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

Photo by Alex Gruber on Unsplash

Which products will Instacart users order again?

Problem Statement

Instacart is a technology company providing a same day delivery grocery service 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 3 million Instacart shoppers presenting researchers with the challenge of predicting which previously purchased products will be in the user's next order. The problem is predicting consumer's future purchases using historical data to predict future behavior.

The dataset for this competition is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, the dataset provides between 4 and 100 of their orders, with the sequence of products purchased in each order.

The findings and presentation of the data analysis can provide insights to stores that want a model on how purchases occur online based on users' experience and purchasing behaviors. Clients in sales and marketing looking for ways to increase ecommerce sales and finetune a more accurate consumer journey path for their clients will also benefit from the research.

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

The resulting predictions from the machine learning models provide valuable information to inventory and supply chain sectors of business.

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

The sequence of processes that I followed to examine the data are outlined below, as well as a link to its 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.

- This is the breakdown of the files provided by Instacart through the Kaggle Instacart Basket Analysis competition:

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 Findings from Exploration of Data

For the initial Exploration of the data, 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)
```

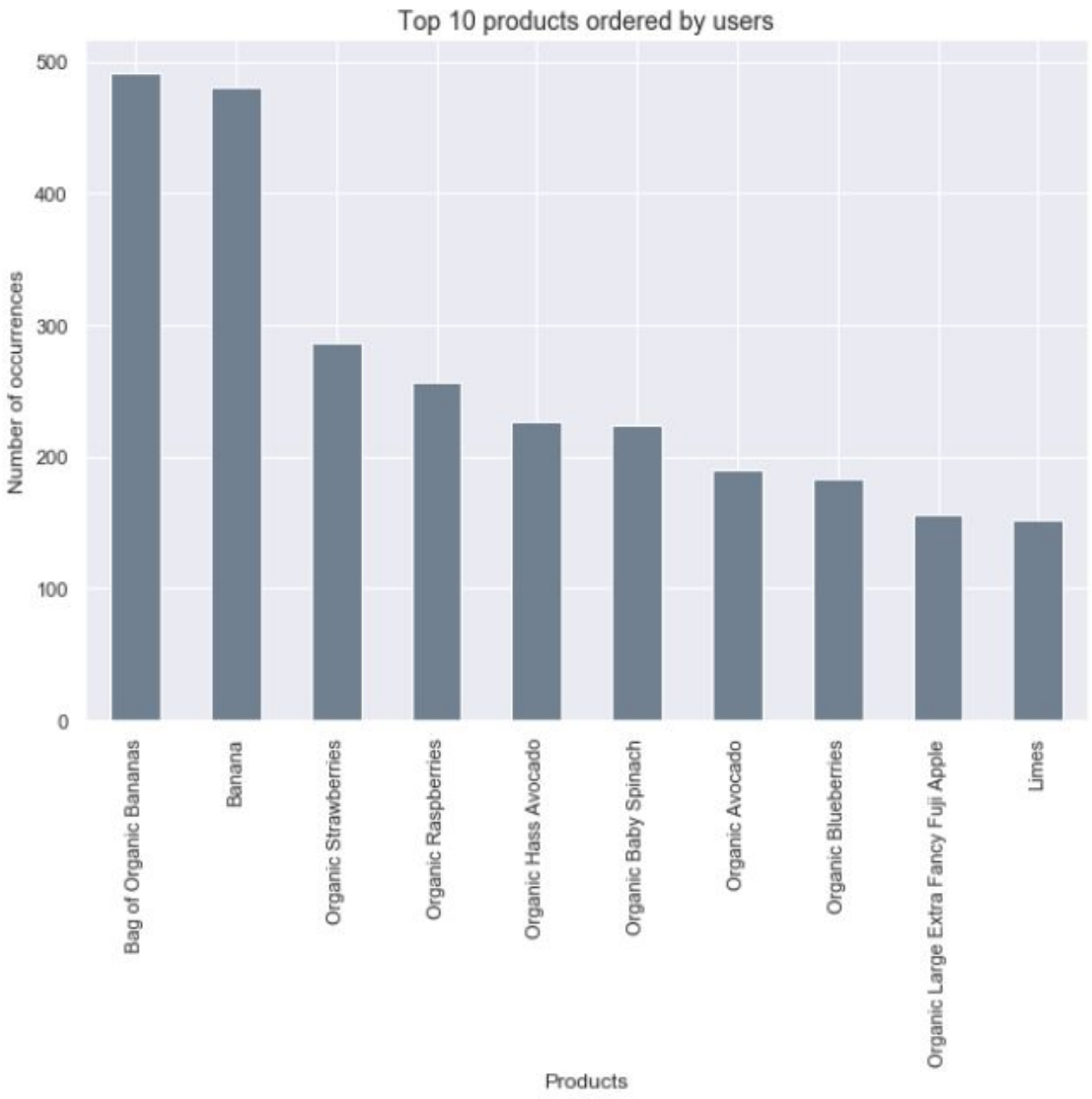
user_id	order_id	order_number	product_id	product_name	aisle	department	eval_set	add_to_cart_order	order_dow	order_hour_of_day
102271	2249	9	11365	Leaf Spinach	packaged produce	produce	prior	1	3	21
102271	2249	9	43352	Raspberries	packaged produce	produce	prior	2	3	21

