Instacart Market Basket Analysis
Final Report by Veronica Ferman
April, 2020

Photo by Alex Gruber on Unsplash

Which products will Instacart users order again?

Product Recommendation for the digital customer

The on-line grocery marketplace is in the middle of a digital disruption as consumers increasingly shop for food, personal care items, and other household items through apps and websites that offer products delivered to their homes. The increasing number of on-line consumers present on-line stores with the opportunity and challenge to curate to their individual taste in a personalized and timely manner.

A product recommendation system that utilizes the power of machine learning can streamline the consumer journey by showcasing the right products at the right time at a mass scale.

An efficient machine learning model can find trends in customers' historical data and predict which products are most relevant. This knowledge can also be applied to personalize the shopping content of other users with similar characteristics.

The Instacart Prediction Challenge

Instacart is one of the companies in the digital grocery space. It provides grocery delivery services to users across the U.S and Canada. Users place their grocery orders through the Instacart app and a personal Instacart shopper delivers the order to their home.

Instacart made public an anonymized data of over 3 million grocery orders from more than 200,000 users and presented the 'Instacart Market Basket Analysis' challenge of predicting which previously purchased products would be in the users next order. The challenge was introduced through Kaggle, a community of data science and data engineers.

The Instacart Market Basket Analysis dataset is a relational set of files describing customers' orders over time. In the next pages, it will be analyzed for historical trends and patterns and will result in machine learning development of a model that can be used as a product recommendation system for the digital grocery shopper.

Best Model Selection

To make predictions of future purchases, the analysis includes the comparison of three machine learning algorithms that will build models based on the sample data. The report will conclude with a sample of a predicted shopping cart for one of Instacart users.

Python Programming Language has been chosen as the programming language for the analysis.

The sequence of processes to examine the data and build the best machine learning model for a product recommendation system are presented next. For the project code, please visit GitHub Repository.

Data Cleaning and Wrangling

Data cleaning and data wrangling were the first steps in the process of preparing, analyzing and constructing a predictive model for the project.

- The breakdown of the files provided by Instacart through the Kaggle Instacart Market Basket Analysis competition are as follows::

File name	**Description**
aisles.csv	Aisles in store by aisles id.
departments.csv	Departments in store by department id.
order_products_prior.csv	Contains previous order contents for all customers. It specifies which products were purchased in each order.
order_products_train.csv	Contains train order contents for some customers.
orders.csv	This file tells to which set (prior, train, test) an order belongs. It contains all orders.
products.csv	Product identifier with names of products

- Files were converted to pandas dataframes and each one was examined for content, outliers and missing values.

```
The data contains a total of 206209 users
who made a number of 3421083 orders categorized as follows:

3214874 orders categorized as prior
131209 orders categorized as train
75000 orders categorized as test
```

```
There are 49688 unique products
```

- The results showed missing values as ‘NaN’ in the ‘days since prior order’ column of the orders.csv file. The other dataframes showed zero missing values.

```
data.isnull().sum()
```

```
order_id                      0
user_id                       0
eval_set                      0
order_number                  0
order_dow                     0
order_hour_of_day             0
days_since_prior_order    206209
dtype: int64
```

At this point of the research, the NaN values were left as such until reaching the modeling stage of the project and consider its relationship to the other variables.

Initial Data Analysis

For the initial data exploration, a subset of the data was created using the 500th user id of order_products_train.csv. Matching user ids from order_products_prior were selected to form the sample. The subset has the following characteristics:

Sample subset characteristics
3969 orders
3714 prior orders
255 unique users
Users placed a mean of 6.42 orders

The files were merged into a dataframe that cross references:

User’s id
Orders
Products in each order
Department to which the product belongs
Aile to which the product belongs

Day of the week when order was placed
Order in which products were added to the shopping cart
Days since previous order
Whether the product is a reorder
Time of day when order was placed

```
product_merged_sample.head(5)
```

user_id	order_id	order_number	product_id	product_name	aisle	department	eval_set	add_to_cart_order	order_dow	order_hour_of_day
102271	2249	9	11365	Leaf Spinach	packaged produce	produce	prior	1	3	21
102271	2249	9	43352	Raspberries	packaged produce	produce	prior	2	3	21

