Instacart Market Analysis Using the Machine Learning Tools

Don Kim 500486770
don.kim@ryerson.ca
CKME136
Ryerson University, July 2020

Ryerson University

Table of Contents

Introduction

Online shopping is a rapidly growing market and the approach to the online shopping market is different from offline markets. Successful market strategies are widely available for traditional shopping; however, online shopping businesses are a very recent movement in the industry. Thus, new analysis is required to make online shopping successful. Thanks to Instacart, I am able to conduct the research on how the online grocery shopping works based on the open data set that is available for us to use academically. There are more than 3 million orders and 50K products on data and these are big enough dataset to find insights of online grocery business.

Research questions are following:

- What are relationships of the products to develop product recommendation system? (association rule)
- What behaviors do customers have in each different segment? (K mean clustering)
- How attributes affect the product to be reordered? (classification)

Product recommendation can be done much easily online and it is the key to attract customers' attention. There are 49688 products and the pattern of buying products can be found using association rules. There are different types of customers: active customers, repeat customers or lapsed customers etc. These patterns can be found using K mean clustering. And some of the products are repurchased and some of them are not. It is important to find the characteristics of reordered items and chances that the product would be reordered.

Ability to answer the research questions can help to guide online shopping businesses to be successful.

Code: https://github.com/donkimc/Instacart_capstone

Literature Review

Many research papers related to online shopping have been released from institutes and I have selected some of the research papers to review the background information and interpretations to find a current online shopping market trend and insights.

In the study of Benn, Yael (2015), the experiment was conducted on online grocery shoppers how they find the product through the website. Each participant's eye movement was examined while they searched through the mock website. The result is over 90% of online shoppers use department categories, 80% of them use a search tab, and 68% use recommended products on the site. It shows that online grocery shoppers have the same pattern as traditional offline shoppers that they find the product through aisle and department.

In the study of Bauerova (2018), there are various factors influencing purchasing groceries in online stores. The total amount of order and frequency of buying has a weak negative correlation. And the key factors in decision-making on online grocery shopping is the delivery time and fee.

Singh, Reema & Soderlund, Magnus (2019) examine many types of customer experiences affect the overall online grocery shopping experiences. And that leads to whether customers would reorder the product. The types of experience are customer service, website experience, product experience, delivery experience, and brand experience. Each feature has a coefficient and a satisfaction rate can be calculated.

In the study of Singh, Reema (2018), online customer experience (OCE) factors influence the customer to come back to an online grocery store to repurchase or switch to another online retail store. Customers always want to get the best experience out of online shopping. Experience rewards are varied from service efficiency to visual appearance on a product.

Fredette, Marc (2018) analyzed that online grocery shopping is very different from other online retail shopping since it involves a variety of products in a single cart than similar types of products in a cart. (such as books or electronics) And online grocery shoppers tend to look for convenience and time-saving. Customers look for the item by searching categories more than searching keywords. There are more perishable goods that require measurement. Measurement of products demands cognitive load on customers; therefore, it would have a negative impact on the decision making of products.

In study Anesbury, Zachary (2016), it is finding the time duration of selecting products and page views in online grocery shopping. Shampoo took the longest time to be purchased and banana or milk took the least amount of time to be purchased. Page view has a similar result. It concluded that this pattern is very similar to the offline shopping experience.

Timothy J. Richards (2017) conducted an experiment that long-tail retail strategies are applied to online grocery shopping. Niche items such as sauces or condiments yield long-tail effects. There, SKU reduction would be effective to produce this effect. Hugh online grocery stores are focusing on supply-chain efficiency, but they should also consider the long-tail effect on online grocery, thus adapting niche items.

In the study of Piroth, Philipp et al. (2020), the German online grocery market is expanding at a pace and the paper is finding the factors that lead to success in the online market. Logistics is the most important factor and it should work flawlessly in order to deliver the product on time and the product's availability is always kept with the customer's demand.

In the study of Khalifa (2007), customer retention is the major factor that businesses can be successful. The satisfaction of a product or service is used to be the measurement to make a business thrive. But when it comes to online shopping, habit is also the factor that influences customer retention. All of these factors have been considered to indicate the rate of customer retention in the online market.

In the study of Pauzi (2017), many factors influence customers' intention for online grocery shopping, and the list of them is Social influences, Facilitating conditions, hedonic motivations, perceived risk, and trust. Customers are not only purchasing goods for necessity but also, they get the pleasure of shopping. It applies to online shopping.

Mackenzie, Adiran (2018) examines the personalized recommendations are implemented using apriori conditional probabilities. There are many combinations of possibilities to combine products and find a correlation between the products gives the recommendation. For instance, rice and sauces are a strong correlation. This recommendation service is a unique feature since big data is available.

Dataset

The dataset is available from the Instacart website (The Instacart Online, 2020) and this contains 2017 online shopping order data for academic purposes. There are more than 3 million grocery orders and it is sufficient data to find insights and answer the research questions. 5 CSV files are provided and table 1 is the details of files. Aisles and department information are available in addition. In order products, two parts are separated: prior and train. This CSV data set has the same attributes but they are separated due to other purposes of the training dataset. However, the merged dataset of two files is used in this project.

Table 1. CSV data files

File	*attributes*	*instances*	*File description*
aisles.csv	2	134	Aisles ID and name
Departments.csv	2	21	Department ID and name
order_products_prior.csv	4	32434489	Each order in cart with every product item. (Major dataset of order)
order_products_train.csv	4	1384617	Each order in cart with every product item. (training dataset of order)
orders.csv	7	3421083	Each order with date and time attributes
products.csv	4	49688	Product ID and name with its location in aisle and department

Table 2. shows that all the features of each CSV file. Orders ID and product ID are starting from 1. Cart order in ‘order_products_train’ can be an infinite number but max cart order can be found to be 140. Days_since_prior_order attribute in ‘order.csv’ has a range from 0 to 30. There is a case that the product can be ordered for more than 30 days. However, anything over 30 days is located under 30.

Table 2. Features and description

File	*Field Name*	*Field description*	*Data type*
aisles.csv	Aisle_id	Aisle ID	Int64
	Aisle	Name of Aisle of product	Object (str)
departments.csv	Department_id	From 0 to 20	Int64
	Department:	Name of department	Object(str)
order_products_prior.csv	Order_id	Order ID (From 1)	Int64
	Product_id	Product ID	Int64
	Add_to_cart_order	Max card order is 140.	Int64
	Reordered	(0 or 1): 0 - not reordered product / 1 - reordered product	Int64
order_products_train.csv	Order_id	Order ID (From 1)	Int64
	Product_id	Product ID	Int64
	Add_to_cart_order	Max cart order is 140	Int64
	Reordered	(0 or 1): 0 - not reordered product / 1 - reordered product	Int64
orders.csv	Order_id	Order ID	Int64
	User_id	User ID	Int64
	Eval_set:	Categorical; Priori or train	object
	Order_number	range from 1 to 100	Int64
	Order_dow (day of week):	range from 0 to 6. 0 is Sunday, 6 is Saturday	Int64
	Order_hour_of_day:	Range from 0 to 23 (24 hours)	Int64
	Days_since_prior_order	range from 0 to 30. (30 days and over are all under 30)	Float64
products.csv	Product_id	Product ID	Int64
	Product_name	Name of product	object
	Aisle_id	Aisle ID (Total 134)	Int64
	Department_id	Department ID (Total 20)	Int64

Cleaning Data

Table 3. is the summary statistics of the selected attributes. Any instances in the attributes cannot be eliminated since they are ordinal values unless they are null values. Null data is found in days_since_prior_order attribute under order.csv and it is 0.58% of the total order dataset instances. Therefore, the null data is removed since it would not influence significantly. Considering replacing with mean is not practical since the data type is an integer and it has a range of 0 to 30 but mean gives a floating number.