After the initial exploration data analysis, the following are the results that answer 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.

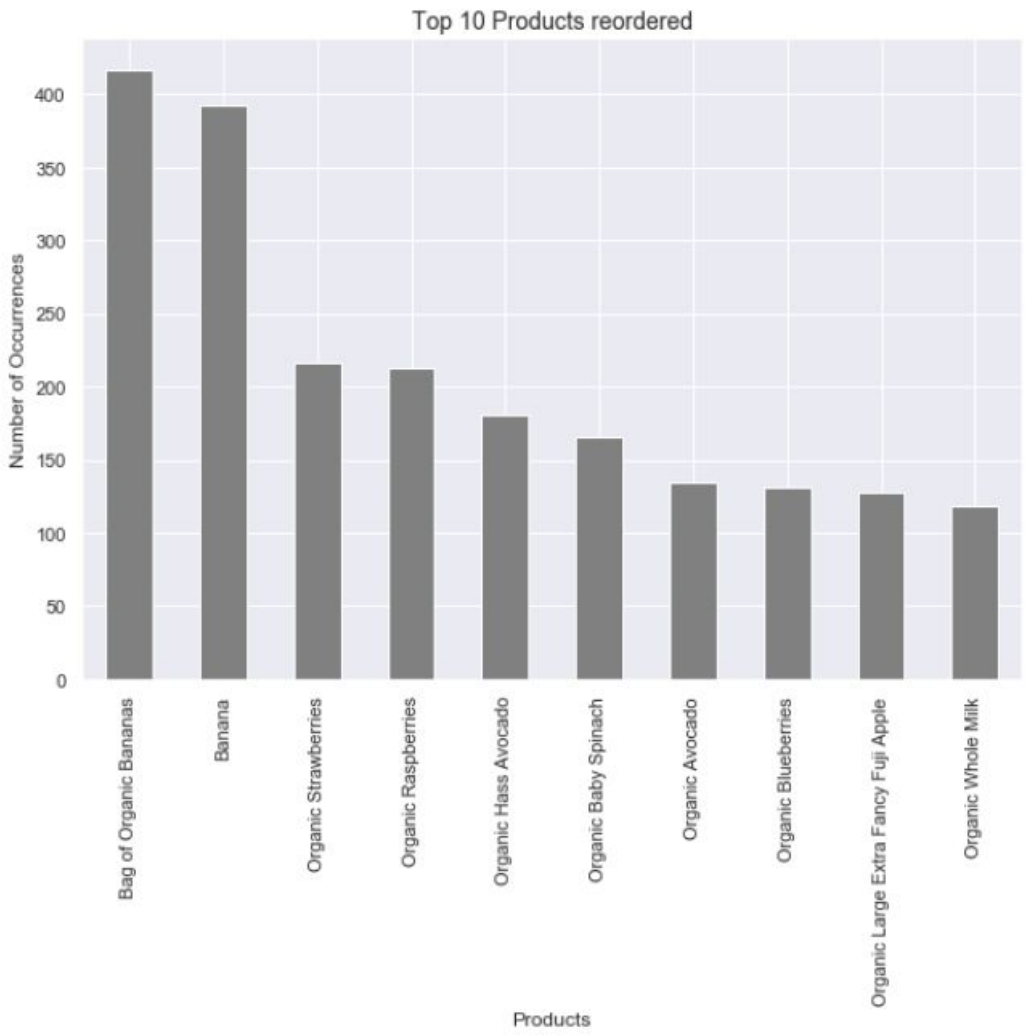
Top 10 products ordered by users
Bag of organic bananas
Bananas
Organic strawberries
Organic raspberries
Organic Hass avocado
Organic baby spinach
Organic avocado
Organic blueberries
Organic large extra fancy fuji apple
Limes



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.

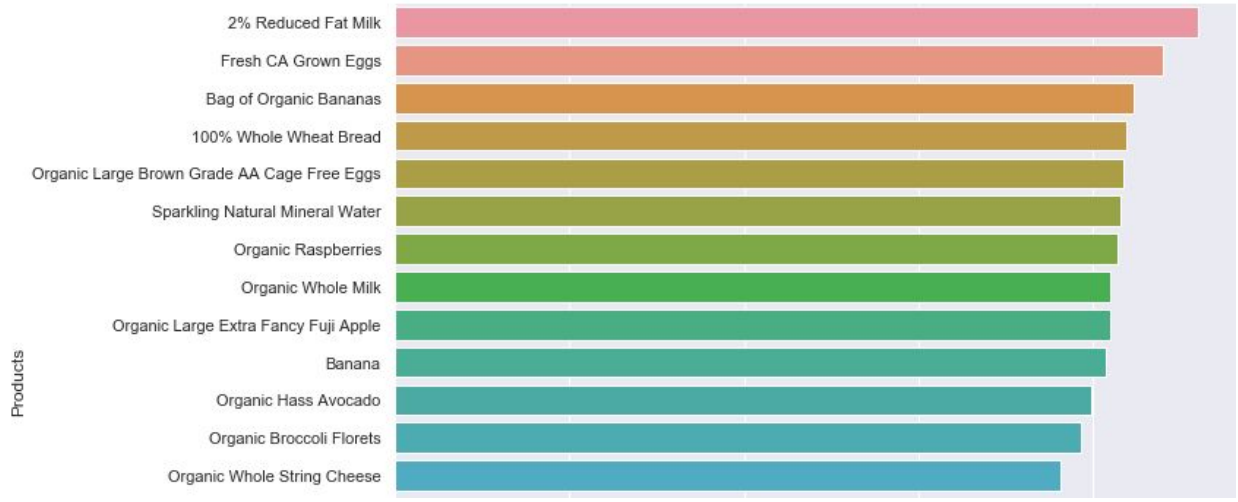
The plot shows that users are consistent with their top organic fruits and vegetables choices and reorder them again. This time, organic whole milk makes the top 10 list.



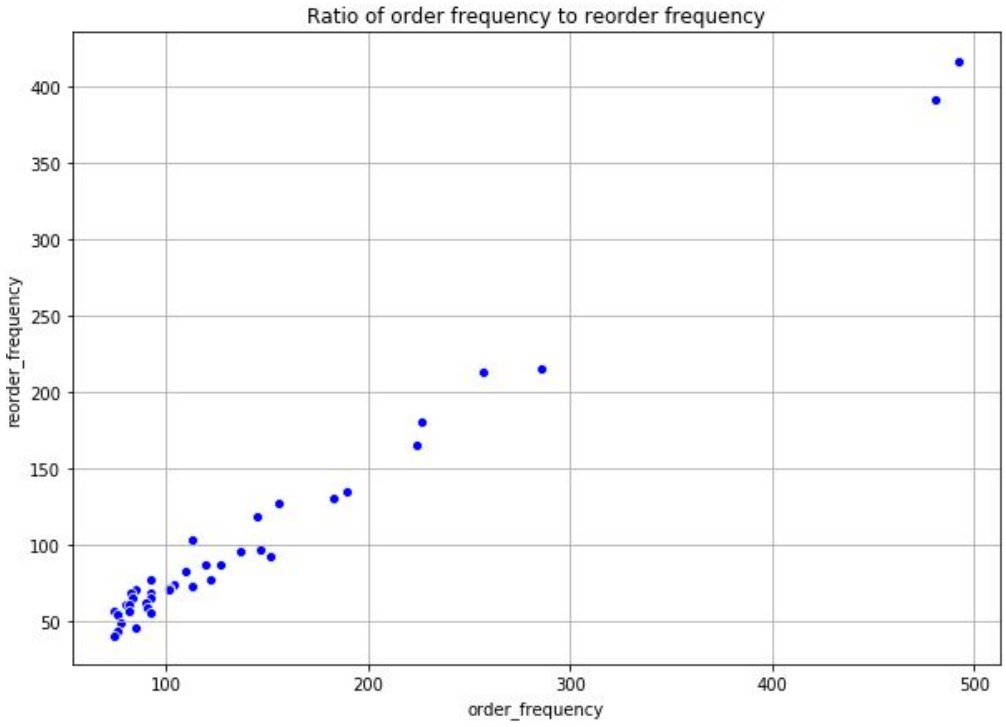