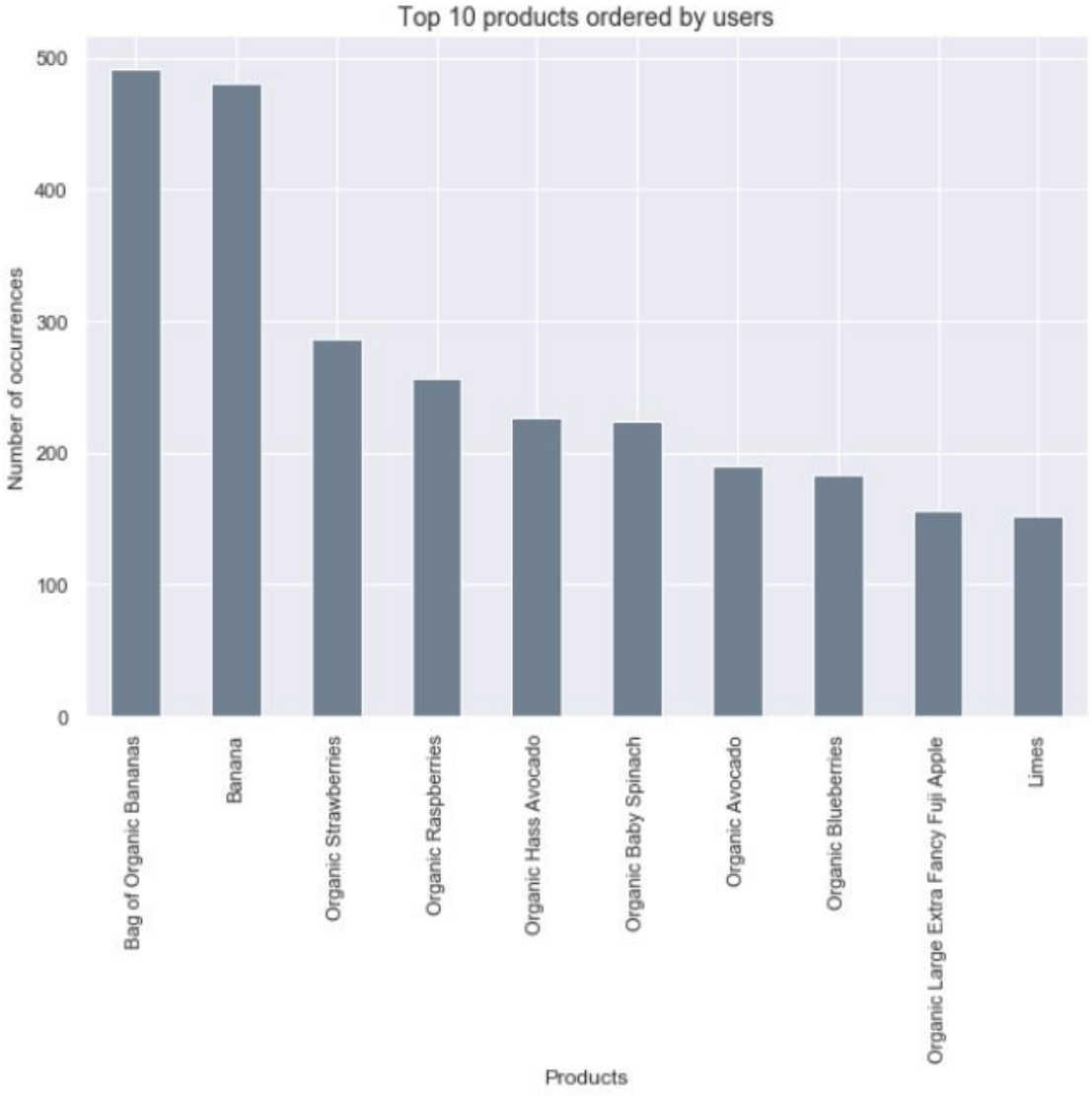
Exploratory Data Analysis

The initial data analysis resulted in the formulations of the following key questions:

Which products were the best sellers?

The following plot shows a graphical representation of the top 10 products ordered by users. Organic fruits is the top category with organic bananas taking the first place among users.

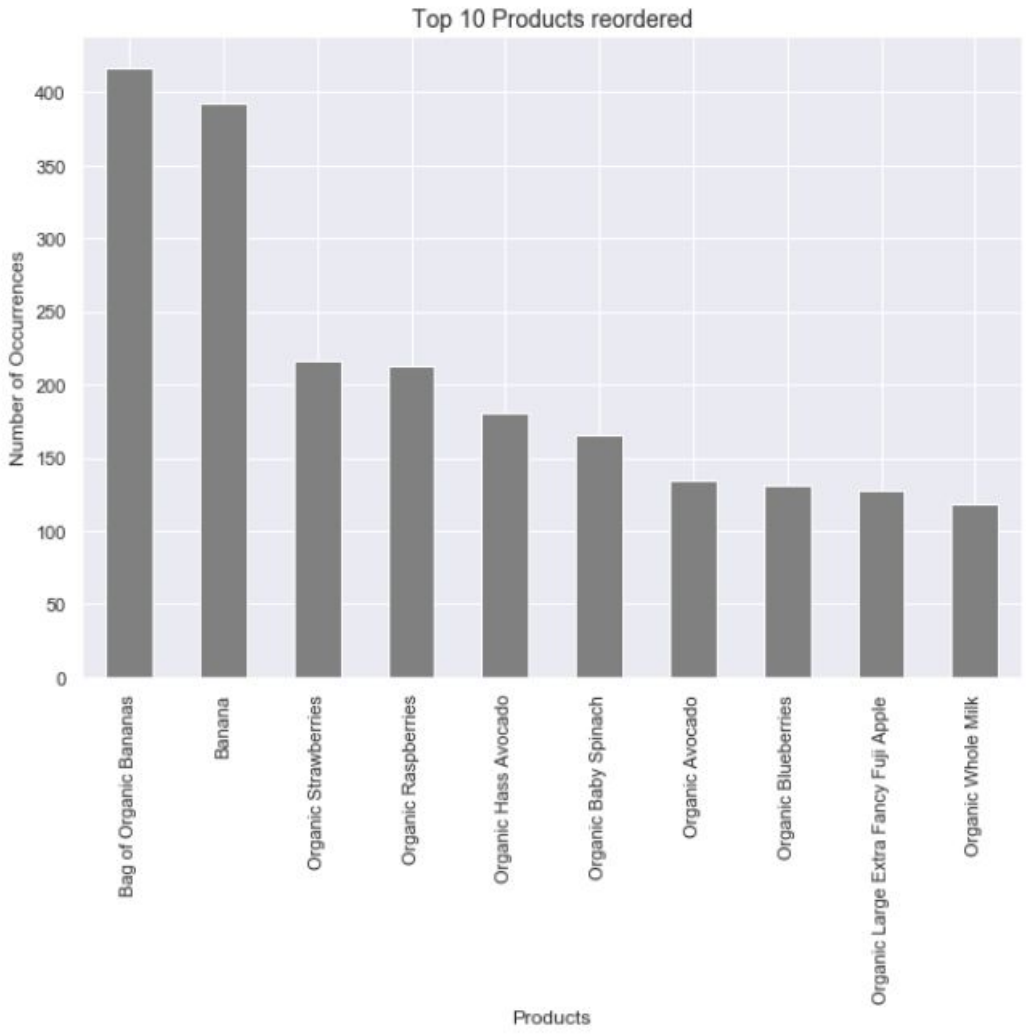
Top 10 products ordered by users
Bag of organic bananas
Bananas
Organic strawberries
Organic raspberries
Organic Hass avocado
Organic baby spinach
Organic avocado
Organic blueberries
Organic large extra fancy fuji apple
Limes