Table 3. Summary statistics of the selected attributes

Attributes	*Mean*	*Median*	*Std*	*25%*	*50%*	*75%*
Add_to cart_order	8.35	6	7.13	3	6	11
Order_number	17.15	11	17.73	5	11	23
Order_dow	2.77	3	2.05	1	3	5
Order_hour_of_day	13.45	13	4.23	10	13	16
Days_since_prior_order	11.11	7	9.2	4	7	15

Data exploration

Online grocery shopping is available 24 hours a day and 7 days a week; therefore, customers can order the product any time of day whenever they have access to the internet. Figure 1 shows the demand for ordering products on the scale of hours. Figure 1. Shows the total order number throughout the week. Order_hour_of_day attributes didn't specify when the start of the day is. The assumption is required that what day the week starts. In figure 1, blue and pink lines have a pick at 14 hours and the rest have a pick at 10 hours. This indicates that blue and pink lines are the weekend and the rest is on a weekday. The starting week is set to be on Sunday and it has a numerical value 0.

In Figure 1, there is a surging demand on Sunday and Monday. It is due to the workers are going back to work on Monday and the demand for online shopping increases. (Tuttle, 2012) In Figure 2, the plot shows the camel graph that there are picks on 10 and 14 hours. The shopping demand is low during the night until morning.

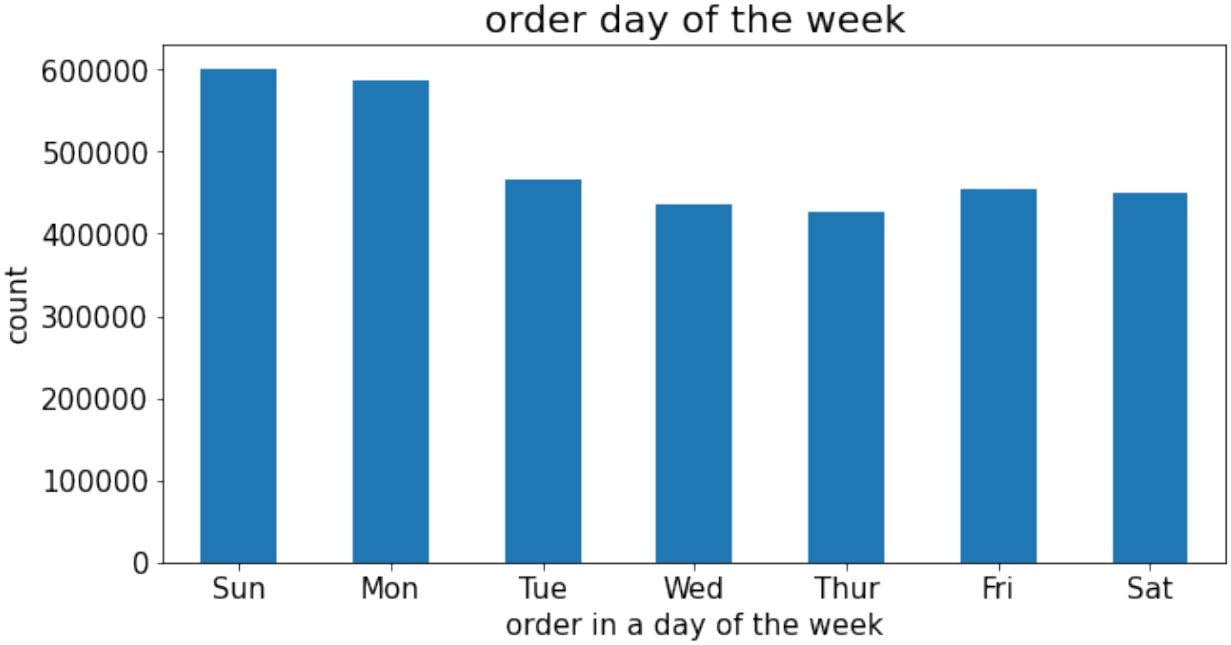


Figure 1. Total order in a week

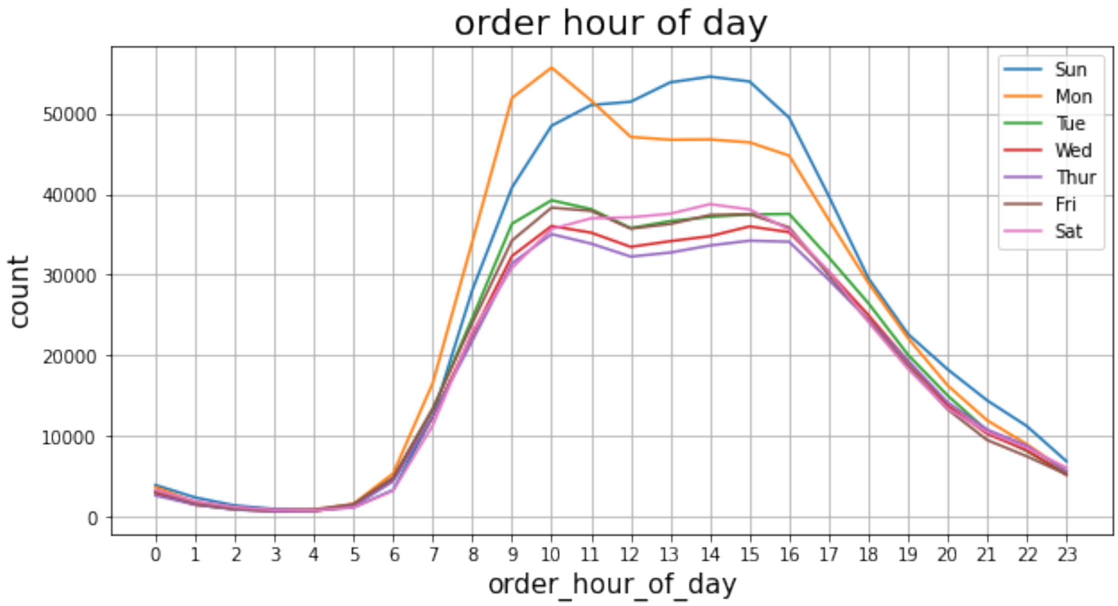


Figure 2. Order hours of day, from 0 to 24 hours on Monday to Sunday

Customers' orders from online grocery shopping in frequencies and Figure 3 show that the duration of days when customers are coming back to shop again. Most of the customers are coming back to the online shopping after 7 days. In the plot, the data doesn't have information after 30 days. It collects the counts in 30 if the customers are coming back after 30 days.