To further explore the relationship between product ordering and reordering, a table with the reorder ratio for the **top 40 products** in the subset was created:

	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

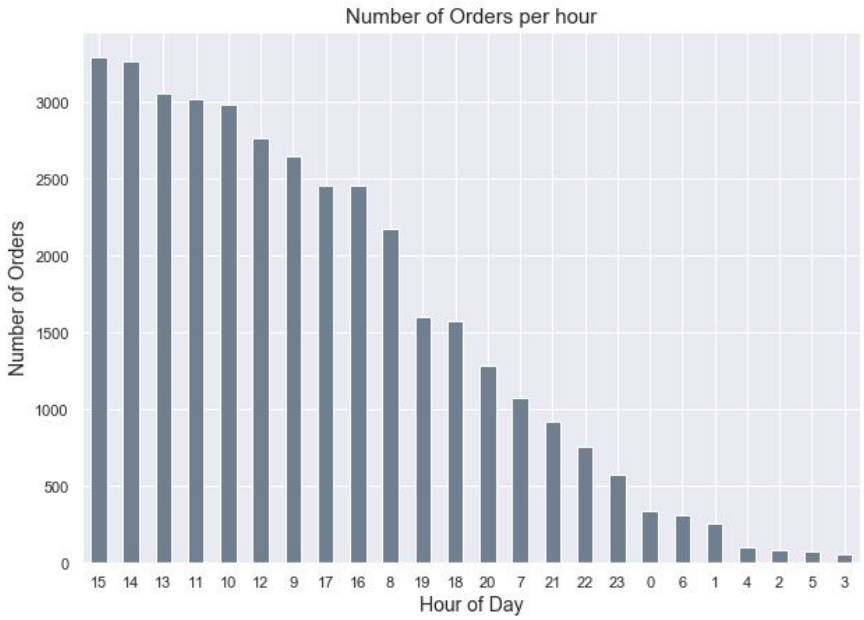


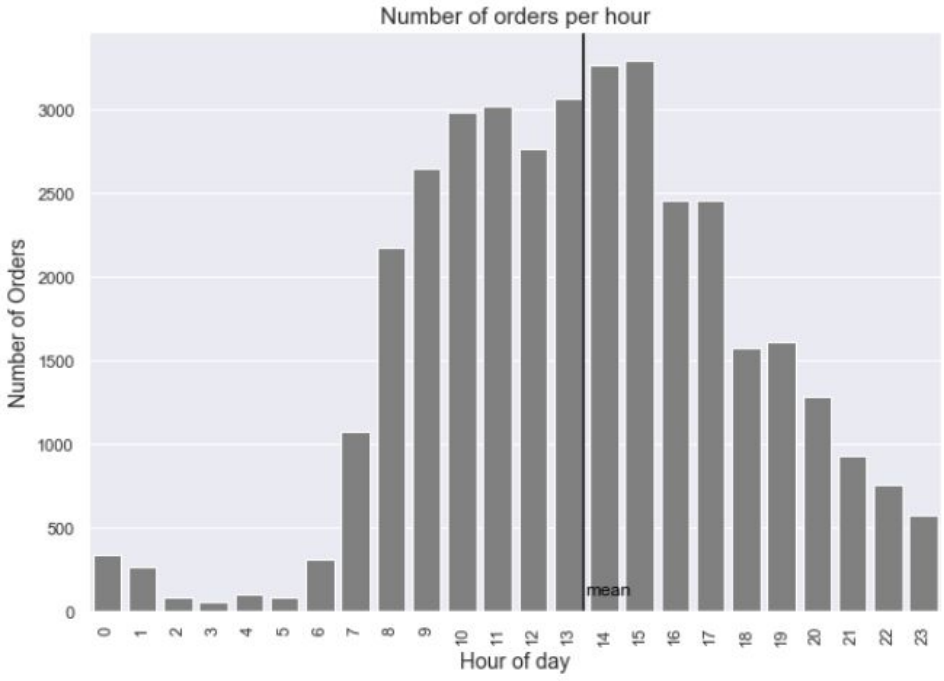
The following scatterplot representation shows a strong correlation between ordered and reordered frequencies. It shows how orders in the shopping cart might be repetitive and customers are loyal to their choice of products.



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.

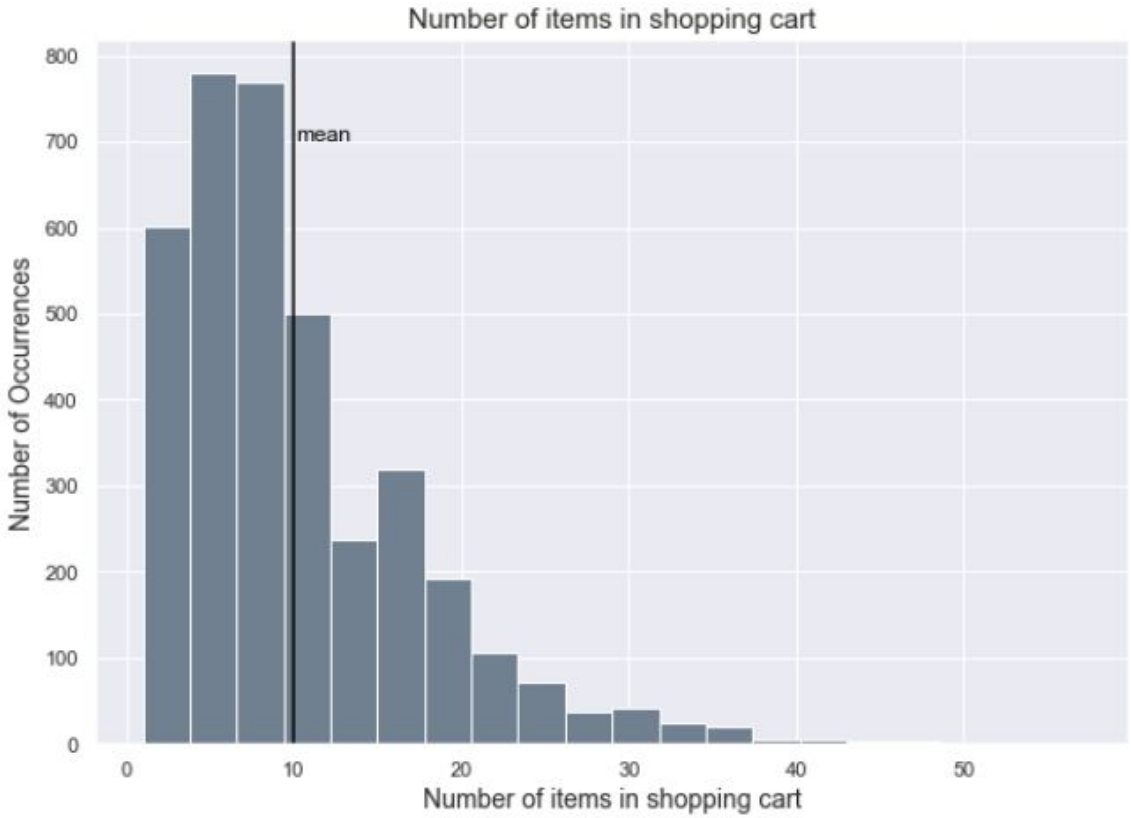




How many items in a shopping cart?

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
```