Which products have a high reorder frequency?

To appreciate the differences or similarities between the products that users are ordering and reordering, the top 10 reordered products are shown in the following plot.

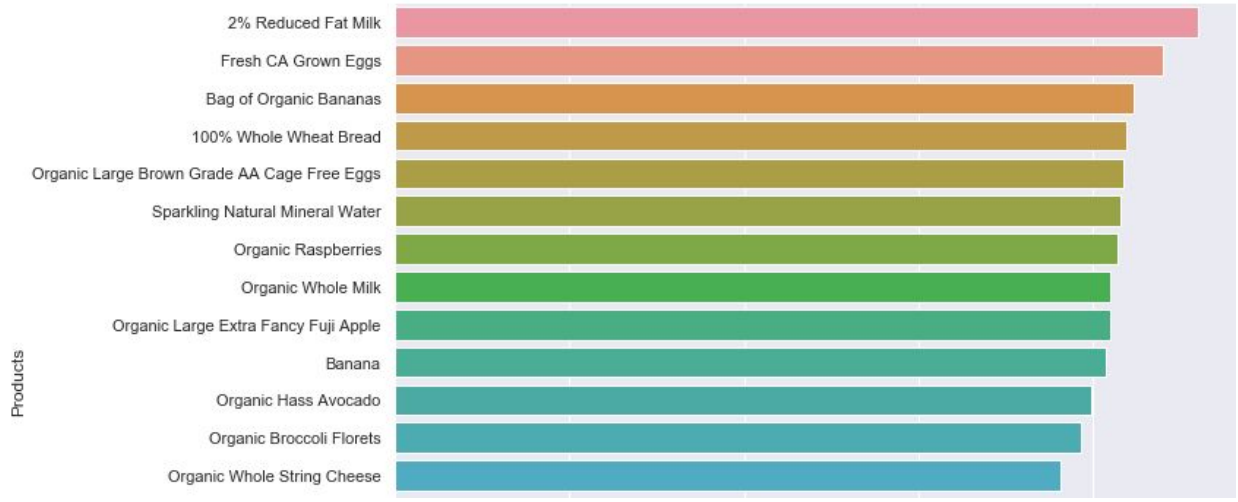
The plot shows that users are consistent with their top organic fruits and vegetables choices and reorder them again. This time, organic whole milk makes the top 10 list.



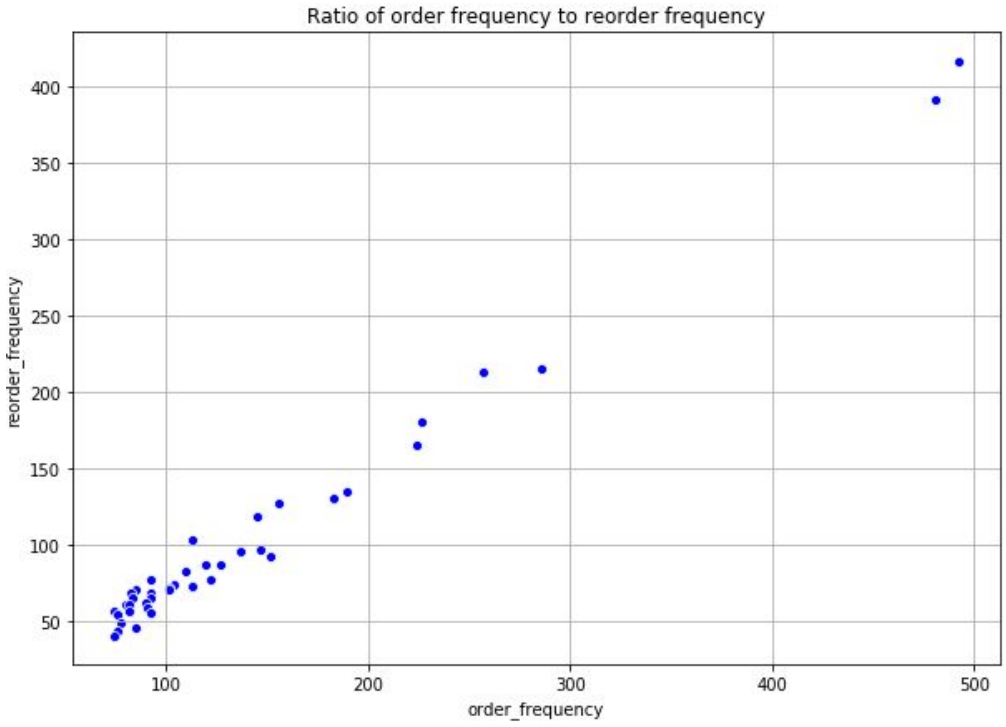
To further explore the relationship between product ordering and reordering, a table with the reorder ratio for the **top 40 products** in the subset was created:

