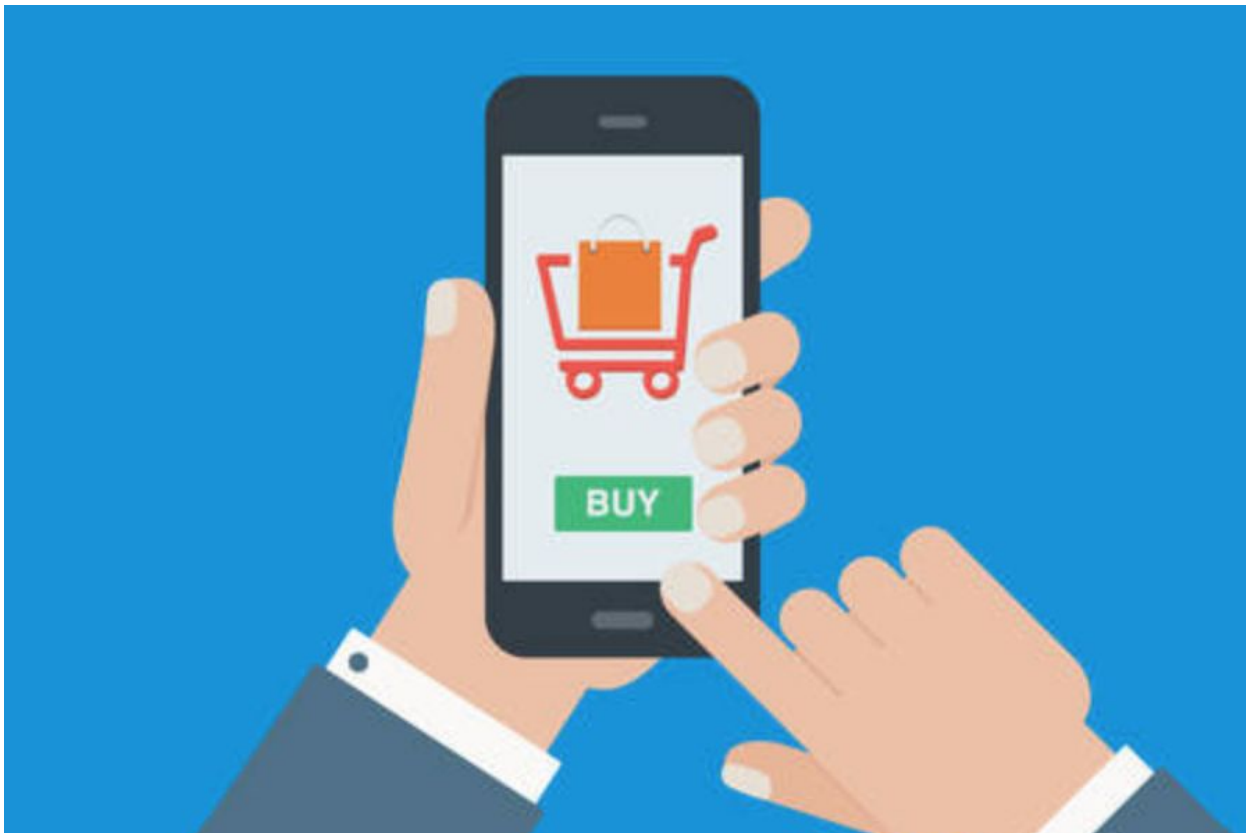


Online Order Prediction

Data & Methodology

31.March.2019

Interim Report

By

Diptarup Dey | Pooja Prakash | Rakesh Jhalani | Thushara TK

Great Lakes PGP BABI | Bangalore Jun'18 | Group 8

Table of Contents

Overview

Instacart is an American company who provides door delivery service from the customer's favourite local shops like Whole Foods, Costco, Petco etc. Customers can use Instacart app, select the store, choose the items they want to purchase, add them to the cart and decide when they want to receive the delivery at their doorstep as early as within 2 hours. A lot of planning goes into the whole buying, packaging and delivering process with sometimes perishable grocery. So predicting what the customers are going to order within the day or hour becomes very critical for Instacart.

Instacart has provided an anonymized dataset with a sample of 3 million grocery orders from 200,000 Instacart users. transaction records that Instacart provides for the public to work on, understand the patterns, train models and predict what an Instacart customer will order next.

Scope & Objectives

The main objective of the project is to predict grocery reorder for the customers, given their past purchase history

Scope of the project :

1. Given a training dataset with future orders of a few customers, predict what rest of the customers are going to order.
2. The dataset contains information like, day-of-the-week, hour-of-the-day etc, which gives a good idea about the traffic during week and day respectively.
3. The dataset also contains information about the day's gap between each order. This gives a time-based nature also. If we know the time since a customer purchased an item, how can that gap be used to predict his next purchase?

Data Source

The dataset is provided as-is for non-commercial use.

