Capstone Project 1: Instacart Market Basket Analysis: Predicting Customer Dietary Preference and Next Order List

By: Dr. Pamela Ubaldo Augustine

I. Introduction

Have you noticed that recommendations given to you by experts, family or friends often influence what you buy? This is because recommendations work! About 30% of Amazon page views are from recommendations (1). In June 2012 amazon has increased sales to 29% from 9.9 B the previous year to 12.83B because of its recommendation system (2).

In this project, I used the "The Instacart Online Grocery Shopping Dataset 2017" open sourced this year with over 3 million orders and 200,000 users in the dataset. Instacart is an online grocery delivery app that made it the top of the list of Forbes most promising company in 2015 (3). Using this dataset, I want to answer to main questions 1) Can customers be grouped using similarities of what they buy? 2) Can their next order list be predicted using just any order or using their entire order history?

Answers to these two main questions will enable me to build a recommendation system for Instacart's anonymized customers. Goals of this project are 1) provide an overall insight from the data using exploratory data analysis 2) identify eating preferences of customers (meat lovers, pescatarian, vegetarian vegan or nonvegan) and 3) predict customer's next order list.

II. Data Wrangling

Dataset was acquired from https://www.instacart.com/datasets/grocery-shopping-2017 composed of seven csv files namely aisles, departments, orders_products_prior, order_products_train, orders, products and sample submission files. All seven files were opened and stored in a dataframe using python. The dataset contains over 3 million orders of 206, 209 anonymized users. Each csv file is inspected and only "day_since_prior_order" column in "orders" has a total of 206, 209 missing data which represent first orders of all the 206,209 users. This is an example of Missing at Random (MAR) data mechanism where missing data is related to observed data.

The "orders" dataframe include three evaluation sets namely prior, train and test. All order history of 206, 209 users are in "prior" while only latest orders of 131,209 users are in "train" and latest orders of 75,000 users are in "test". These three evaluation sets were separated into three dataframes namely oprior, otrain and otest.

Dataframes oprior, order_products_prior, products, aisles and department were all merged together into "allprior" . While dataframes otrain, order_products_train, products, aisles and department were merged together into "alltrain" before doing any data extensive data analysis.

III. Design of Experiment

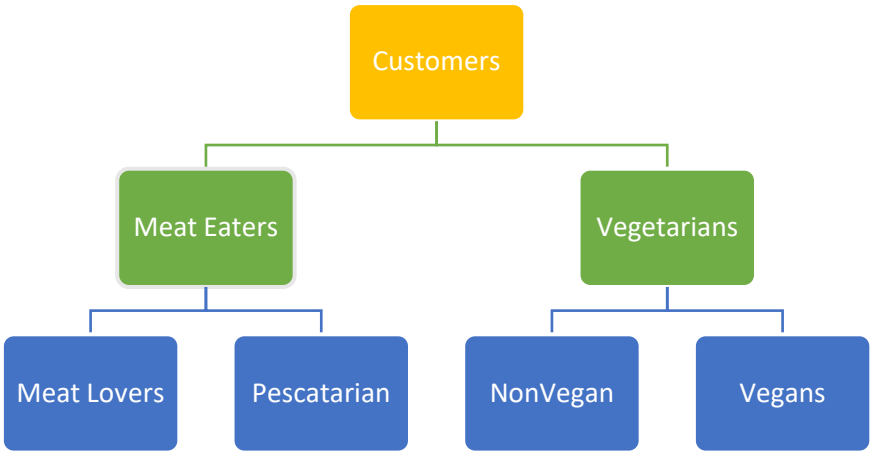
The next steps were done after cleaning and merging the dataframes

A. Exploratory Data Analysis

- A list of questions about the data set were answered. There are 18 questions answered that aims to help achieve the goals of this project

B. Customers are segmented according to dietary preference

Figure 1. Customer Dietary Preference



- Each diet preference is segmented according to the aisles customers frequently order from. Meat Lovers are those that order from aisles in Table 1 while Pescatarians order only on seafood aisles (bold in Table 1). NonVegans vegetarians never order from aisles in Table 1 but have orders on aisles in Table 2 while Vegans never order from aisles in both Table 1 and 2.