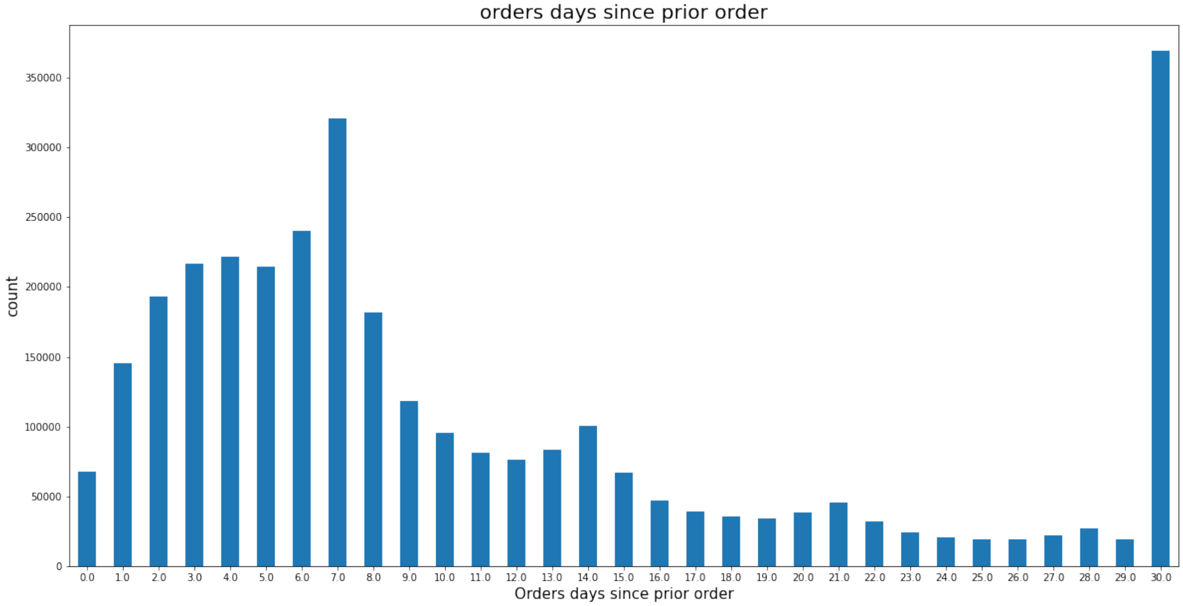


Figure 3. Order days since the order

Products are reordered again and Figure 4 shows that the ratio of the reordered product. 1 represents the reordered and 0 is not. 59% of products are reordered and 41% of products are newly ordered products.

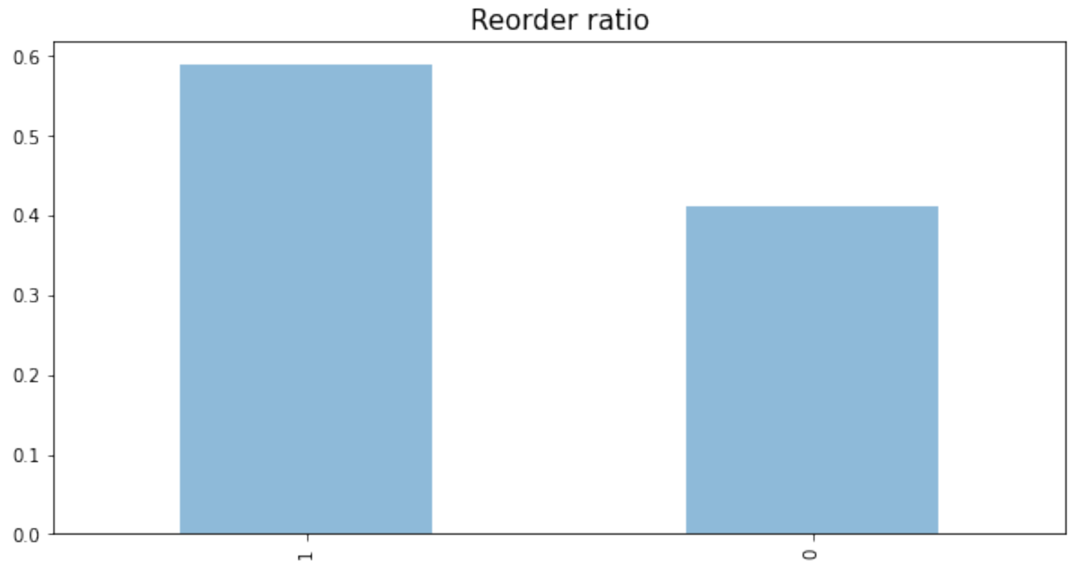


Figure 4. Reordered product ratio

Customers can add as little as one item and max number of products in cart goes up to 145 items. Figure 5 shows the plot of the number of orders in the cart. The median size of the cart is 8. It also shows that 25% and 75% percentile of customers are adding the products in the cart from 5 to 14 items.

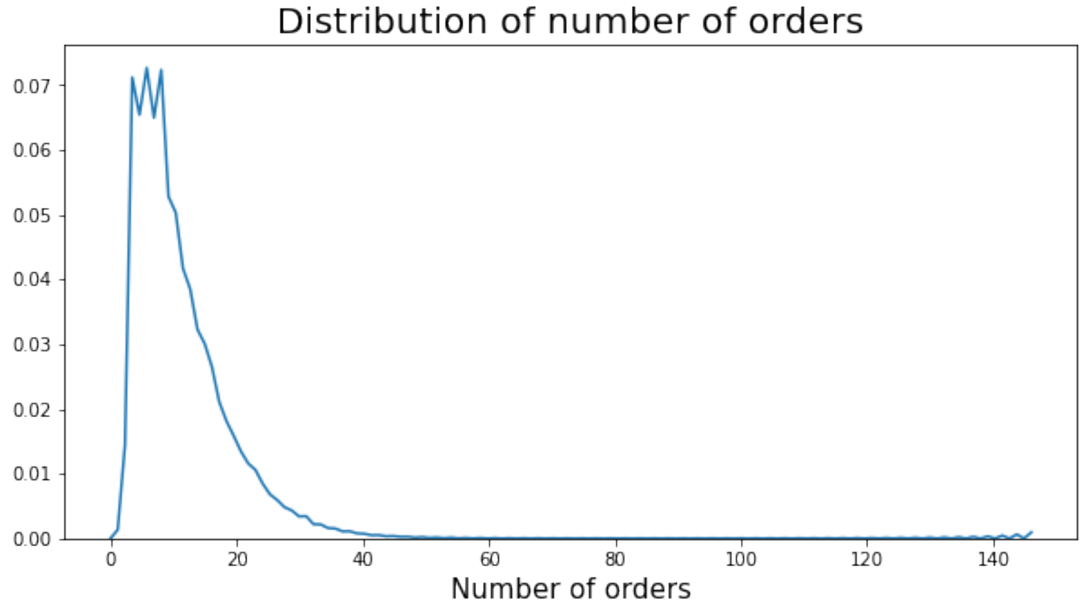


Figure 5. Distribution of number of orders

Customer retention is the key factor in the success of business and customer in Instacart is coming back at least 3 times according to Figure 6. After 4 number of orders, it is decreased gradually.