Instacart Dataset

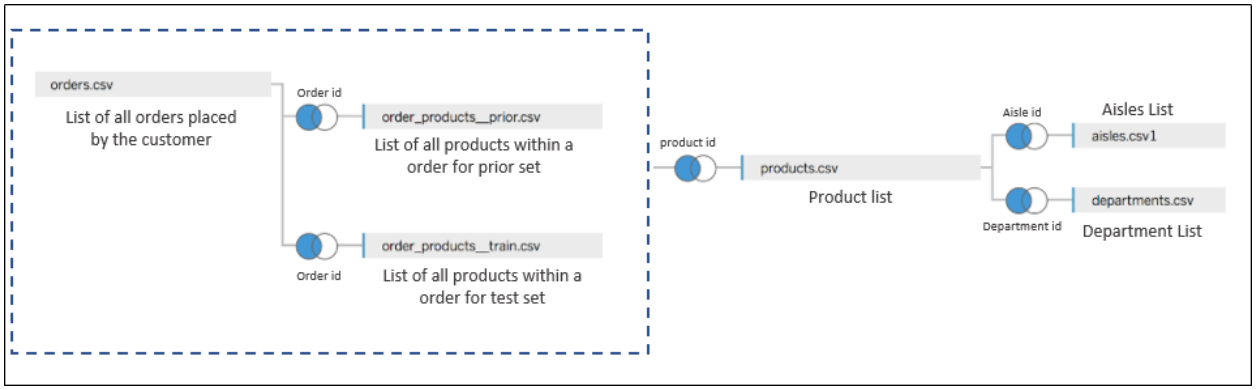
Terms & Conditions

Data Dictionary

Instacart market basket analysis is also part of Kaggle competitions.
https://www.kaggle.com/c/instacart-market-basket-analysis

Data Cleaning and Preparation

There are 7 datasets. Below is the detailed view of the joins of these datasets and basic description of each dataset



Customer and Order base

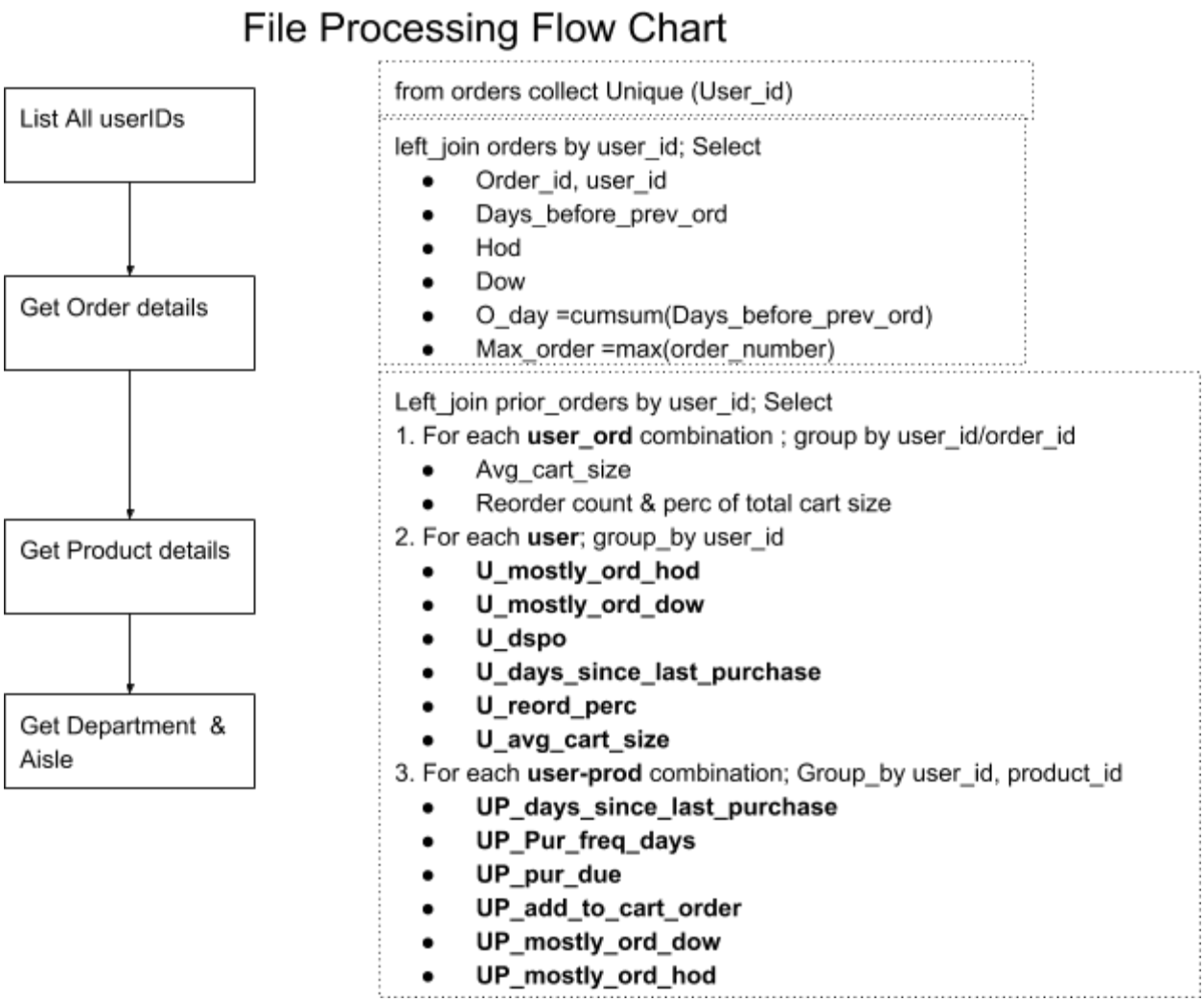
	Total	Prior	Train	Test
# of Orders	3421083	3214874	131209	75000
# of Customers	206209	206209	131209	75000

The order base is already split into Prior, Train and Test. The Products for order within the test set is unavailable. Test and train sets will be recreated using the remaining data available.

File Processing Flow

Data is available in multiple files. A training dataset is prepared by joining the Orders, Prior_orders, Products, Aisles & Departments files. Data augmentation has been done in order to make it more appropriate for the model building. The methodology has been explained in further sections.

Below is the file processing flow chart.