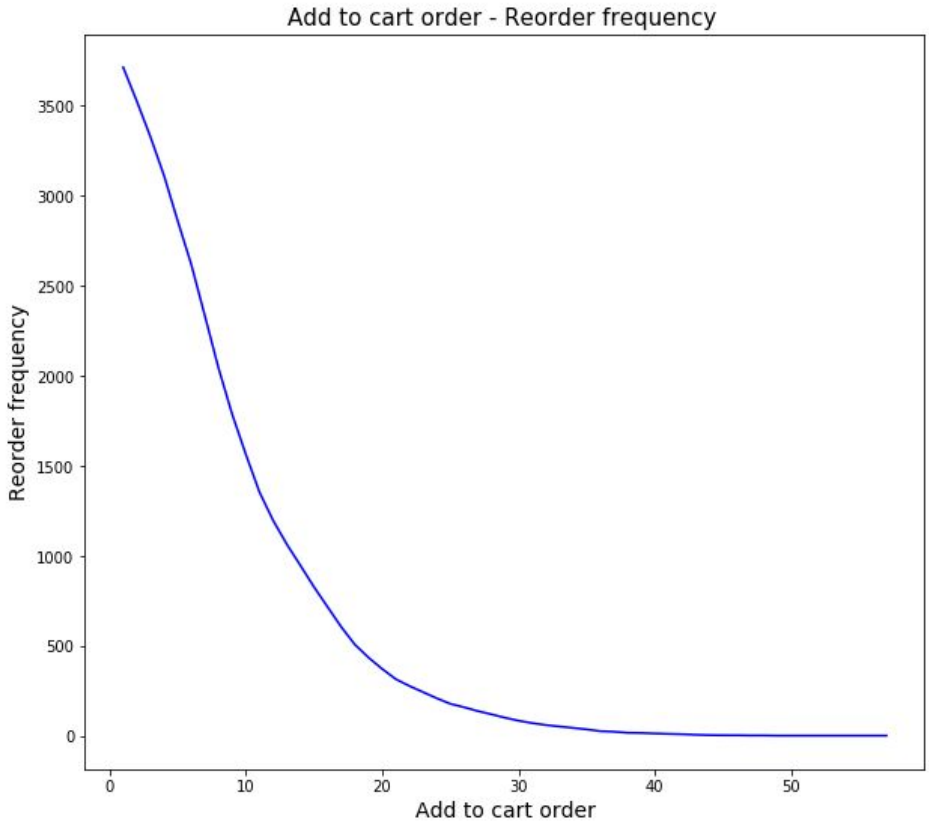


Is there a relationship between the order in which users add their products to the shopping cart and the products they reorder?

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.

