Marketing strategies proposed by ‘Cuse Miners

Buffie Holley
Department of Computer Science
Syracuse University
bholley@syr.edu

Bharath Karumudi
Department of Computer Science
Syracuse University
bhkarumu@syr.edu

Donald Redden
Department of Computer Science
Syracuse University
dredden@syr.edu

Gretchen Reeves
Department of Computer Science
Syracuse University
gdreeves@syr.edu

Abstract

The online grocery retailer Instacart has made public a large data set of their customer transactions in the year 2017 for the purposes of furthering the field of data science. The data consists of millions of data instances representing customer transactions. The data will be analyzed using several methods including neural network modeling, apriori modeling, and general data visualization. The immense amount of data makes the processing cumbersome, but several conclusions are able to be drawn, with the purpose of suggesting marketing ideas to the company to increase future overall revenue. By visualizing the raw data, we are able to conclude that shopping occurs steadily throughout the week mainly between the hours of eight in the morning and six in the evening with heavier order volume on Saturday and Sunday. The frequency of orders spikes at roughly weekly intervals with the largest reorder frequency of thirty days. Upon closer inspection using neural network modeling we are able to classify customers into three frequency categories; weekly, bi-weekly, and monthly. Next, the data is joined into complete transactions and analyzed using apriori modeling to determine shopping item associations and add to cart ordering. The market basket associations clearly identify opportunities for ways in which Instacart should consider making modifications to their marketing plans to increase revenue. The particular marketing and staffing suggestions are included in the full report.

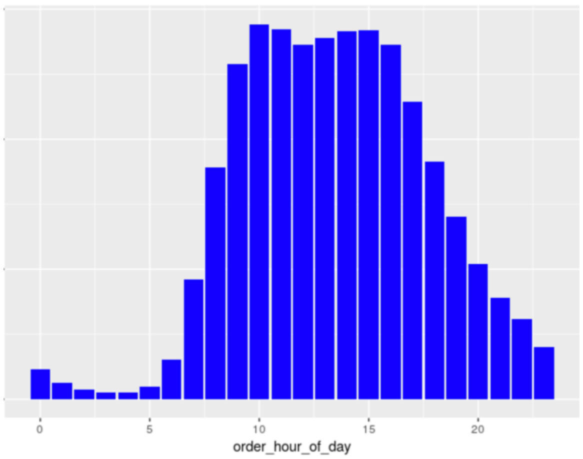
Dataset

Instacart provided the data they collected for almost 200,000 customers shopping in 2017. The data set is provided with over 3,400,000 individual transactions including the items purchased, order placed into the cart, frequency of ordering and many other attributes. Instacart shared this data with the purpose of allowing researchers and students to train and test known and new algorithms for data prediction. By sharing this data, the company is allowing research and conclusions to be drawn for non-commercial use in hopes that it will help to further enhance the techniques used currently as well as help in development of new models for marketing prediction. [2]

Initial Findings

Each order of the 3.4 million transactions provided in the dataset includes the hour of the day the order was placed. Taking a look at the time of each order generates the following visualization.

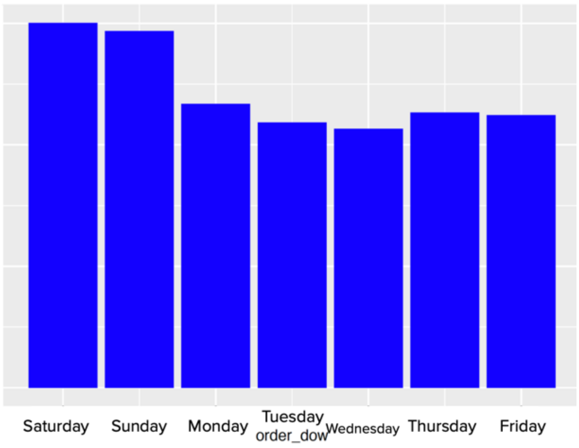
```
orders %>%
  ggplot(aes(x=order_hour_of_day))
+
  geom_histogram(stat="count",fill=
"blue")
```



The majority of orders occur between the hours of 8:00 am and 6:00 pm. Based upon this information it is clear that most grocery shopping occurs during traditional business hours, and it is critical that Instacart have their site running well during these times. Due to the heavy volume of purchasing during the workday it would also be important to have plenty of staff at the warehouse to begin filling these orders. It would be best to do any site maintenance during their lowest hours of business between 03:00 am and 4:00 am as this would cause the least disturbance in customer purchasing.

The orders data also included the day of the week that the orders were placed.