Exploratory Data Analysis

Number of orders per User

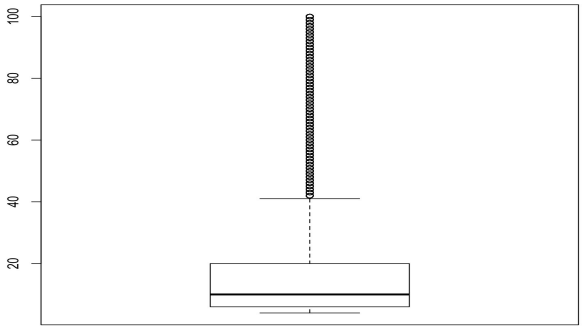


Fig: Boxplot of number of orders placed by each customer from prior transactions.

Findings: 50% of customers have placed less than ~15 orders

Peak order hours

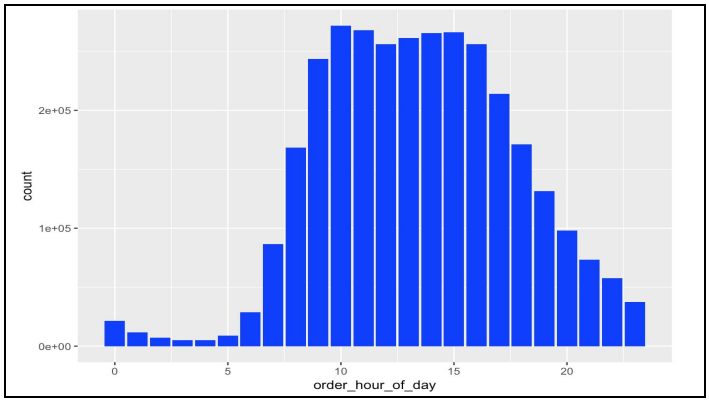


Fig: Distribution of orders across the each hour of the day

Findings: The peak hours are from 9am to 4pm. The order gradually decreases over the night.

Peak order Days

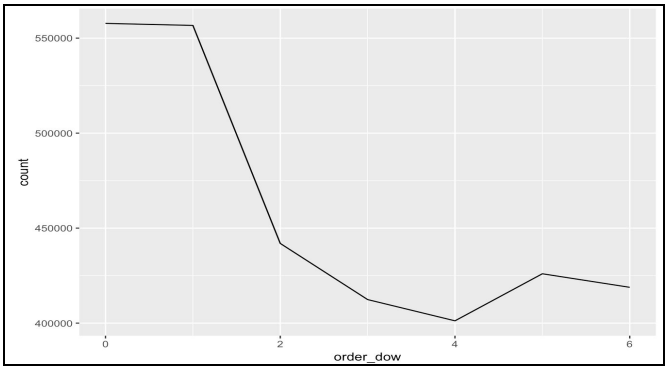


Fig: Distribution of orders across the each day of the day

Findings: The orders peak during the Weekends(0-Saturday, 1-Sunday) and has a sharp drop as the weekday starts (2-Monday)

Most frequently ordered/reordered products

Looking at the plot below, Bananas, organic bananas, organic strawberries etc are most frequently ordered by most users.

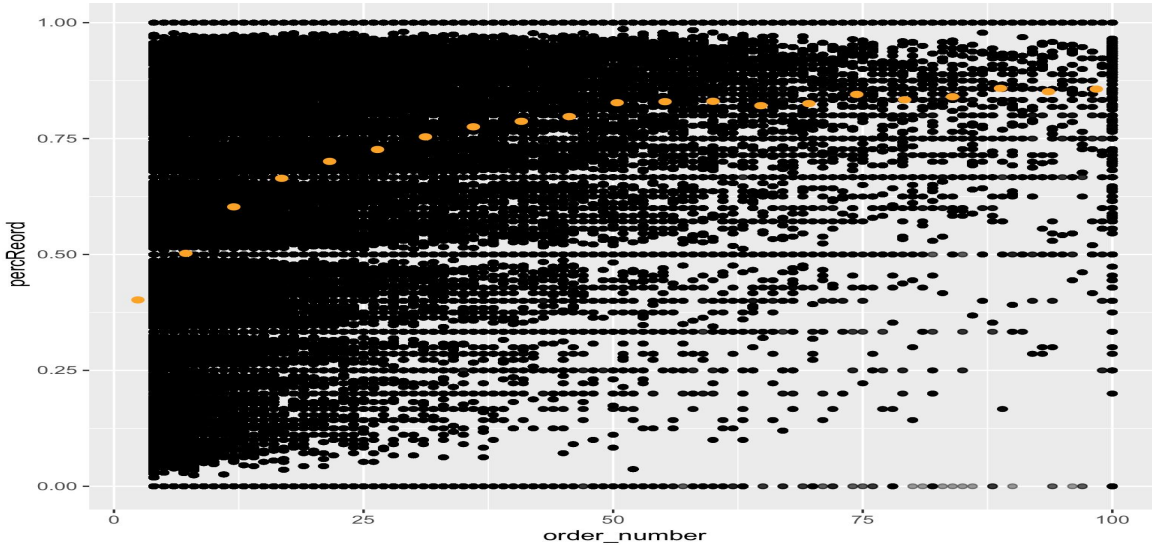
Most frequently Ordered Products/ Most reordered

