

Instacart Market Basket Analysis Project

Pamela Ubaldo Augustine

Instacart, #1 in Forbes most promising company list in 2015, is conducting an Instacart Market Basket Analysis Kaggle competition to predict products that existing customers will purchase again. This capstone project aims to 1) identify customer if meat eater, pescatarian, vegetarian, vegan, 2) identify customer if they are weekly, bimonthly or monthly buyers and 3) predict customer's next order list. Below is a list of questions that will help achieve these three goals.

QUESTIONS TO ANSWER USING THE DATA

1. How many products?
2. How many aisles?
3. How many department?
4. How many customers?
5. How many total orders?
6. Are there missing data? What type of missing data?
7. When are the peak hours (orders>100,000)? When is the orders highest and lowest?
8. What day of the week has the highest and lowest order volume?
9. What is the probability of each product being ordered?
10. What is the probability of each department being ordered from?
11. What is the probability of each aisle being ordered from?
12. Can I identify meat eaters, vegetarian, vegan their percentage in the entire customer list?
13. What is the probability of customers being meat eater, vegetarian or vegan
14. From what aisle will vegan usually get from?
15. What products appear in all customer A orders? – These products will have high probability being reordered by customer A
16. How many orders for each customer?
17. What is the average number of products for across all orders for each customer?
18. Can I Identify if the customer is a daily, weekly, bimonthly or monthly buyer?
19. What is the probability that customer is daily, weekly, bimonthly or monthly buyer
20. Is there a pattern of what day customer buy from Instacart?
21. Using average number of product per order for each customer, and probability of product to be reordered by customer, can I predict products that will be reordered by customer?
22. Can I predict the right which of the predicted products to reorder will be added to cart from 1st to last?

Data Wrangling

Data was acquired from <https://www.instacart.com/datasets/grocery-shopping-2017> and also available in <https://www.kaggle.com/c/instacart-market-basket-analysis/data>. The data is composed of seven csv files namely aisles, departments, order_products_prior, order_products_train, orders, products and sample submission files.

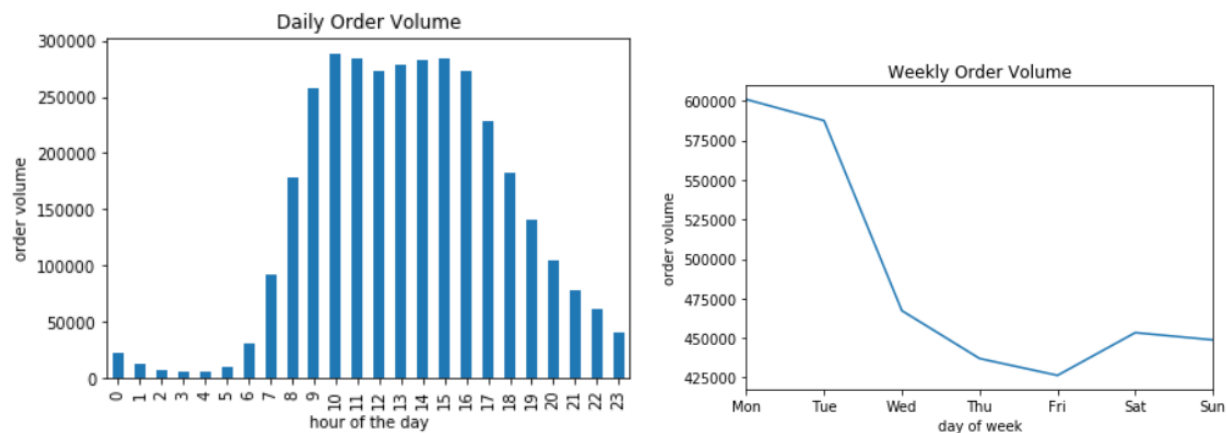
All the seven csv files were read using python pandas and stored in a data frame for data analysis. Among all seven only orders data frame has 206, 209 missing data in the “day_since_prior_order” column. There are a total of 206, 209 customers in the data set and the missing data or NaN in the “day_since_prior_order” represents the first time customers purchased from Instacart. This is an example of Missing At Random (MAR) data mechanism where missing data is not related to missing data but observed data.

After analyzing missing data, I separated order data frame into three categories namely “prior, train and test” in the “eval_set” column. These are saved into three data frames namely oprior , otrain and otest. The otrain data frame is then merged with the order_products_train, products, departments and aisles data frames. The merged data frames are saved to a new data frame named Alltrain which is now ready for data analysis.

The same was done with oprior dataframe which is merged with order_products_prior, products, department and aisles data frames which resulted to data frame named Allprior which is ready for data analysis.