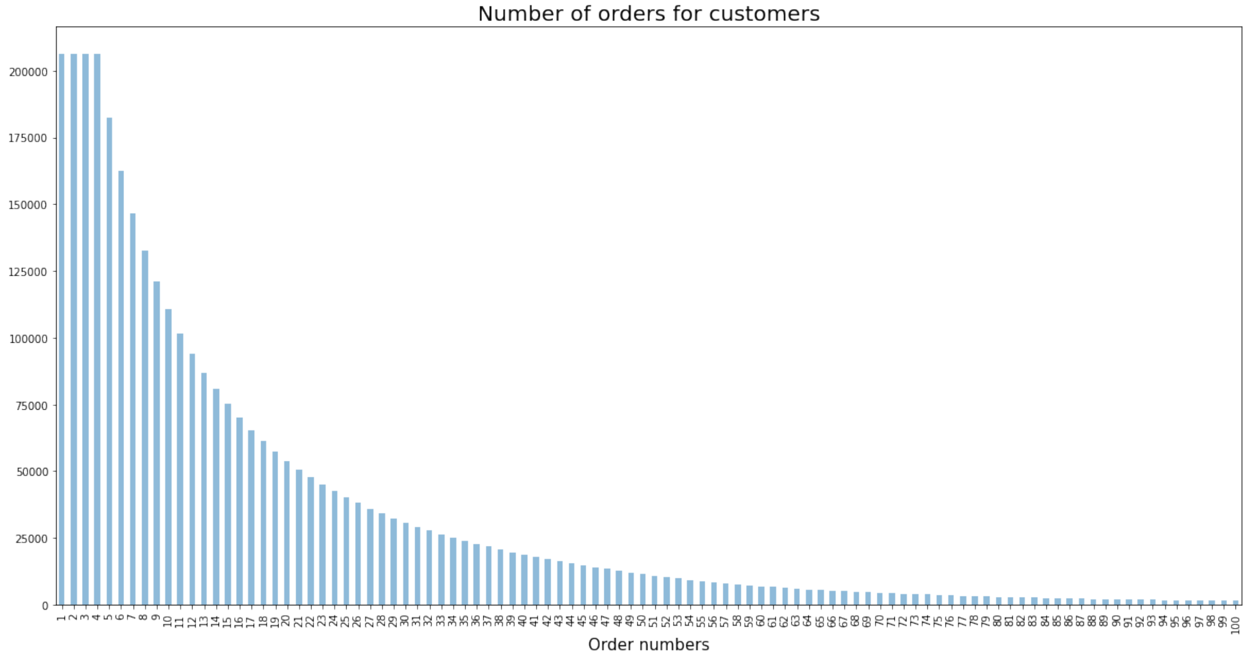


Figure 6. Number of orders for customers

There are 21 departments in Instacart online grocery shopping. The majority purchased goods are from produce, dairy eggs, snacks, and beverages in Figure 7. The top-selling products are shown in Figure 8. It is a mostly perishable item and there is a high demand for Banana from customers.

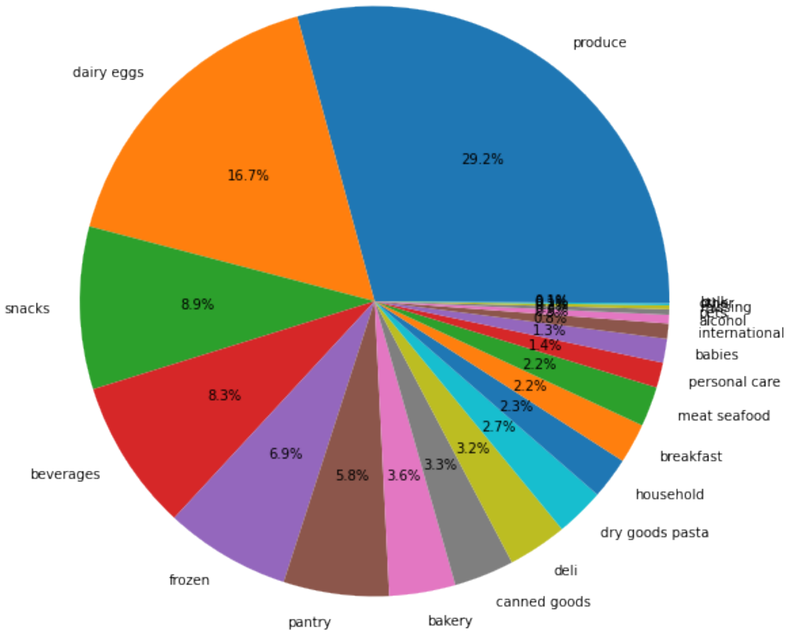


Figure 7. Department distribution

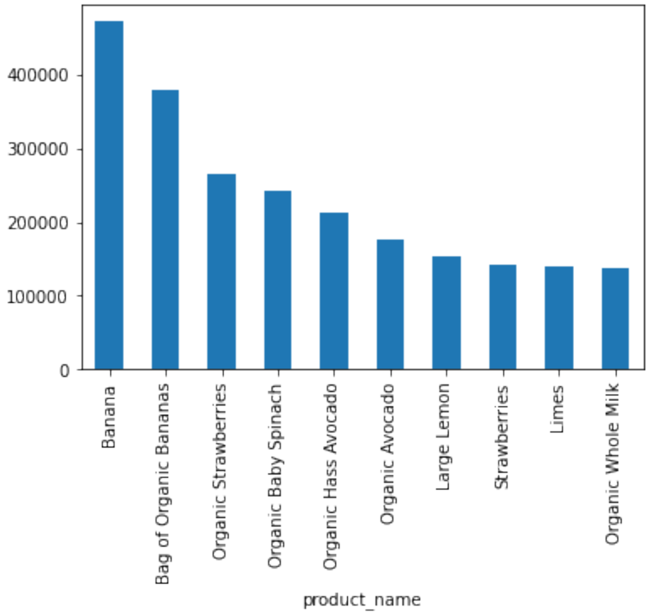


Figure 8. Top 10 most selling products

Spearman correlation is conducted on the dataset to find the relationship. Reordered and add_to_cart_roder attributes have a negative weak correlation and the coefficient is -0.133024. The correlation is explained in Table 4.

Table 4. Spearman correlation of selected attributes

	product_id	*aisle_id*	*department_id*	*order_id*	*add_to_cart_order*	*reordered*
product_id	1.000000	0.002254	-0.028503	-0.000082	0.005529	0.003718
aisle_id	0.002254	1.000000	0.062203	-0.000063	0.009451	0.003924
department_id	-0.028503	0.062203	1.000000	-0.000229	0.029437	-0.039371
order_id	-0.000082	-0.000063	-0.000229	1.000000	-0.000320	-0.000253
add_to_cart_order	0.005529	0.009451	0.029437	-0.000320	1.000000	-0.133024

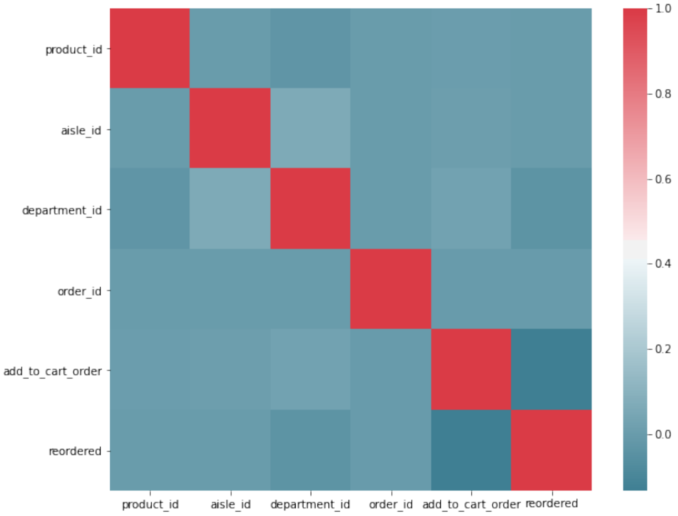
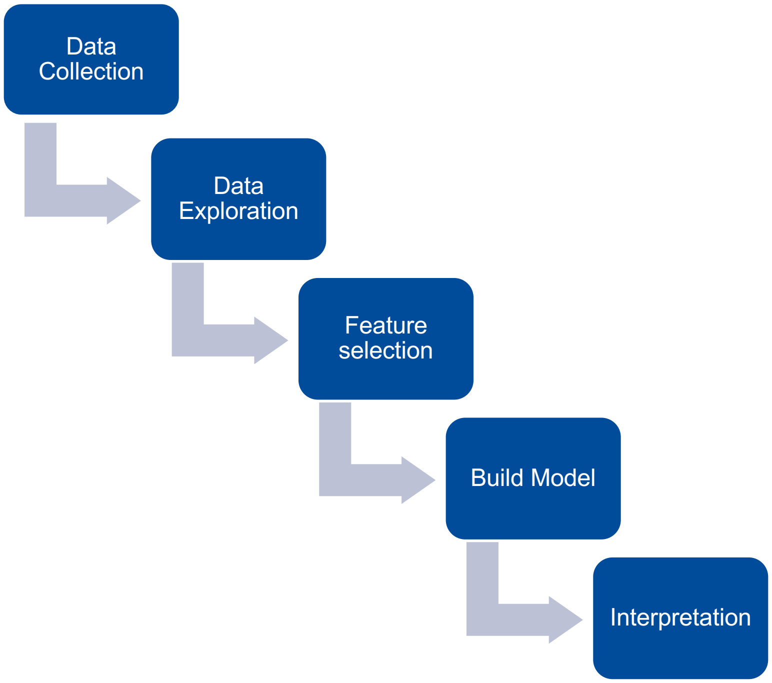


Figure 9. Correlation visualization

Approach

5 steps approach is adopted in this project. Each step is clearly explained below.



