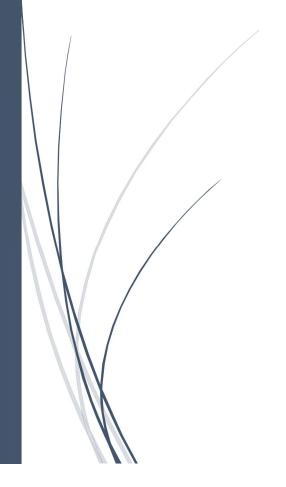
6/21/2020

Instacart Market Basket Analysis

Springboard Capstone Project # 2



Anu Kashyap SPRINGBOARD

Contents

Backg	round	2
Proble	em Statement & Objective	2
Data		2
Explor	atory Data Analysis	3
1.	Orders by Order Number	3
2.	Order Frequency	3
3.	Order Distribution by Days Since Prior Order	3
4.	Order Distribution by Order Size	4
5.	Products by Department	4
6.	Top 20 Products by # Orders	5
7.	Product by Order Day of Week	6
8.	Product by Order Hour of Day	7
9.	Product Reorder Ratio	7
10.	Reorder Ratio by Add to Cart Order	8
11.	Reorder Ratio by Order Day of Week and Hour of Day	8
Indep	endent Variables	9
Comp	arison Between Prediction Models	10
Variab	le Importance	11
Insigh [.]	ts	12

Background

Instacart operates an online grocery delivery and pick-up service. Orders are fulfilled and delivered by an Instacart personal shopper, who picks, packs, and delivers the order within the customer's designated time frame

Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session

Through a competition, Instacart is challenging the Kaggle community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order.

Problem Statement & Objective

Problem statement: Which products will an Instacart consumer purchase again?

The overall objective is to predict products that a user will buy again, try for the first time or add to cart next during a session

- Instacart currently uses XGBoost, word2vec and Annoy in production on similar data to sort items for users to "buy again"
- This data, and the algorithms trained upon it, are enabling Instacart to revolutionize how consumers discover and purchase groceries
- This helps Instacart make the right product recommendations to the customer, thereby, making the shopping experience more convenient for consumer

Data

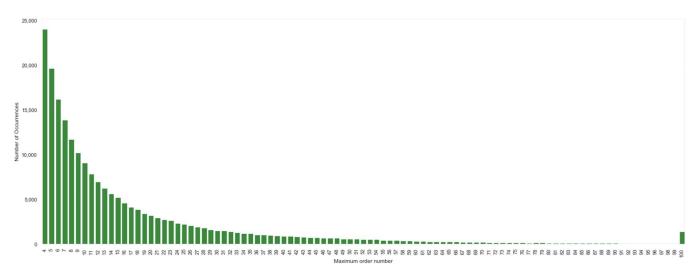
The dataset for this competition is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. The datasets are:

- 1. aisles.csv: contains the aisle_id and aisle_name of a product
- 2. departments.csv: contains the department_id and department_name of a product
- 3. order_products__prior.csv: contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product
- order_products__train.csv:
- 5. orders.csv: Contains information about which set (prior, train, test) an order belongs. Need to predict reordered items only for the test set orders. 'order_dow' is the day of week
- 6. products.csv: contains mapping between product, aisle and department