	product_name	order_frequency	reorder_frequency	reorder_ratio
16	2% Reduced Fat Milk	113	104	92.04
38	Fresh CA Grown Eggs	75	66	88.00
0	Bag of Organic Bananas	492	417	84.76
23	100% Whole Wheat Bread	93	78	83.87
29	Organic Large Brown Grade AA Cage Free Eggs	85	71	83.53
31	Sparkling Natural Mineral Water	83	69	83.13
3	Organic Raspberries	257	213	82.88

Top products with high reordered ratios

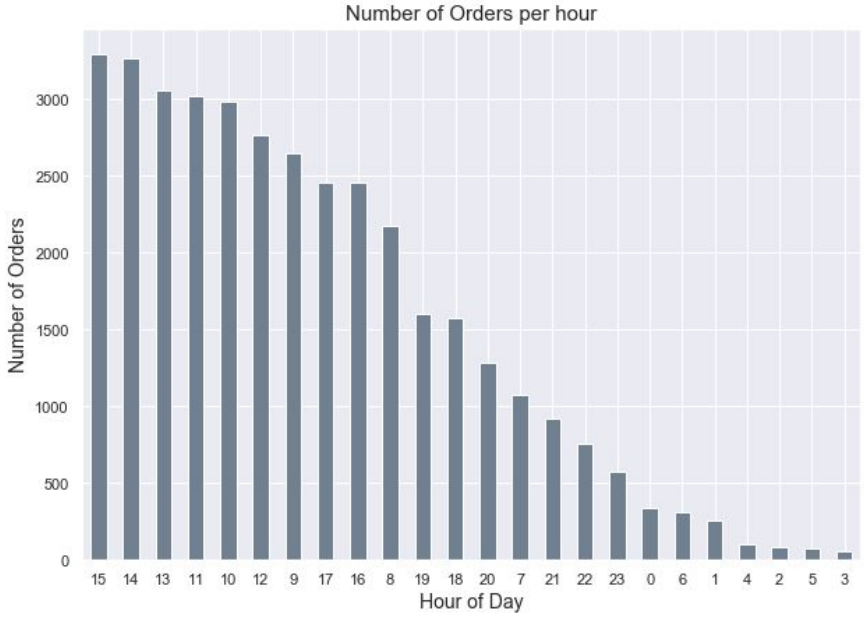


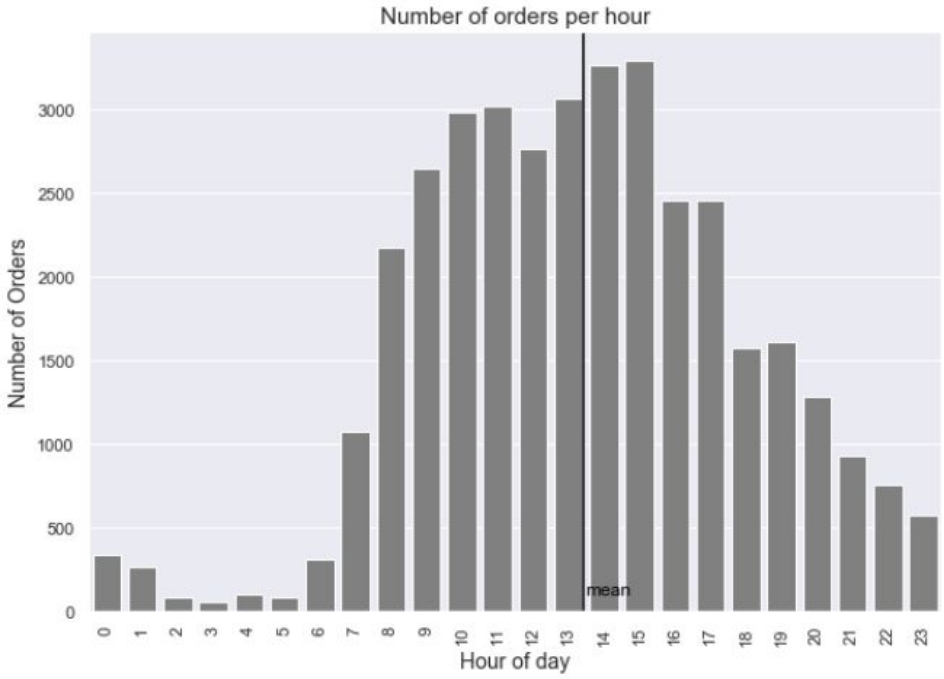
The following scatterplot representation shows a strong correlation between ordered and reordered frequencies. It shows how orders in the shopping cart might be repetitive and customers are loyal to their choice of products.