Step 1: Data Collection

Define research questions
Research questions are defined and what possible machine learning methods can be used to solve the research questions are studied. The RQ is related to online grocery shopping carts and research on market analysis and customer segmentation.
Data source selection
Instacart provides the open dataset available to download for academic purposes. It is a big data that there are more than 3 million orders. It is enough to conduct the machine learning to find the pattern on orders and customer.

Step 2: Data Exploration

Data review and cleaning
Attributes and instances are reviewed. Any possible null value and outlier are removed or replaced with mean. Add_to_cart_order attribute has a null value and they are removed since it is a small portion. Data descriptions are examined to find patterns.

Data visualization and exploration
The number of orders during week and day are plotted to see the pattern of customer purchase. Sunday and Monday have the highest order number of buying groceries. The average cart size is 8 and the most popular product is Banana. Product and daily are the big 2 departments in Instacart.

Step 3: Feature selection

Identify key attributes and feature elimination
In this step, key attributes are found to be used in the model's independent variables or for the machine learning process. A feature elimination technique will be used to find attributes. Any attributes that have low variation will be removed.

Step 4: Build Model

Approaches
Association rules or Apriori algorithms are used to implement product recommendations. Attributes in Orders and Order_product tables are exploited to find a relation to each product. K mean clustering can find the customer segmentation. (Kim et al. 2007) Decision tree, Naïve Bayes, and Logistic regression techniques are performed and their performances are compared to see the prediction of reordered products.

Evaluation
Classification algorithms are conducted, and its performance measurements are compared. And the stability of algorithms is examined.
Improvement
Exploration of possible improvements in algorithms will be discussed in this stage. And other python packages will be tested to see if there is any improvement.

Step 5: Interpretation

Once all the satisfactory output is concluded and the research questions can be explained. And another application of this model is reviewed. A final report will be written based on the output.

Results

Machine models are built to answer research questions: Product recommendation, customer segmentation, and reordered item prediction. As mentioned in the introduction, many prediction models are tried to predict the best results.

Product recommendation

Apriori association rule is applied to predict a product recommendation. The followings are the formula for support, confidence, and lift. The lift should be larger than 1 to have a positive association. If the lift score is less than 1, it is a negative association between items which means product B in the cart is discouraging product A to be added in cart.

$$\text{Support}(\{A\}{-}{>}\ \{B\}) = \frac{\text{Transaction containing both A and B}}{\text{Total number of transactions}}$$

$$\text{Confidence}(\{A\}\ {-}{>}\ \{B\})\ = \frac{\text{Transaction containing both A and B}}{\text{Transactions containing B}}$$

$$\text{Lift}(\{A\}\ {-}{>}\ \{B\})\ = \frac{\text{Confidence}(\{A\}\ {-}{>}\ \{B\})}{\text{Fraction of transactions containing B}}$$

Three sets of parameters are used on the test. This is the first test parameter. Here are the parameters used for association rule.

- Minimum length: 2
- Minimum support is 0.01
- Minimum confidence is 0.1
- Minimum lift is 2

This setup is for only popular items since support is set to 1%. It gives 7 rules based on the prior dataset in Table 5. Most of the items are the top 10 selling items according to Figure 8. When the customer buys the item in A column, he/she is highly to purchase in the item in the B column. The highest confidence is on Banana and Organic Fuji Apple. In the customer's cart, a customer who buying Banana is also buying Organic Fuji Apple and it is 38% chances. The highest lift is on Organic Strawberries and Organic Raspberries. There is a strong association that customer who buying the Organic Strawberries is also buying Organic raspberries. The lift indicator is a better measurement if all the items on the list are popular. Items happen to be purchased together since they are popular items.

Table 5. Association Rule in first scenario

A ->	*B*	*Support*	*Confidence*	*Lift*
Bag of Organic Bananas	*Organic Raspberries*	0.012599	0.106741	2.503775
Bag of Organic Bananas	*Organic Hass Avocado*	0.019391	0.164293	2.472945
Organic Strawberries	*Organic Raspberries*	0.010533	0.127938	3.000973
Organic Strawberries	*Organic Hass Avocado*	0.012689	0.154124	2.319880
Organic Strawberries	*Organic Hass Avocado*	0.012689	0.154124	2.319880
Organic Hass Avocado	*Organic Baby Spinach*	0.010856	0.144266	2.171499
Banana	*Organic Fuji Apple*	0.010558	0.378693	2.576259
Banana	*Organic Avocado*	0.016609	0.112990	2.054395

This is a second scenario. Here are the parameters used for association rule.

- Minimum length: 2
- Minimum support is 0.005
- Minimum confidence is 0.05
- Minimum lift is 3

It gives 6 rules. This setup covers items from unpopular. The lift is set to be 3 to have positive associations. Organic garlic has a strong association with Organic yellow onion since lift is 5.7.

Table 6. Association Rule in second scenario

A ->	*B*	*Support*	*Confidence*	*Lift*
Organic Hass Avocado	*Organic Lemon*	0.006609	0.242131	3.644560
Organic Strawberries	*Organic Raspberries*	0.010533	0.127938	3.000973
Organic Garlic	*Organic Yellow Onion*	0.006866	0.194603	5.698983
Limes	*Organic Cilantro*	0.005464	0.124905	5.775753
Limes	*Large Lemon*	0.008524	0.194863	4.103710
Organic Hass Avocado	*Organic Cucumber*	0.005430	0.217136	3.268339

The third scenario is the following. Here are the parameters used for association rule.

- Minimum length: 2
- Minimum support is 0.001
- Minimum confidence is 0.01
- Minimum lift is 45

This setup is for finding association among unpopular items but very high lift value. There are 12 rules provided and the highest lift is on Icelandic Style Skyr Blueberry Non-fat Yogurt and Non fat raspberry yogurt. Lift value is 75.64

Table 7. Association Rule in third scenario

A ->	B	Support	Confidence	Lift
Total 2% with Strawberry Lowfat Greek Strained...	*Total 2% Lowfat Greek Strained Yogurt With Blu...*	0.002902	0.449513	48.343394
Total 2% Lowfat Greek Strained Yogurt with Peach	*Total 2% Lowfat Greek Strained Yogurt With Blu...*	0.001954	0.302630	48.870721
Total 2% Greek Strained Yogurt with Cherry 5.3 oz	*Total 2% Lowfat Greek Strained Yogurt With Blu...*	0.001646	0.254938	45.421918
Nonfat Icelandic Style Strawberry Yogurt	*Vanilla Skyr Nonfat Yogurt*	0.001212	0.365563	64.780041
Icelandic Style Skyr Blueberry Non-fat Yogurt	*Nonfat Icelandic Style Strawberry Yogurt*	0.001418	0.427553	71.378186
Non Fat Raspberry Yogurt	*Nonfat Icelandic Style Strawberry Yogurt*	0.001189	0.358342	70.344028
Icelandic Style Skyr Blueberry Non-fat Yogurt	*Non Fat Acai & Mixed Berries Yogurt*	0.001221	0.453129	75.647888
Icelandic Style Skyr Blueberry Non-fat Yogurt	*Vanilla Skyr Nonfat Yogurt*	0.002069	0.366718	61.222024
Non Fat Raspberry Yogurt	*Vanilla Skyr Nonfat Yogurt*	0.001585	0.280950	55.151722
Icelandic Style Skyr Blueberry Non-fat Yogurt	*Non Fat Raspberry Yogurt*	0.002247	0.375136	73.640837
Grapefruit Sparkling Water	*Lemon Sparkling Water*	0.001037	0.351201	75.634068
Total 2% with Strawberry Lowfat Greek Strained...	*Total 2% Lowfat Greek Strained Yogurt with Peach*	0.001161	0.179834	72.141821

Customer segmentation

There are patterns on customer's shopping habits and these patterns can be found using K mean clustering. Relevant features are selected and they are related to customer's patterns. 4 features are chosen to be used on K mean clustering:

- Total number of orders
- Average of days since order
- Average size of orders per customer
- Total reordered

The Elbow method is used to find the optimal cluster size. Based on Figure 10, cluster size is 5.