Exploratory Data Analysis

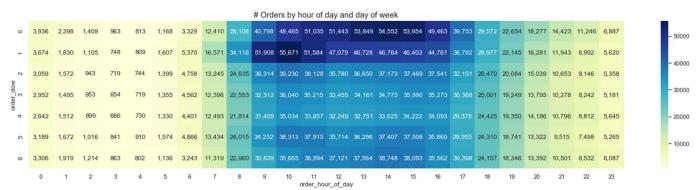
1. Orders by Order Number

• All users have at least 4 orders and at most 100 orders



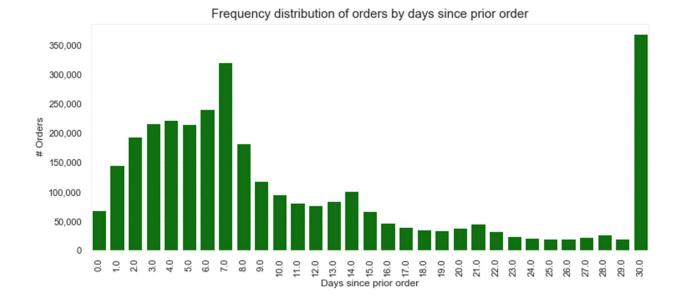
2. Order Frequency

Saturday afternoon and Sunday mornings/afternoons have high order frequencies



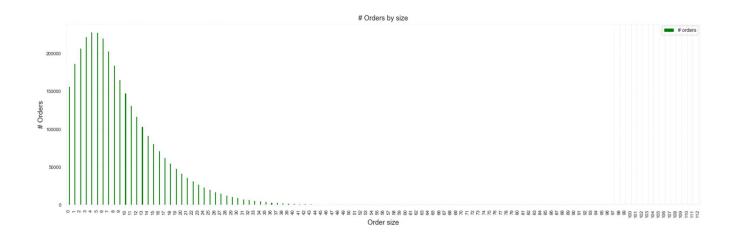
3. Order Distribution by Days Since Prior Order

- Customers order once in every week (peak at 7 days) or once in a month (peak at 30 days)
- Also observed smaller peaks at 14, 21 and 28 days (weekly intervals)



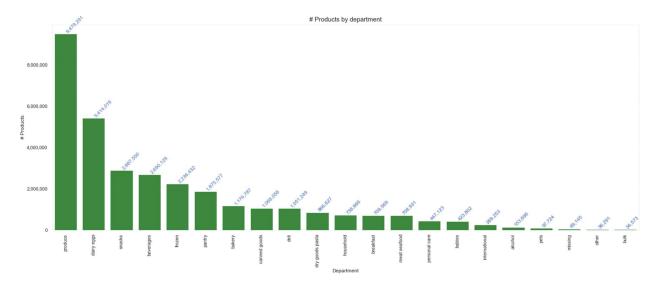
4. Order Distribution by Order Size

• Maximum order size is 5, with most orders with 3-7 products



5. Products by Department

• Most products are ordered from the Produce department (Fruits and vegetables)



6. Top 20 Products by # Orders

• Most of the top 20 ordered products are either fruits and vegetables

#	Aisle	Department	Product	# Orders	Reorder Ratio
1	fresh fruits	produce	Banana	472,565	84.4%
2	fresh fruits	produce	Bag of Organic Bananas	379,450	83.3%
3	fresh fruits	produce	Organic Strawberries	264,683	77.8%
4	packaged vegetables fruits	produce	Organic Baby Spinach	241,921	77.3%
5	fresh fruits	produce	Organic Hass Avocado	213,584	79.7%
6	fresh fruits	produce	Organic Avocado	176,815	75.8%
7	fresh fruits	produce	Large Lemon	152,657	69.6%
8	fresh fruits	produce	Strawberries	142,951	69.8%
9	fresh fruits	produce	Limes	140,627	68.1%
10	milk	dairy eggs	Organic Whole Milk	137,905	83.0%
11	packaged vegetables fruits	produce	Organic Raspberries	137,057	76.9%
12	fresh vegetables	produce	Organic Yellow Onion	113,426	69.7%
13	fresh vegetables	produce	Organic Garlic	109,778	68.0%
14	fresh vegetables	produce	Organic Zucchini	104,823	68.8%

15	packaged vegetables fruits	produce	Organic Blueberries	100,060	62.9%
16	fresh vegetables	produce	Cucumber Kirby	97,315	69.2%
17	fresh fruits	produce	Organic Fuji Apple	89,632	71.2%
18	fresh fruits	produce	Organic Lemon	87,746	69.0%
19	fresh fruits	produce	Apple Honeycrisp Organic	85,020	73.5%
20	packaged vegetables fruits	produce	Organic Grape Tomatoes	84,255	65.6%