Product Name	# of User	Reord Count	Top # of users
Banana	73,956	55,166	1
Bag of Organic Bananas	63,537	45,231	2
Organic Strawberries	58,838	38,131	3
Organic Baby Spinach	55,037	36,557	4
Organic Hass Avocado	43,453	29,031	7
Organic Avocado	42,771	27,429	9
Large Lemon	46,402	27,195	5
Limes	44,859	25,021	6
Strawberries	43,149	24,449	8
Organic Garlic	35,115	20,267	11
Organic Yellow Onion	34,354	19,999	12
Organic Raspberries	31,648	19,341	14
Organic Blueberries	37,138	19,091	10
Organic Zucchini	32,658	18,772	13
Cucumber Kirby	30,002	17,002	15
Yellow Onions	29,898	15,309	16
Organic Grape Tomatoes	29,025	15,278	17
Organic Lemon	27,210	15,133	19
Seedless Red Grapes	27,512	14,395	18
Organic Baby Carrots	26,424	14,093	20

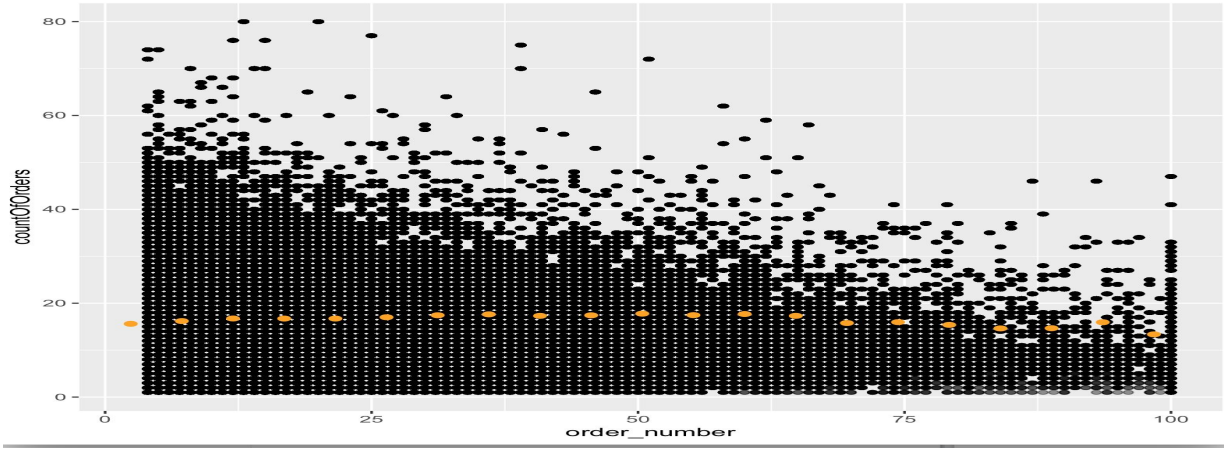
Orders, Reorder and cart size patterns

Below graph shows, as more and more orders are placed by the user the reorder percentage of the cart goes up and the list of products the customer would buy becomes more predictable.

Pic. Reorder percentage as more orders are placed.



Below is a plot of the maximum orders placed by a user and his average cart size.



When user is a frequent buyer(denoted by higher order number) his cart size is smaller. Bigger order number shows more frequent purchases. CountOfOrders is the count of items in the last purchase.

Customer Profiling

Profile customer based on the products purchased over the last 90 days. This will enable us to identify the current relevant product of each user.

Customer and order base

Random sample of 1000 user and their latest 90 days orders is selected to execute this exercise.

Product categories

There are a total of ~50k unique products. These products belong to 21 departments. Broader product categories is created based on the department these product belong to.

Below are the final set of categories:

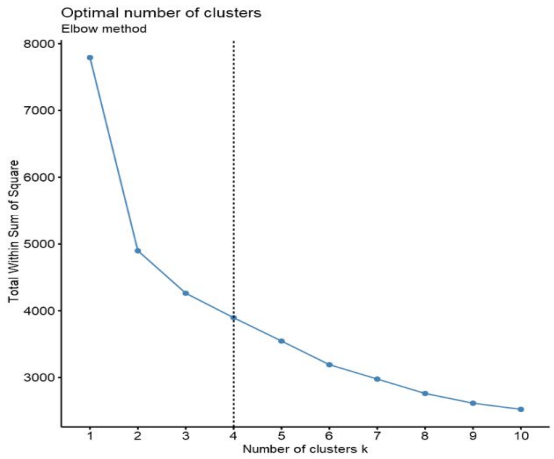
1. **Cooking Essentials:** Includes deli and pantry
2. **Packed food:** Includes frozen items, dry goods pasta ,canned goods,breakfast
3. **Household & Pets:** Includes Household items and pet related products
4. **Personal & Baby Care:** Includes Personal and Baby Care products
5. **Dairy & Egg_Meat:** Includes dairy products, egg, meat, fish etc
6. **Snacks and Beverages:** IncludeS snacks and beverage items
7. **Fresh Veggies:** Include the Fresh produce
8. **Other:** Rest of the remaining items belong here

Methodology: Clustering Analysis

Clustering analysis is done to group these users based on products purchased. The above set of product categories is used to identify these groups.