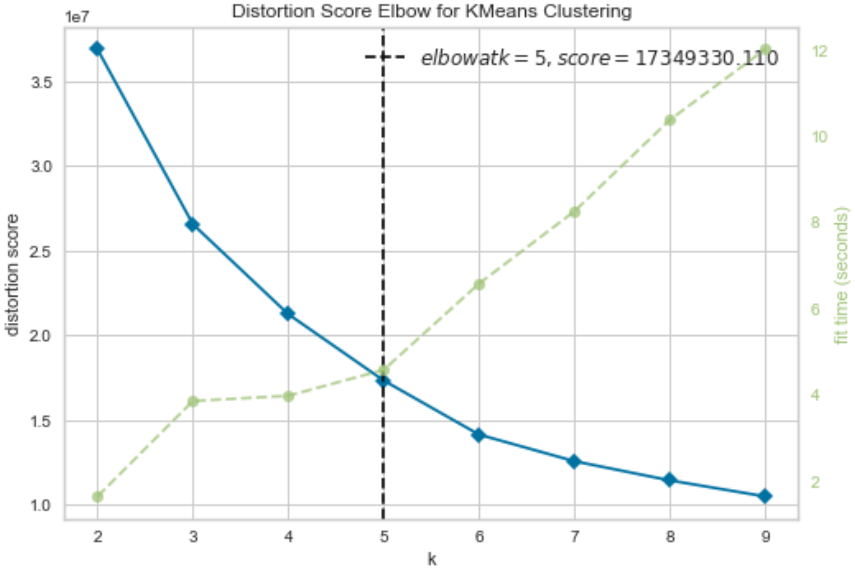


Figure 10. Elbow for KMeans Clustering

After running the KMean clustering, characteristic of 5 clusters are shown in Table 8. Better representation of characteristics is plotting on the graph and it is shown in Figure 11. *Values in total orders and average_days_since_order are disproportional, therefore, average_days_since_order value is omitted in Figure 11.

Table 8. Values in 5 clusters

CLUSTER	*TOTAL_ORDERS*	*AVERAGE_DAYS_SINCE_ORDER*	*AVERAGE_SIZE_OF_ORDERS*	*TOTAL_REORDERED*
1	10.418738	11.793417	7.716277	3.554531
2	34.012697	8.928211	10.364182	7.086662
3	8.985992	16.306226	19.622553	10.298292
4	73.085718	4.704624	9.724681	7.426370
5	4.664466	23.053657	7.733388	2.693712

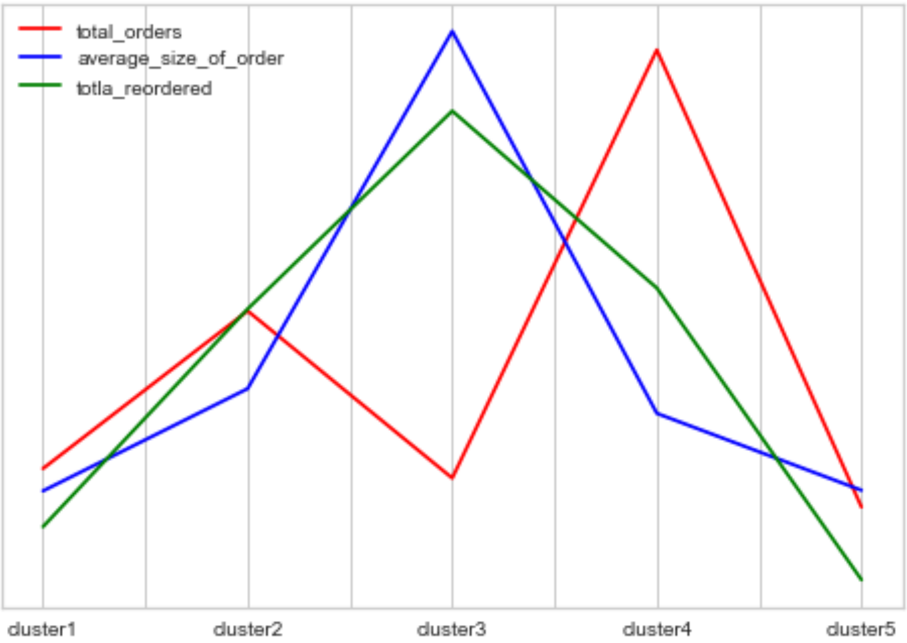


Figure 11. Graph of customer segmentation clusters

A brief explanation of each cluster is below:

Cluster 1: This customer is not a frequent shopper. Quantities of purchasing are small as well.

Cluster 2: This customer is similar to customers in cluster 1. But the amount of order and frequency is higher. And there are a significant number of reordered items in the shopping cart.

Cluster 3: This customer is not using online shopping often. Once they go on shopping, they buy a lot of items and most of the items are reordered items.

Cluster 4: The order size is not big, but this customer is using online shopping very frequently.

Cluster 5: This cluster is similar to cluster 1. But in terms of the number of items and frequency of shopping, this customer is barely using the online shopping and almost no reordered item in the shopping cart.

Cluster 1,2 and 5 have very similar characteristics. The customers in these segments are not doing online shopping much. However, clusters 3 and 4 are showing very different characteristics of segmentation. These customers are loyal customers that the company should focus on.

Prediction of reordered items

3 Classification models are used to predict reordered items: Decision tree, logistic, and gaussian Naïve Bayes. Given features are not sufficient to run the model, therefore, aggregated features are added to improve the models. Total reordered ratio and reordered_rate are created based on the following formula. The total reordered ratio is found for each order ID. The reordered rate is found for each product ID.

$$\text{Total reordered ratio} = \frac{\text{total reordered}}{\text{total order}}$$

$$\text{Reordered rate} = \frac{\text{number of reordered in product}}{\text{total number ordered in product}}$$

7 features are selected because they are related to high correlation coefficients. Table 9 shows the correlation chart.

- Reordered_rate : rate of a number of reordered over the total number of order in the product. The scale of 0 to 1. Each product will have a rate. Some of products will have 0 since the product is never reordered by anyone.
- Total_reordered: Total number of reordered items in each order
- Total_reordered_ratio : Total reordered item over total order
- Order_number : number of order for each customer
- Days_since_prior_order : Days since the order is placed
- Add_ to_cart_order : The sequence of an item is ordered in cart
- Reordered:

Table 9. Correlation chart of 7 features before reorder prediction model.