Inferential Statistics

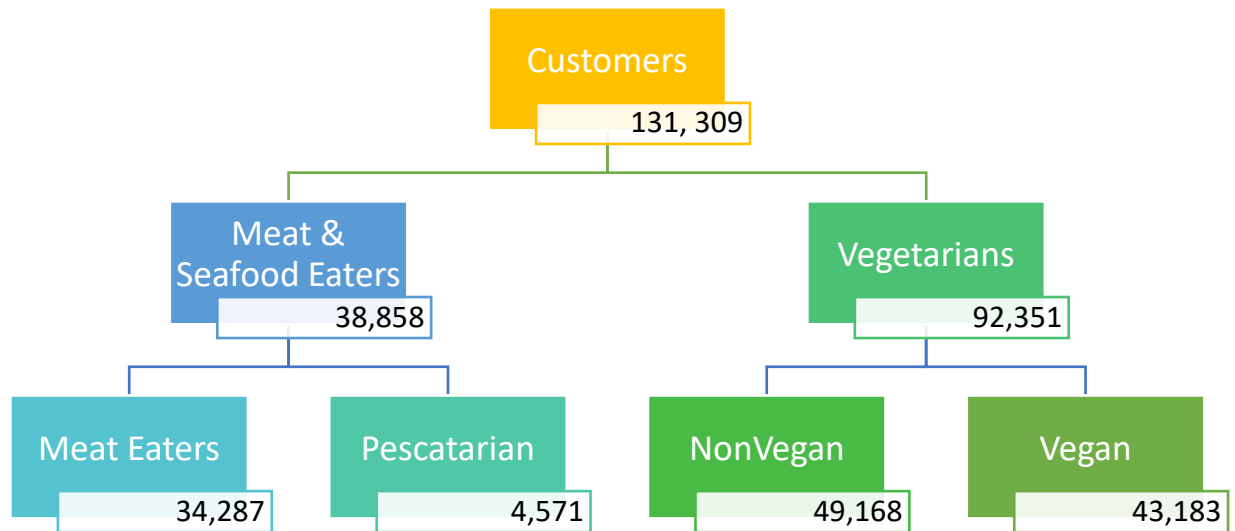
Each data frames where inspected using python pandas that helped answer questions in the list above. Results show that there are a total of 49, 688 total products, 134 total aisle, 21 departments, 206, 209 customers, and 3,421,083 total orders. Alltrain, the training data set was used in creating the model. Alltrain data shows that 8am-10pm are peak hours where order volume is greater than 100,000. Order volumes are at maximum at 10am and at minimum at 3am.



During the week order volume is at highest on Mondays, declines as weekend comes and a lowest on Fridays.

Probabilities of being ordered from each product, department and aisle were calculated. In addition, customers present in the Alltrain data frame were categorized into Meat Eaters, Pescatarian, Vegan and

NonVegan (Goal1).



“Meat & seafood eaters” are users that ordered from the following aisles. Meat Eaters were users that ordered from the unhighlighted aisles below while Pescatarian are users that ordered from the blue highlighted aisles below given that these users are not present in the “Meat Eaters” users.

Meat& Seafood Eaters Table

aisle_id	aisle
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

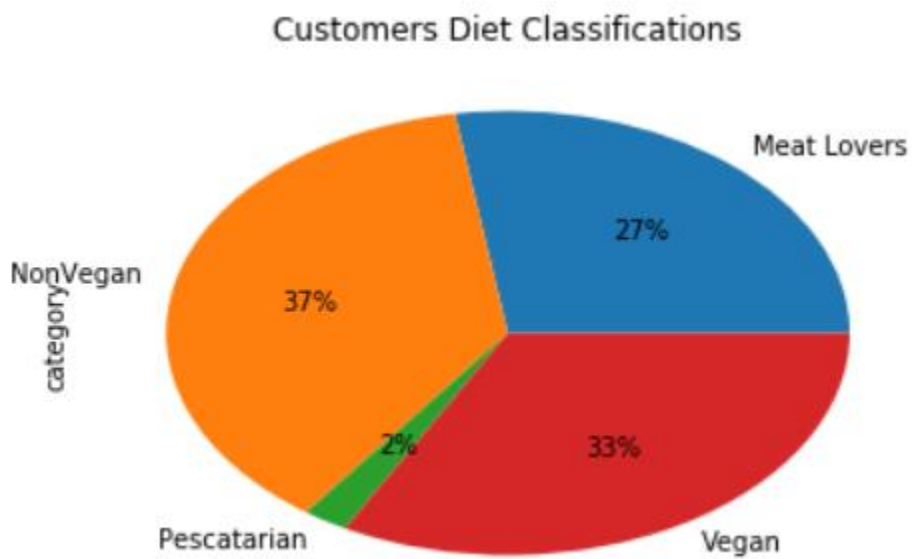
The Vegetarian groups were identified using the process of elimination. Users present in “Meat & Seafood Eaters” were eliminated in the Alltrain data frame where all user_id were present. Aisles where Vegan possibly will not order from were identified below. However, all these aisles except for the egg aisle were inspected for vegan alternatives.

In the yogurt, cheese, milk and cream aisles, products that are non-dairy or made of almond, soy, cashew, coconut, hemp, rice, and other vegan alternatives were eliminated and whatever is left in the data frame has only “NonVegan” customers. Vegan customers are then identified by eliminating “NonVegan” customers on the data frame containing all “Vegetarian” customers.

Aisles Vegan Will not Eat From Table

aisle_id	aisle
2	specialty cheeses
21	packaged cheese
53	cream
84	milk
86	eggs
108	other creams cheeses
120	yogurt

Percentage of meat eaters, pescatarian, nonvegan and vegan is illustrated below using a pie chart.



This model will then be cross validated using “prior” and “test” evaluation sets.

IPython NOTEBOOK SOLUTIONS

All solutions can be viewed in IPython Notebook in my github below.

<https://github.com/DrAugustine1981/InstacartCapstoneProject/blob/master/EDA.ipynb>