Optimal Number of Clusters

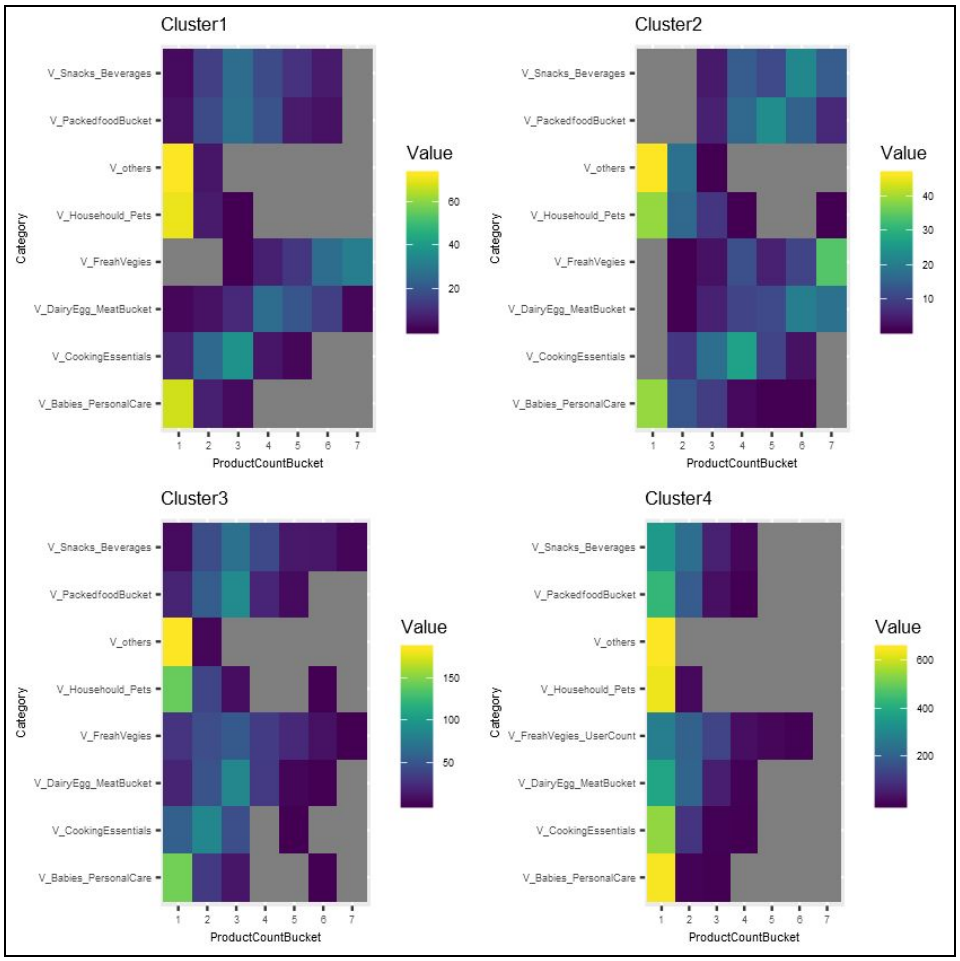
Based on the Elbow curve 4 is the ideal number of clusters



K Means clustering to obtain 4 clusters

User Count across Product Categories and quantity Bought

Value: User Count; **ProductCountBucket:** Higher the ProductCountBucket higher is the total purchase per user



Defining the Clusters

Cluster1: Medium Purchase customers. High veggies and medium quantity of DairyEgg, Cooking, packed food and beverages

Cluster2: High Purchase customers. Who buy regular essentials like Veggies, DairyEgg, Cooking, packed food and beverages

Cluster3: Medium Purchase customers. No veggies and medium quantity on DairyEgg;Cooking;packed food and beverages

Cluster4: Low purchase customer. They are low on most categories

Key Findings Summary

Customers repeat orders often. At regular intervals, the same product is purchased and certain products are often brought along with other products. These patterns provide us with a hint as to what the customer will purchase next.

Max purchased items - Milk

Items purchased at **intervals -** Cheese

Change in product use/change in behaviour: Lays to fresh vegetable

Differentiate between **a missed out item Vs change in behaviour** - Cheerios purchased in every other order, but recent history is inconsistent, but the meat remains.

Association - Everytime bread is bought cheese is also bought

Variation within product category - Yogurt - customer has a history of trying out different flavours

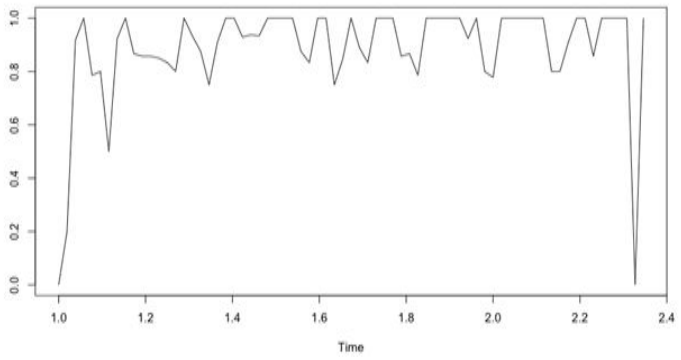
Give preference to the latest data but outlier data should be handled - **one off purchase** of pepsi

Clustering customers based on Aisles/Dept - Customer eats meat but never made purchases in Baby Aisles or organic category

Finding 1: Repeat purchase percentage(user ID #90)

Below is a Repeat purchase percentage plot for user ID #90. This user has a high repeat purchase percentage. This means the user buys products from a fixed list of products. There is an exception in recent history. Below is a time-series plot.