Table 1. Meat Lover and Pescatarian Aisles

aisle_id	aisle name
5	marinades meat preparation
7	packaged meat
15	***packaged seafood***
34	***frozen meat seafood***
35	poultry counter
39	***seafood counter***
49	packaged poultry
95	***canned meat seafood***
96	lunch meat
106	hot dogs bacon sausage
122	meat counter

Table 2. NonVegan Vegetarians Aisles

aisle_id	aisle name
2	specialty cheeses
21	packaged cheese
53	Cream
84	Milk
86	Eggs
108	other creams cheeses
120	Yogurt

- Three sample set with 30,000 users each were obtained from “alltrain” dataframe and stored in three separated dataframes E1, E2, E3.

- Each user and order in E1, E2 and E3 were labeled according to diet preference and compared to each other. Labeled users and orders are stored on separate dataframes described in Table 3.

Table 3. DataFrame Names Segmented by USERS/ORDERS

30,000 users with order history	Segmented by USERS	Segmented by ORDERS
E1	P_user1	P_order1
E2	P_user2	P_order2
E3	P_user3	P_order3

- Hypothesis testing was done see if distributions of dietary preferences are significantly different (or not) when segmented by users or by orders. P_user1 and P_order1 dataframes were used in this part.

- Resampling by simulating the aisles each order is from using users in E1 and probabilities of each aisle being ordered from and stored in Simu1 dataframe.

- Simu1 is also segmented into dietary preference according to users and orders and stored in Psimulated_user and Psimulated_order dataframes. Hypothesis testing is done to see if there is significant difference between dietary preference distribution in the simulated and empirical samples. Psimulated_user vs P_user1 and Psimulated_order vs. P_order1.

C. <u>Calculation of P(Diet|Aisle) and P(Aisle|Diet) using Bayes Theorem</u>

D. <u>Prediction of customer's next order list</u>

- Next order list is predicted using 1) average number of products per order of each user and 2) probabilities of each product being ordered by user using order history
- Cosine similarities and F1 score were used as metrics

IV. Information from the Data

A. <u>Exploratory Data Analysis</u>

Inspecting the dataset, I learned that there are ***49,688*** products, ***134*** aisles, ***21*** departments, ***206, 209*** users and ***3,421, 083*** orders. Peak hours happen from 8am to 10 pm where orders are over 100,000 orders per hour. Highest order is at 10am with ***288,418*** orders and lowest order is at 3 am with only ***5,474*** orders as shown in Figure 2.

Orders are at peak on Mondays and Tuesdays are where orders are over ***500,000*** orders. Orders slows down on Wednesdays and hit a low of ***426,339*** orders on Friday. Orders starts picking up a little on weekends Saturdays and Sundays as shown in Figure 3.

This means that website must be at optimum performance on peak hours 8am to 10 pm every day. Website maintenance that requires a downtime is best done on a Friday during off peak hours 12am to 6am when order volume is at low.

Looking at each products, aisles and department, the top 10 highest probability of being ordered are summarized in Table 4. Produce and fresh fruits are the most ordered department and aisle. Most ordered product is banana and 6 of the top 10 products are organic produce. These top 10 products, department and aisles are those that needed to have the most stocked for future orders.

