CSCI-P556: Applied Machine Learning

Instacart – Customer Cart Prediction and Recommendation

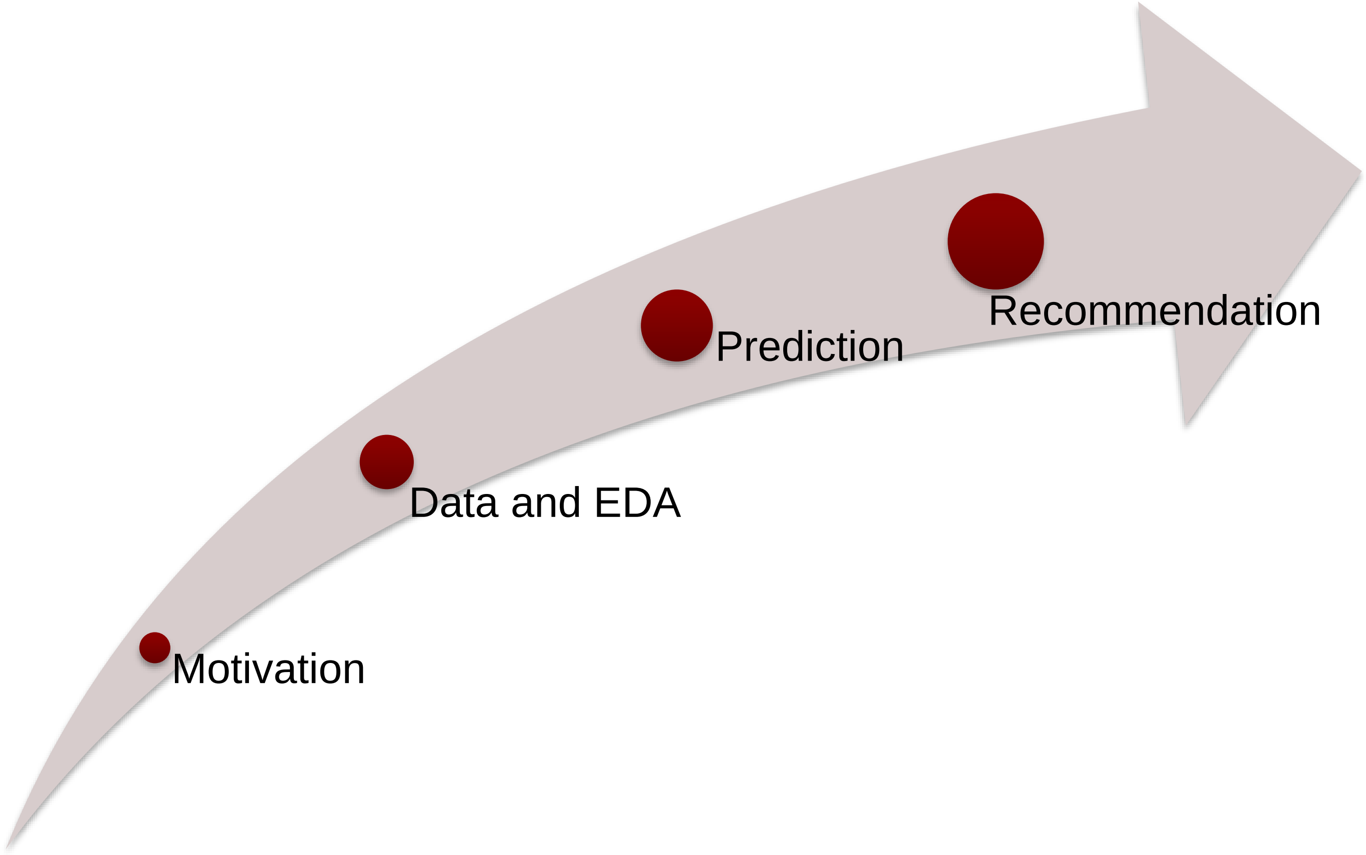
Siddartha Rao, Siddharth Kothari, and Vishal Singh