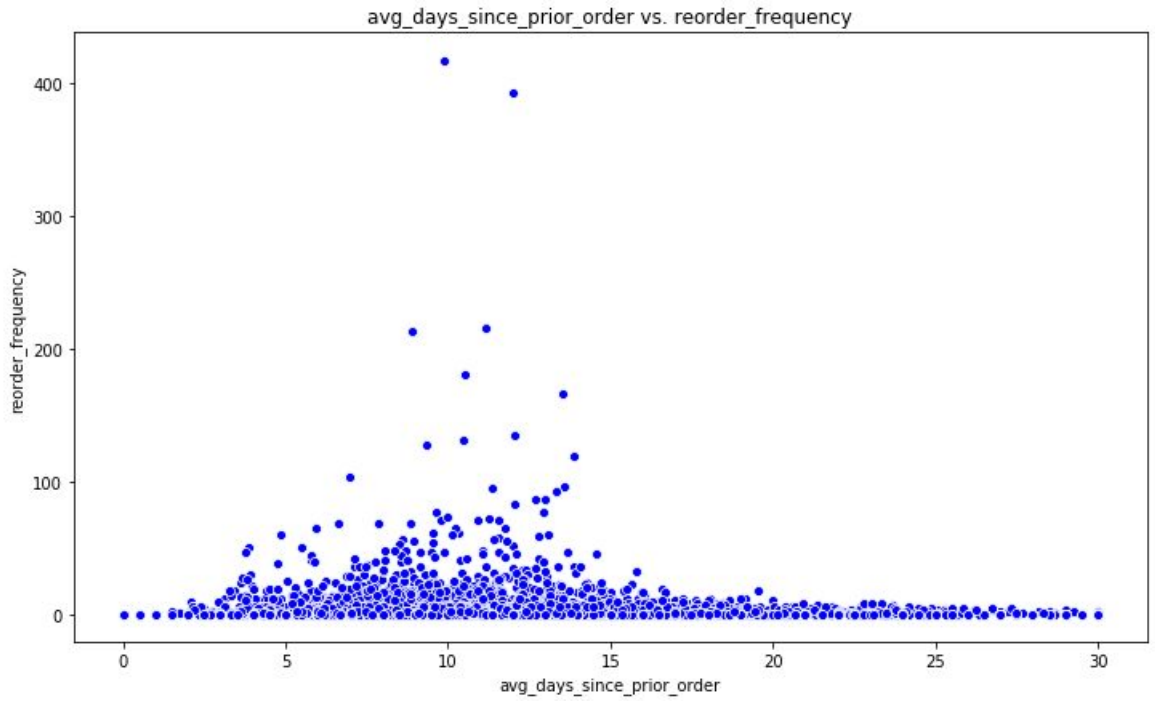
Products that are placed first in the shopping cart have a higher reorder frequency. The reorder frequency diminishes as products move down in their add to cart order.

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 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 26 data features were used to train the machine learning models after considering curse of dimensionality factors:

```
'order_id' 'user_id' 'order_number' 'order_dow' 'order_hour_of_day'
'days_since_prior_order' 'product_id' 'add_to_cart_order' 'reordered'
'aisle_id' 'department_id' 'number_orders' 'number_reorders'
'reordered_ratio' 'avg_days_since_prior_order' 'user_total_items'
'user_total_distinct_items' 'user_average_days_between_orders'
'user_number_orders' 'user_average_basket' 'user_add_to_cart_order'
'user_day_of_week' 'user_order_hour_of_day' 'number_reorders_by_cart'
'avg_reorders' 'product_avg_day_of_week' '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

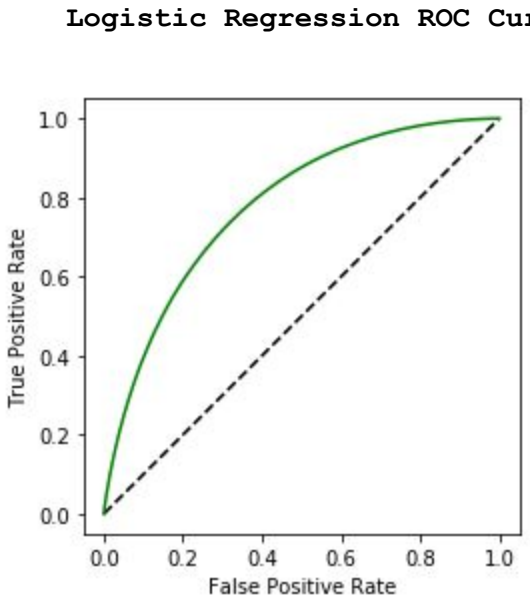
- An accuracy score, ROC curve,confusion matrix and classification reports were used to gauge performance of models.

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

Results

- Logistic Regression Model

  Best Accuracy Score: 72.25% with a C value of .1
  Cross Validation Score: 71.20% in 5 folds



- XGBoost Model

Accuracy score: 78.68% without hyperparameter tuning

Accuracy score: 80.55%
with Best params: learning_rate: 0.1, max_depth:6, silent: False

```
Confusion Matrix:
[[ 772905  210241]
 [  94805 1304506]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.59      0.71    983146
           1       0.77      0.96      0.85   1399311

    accuracy                           0.81   2382457
   macro avg       0.84      0.77      0.78   2382457
weighted avg       0.83      0.81      0.80   2382457