Figure 2. Hourly Order Volume

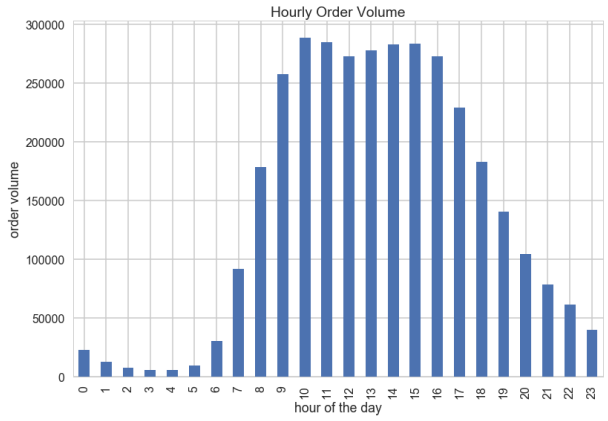


Figure 3. Daily Order Volume

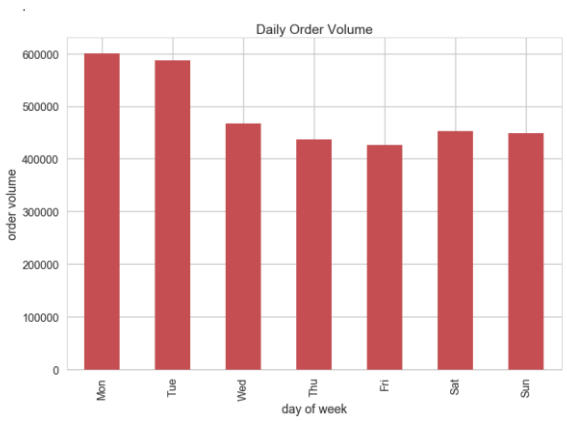


Table 4. Top 10 Most Ordered Products, Aisles and Departments

Top	Products	Departments	Aisles	aisle_id
1	Banana	produce	fresh fruits	24
2	Bag of Organic Banana	dairy eggs	fresh vegetables	83
3	Organic Strawberries	Snacks	packaged vegetables fruits	123
4	Organic Spinach	beverages	yogurt	120
5	Organic Hass Avocado	frozen	packaged cheese	21
6	Organic Avocado	pantry	milk	84
7	Large Lemon	bakery	water seltzer sparkling water	115
8	Strawberries	canned goods	chips pretzels	107
9	Lime	deli	soy lactosefree	91
10	Organic Whole Milk	dry good pasta	bread	112

I also looked at aisle distribution of orders using “alltrain” with latest orders and “allprior” with order history of customers. Figure 4 shows the aisle distribution for both where each aisle number is plotted against probability of being ordered from. There are some differences in aisle distribution looking at just latest order (train) and considering order history (prior). However, most ordered aisles are still prevalent for both represented by a long horizontal bar such as aisle 24, 83, 123, 120 and 21.

These calculated probability is used in resampling or generating a simulated dataset used for hypothesis testing of whether calculated % of Dietary preference are just random or not.

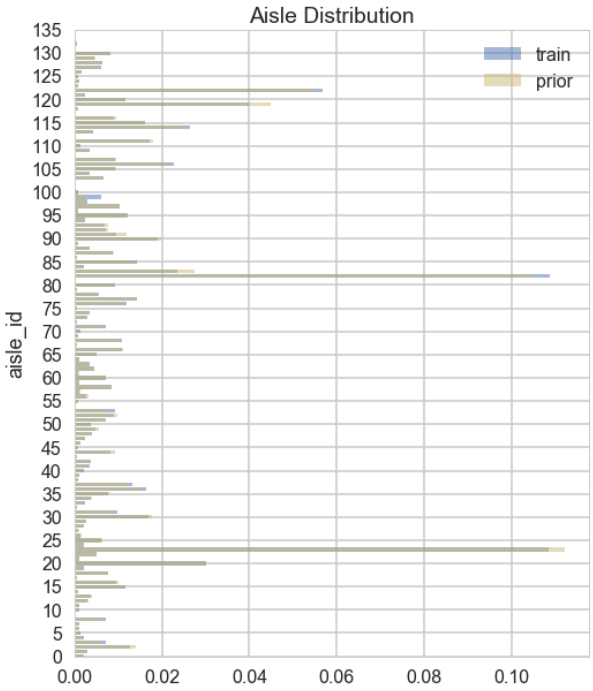


Figure 4. Aisle Distrubution of Orders

GOAL 1. Dietary Preference of Customers

Dietary preference distribution is very similar across all three datataframes P_user1, P_user2 and P_user3 each with 30,000 users as shown in Table 5. This means this is a good representation of dietary distribution by users of the entire population. About 65% of users are Meat Lovers, 23% NonVegans, 7% Vegans and 5% Pescatarians. This sounds about right since prior to calculating these numbers I would assume that most people in the US are meat eaters especially in the Midwest region. Seafood is not really so popular except in the coast so I would expect that meat lovers also eat seafood but smaller population of customers eat only seafood.