December 14, 2018

SCHOOL OF INFORMATICS
AND COMPUTING

INDIANA UNIVERSITY

Project pipeline



Motivation

For the brand-

- Gain more users
- Provide delightful shopping experience to increase customer retention

For the users-

- Save time and effort in shopping
- Discover new and better products through recommendations

Data

- The Instacart Online Grocery Shopping Dataset 2017
  - Relational Datasets describing customer orders
  - 3.3 million orders for ~50k products

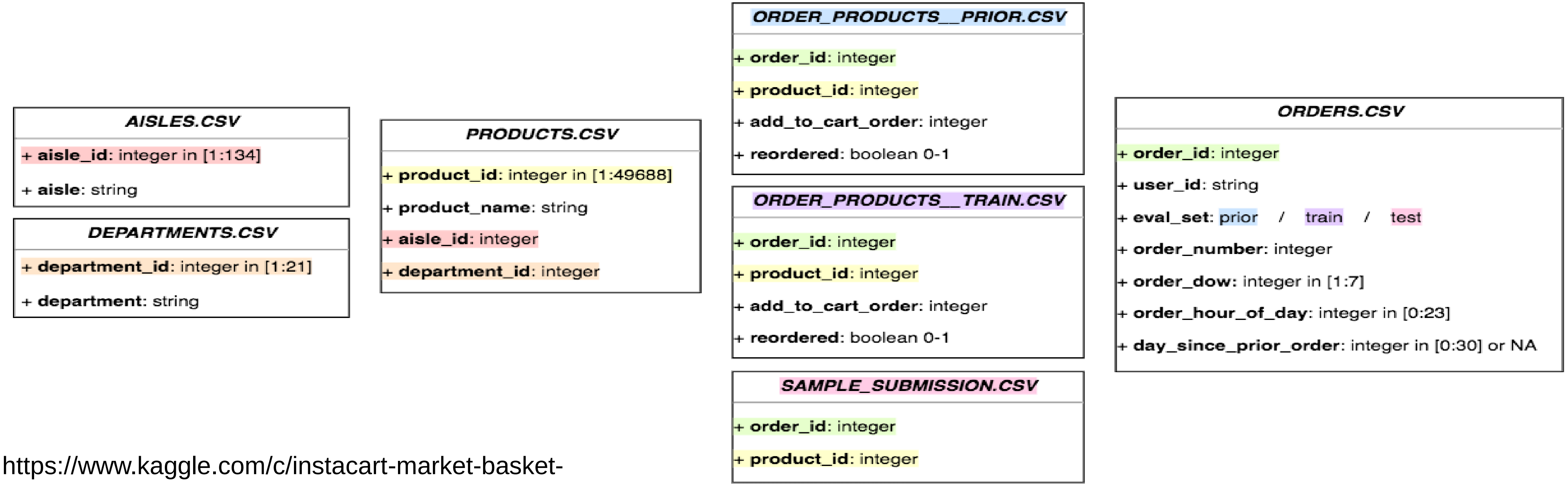
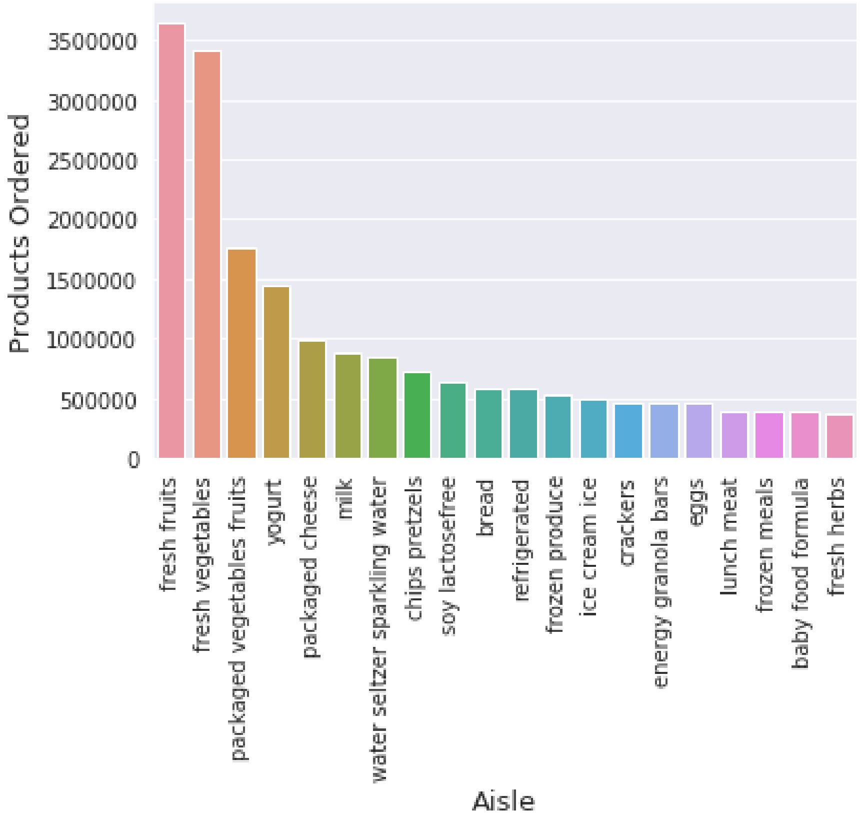