Accuracy: 0.8054596578238348
```

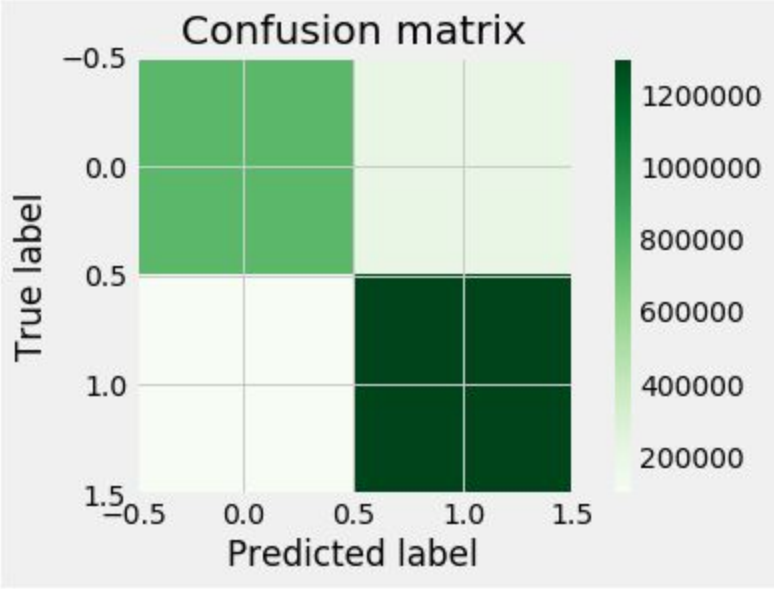
ROC Curve for XGBoost Model

- Random Forest Model - Best model

Accuracy score: 87%

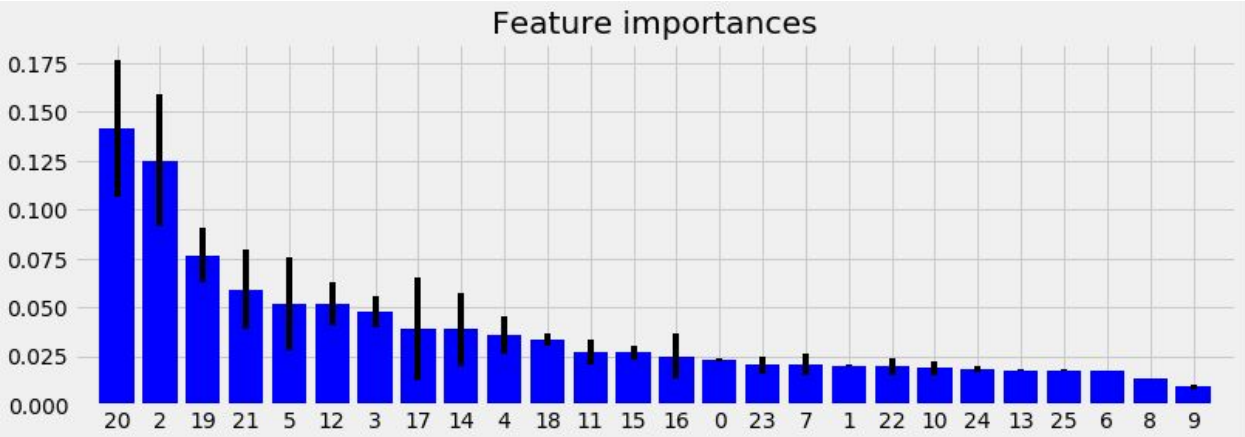
```
Confusion Matrix:
[[ 772905  210241]
 [  94805 1304506]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.79      0.84    983146
           1       0.86      0.93      0.90   1399311

    accuracy                           0.87   2382457
   macro avg       0.88      0.86      0.87   2382457
weighted avg       0.87      0.87      0.87   2382457

Accuracy: 0.8719615925911779
```

Random Forest Confusion Matrix



ROC Curve for Random Forest Model

Feature Importance - Random Forest Model



Important features

1)user_add_to_cart_order
2)order_number
3)user_average_basket
4)user_day_of_week
5)days_since_prior_order

- 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 86% which represents the highest of the three models tested.