Prior to calculating these number my guess is from highest to lowest percentage is Meat Lovers>Pescatarians> NonVegans> Vegans but I only got one right. Not a lot of people like only vegetables so Vegans will be the smallest percentage but I was wrong. Pescatarian is the smallest.

Dietary preference distribution by order is also compared across three dataframes P_order1, P_order2 and P_order3 each with 30,000 users shown in Table 6. These distribution is also very similar which means these can be a good representation of diet distribution by order of the entire population. About 40% of orders are NonVegans, 32% Vegan, 25% Meat Lovers and 33% Pescatarians.

Table 5. Dietary Preference Distribution by USERS

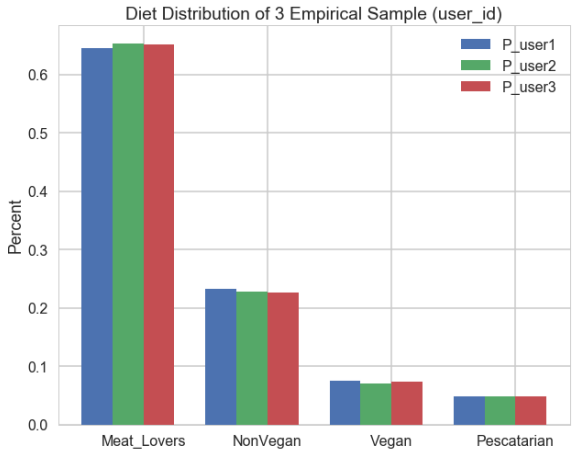
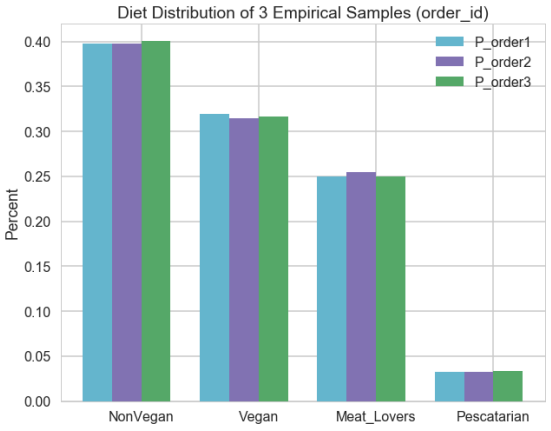


Table 6. Dietary Preference Distribution by ORDERS



HYPOTHESIS TESTING PART I.

I also did a hypothesis testing to see if there is significant difference between dietary preference distribution by USERS and by ORDERS. Below are my hypotheses.

Ho: There is no significant difference in classifying diet preference by USERS versus by ORDERS

H1: There is significant difference in classifying diet preference by USERS versus by ORDERS

Looking at Figure 5, we can already see the difference between the two. But let us calculate p_values for each diet preference for sure which is shown in Table 7. All p_values are less than 0.05 which means we can reject Ho and accept H1: There is significant difference in classifying diet preference by USERS versus by ORDERS.

Figure 5.. Comparison of Diet Distribution by USERS vs by ORDERS

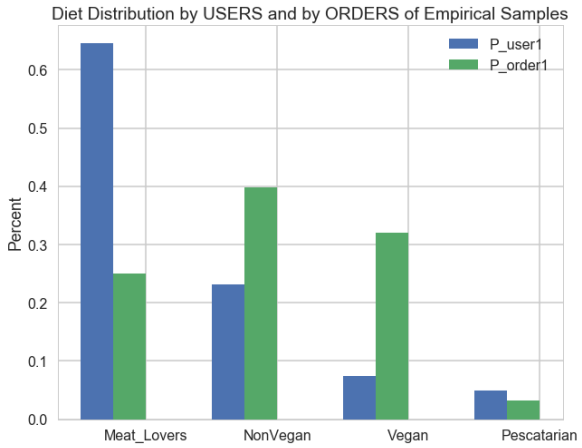


Table 7. P- values for difference in Diet Distribution by USERS and by ORDERS

Diet Preference	p_values
Meat Lovers	0.00
NonVegan	1.84×10^{-210}
Vegan	1.42×10^{-296}
Pescatarian	5.95×10^{-3}

HYPOTHESIS TESTING PART II.