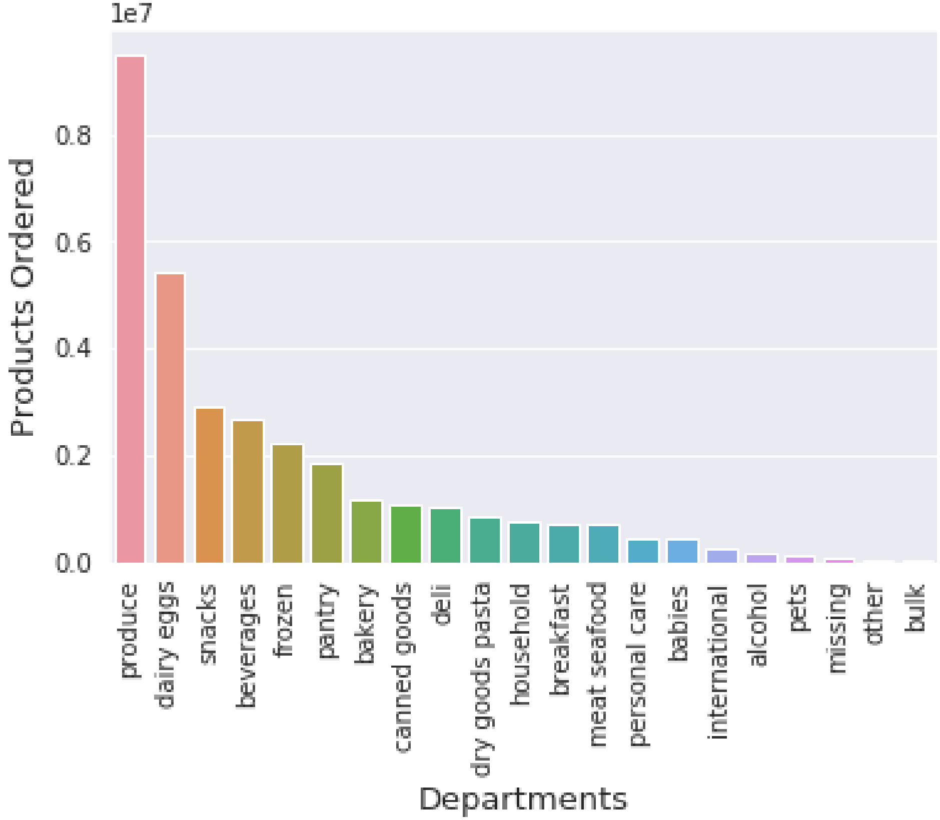
Image source: https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33128#183176