Pic. Repeat purchase percentage(User ID#90)



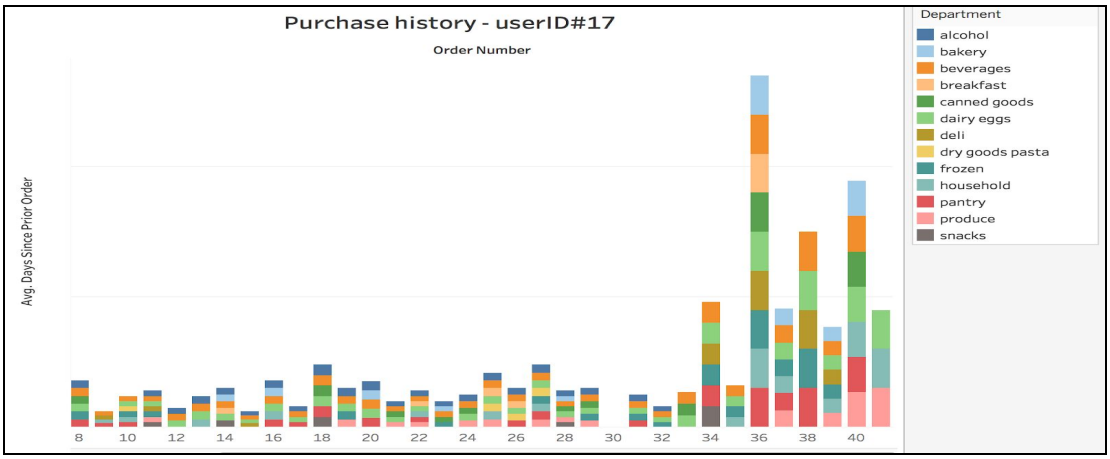
Mean repeat % = 89%

High repeat percentage indicates a lesser probability of ordering new products

Finding 2: Significance of frequency in product purchase (User ID #17)

Below is his purchase pattern in a cumulative graph. The height of the graph shows days between orders. Initially, there are regular orders at short intervals and towards the end, the frequency has changed. Orders are mostly from departments like Dairy-eggs, household, produce and beverages all throughout. There are frequent alcohol purchases in the beginning, but none in recent history. The same trend is reflected in the final order (41st order).

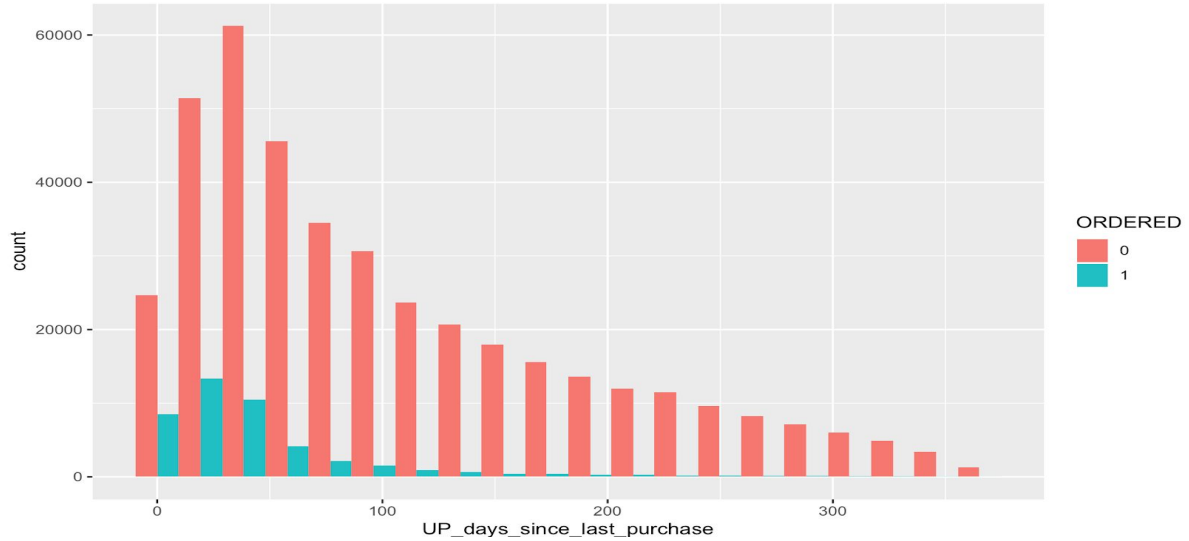
Pic. purchase pattern in a cumulative graph(user ID#17).



Findings 3: Significance of Recency in product purchase

The below graph shows the significance of Recency factor. The histogram shows a lot of products are ordered within the last 100 days. Product purchased more than say,100 days back, are less likely to be reordered. So recency of purchase is an important factor.

Pic. Histogram of purchase recency of each user product combination



Methodology

Based on the problem description, predicting the list of products a user would order would be a tedious task if it's looked at as a multi classification problem with 50K products and different combinations of it. To make it simpler this needs to be reformed as a binary classification problem. Consider each pair of user_id, product_id and predict the label as reorder = 1 or 0. Use prior orders to build user-specific features and user-product cross-relational features and use that to build a model to predict the reorder probability of each user-product combination. Determine the cut-off probability at which the model gives the best accuracy and every user_id,product_id combination above that probability level will be the final list of products customer would order.

There are 2 scenarios to predict here :

1. Order from a previously purchased list of products
2. Order products not previously ordered

The former is a relatively smaller set for most of the users and looking at features like the number of times the product was ordered, previous purchase frequencies etc we can arrive at a score or a probability. Recency and Frequency are 2 factors driving the final decision mainly.

But the latter is a bigger set and we do not have enough information to predict if a customer would order a product he never purchased before. But we can arrive at a score determining the probability of not ordering it. So for example, if the customer never made a purchase from baby products department, there is a higher confidence level that customer wouldn't order from that department so very low probability. Aisle & Department purchase habits, the percentage of new product orders in the cart, additional product features based on the type; all this information comes handy in determining the second scenario.

The second scenario will be attempted as part of the final phase of the project.