I also used hypothesis testing to compare if calculated dietary distribution by USERS and by ORDERS are not just because of RANDOM chance so resampled by creating a simulated dataframe "Simu1" using aisle probability and number of products in each orders in "E1" dataframe.

A. Dietary distribution by USERS using Empirical vs Simulated dataset

Ho: There is no significant difference between Dietary distribution by USERS using Empirical dataset vs Simulated dataset

H1: There is no significant difference between Dietary distribution by USERS using Empirical dataset vs Simulated dataset

Looking at Figure 6, we can see quite a difference in percentages of the four diet preferences but again we need to calculate p values to be sure. P-values in Table 8 are all less than 0.05 which means we can again reject Ho and accept **H1:** There is no significant difference between Dietary distribution by USERS using Empirical dataset vs Simulated dataset. This also means that dietary distribution by USERS in the empirical dataset is not just random and customers really have diet preference.

Figure 6. Dietary Distribution by USERS between Empirical and Simulated datasets

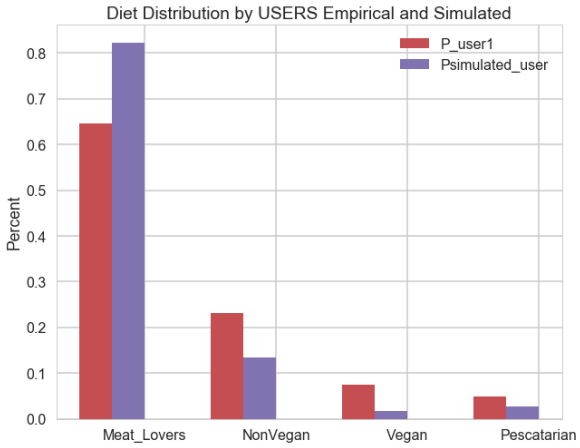


Table 8. P-values of difference in Diet Preference by USERS between Empirical and Simulated datasets

Diet Preference	p_values
Meat Lovers	0.00
NonVegan	4.94×10^{-39}
Vegan	1.11×10^{-12}
Pescatarian	6.80×10^{-3}

B. Dietary distribution by ORDERS using Empirical vs Simulated dataset

Ho: There is no significant difference between Dietary distribution by ORDERS using Empirical dataset vs Simulated dataset

H1: There is no significant difference between Dietary distribution by ORDERS using Empirical dataset vs Simulated dataset

Figure 7. Dietary Distribution by ORDERS between Empirical and Simulated datasets

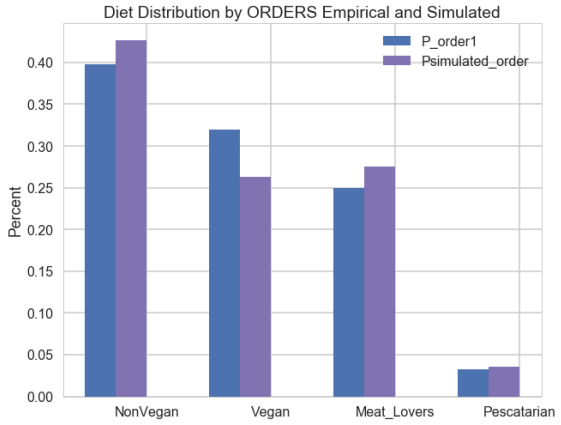


Table 9. P-values of difference in Diet Preference by ORDERS between Empirical and Simulated datasets

Diet Preference	p_values
Meat Lovers	8.44×10^{-47}
NonVegan	1.37×10^{-70}
Vegan	3.84×10^{-232}
Pescatarian	1.40×10^{-1}

Looking at Figure 7, we can see quite a difference in percentages of the diet preferences except for Pescatarian. P-values in Table 9 are all less than 0.05 except for Pescatarian which means for all other dietary preference we can reject Ho and accept **H1:** There is no significant difference between Dietary distribution by ORDERS using Empirical dataset vs. Simulated dataset. However, for Pescatarian we cannot reject Ho.