Exploratory Data Analysis

Exploratory Data Analysis

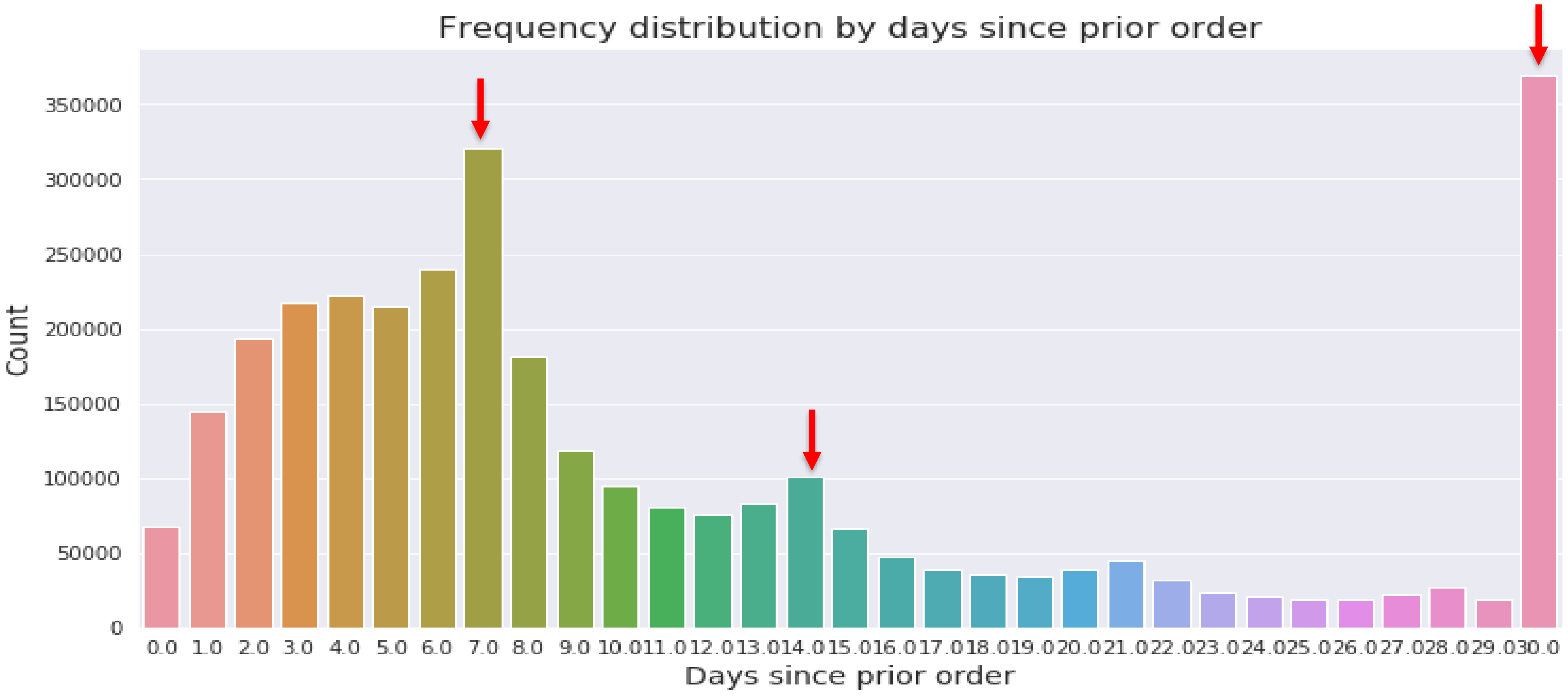
What do the users order?





