13	2424	1193	0.492162	3	0.230769	2.666667	0.230769	6	0	0.6032769194099571
2774568	40604	13	88	33	12	6.769231	15	11	0.916667	21	16	32351	18069	0.55853	1	0.076923	4	0.076923	11	4	0.23489264611139693
2774568	16797	13	88	33	12	6.769231	15	11	0.916667	24	4	142951	99802	0.698155	3	0.230769	4	0.230769	4	1	0.6333409444561171
2774568	18599	13	88	33	12	6.769231	15	11	0.916667	4	9	6204	2581	0.416022	4	0.307692	3.75	0.307692	1	0	0.8280399916250741
2774568	15143	13	88	33	12	6.769231	15	11	0.916667	24	4	3447	1696	0.492022	1	0.076923	3	0.076923	12	1	0.12320866742365406
2774568	9387	13	88	33	12	6.769231	15	11	0.916667	24	4	36187	23537	0.650427	5	0.384615	3.6	0.384615	6	0	0.538132
2774568	12845	13	88	33	12	6.769231	15	11	0.916667	117	19	10027	3639	0.36292	1	0.076923	2	0.076923	9	3	0.1886540990929859
2774568	43961	13	88	33	12	6.769231	15	11	0.916667	123	4	55371	34916	0.630583	4	0.307692	3.75	0.307692	2	3	0.8137307693481204
2774568	42557	13	88	33	12	6.769231	15	11	0.916667	88	13	8324	4759	0.57172	1	0.076923	7	0.076923	5	2	0.3712789798179805
2774568	18370	13	88	33	12	6.769231	15	11	0.916667	21	16	18449	11140	0.603827	1	0.076923	10	0.076923	8	2	0.24977529288400738
2774568	38596	13	88	33	12	6.769231	15	11	0.916667	108	16	3948	1376	0.348531	1	0.076923	1	0.076923	11	4	0.1861054275645059
2774568	16965	13	88	33	12	6.769231	15	11	0.916667	37	1	13273	7146	0.538386	2	0.153846	4.5	0.153846	8	2	0.3622060215566743
2774568	24010	13	88	33	12	6.769231	15	11	0.916667	52	1	6562	4113	0.626791	2	0.153846	5.5	0.153846	6	0	0.4324113917686982
2774568	7503	13	88	33	12	6.769231	15	11	0.916667	117	19	12474	6905	0.553551	1	0.076923	6	0.076923	10	1	0.2520541382291765
2774568	8021	13	88	33	12	6.769231	15	11	0.916667	54	17	27864	16472	0.591157	1	0.076923	5	0.076923	11	4	0.28775331038345886

Figure 23 Prediction of user purchasing behavior

Figure 23 illustrates an example of LGBM prediction on customer behavior. In this scenario, we aim to predict whether a user with id 13 will reorder a product listed in the column **product_id**. The predicted probabilities are provided in the column **pred**. For instance, examining the products within **aisle_id** 123, we observe that most products, such as those with ids 21903 and 43961, have a high probability of being reordered, except for the product with id 42265. This phenomenon can be explained by considering the third most influential feature: **UP_orders_since_last**. Both products 21903 and 43961 have a low **UP_orders_since_last**, while product 42265 has a high **UP_orders_since_last**, which accounts for the difference in their predicted probabilities.

Overall, the analysis demonstrates that the prediction model effectively utilizes various features to differentiate the likelihood of reordering. By examining these key features, we can better understand the factors influencing customer behavior and improve the accuracy of predictive models. This insight is crucial for enhancing personalized marketing strategies and inventory management, ultimately leading to improved customer satisfaction.

6 Conclusion and future work

In conclusion, we have demonstrated a slightly better way to survey customers' habits by Association rules and some classification techniques. Furthermore, we have shown some new insight about customer's habit when applying these algorithms. In the future, about our first strategy, we are going to apply some feature engineering techniques like removing outliers, edit features, scale, check correlation, etc. to improve our result. Moreover, we will combine two of our tactics together. That means, we use apriori algorithm to check the customer habits and take insights in RandomForest, XGBoost and LGBM in order to predict the user's next product.

References

[1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, jun 2002.

[2] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[3] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Neural Information Processing Systems*, 2017.