Calculation of P(Diet|Aisle) and P(Aisle|Diet) using Bayes Theorem

I used Bayes theorem in calculating what is the probability that customer has diet preference X given that they bought from aisle Y (P(Diet|Aisle)) or what is the probability that customer bought from aisle Y given that he has diet preference X.

It is interesting to see that P(Diet|Aisle) across each aisle is very similar to P(diet preference) by USERS. Meat Lovers are at top across the aisles and least is Pescatarians.

Figure 8. P(Diet|Aisle) vs aisle_id

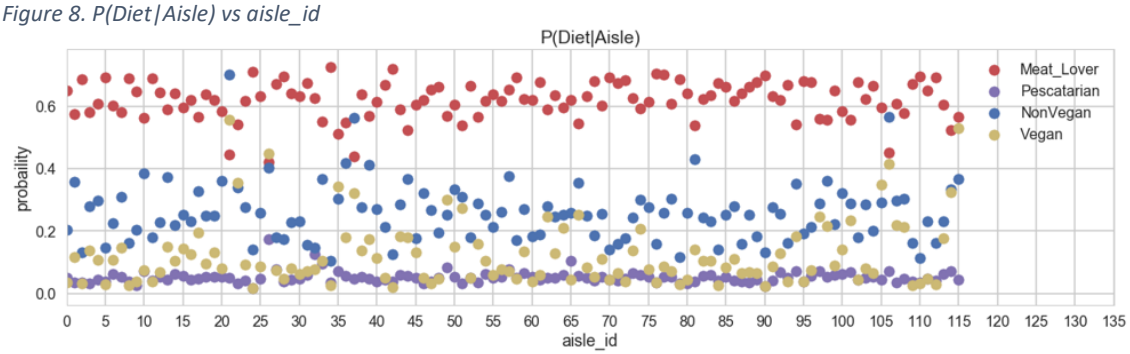


Figure 9. P(Aisle|Diet) vs aisle_id

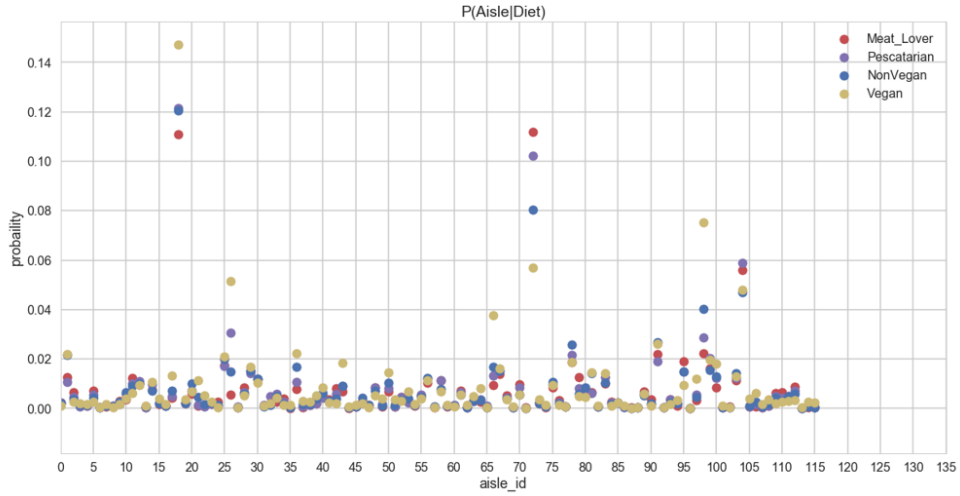


Figure 9 distinguishes P(Aisle|Diet) for the four different diet preferences across some aisles. Example aisle 18: bulk dried fruits and vegetable aisle, P(Aisle|Diet) is highest for Vegans>NonVegans> Meat Lovers>Pescatarians. Another example aisle 72: condiments aisle, P(Aisle|Diet) is highest for Meat Lovers>Pescatarian>NonVegan>Vegan.