Feature Engineering

There are 5 types of features

User Level Features

- Hour of the day the user mostly places an order
- Day of the week the user mostly places an order
- Average of days between orders
- How many days since the last purchase?
- Percentage of the cart that are reorders
- Average number of products purchased by the user

User Product Cross-relationship features

- How many days since the last order of the product?
- Product purchase frequency
- Purchase_due
- Average of Ratio of add to cart order and cart size; shows the importance of the item in the order, lower the value more the importance
- Mostly ordered day of the week
- Mostly ordered Hour of the day

User Aisle Cross-relationship features

- Total number of orders from the aisle
- Average number of products from the aisle
- Days since last purchase from the aisle
- Purchase duration
- Mostly ordered day of the week

- Mostly ordered Hour of the day
- Average add to cart order
- The frequency of purchase from the aisle
- Purchase due

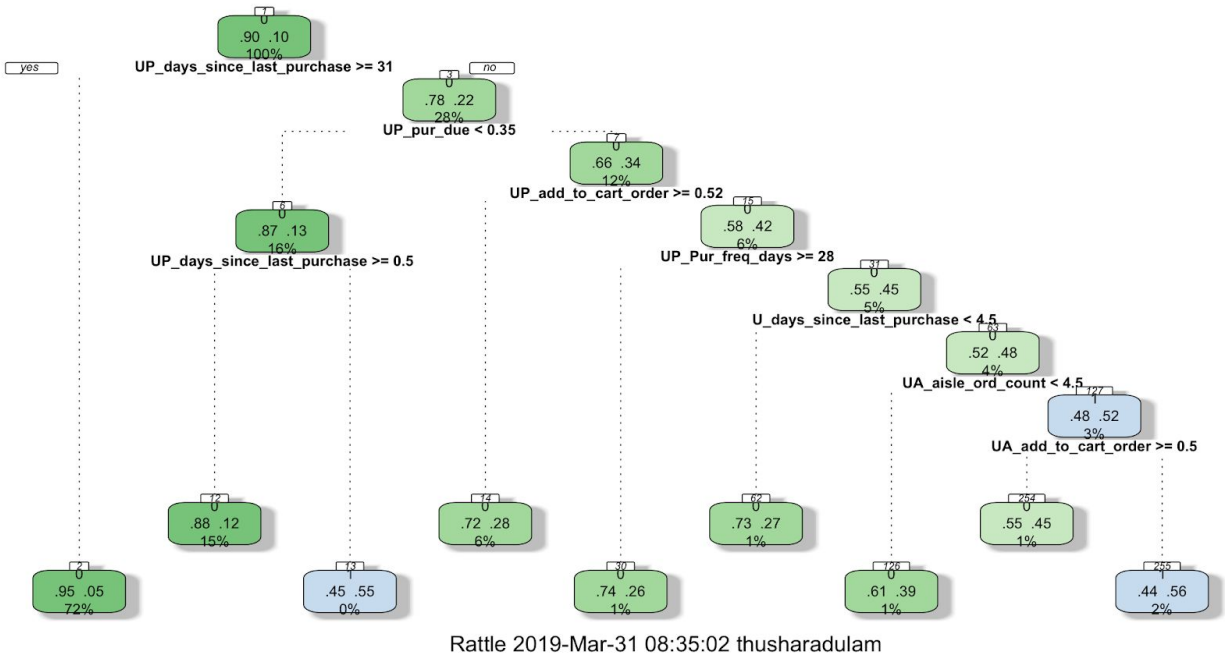
User Department Cross-relationship features

- Total number of orders from the department
- Average number of products from the department
- Days since last purchase from the department
- Purchase duration
- Mostly ordered day of the week
- Mostly ordered Hour of the day
- Average add to cart order
- The frequency of purchase from the department
- Purchase due

Significance of features

A basic cart model shows features that significantly separated products that will not get reordered. Below Decision Tree shows if a customer purchased a product more than a month back, 95% of data tells that the same won't be reordered. But 5% of orders with REORDER probability is significant considering the highly unbalanced data.

Pic. Cart Model showing the top fields in information gain



Build Model

Build a model to accurately predict if a user will reorder any of previously purchased product in their next order.

Customer and order base

Random sample of 1000 user and all the orders till date

Model Dataset:

Level of the dataset: User_id,Product_id

Dependent Variable: 1- User reordered the product in the final purchase

0- User did not reorder the product in the final purchase

Independent Variable: 35 variables derived by feature engineering

Model Type: Simple logistic regression

Model Summary:

After multiple iteration, 21 independent variables were narrowed as significant variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.554e+00	1.120e-01	-13.870	< 2e-16	***
U_max_order.x	-2.990e-02	6.396e-04	-46.757	< 2e-16	***
UP_days_since_last_purchase	-1.867e-02	1.924e-04	-97.067	< 2e-16	***
UP_Pur_freq_days	2.338e-03	3.011e-04	7.765	8.18e-15	***
UP_pur_due	1.407e-02	8.151e-04	17.263	< 2e-16	***
UP_add_to_cart_order	-8.726e-01	2.389e-02	-36.518	< 2e-16	***
UP_mostly_ord_dow	-1.089e-02	2.523e-03	-4.317	1.58e-05	***
aisle_id2	-5.754e-01	1.551e-01	-3.710	0.000207	***
aisle_id3	-2.893e-01	1.170e-01	-2.474	0.013359	*
aisle_id4	-4.702e-01	1.275e-01	-3.687	0.000227	***
aisle_id5	-1.407e+00	2.105e-01	-6.685	2.30e-11	***