When do shoppers buy?

The barplot and countplot below show that many shoppers buy between 8am and 5pm. The countplot shows a steady increase that peaks at 3:00 p.m.

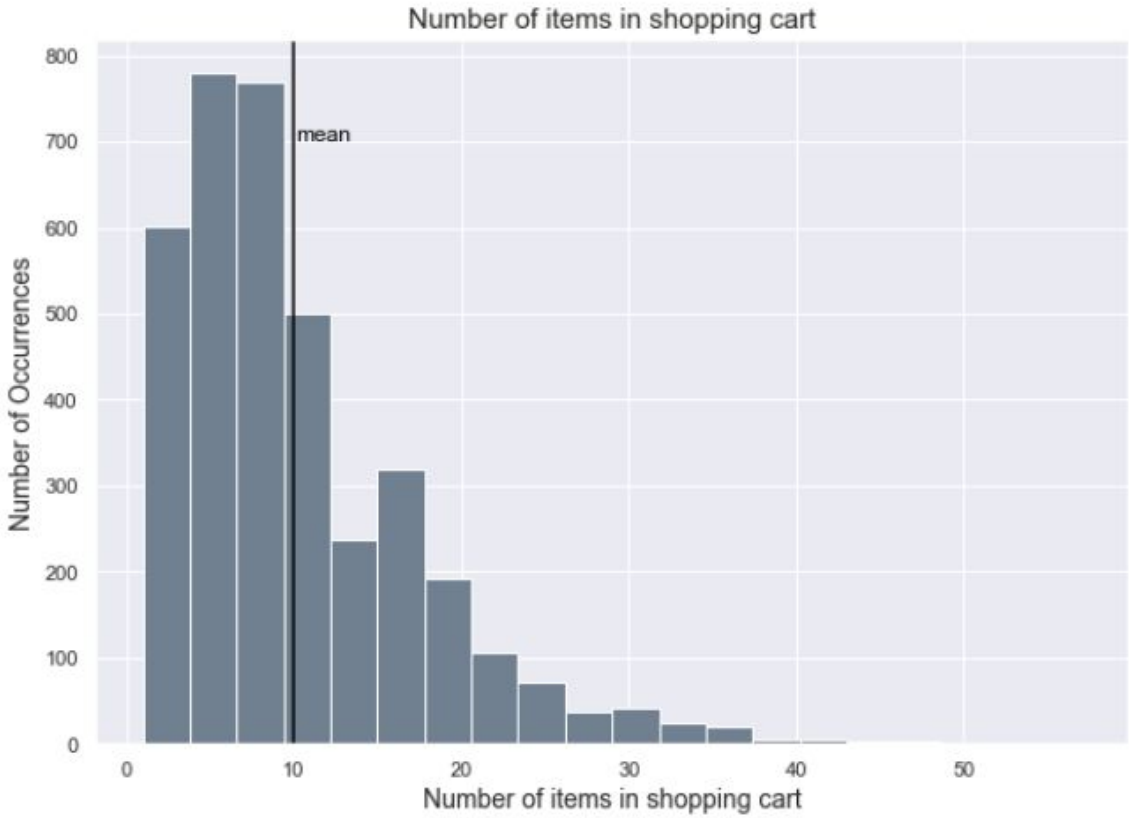




How many items in a shopping cart?

The minimum number of items in a shopping cart is 1 and the maximum number is 57. 75% percent of shoppers have 14 items or less in their shopping cart. The histogram below shows that the average is 10 .

```
count    3714.000000
mean        9.999731
std         7.245809
min         1.000000
25%         5.000000
50%         8.000000
75%        14.000000
max        57.000000
Name: items_in_shopping_cart, dtype: float64
```