7. Product by Order Day of Week

• Banana is the most ordered and reordered product and mostly ordered on Saturday and Sunday

Top 20 reordered products by day of week

				a producti	, 2, 44, 4			
Banana	81,347	74,425	50,188	44,925	44,479	50,201	53,044	
Bag of Organic Bananas	59,420	58,202	41,666	37,780	37,203	40,832	40,810	
Organic Strawberries	42,129	36,015	25,399	23,725	23,963	26,512	28,102	
Organic Baby Spinach	42,418	32,494	22,974	19,804	19,786	22,167	27,241	
Organic Hass Avocado	34,867	30,711	22,058	19,476	19,271	21,167	22,581	
Organic Avocado	30,230	24,425	16,231	14,389	13,998	15,774	18,997	
Organic Whole Milk	21,788	19,482	14,608	13,762	13,801	14,957	16,112	
Large Lemon	23,433	18,802	12,766	11,027	11,321	13,393	15,513	
Organic Raspberries	20,092	18,729	13,562	12,540	12,395	14,039	14,052	
Organic Raspberries Strawberries Limes Organic Yellow Onion	19,138	17,965	12,698	11,673	11,992	13,189	13,147	
Limes	20,622	15,954	11,222	10,126	10,372	12,363	15,109	- 4
Organic Yellow Onion	18,135	13,668	9,683	8,503	7,997	8,983	12,103	
Organic Garlic	17,063	12,396	8,960	7,884	7,738	8,833	11,789	
Organic Zucchini	16,335	12,649	8,809	7,821	7,559	8,306	10,686	
Cucumber Kirby	15,295	12,065	8,199	7,213	7,102	7,711	9,728	
Organic Fuji Apple	11,832	13,147	8,497	7,111	6,970	8,512	7,742	
Organic Blueberries	11,870	10,969	7,943	7,389	7,552	8,592	8,607	
Apple Honeycrisp Organic	12,789	11,825	7,835	7,048	6,820	7,862	8,331	
Organic Lemon	12,205	10,960	7,780	6,805	6,579	7,775	8,432	
Organic Grape Tomatoes	12,880	9,766	6,601	5,604	5,807	6,551	8,021	
	0	1	2	3 order_dow	4	5	6	

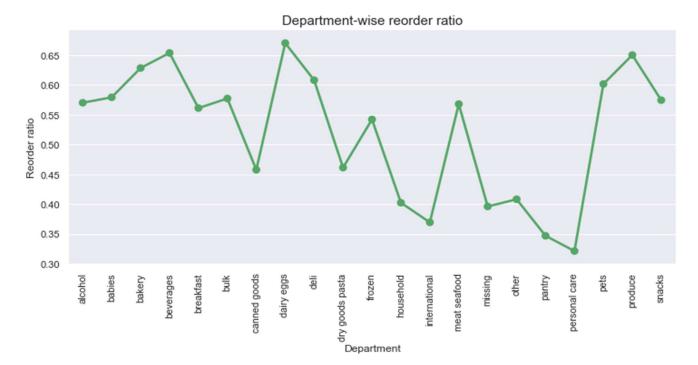
8. Product by Order Hour of Day

 Banana is the most ordered and reordered product and mostly ordered during afternoon hours