	REORDERED_RATE	*TOTAL_REORDERED*	*TOTAL_REORDERED_RATIO*	*ORDER_NUMBER*	*DAYS_SINCE_PRIOR_ORDER*	*ADD_TO_CART_ORDER*	*REORDERED*
REORDERED_RATE	1.000000	0.062083	0.200248	0.05985	-0.036800	-0.143393	0.325
TOTAL_REORDERED	0.062083	1.000000	0.510675	0.23820	-0.072660	0.534086	0.295
TOTAL_REORDERED_RATIO	0.200248	0.510675	1.000000	0.43317	-0.229236	0.006018	0.579
ORDER_NUMBER	0.059858	0.238205	0.433179	1.00000	-0.358422	-0.004921	0.250
DAYS_SINCE_PRIOR_ORDER	-0.036800	-0.072660	-0.229236	-0.35842	1.000000	0.053951	-0.132
ADD_TO_CART_ORDER	-0.143393	0.534086	0.006018	-0.00492	0.053951	1.000000	-0.145
REORDERED	0.325475	0.295873	0.579376	0.25097	-0.132814	-0.145232	1.000

25% of the dataset is split to a training set from the prior dataset. In Figure 4, 59% of products in the order are reordered items and the other 41% of products are not reordered items. These datasets are not perfectly balanced; however, it is not hard to say uneven distribution, so it is not rebalanced.

Decision Tree

The decision tree gives a 74% accuracy rate. Precision and recall are 0.743 and 0.742. This model gives a reasonable prediction since the score is high. The confusing matrix is below.

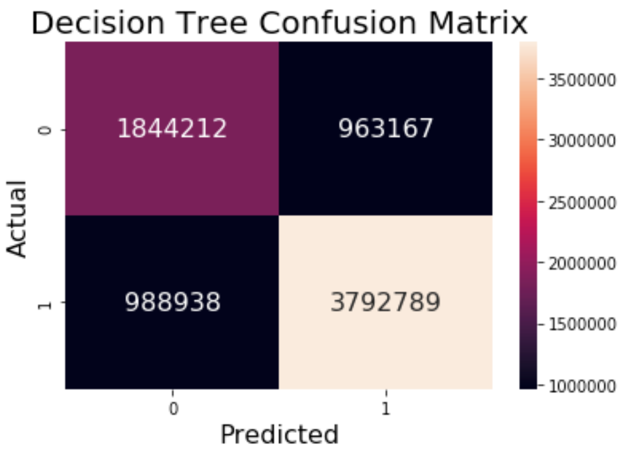


Figure 12. Confusion matrix of decision tree

Table 10. Precision, Recall and F1 score for decision tree

	PRECISION	*RECALL*	*F1 SCORE*
0	0.650940	0.656916	0.653915
1	0.797482	0.793184	0.795327
ACCURACY	-	-	0.742775
WEIGHTED AVG	0.743273	0.742775	0.743015

Logistic Regression

Logistic regression gives 80.4% accuracy and precision and recall score is above 80%. This model predicts the reordered item beyond the expectation. Cross-validation is done in 10 folds. It gives 80.51% accuracy with a standard deviation of 0.05%.

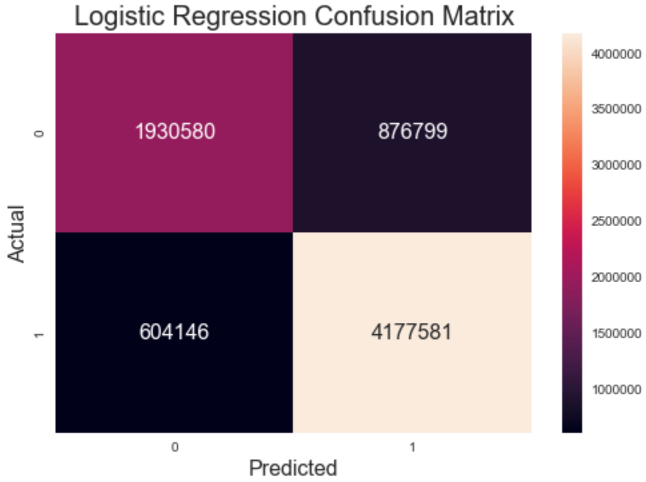


Figure 13. Confusion matrix of Logistic regression

Table 11. Precision, Recall and F1 score for Logistic regression

	PRECISION	*RECALL*	*F1 SCORE*
0	0.761652	0.687681	0.722779
1	0.826527	0.873655	0.849438
ACCURACY	-	-	0.804859
WEIGHTED AVG	0.802528	0.804859	0.802584

Naïve Bayes

Gaussian Naïve Bayes model is used for prediction. Accuracy is 77.9% and precision and recall are 78% and 77.9%. This model is a good prediction model since the scores are above 0.5.

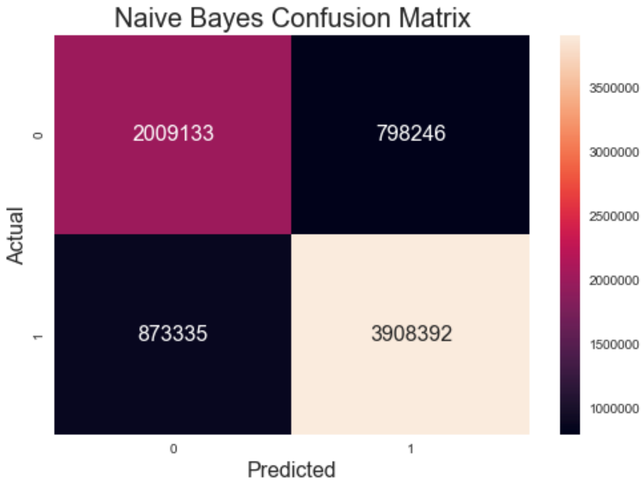


Figure 14. Confusion matrix of Naïve Bayes

Table 12. Precision, Recall and F1 score for Naïve Bayes

	PRECISION	*RECALL*	*F1 SCORE*
0	0.697018	0.715661	0.706217
1	0.830400	0.817360	0.823828
ACCURACY	-	-	0.779739
WEIGHTED AVG	0.781059	0.779739	0.780321

ROC curve

ROC curve is plotted between true positive rate and false positive rate in Figure 15. The model is better when the curve is plotted toward the top left. All three models show a good indication of prediction. Logistic regression is the best among the three models.