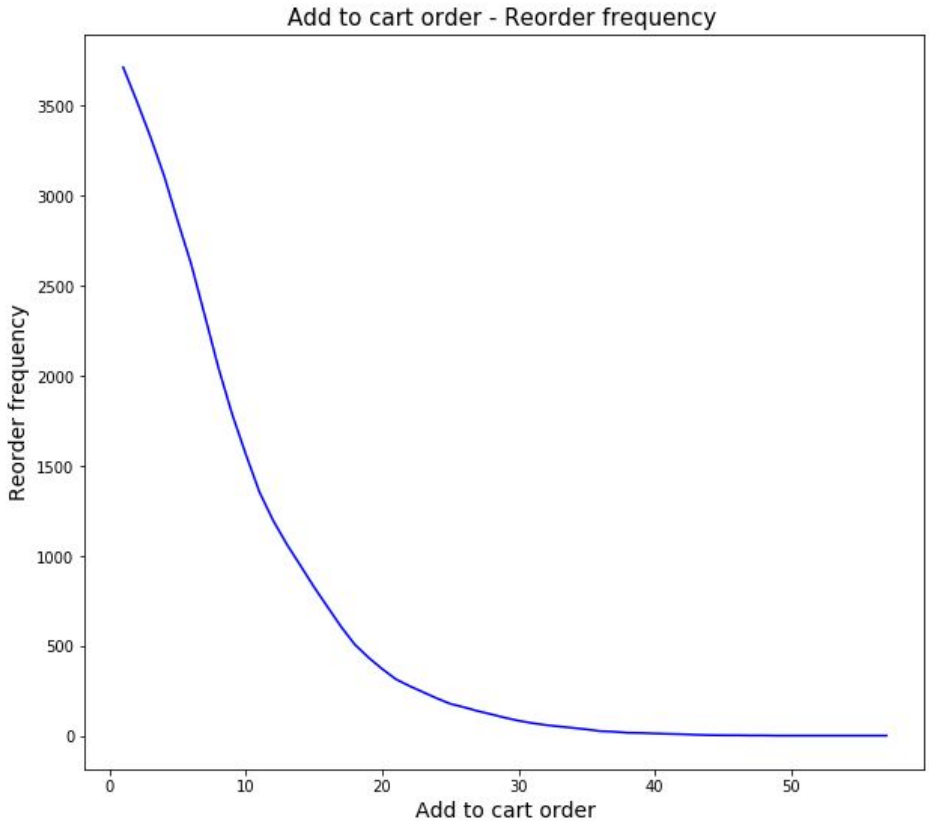


Is there a relationship between the sequence in which users add products to their shopping carts and the products they reorder?

The following plot shows that there is a relation between the order in which users placed their product in the shopping cart and the frequency that the product is reordered.

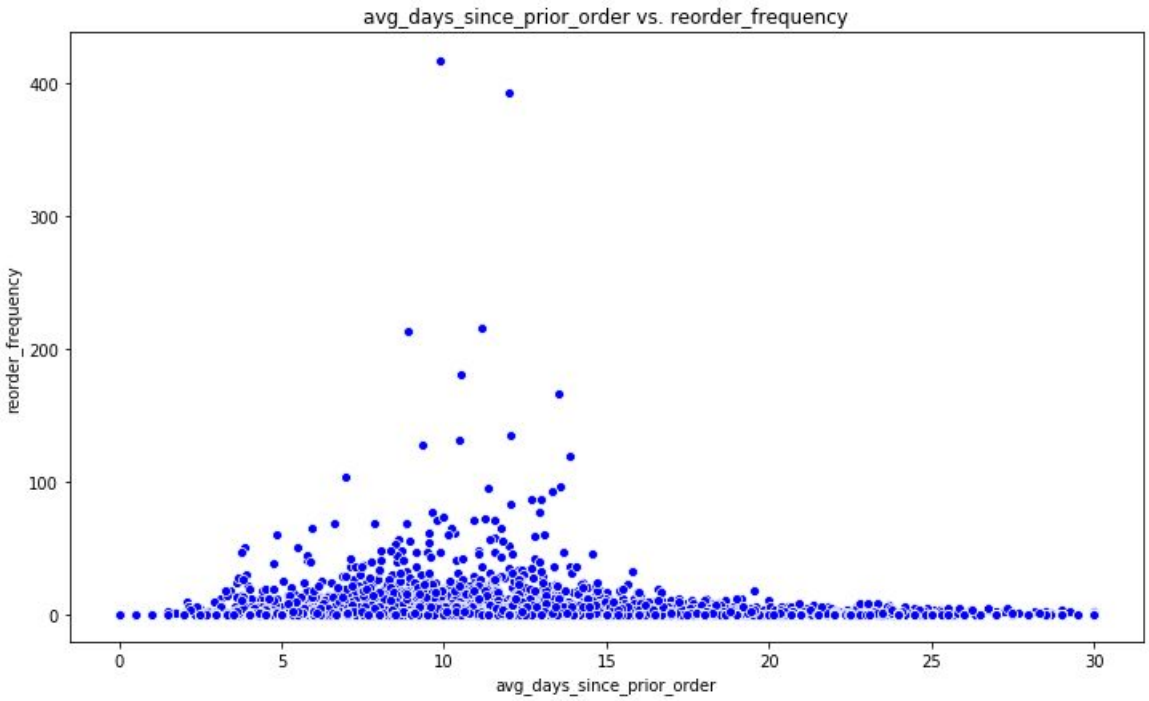
Products that are placed first in the shopping cart have a higher reorder frequency. The reorder frequency diminishes as products move down in their add to cart order.

add_to_cart_order	number_reorders
1	3714
2	3526
3	3328
4	3112
5	2861



How often do users reorder products?

The following scatterplot shows that there is a higher level of reordering activity between 2 and 15 days since the previous order.



Summary of Exploratory Data Analysis

- The analysis of a data subset of Instacart users has shown that shoppers use the Instacart app to buy many of their fruits, vegetables and dairy items.
- Organic produce is a top category.
- There is consistency when reordering items and there is a strong positive correlation between orders and reorders, especially among the products of higher order frequencies. Customers repeat their choices.

- There is a correlation between the order in which users place their products in the shopping cart and reorder frequency. Items that are placed first have a higher reorder frequency.

- There are certain times of the day and certain number of days since prior orders in which many of the Instagram users place their orders.

Based on the data analysis findings, the following approach was followed to create the most relevant features of the machine learning model that would yield a high level of accuracy in the prediction of products.

- Train the machine learning models on data from 50,515 randomly selected Instacart users.
- Refine model features according to the following levels:
    - Product level features
    - User behavior features
    - Time related features
- For time related features and add to cart order independent variables, data was aggregated at a product and user level to reveal patterns in the data.