										Тор	20 re	order	ed pro	ducts	by h	our of	day								
	Banana	2,315	1,136	681	477	563	1,108	4,125	12,607	23,764	32,844	34,691	32,511	30,366	31,072	32,157	32,015	30,785	25,397	20,271	15,490	12,236	9,973	7,539	4,486
	Bag of Organic Bananas	2,240	1,136	650	464	489	863	3,021	10,143	18,812	25,595	26,736	25,568	24,314	25,069	25,613	25,676	24,148	19,668	15,400	12,178	9,646	7,941	6,464	4,079
	Organic Strawberries	1,366	652	346	254	299	582	2,226	7,312	12,410	15,648	16,492	16,124	15,542			16,537	15,924	12,448	10,164	8,084	6,746	5,786	4,632	2,770
	Organic Baby Spinach	1,348	727	381	280	284	552	1,796	5,797	10,516	13,655	15,414	15,222	15,034	15,207	15,572	15,306	14,334	11,715	9,355	7,193	5,885	4,909	3,971	2,431
	Organic Hass Avocado	1,102	519	305	239	243	480	1,626	5,361	9,615	12,621	13,908	13,874	13,519	13,756	14,195	14,337	13,319	10,932	8,367	6,656	5,384	4,298	3,397	2,078
	Organic Avocado	839	415	260	185	187	315	1,167	3,805	7,226	9,852	11,028	11,156	10,919	11,230	11,293	11,166	10,741	8,635	6,697	5,182	4,000	3,399	2,674	1,673
	Organic Whole Milk	730	328	187	110	132	317	1,296	4,262	7,495	9,116	9,429	8,850	8,490	9,068	9,486	9,113	8,205	6,563	5,363	4,350	4,006	3,471	2,664	1,479
	Large Lemon	663	305	210	154	177	310	933	2,893	5,833	8,121	9,185	9,333	9,090	8,830	9,078	8,926	8,228	6,762	4,968	3,601	2,835	2,528	2,061	1,231
۵	Organic Raspberries	711	316	195	125	144	303	1,132	3,733	6,788	8,771	8,919	8,347	7,906	8,247	8,288	8,329	7,856	6,330	4,933	3,990	3,385	2,905	2,371	1,385
nam	Strawberries	587	320	169	118	167	318	1,058	3,427	6,220	8,284	8,735	8,302	7,784	7,770	8,076	7,876	7,419	5,972	4,680	3,678	2,955	2,588	2,036	1,263
product_name	Limes	617	309	168	127	136	217	822	2,478	4,931	7,003	8,239	8,291	8,161	8,283	8,169	8,221	7,622	5,965	4,641	3,318	2,579	2,380	1,900	1,191
bio	Organic Yellow Onion	477	275	125	93	113	221	673	2,157	4,223	5,899	6,776	6,873	6,724	7,061	6,947	6,873	6,087	4,682	3,414	2,608	2,166	2,046	1,594	965
	Organic Garlic	511	269	150	106	121	202	590	1,876	3,703	5,266	6,390	6,530	6,505	6,507	6,726	6,392	5,915	4,578	3,374	2,559	1,996	1,884	1,560	953
	Organic Zucchini	494	238	140	107	111	203	692	2,160	3,936	5,360	5,852	6,052	5,830	5,983	6,124	5,954	5,525	4,428	3,369	2,602	2,227	2,093	1,660	1,025
	Cucumber Kirby	497	248	122	91	120	231	691	1,927	3,439	4,831	5,472	5,550	5,488	5,532	5,621	5,429	5,343	4,399	3,410	2,667	2,146	1,770	1,444	845
	Organic Fuji Apple	390	163	89	80	86	164	634	2,009	4,062	6,059	6,284	5,450	4,735	4,818	4,863	4,930	4,749	3,922	2,988	2,264	1,788	1,480	1,147	657
	Organic Blueberries	458	221	123	93	90	174	631	2,164	3,745	5,018	5,295	5,032	4,706	4,917	4,957	4,900	4,608	3,761	3,142	2,403	2,175	1,889	1,490	930
А	pple Honeycrisp Organic	400	151	103	75	81	198	642	2,135	3,961	5,277	5,461	5,206	4,978	5,044	5,022	5,112	4,639	3,650	2,769	2,219	1,796	1,575	1,331	685
	Organic Lemon	415	210	144	92	99	161	576	1,684	3,390	4,577	5,150	5,076	4,824	5,014	4,991	5,005	4,748	3,779	2,974	2,320	1,864	1,448	1,227	768
(Organic Grape Tomatoes	405	187	120	78	81	143	537	1,658	2,956	4,132	4,610	4,449	4,441	4,371	4,560	4,609	4,246	3,414	2,745	2,158	1,773	1,619	1,238	700
		0	1	2	3	4	5	6	7	8	9	10 ore	11 der_ho	12 ur_of_d	13 ay	14	15	16	17	18	19	20	21	22	23