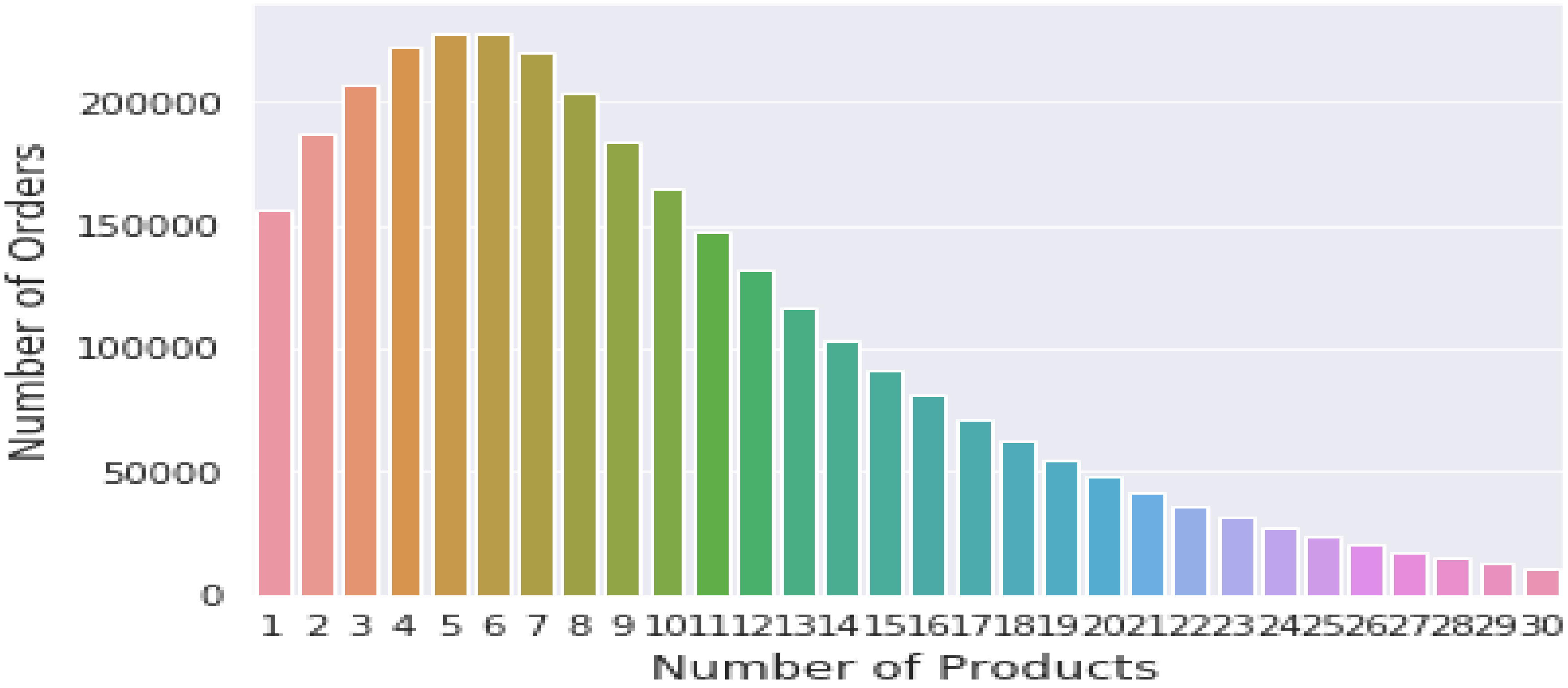
Exploratory Data Analysis

When do the users order?



Exploratory Data Analysis

How much do the users order?