Examples:

Average order hour of day - product level

product_id	avg_order_hour_of_day
1	12.678571
2	12.678571

Average order hour of day - product and user level

	product_id	user_id	user_order_hour_of_day
0	1	1540	14.529412
1	1	8703	13.000000

- The following 31 data features were used to train the machine learning models after considering curse of dimensionality factors:

    1. order_dow (day of week)
    2. order_hour_of_day
    3. days_since_prior_order
    4. product_id
    5. add_to_cart_order
    6. aisle_id
    7. department_id
    8. number_orders
    9. number_reorders
    10. reordered_ratio
    11. avg_days_since_prior_order
    12. user_reorder_ratio
    13. user_reorder_sum
    14. user_aisle_reorder_ratio
    15. user_aisle_reorder_sum
    16. user_department_reorder_ratio
    17. user_department_reorder_sum
    18. user_product_orders
    19. user_products_order_rate
    20. user_total_items
    21. user_total_distinct_items
    22. user_average_days_between_orders
    23. user_number_orders
    24. user_average_basket
    25. user_add_to_cart_order
    26. user_day_of_week
    27. user_order_hour_of_day
    28. number_reorders_by_cart
    29. avg_reorders
    30. product_avg_day_of_week
    31. avg_order_hour_of_day

- The target variable is a 0 if the product is not a reorder and a 1 if the product is a reorder.

    50,515 unique users were split into 70% train and 30% test data. Test data is the portion of data not seen by the model that will allow us to test the accuracy of the model.

- Because of the binary classification nature of the problem, the following machine learning models were selected to yield the prediction of products:

    - Logistic Regression
    - XGB Boost
    - Random Forest

- Accuracy scores, confusion matrix information and classification reports were used to gauge performance of models.

- Hyperparameter tuning was employed in the case of the Logistic Regression and XGB Boost models

Machine Learning Prediction Results

Logistic Regression Model

Test data accuracy score: 84% with a C value of .1
Train data accuracy score: 84%
The accuracy score is the ratio of correctly predicted observations to the total number of observations.

The accuracy scores show that the model can adapt to new unseen data. The model can be used to predict reorder items in new data with new users.

Confusion Matrix

```
Confusion Matrix:
[[ 725253  256620]
 [ 109429 1293031]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.75      0.80    981873
           1       0.84      0.91      0.87   1402460

    accuracy                           0.84   2384333
   macro avg       0.84      0.83      0.83   2384333
weighted avg       0.84      0.84      0.84   2384333
```

The confusion matrix gives us insights about the model and the way it is classifying the classes. The 91% recall shows that most of the reorder class samples (class 1) are being correctly classified as such (true positives), however, the model tends to assign the reorder class to samples that are not reorders (false positives).

The Instacart dataset contains more samples class 0 (orders) than samples that are class 1 (reorders). This imbalance in the data helps explain why the model miss-classifies class 0 items at a higher rate.

XGBoost Model

Test data accuracy score: 87% with hyperparameters: learning_rate: 0.001, max_depth:6, silent: False
Training data accuracy score: 87%

The XGBoost model test data accuracy score shows that the model can generalize well on new data. It correctly predicted the reorder class in the unseen test data 87% of the time.

The confusion matrix shows a 96% recall for the reorder class which means that the model is very good at recognizing the items that the user will reorder.

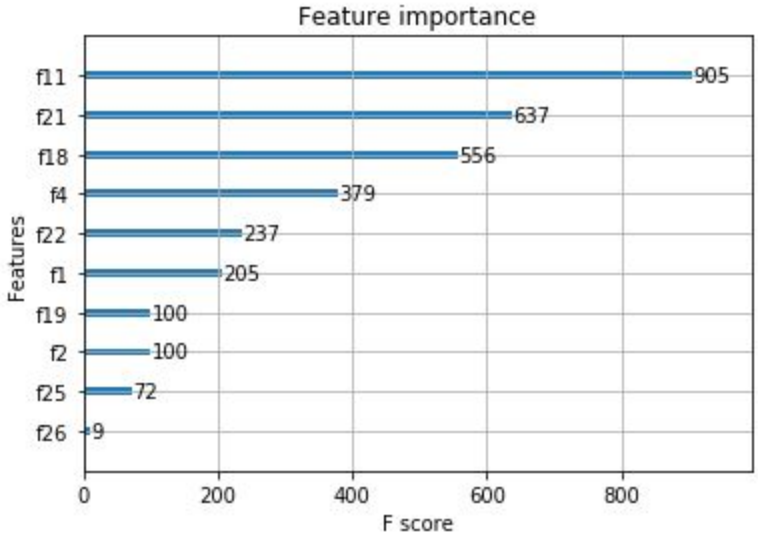
It tends to classify samples as reorders when they are not reorders (false positives) but when it does recognize an item that is not a reorder, the model is 92% correct in its prediction.

```
Confusion Matrix:
[[ 725253  256620]
 [ 109429 1293031]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.74      0.82    981873
           1       0.84      0.96      0.89   1402460

    accuracy                           0.87   2384333
   macro avg       0.88      0.85      0.86   2384333
weighted avg       0.87      0.87      0.86   2384333
```

The feature importance plot below shows that the features most important to the model prediction capabilities are:

Feature 11: avg_days_since_prior_order
Feature 21:user_total_distinct_items
Feature 18: user_product_orders (orders by product)
Feature 4: product_id
Feature22:user_average_days_between_orders

