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 predict Instacart customer's next order list. But before doing that below is a list of questions that I want to answer using the data.

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. What is the probability of each product being ordered?
7. What is the probability of each department being ordered from?
8. What is the probability of each aisle being ordered from?
9. Can I identify meat eaters, vegetarian, vegan their percentage in the entire customer list?
10. What is the probability of customers being meat eater, vegetarian or vegan
11. From what aisle will vegan usually get from?
12. What products appear in all customer A orders? – These products will have high probability being reordered by customer A
13. How many orders for each customer?
14. What is the average number of products for across all orders for each customer?
15. Can I Identify if the customer is a daily, weekly, bimonthly or monthly buyer?
16. What is the probability that customer is daily, weekly, bimonthly or monthly buyer
17. Is there a pattern of what day customer buy from Instacart?
18. What products are bought the most at each hour of the day
19. Can I identify shoppers with children or babies? Or pregnant?
20. Can I identify if shopper is male of female?
21. Can I identify if customer is single or married?
22. Can I identify if customer has a pet? Dog or cat?
23. Can I identify ethnicity of customer? Asian, Indian, Chinese, Italian, American, International?
24. Can I identify if customer is a healthy conscious, mediocre eater or not so health conscious?
25. Can I identify if customer has intolerance with gluten and lactose?
26. 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?
27. Can I predict the right which of the predicted products to reorder will be added to cart from 1st to last?
28. Can I predict a frequent, moderate, occasional alcohol drinker or nonalcoholic drinker?
29. Can I predict age group of customers?
30. Can I predict the season /months order was made -winter, spring, summer, fall?
31. Can I predict if customer is diabetic? Has arthritis? Has allergies? Digestive problems?

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.