U_dspo	3.014e-02	1.583e-03	19.034	< 2e-16	***
U_days_since_last_purchase	2.070e-02	7.093e-04	29.178	< 2e-16	***
U_reord_perc	2.199e+00	4.019e-02	54.710	< 2e-16	***
UA_aisle_ord_count	2.147e-02	1.029e-03	20.861	< 2e-16	***
UA_aisle_cart_perc	2.609e-01	7.399e-02	3.526	0.000422	***
UA_days_since_last_pur	3.732e-03	3.172e-04	11.766	< 2e-16	***
UA_pur_duration	9.665e-04	1.071e-04	9.021	< 2e-16	***
UA_Pur_freq_days	-2.052e-03	5.149e-04	-3.985	6.74e-05	***
UD_aisle_cart_perc	1.957e-01	5.889e-02	3.323	0.000889	***
UD_days_since_last_pur	-3.732e-03	4.367e-04	-8.547	< 2e-16	***
UD_add_to_cart_order	-7.466e-02	1.219e-02	-6.126	9.02e-10	***
UD_Pur_freq_days	-5.826e-03	7.099e-04	-8.206	2.29e-16	***
UP_dow_diff	-5.347e-02	5.044e-03	-10.601	< 2e-16	***

Note: independent variables with vif > 5 were removed one after the other. Also variable with very low significance are not included in the model

Key features to predict the re-ordered items based on the above model:

Overall	names
97.06661	UP_days_since_last_purchase
54.70957	U_reord_perc
46.75677	U_max_order.x
36.51799	UP_add_to_cart_order
29.17768	U_days_since_last_purchase

1. Days since last purchase of this specific product
2. User Reorder Percentage
3. Average orders placed by the user
4. Add to cart order of the product
5. Days since the user last purchased any product

Next Steps

1. Improve Model Accuracy
2. Address the performance issues because of multiple joins of huge files
3. Below Feature engineering alternatives can be incorporated to improve the prediction of new product purchases.
    a. Add more product features based on types like “Organic”, “Asian” etc
    b. Apply association rules and add scores as features
    c. Customer profiling and model group behaviours

4. Identify feature significance and remove less significant ones

Appendix

Data Dictionary

Orders
Order_id
user_id
eval_set
order_number
order_dow
order_hour_of_day
days_since_prior_order

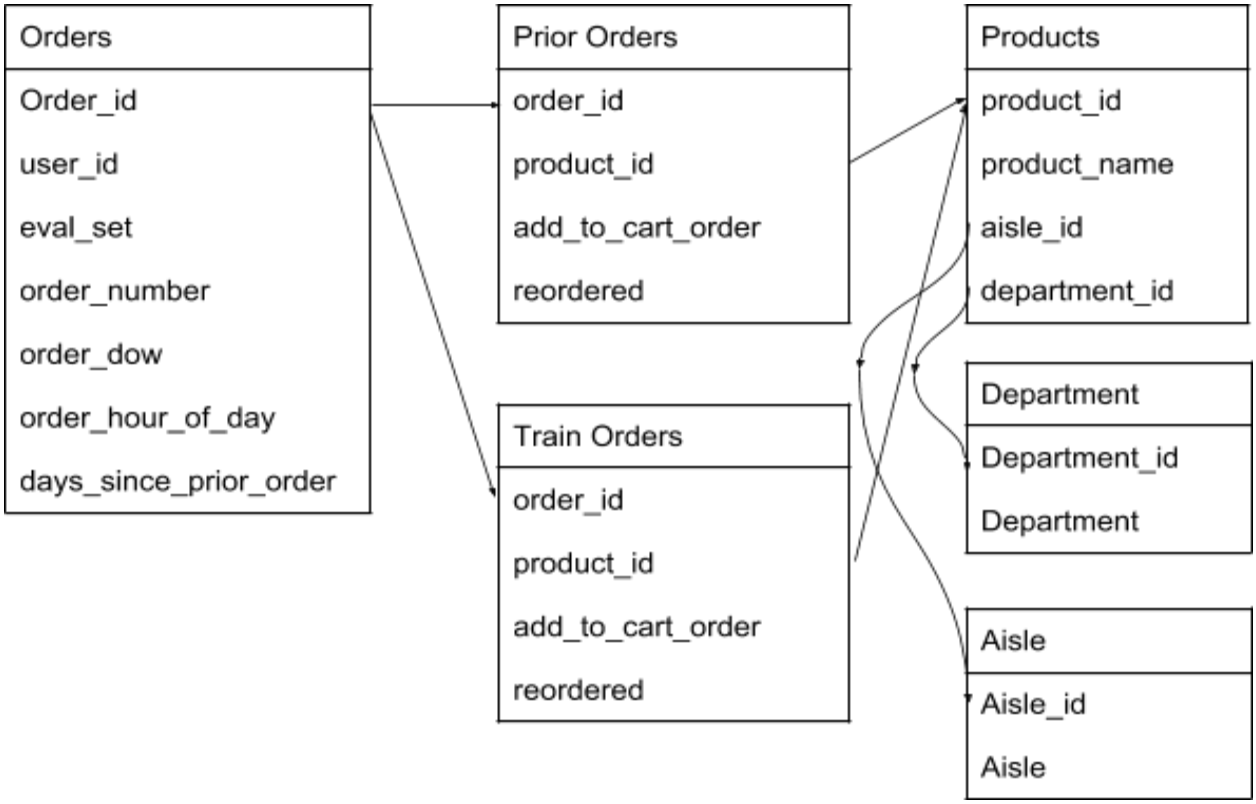
Prior Orders
order_id
product_id
add_to_cart_order
reordered

Train Orders
order_id
product_id
add_to_cart_order
reordered

Products
product_id
product_name
aisle_id
department_id

Department
Department_id
Department

Aisle
Aisle_id
Aisle



Attaching code for data file preparation here

References & Bibliography

“The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 1/Jan/2019