Feature Importance XGB Model



Random Forest Model

Test data accuracy score:85%

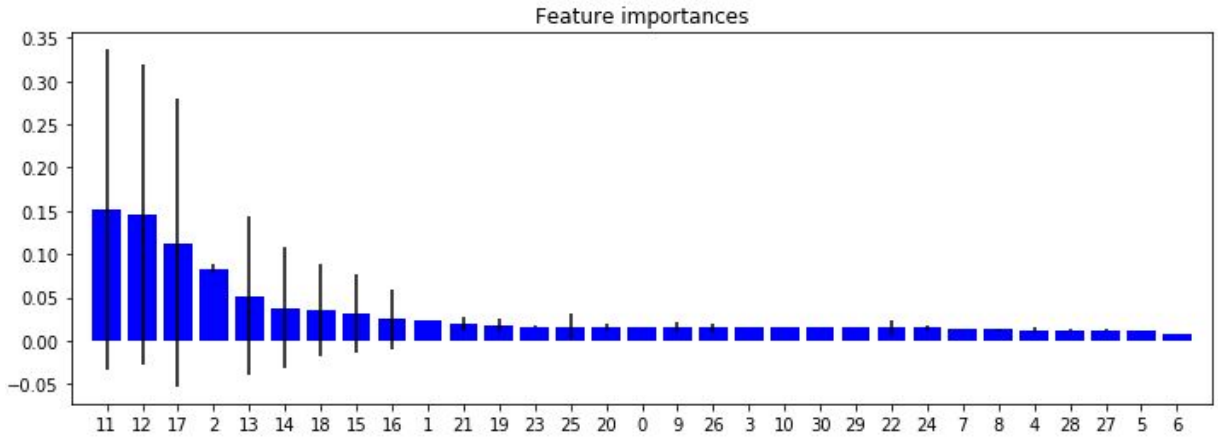
The Random Forest Model generalized well on test data with a test accuracy score of 85%.

```
Confusion Matrix:
[[ 725253  256620]
 [ 109429 1293031]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.74      0.82    981873
           1       0.84      0.96      0.89   1402460

    accuracy                           0.87   2384333
   macro avg       0.88      0.85      0.86   2384333
weighted avg       0.87      0.87      0.86   2384333
```

The most important features for the model are:

Feature 11: avg_days_since_prior_order

Feature12: user_reorder_ratio
Feature17: user_department_reorder_sum
Feature2: order_hour_of_day
Feature13: user_reorder_sum



Shopping Cart Sample

user_id	order_number	product_name	Actual	Predicted
3855	2	Protein & Greens Vanilla Flavor Drink Mix	0	0
3855	2	Soda	0	1
3855	2	Cauliflower Florets	1	1
3855	16	Fresh Asparagus	0	0
3855	16	Crumbled Bacon	0	0
3855	16	Raspberries	1	1
3855	6	Wheat Thins Original	0	0
3855	6	Fat Free Skim Milk	1	1
3855	6	Raspberries	1	1
3855	6	Soda	1	1
3855	6	Roasted Pine Nut Hummus	0	0
3855	6	Hass Avocados	1	1
3855	14	Uncrustables Peanut Butter & Grape Jelly Sandwich	1	1
3855	14	Soda	1	1
3855	14	Fat Free Skim Milk	1	1

The previous table is an example of 4 orders placed by one Instacart user. The predictions are the results of the XGBoost machine learning model which received the highest accuracy score.

The table shows that the model predicted with a high level of accuracy the reorder items in the shopper's list. By utilizing Machine learning knowledge, we can recommend products that users might be enticed to order together such as:

Fat Free Skim Milk and **Uncrustables Peanut Butter & Grape Jelly Sandwich**

Or, bundle products that make an easy buy for users such as a healthy package of zero calorie cola, fruit snacks, apples and original popcorn:

user_id	order_number	product_name	Actual	Predicted
2253	18	Original Popcorn	1	1
2253	18	Apples	1	1
2253	18	Fruit Snacks	1	1
2253	18	Zero Calorie Cola	1	1

The product recommendation system can be scaled to include on-line shoppers in scale and find users that have similar characteristics to provide them with personal shopping experiences with the right products at the right time.

The performance of the machine learning model can be enhanced to deliver results at scale across a number of users with similar user behavior. The Instacart Market Basket Analysis challenge is a practical example of what can be accomplished.

Predicted shopping cart sample

order_id	user_id	product_id	product_name	Actual	Predicted
130554	9431	21938	Green Bell Pepper	1	1
130554	9431	23383	Super Natural Organic Whole Milk	1	1
130554	9431	27845	Organic Whole Milk	1	1
130554	9431	8518	Organic Red Onion	0	0
130554	9431	36737	Pizza Uncured Pepperoni Gluten Free	0	0
130554	9431	30169	Total 2% All Natural Plain Greek Yogurt	1	1
130554	9431	28038	Soft Baked Double Chocolate Brownie Cookies	0	1
130554	9431	42307	Organic Reduced Fat 2% Cottage Cheese	1	1
130554	9431	49683	Cucumber Kirby	1	1
130554	9431	48679	Organic Garnet Sweet Potato (Yam)	1	1

In this shopping cart, the random forest model predicted with very good accuracy which products would be reordered by the user. The target variable shows a 1 when the product is a reorder.

Conclusion

The results show that model features at a user level were very important in the product classification as they relate to consumer purchasing habits.

The random forest model had an accuracy of 87% which represents the highest of the three models tested.