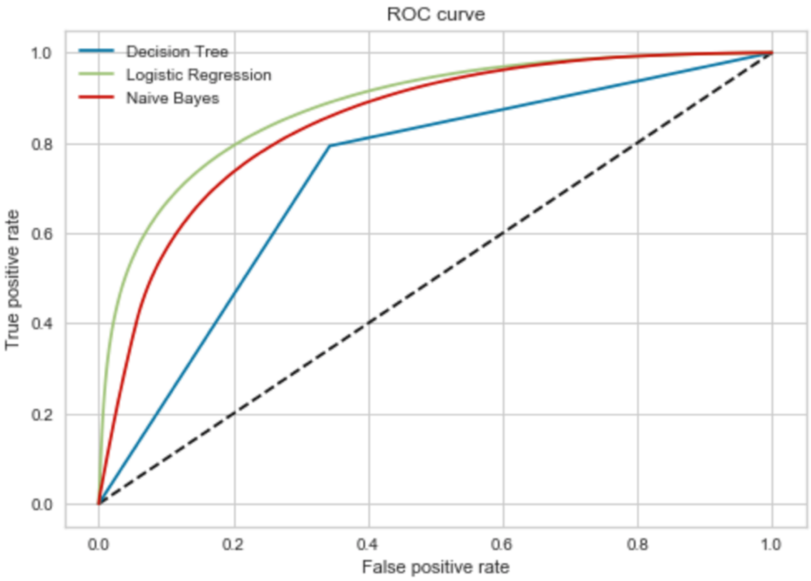


Figure 15. ROC Curve

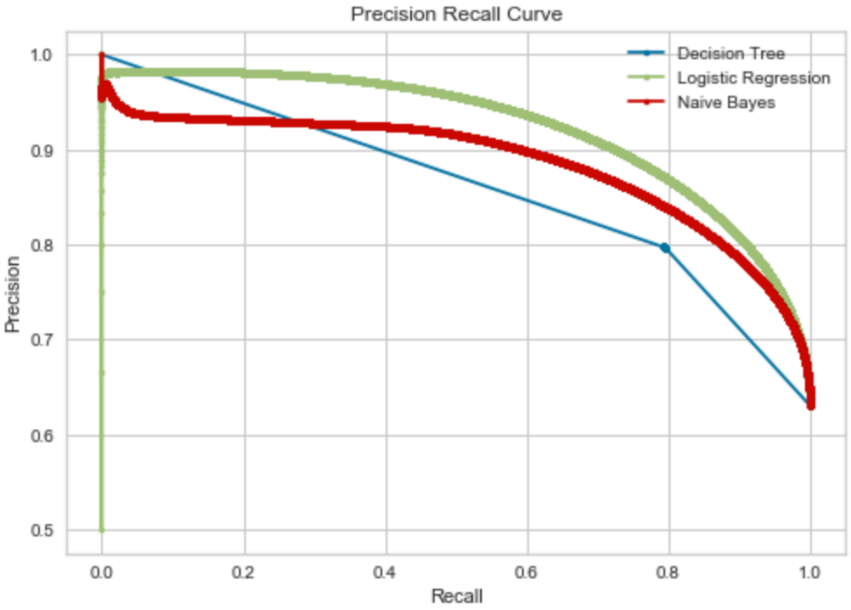


Figure 16. Precision and Recall graph

In Figure 16, the precision and recall graph is plotted. All three models have a good precision and recall scores, but the logistic regression is the best among the three models.

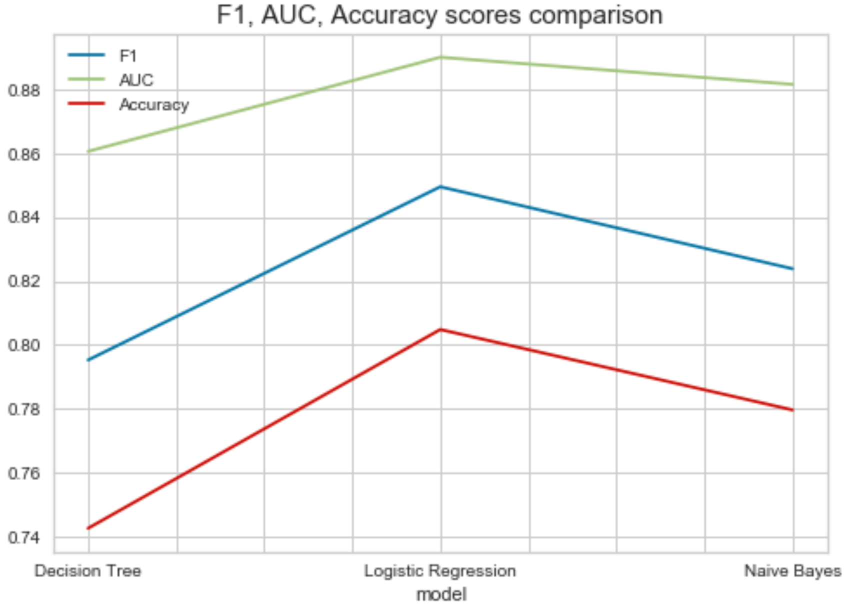


Figure 17. F1, AUC, Accuracy scores comparison

F1 score is the average of precision and recall. AUC is the degree of distinguishable between 0 and 1. The higher scores are better at predicting 0 and 1 (not-reordered or reordered). In Figure 17, the Logistic regression model has the highest score of F1, AUC, and accuracy among the three models.

Conclusions

Order transactions of online grocery shopping sites from Instacart are analyzed to find the insights of customer segmentation, product recommendation, and predicting reordering products. First, data analysis is conducted to understand the background of the dataset. Useful features extracted to be used in building models. K Mean clustering, Association rule, and regression machine learning models are used to answer research questions.

Many combinations of products are associated using the Apriori rule. Lift is high as 75 between products. Any products with a high Lift score can be introduced in the recommendation section when customers add the products in the cart. The minimum lift is 2, thus, there is a positive association between products.

There are 3 types of customers: frequently visited customers with a small cart size, customers with huge cart size, and infrequently visited customers with small cart size. The company can target each customer segment differently. Most royal customer segmentation clusters are frequently visited customers and customers with many items in carts. Once they are identified, this data can be helpful in logistic service in the warehouse to forecast the demand and the company can provide the royalty program to customers to increase customer retention.

The logistic regression model is the best machine learning model and it predicts the reordered items with more than 80% accuracy and 80% F1 score. Prediction of the reordered item can help a company's marketing strategies to find the demanded products.

Note: code link https://github.com/donkimc/Instacart_capstone

Reference

[1] Benn, Y., Webb, T. L., Chang, B. P. I., & Reidy, J. (2015). What information do consumers consider, and how do they look for it, when shopping for groceries online? Appetite, 89, 265-273. doi:10.1016/j.appet.2015.01.025

[2] Bauerová, R. (2018). Consumers' Decision-Making in online grocery shopping: The impact of services offered and delivery conditions. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis, 66(5), 1239-1247. doi:10.11118/actaun201866051239

[3] Singh, R., & Söderlund, M. (2020). Extending the experience construct: An examination of online grocery shopping. European Journal of Marketing, ahead-of-print(ahead-of-print) doi:10.1108/EJM-06-2019-0536

[4] Singh, R. (2019). Why do online grocery shoppers switch or stay? an exploratory analysis of consumers' response to online grocery shopping experience. International Journal of Retail & Distribution Management, 47(12), 1300-1317. doi:10.1108/IJRDM-10-2018-0224

[5] Desrochers, C., Léger, P., Fredette, M., Mirhoseini, S., & Sénécal, S. (2019). The arithmetic complexity of online grocery shopping: The moderating role of product pictures. Industrial Management & Data Systems, 119(6), 1206-1222. doi:10.1108/IMDS-04-2018-0151

[6] Anesbury, Z., Nenycz-Thiel, M., Dawes, J., & Kennedy, R. (2016). How do shoppers behave online? an observational study of online grocery shopping. Journal of Consumer Behaviour, 15(3), 261-270. doi:10.1002/cb.1566

[7] Richards, T. J., & Rabinovich, E. (2018). The long-tail of online grocery shopping. Agribusiness, 34(3), 509-523. doi:10.1002/agr.21553

[8] Piroth, P., Rüger-Muck, E., & Bruwer, J. (2020). Digitalisation in grocery retailing in germany: An exploratory study. The International Review of Retail, Distribution and Consumer Research, , 1-19. doi:10.1080/09593969.2020.1738260

[9] Khalifa, M., & Liu, V. (2007). Online consumer retention: Contingent effects of online shopping habit and online shopping experience. European Journal of Information Systems: Including a Special Section on Healthcare Information Systems Research, Revelations and Visions, 16(6), 780-792. doi:10.1057/palgrave.ejis.3000711

[10] Pauzi, S., Thoo, A., Tan, L., Muharam, F., & Talib, N. (2017). Factors influencing consumers intention for online grocery shopping – A proposed framework. IOP Conference Series: Materials Science and Engineering, 215, 12013. doi:10.1088/1757-899X/215/1/012013

[11] Mackenzie, A. (2018). Personalization and probabilities: Impersonal propensities in online grocery shopping. Big Data & Society, 5(1), 205395171877831. doi:10.1177/2053951718778310

[12] Kim, K. j., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. *Expert Systems With Applications*, *34*(2), 1200–1209. https://doi.org/10.1016/j.eswa.2006.12.025

[13] “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on May 19th, 2020

[14] Tuttle, B. (2012, January 9). Why Monday Is E-Retailers' Favorite Day of the Week. Retrieved from https://business.time.com/2012/01/09/why-monday-is-e-retailers-favorite-day-of-the-week/#:~:text=The reason, he said, is,get to over the weekend.