GOAL 2: Predicting Customer's Next Order List

Now let's go to the final goal which is predicting Customer's next order list. In this section, I calculated average products per order per user and used this as the number of products on the predicted next order. I also calculated probabilities of product being ordered specific to each users according to their order history and employed this probability in generating customers next order list using the np.random.choice(a=product_id, size=n, p=probability).

After generating predicted next order list, I compared this against the alltrain dataset which has the last order of users and calculated for cosine similarity and f1 score.

Table 8. Predicted and Actual Next Order List of five users with cosine similarity and F1 score

	user_id	product_id	product_id_train	cosine_similarity_score	F1_score
0	27	33370 10312 27966 33736 42987 30776 18523 1733...	16290 16290 16290 16290 16290 16290 16290 1629...	0.268010	0.078603
1	34	49248 7054 34310 26751 25146 38273	25146 3957 3957 3957 33731 7054 7054 15604 156...	0.481543	0.375000
2	44	28378 38741 27156 6617 23341 21616 18653 47962...	47141 1263 43632 26172 26172 23341 26505 44560...	0.713746	0.571429
3	62	8518 21137 16797 47209 26209 37947 30463 24852	45223 26620 26620 26620 31618 12745 37947 3794...	0.433861	0.262295
4	64	44848 18811 21903 36772 5473 47209 8022 9662 1...	18811 18811 18811 18811 25146 49052 21475 1494...	0.602970	0.394366

For a sample of 30,000 users the average cosine similarity is 53% while average F1 score is 35%.

V. Importance of Customer Segmentation by Dietary Preference

Knowing customer's dietary preference will help pair the right products to recommend to the right users and increase the chance of it being bought. For example, a new meat product will have higher probability of being bought by Meat Lovers compared to Vegans. Another example is a new fish scaler will have higher probability of being bought by both Meat Lovers and Pescatarians.

Email promotions can also be tailored according to dietary preferences for higher chance of being availed.

Therefore, it is important to customized user experience by looking at their preferences and order histories to for more effective marketing strategy and increased sales.

VI. Importance of Predicting Customer's Next Order List

Having an accurate recommendation system will increase online sales up to 29% per year just like that of Amazons (2) simply because they work and about 35% of online sales can be attributed to it (4). It takes out the hassle of planning what to buy for each customer and provides an environment where they can just modify or approve the predicted list for reordering.

VII. Recommendations

1. Test the predicted next order list and measure cosine similarity and F1 score. These scores are calculated by using the last order of customers and not recommending the predicted products to customers.
2. Use the Dietary Preference by USERS in marketing promotions and emails and measure if sales increase or decrease.

VIII. Limitations of the Algorithm

This algorithm has some limitations and can still be improved. In predicting dietary preferences Vegetarians who purchase meat occasionally for a visitor are still identified as Meat Lovers. There could be some noise in the model that can be improved by using unsupervised machine learning such as K Means clustering, DBSCAN or other segmentation and clustering method.

The algorithm for predicting next order list can also be improved with higher cosine similarities and f1 score by using recommendation system libraries in python such as PySpark and Python-Recsys.

IX. References

1. A. Sharma, J.M. Hofman, D.J. Watts, "Estimatingthe Causal Impact of Recommendation Systems fromObservational Data," *Proc. 16th ACM Conf. Economicsand Computation*, 2015, pp. 453–470.
2. http://fortune.com/2012/07/30/amazons-recommendation-secret/
3. https://www.forbes.com/sites/briansolomon/2015/01/21/americas-most-promising-company-instacart-the-2-billion-grocery-delivery-app/#2756ea8242dc
4. https://www.martechadvisor.com/articles/customer-experience/recommendation-engines-how-amazon-and-netflix-are-winning-the-personalization-battle/

X. IPython Notebook

https://github.com/DrAugustine1981/InstacartCapstoneProject/blob/master/CapstoneProject1.ipynb