Prediction

Prediction – Feature Engineering

* Existing features

User Related
- user_total_orders
- User_total_items
- Total_distinct_items
- User_avg_days_bw_orders
- user_avg_basket
- User_total_buy_max

Product related features
- Product_orders
- Product_reorders
- Product_reorder_rate
- Aisle_id*
- Department_id*

Order related features
- Order_dow*
- Order_hour_of_day*
- Days_since_prior_order*
- Days_since_ratio

User_X_Product related features
- UP_chance_ratio
- UP_chance
- UP_chance_vs_bought
- UP_drop_chance
- UP_orders
- UP_orders_ratio
- UP_reorder_rate
- UP_orders_since_last

Prediction – What will the user order?

- **Algorithms**: XGBoost and Light GBM
  1. Faster training speed and higher efficiency.
  2. Lower memory usage.
  3. Better accuracy.
  4. Support of parallel and GPU learning.
  5. Capable of handling large-scale data.
- **Model Building:** We have trained model on user’s last order (eval_set = train). However, the featured created used the data from prior data set too.
- **Feature selection**: We have used in-build method of light gbm and xgboost to find feature importance

Prediction – What will the user order?

- **Output Format:**

	order_id	products
0	2774568	17668 21903 39190 47766 18599 43961 23650 24810
1	1528013	8424 21903 38293
2	1376945	33572 17706 28465 27959 44632 24799 34658 1494...
3	1356845	11520 14992 49683 30489 7076 22959 37687 28134...
4	2161313	11266 196 10441 12427 37710 14715 27839

- **Results:**

Model	Light GBM	XGBoost
Baseline Model CV Accuracy	0.1599	0.2015
Tuned CV Accuracy	0.4412	0.3965
Kaggle Accuracy	0.3809	0.3786

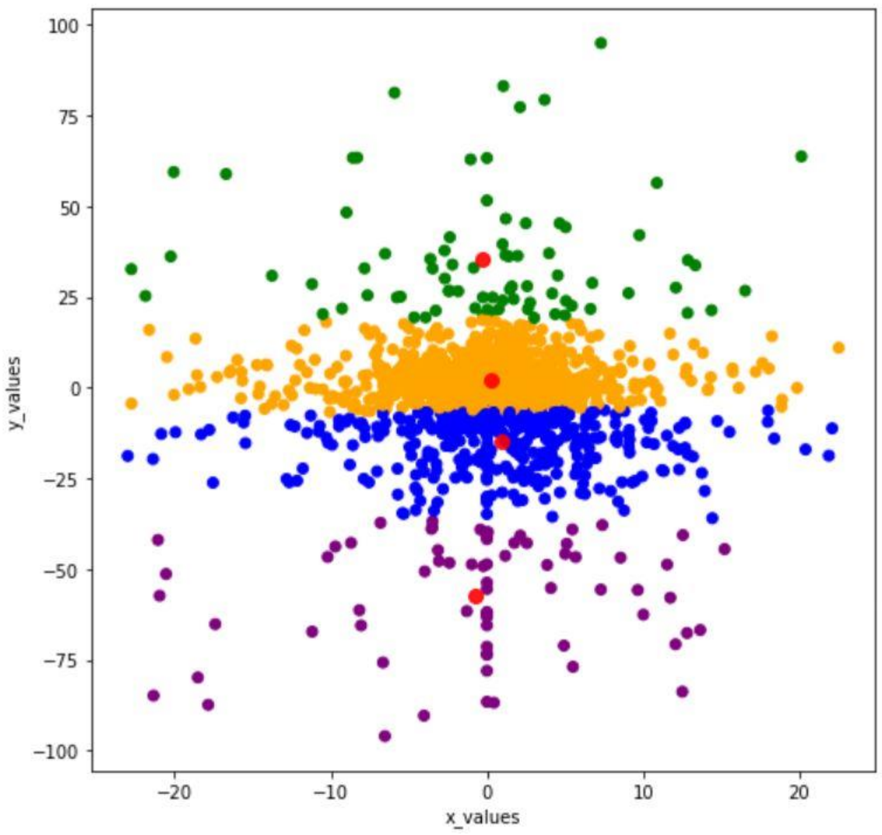
The highest accuracy in Kaggle was 0.4091