9. Product Reorder Ratio

Dairy eggs department has highest reorder ratio and Personal care has the lowest



- 30000

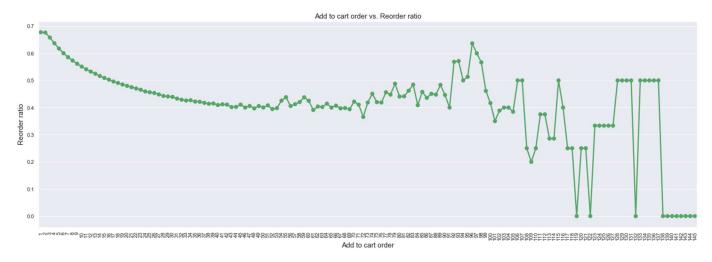
- 25000

- 10000

- 5000

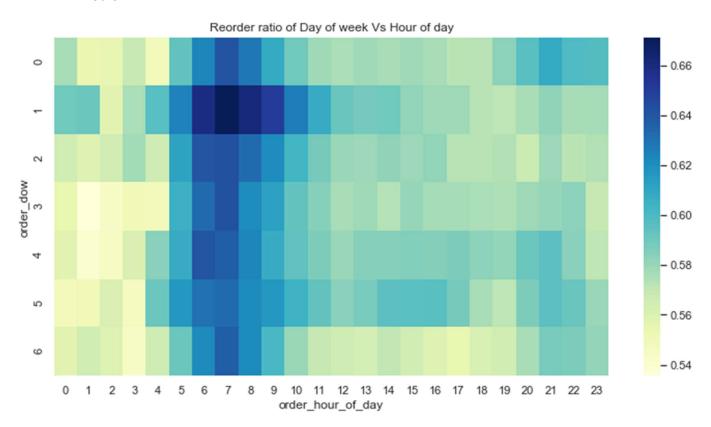
10. Reorder Ratio by Add to Cart Order

• The products that are added to the cart initially are more likely to be reordered again compared to the ones added later which makes sense as we tend to first order all the products we buy frequently and then search for new products



11. Reorder Ratio by Order Day of Week and Hour of Day

 Most of the reordered or frequently ordered products are ordered in the early morning hours



Independent Variables

#	Variable name	Description	Туре
1	user_tot_orders	Number of total orders for a user (user level)	Numeric
2	user_tot_prods	Total number of items a user bought (user level)	Numeric
3	user_tot_dist_prods	Number of distinct products a user bought (user level)	Numeric
4	user_avg_days_bet_orders	Average number of days between orders for a user (user level)	Numeric
5	user_avg_order_size	Average number of items in a user's order (user level)	Numeric
6	order_hour_of_day	Hour of the day for the order for which we need to predict products (order level)	Numeric
7	days_since_prior_order	# of days since the prior order for the order for which we need to predict products (order level)	Numeric
8	days_since_ratio	days_since_prior_order/user_avg_days_bet_orders (order level)	Numeric
9	Department	Department name of the product (order level)	Categorical
10	prod_ordered	# of times the product has been ordered (product level)	Numeric
11	prod_reordered	# of times the product has been reordered (product level)	Numeric
12	prod_reorder_rate	prod_reordered / prod_ordered	Numeric
13	uxp_tot_orders	Total number of orders at user X product level	Numeric
14	uxp_order_rate	uxp_tot_orders / user_tot_orders	Numeric
15	uxp_avg_pos_in_cart	Average position in cart at user X product level	Numeric