Clustering

Clustering - KMeans

- Segmentation of customers is performed on the frequency of products bought from an aisle
- Original idea of segmenting on frequency of pairs of products did not work because of high number of product pairs(large data)
- Cluster found to give the most popular products for a customer, this is used for the recommendations given

Cluster 0 ●	Fresh Fruits
Cluster 1 ●	Fresh Vegetables
Cluster 2 ●	Fresh Vegetables
Cluster 3 ●	Yogurt



Recommendation

Apriori

Support : This is the percentage of orders that contains the item set

Confidence : It measures the percentage of times that item B is purchased, given that item A was purchased.

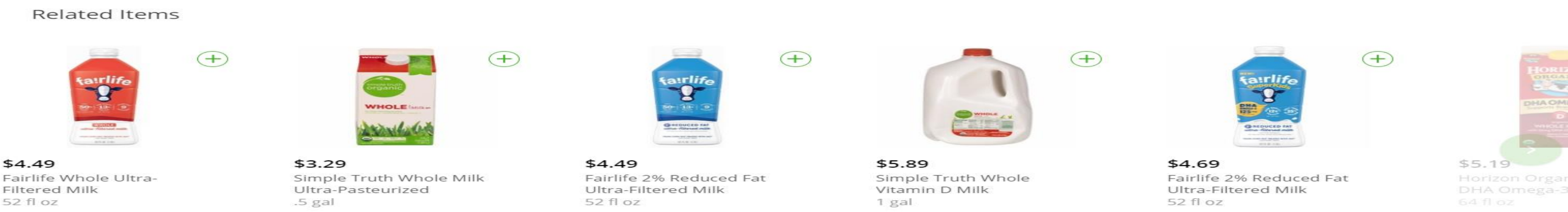
Lift : Lift indicates whether there is a relationship between A and B, or whether the two items are occurring together in the same orders simply by chance

	itemA	itemB	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	Yogurt, Sheep Milk, Strawberry	Blueberry Sheep Milk Yogurt	5	0.012986	5	0.012986	6	0.015583	1.000000	0.833333	64.173333
2	Bamba Peanut Snack	Bissli Pizza Flavor Snack	4	0.010389	5	0.012986	5	0.012986	0.800000	0.800000	61.606400
213	Iced Bhakti Chai Coffee Blend	Apple Mango Passion Fruit Fruit Snack	6	0.015583	7	0.018180	6	0.015583	0.857143	1.000000	55.005714
232	Tai Pei Chicken Chow Mein	Chicken Egg Rolls	6	0.015583	7	0.018180	6	0.015583	0.857143	1.000000	55.005714
241	Filet Mignon Canine Cuisine Wet Dog Food	Dog Food With Beef in Meaty Juices	5	0.012986	7	0.018180	5	0.012986	0.714286	1.000000	55.005714

Recommendation

We are using Light GBM, Apriori and Clustering to recommend products

- Step 1 - When user logs in : When user logs in, we can recommend products based on their transactional history. We are using LGBM to find products to recommend
- Step 2 – When user add products to cart: After user has added product to the cart, we can suggest products which they are likely to buy with the current product. For this we are using apriori and customer segmentation. As Instacart suggest 11 products, we are also recommending 11 products – 9 from apriori and 2 from clusters



Recommendation

- **Accuracy:** For every product we are recommending 11 other products. If user actually bought one or more recommended products then we will consider the recommendation as success. Accuracy for an order will be - total success/total products ordered. Final accuracy is mean of accuracy for all the orders

- **Results:** We are a accuracy of 9.6% which means that 1 in every 10 recommendation will have a product which is actually purchased by the user

- **Improvement:** Running apriori algorithm on entire dataset and clustering users by product pairs can improve the accuracy significantly