16	uxp_reorder_rate	uxp_reordered/user_reorderd_prods	Numeric
17	uxp_orders_since_last_order	# orders since the user ordered this product (Total # of orders for a user – order number of the product when it was ordered by the user)	Numeric
18	uxp_delta_hour_vs_last	order_hour_of_day -	Numeric

Comparison Between Prediction Models

	XGBoost	Light GBM	Random Forest
Dependent variable	Predicts whether a product will be ordered in the user's next ordered	Predicts whether a product will be ordered in the user's next ordered	Predicts whether a product will be ordered in the user's next ordered
Hyperparameter Tuning Method	Randomized Search	Randomized Search	Grid Search
Optimal Hyperparameters	objective: 'binary:logistic' n_estimators: 50 max_depth: 5 learning_rate: 0.2	boosting_type: 'gbdt' objective: 'binary' metric: {'binary_logloss'} num_leaves: 96 max_depth: 10 feature_fraction: 0.9 bagging_fraction: 0.95 bagging_freq: 5	max_dept: 7 n_estimators: 70 max_features: 'auto'
Accuracy (Train set)	90.1%	-	90.7%
Accuracy (Hold out set on Kaggle)	37.4%	37.8%	17.4%

Variable Importance

• uxp_order_rate and uxp_orders_since_last_order are the top two predictors

#	Variable name	Importance	Description
1	uxp_order_rate	44.7%	uxp_tot_orders / user_tot_orders
2	uxp_orders_since_last_order	17.9%	# orders since the user ordered this product (Total # of orders for a user – order number of the product when it was ordered by the user)
3	uxp_reorder_rate	9.9%	uxp_reordered/user_reorderd_prods
4	uxp_tot_orders	7.3%	Total number of orders at user X product level
5	prod_reorder_rate	4.9%	prod_reordered / prod_ordered
6	department	3.8%	Department name of the product (order level)
7	days_since_prior_order	2.6%	# of days since the prior order for the order for which we need to predict products (order level)
8	days_since_ratio	1.4%	days_since_prior_order/user_avg_days_bet_orders (order level)
9	prod_reordered	1.3%	# of times the product has been reordered (product level)
10	uxp_delta_hour_vs_last	1.2%	order_hour_of_day -
11	user_tot_prods	0.8%	Total number of items a user bought (user level)
12	user_avg_days_bet_orders	0.7%	Average number of days between orders for a user (user level)
13	user_avg_order_size	0.7%	Average number of items in a user's order (user level)
14	user_tot_dist_prods	0.7%	Number of distinct products a user bought (user level)
15	user_tot_orders	0.7%	Number of total orders for a user (user level)

16	prod_ordered	0.6%	# of times the product has been ordered (product level)
17	uxp_avg_pos_in_cart	0.6%	Average position in cart at user X product level
18	order_hour_of_day	0.2%	Hour of the day for the order for which we need to predict products (order level)

Insights

- Out of the three classification models used, XGBoost, Light GBM and Random Forest, Light GBM classification model gives the best accuracy at 37.8% on the Kaggle hold out set
- Light GBM hyperparameters were determined using Randomized Search and the best hyperparameters are:
 - Boosting_type: 'gbdt'
 - o Objective: 'binary'
 - o Metric: {'binary_logloss'}
 - o Num_leaves: 96
 - o Max_depth: 10
 - o Feature_fraction: 0.9
 - o Bagging_fraction: 0.95
 - o Bagging_freq: 5
- The top 5 variables as per the variable importance are:
 - 1. uxp_order_rate (44.7%)
 - 2. uxp_orders_since_last_order (17.9%)
 - 3. uxp_reorder_rate (9.9%)
 - 4. uxp_tot_orders (7.3%)
 - 5. prod_reorder_rate (4.9%)