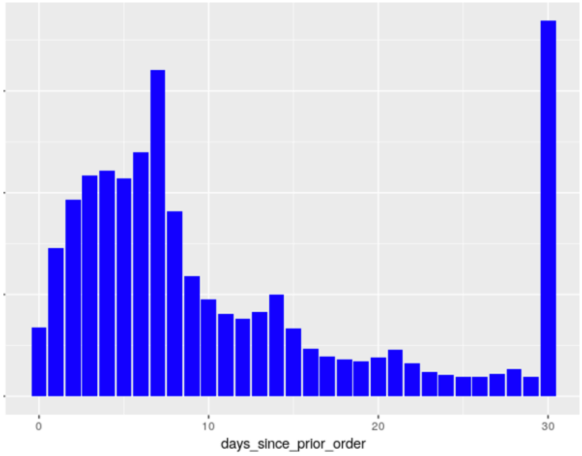
```
Orders %>%
  ggplot(aes(x=order_dow)) +
  geom_histogram(stat=”count”,fill=
”blue”)
```



Looking at all 3.4 million transactions, it is clear that more shopping occurs over the weekend, but only marginally more than during the week. This visualization makes it clear that Instacart should maintain a full staff seven days a week and should have extra staff on hand during the weekend days to fulfill the larger order volume. There does not appear to be a clear day on which there is a low volume of shopping that could use additional marketing to increase sales. Based on this information it would be prudent for Instacart to begin their weekly sales changes on Wednesday so that weekend shoppers are able to peruse the deals before placing their order.

Next let us look at the frequency of ordering by all customers.

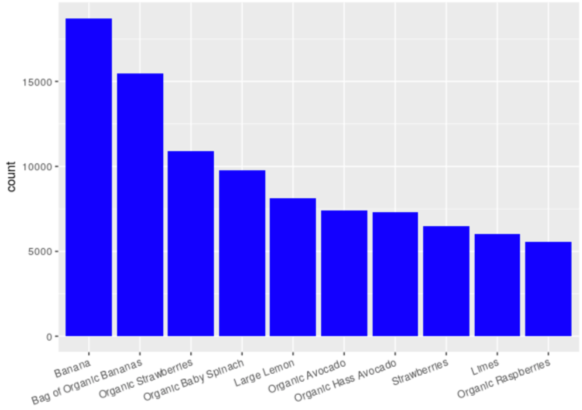
```
Orders %>%
  ggplot(aes(x=days_since_prior_ord
er)) +
  geom_histogram(stat=”count”,fill=
”blue”)
```



An important note about this visualization is that the Instacart data limited the maximum number of days since prior order to thirty days. Meaning that if a customer ordered three months later or six months later they would maintain their customer information, but would be listed as shopping just thirty days later. For the purposes of our investigation and conclusions we will consider these transactions to have happened at the 30-day mark. The data shows three spikes in transaction frequency; the first at 7 days, then 14 days, 21 days, and finally at 30 days since prior order. These spikes indicate that transactions seem to occur on weekly intervals whether they are each week, bi-weekly or monthly. Though is view does not identify which day of the week the transaction took place.

Having exhausted shopping habits involving time, day of week, and frequency let’s look at what products are being purchased the most.

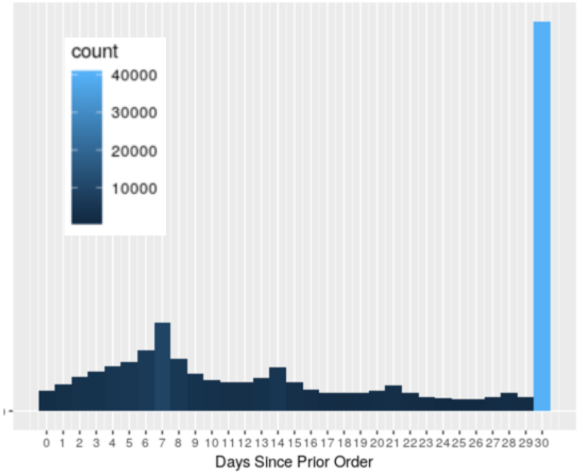
```
Best_sellers <- order_products %>%
  group_by(product_id) %>%
  summarize(count = n()) %>%
  top_n(10, wt = count) %>%
  left_join(dplyr::select(products,
product_id,product_name),by=”produc
t_id”) %>%
  arrange(desc(count))
kable(best_sellers)
best_sellers %>%
  ggplot(aes(x=reorder(product_name
,-count), y=count))+
  geom_bar(stat=”identity”,fill=”bl
ue”)+
  theme(axis.text.x=element_text(an
gle=20, hjust=1),axis.title.x = ele
ment_blank())
```



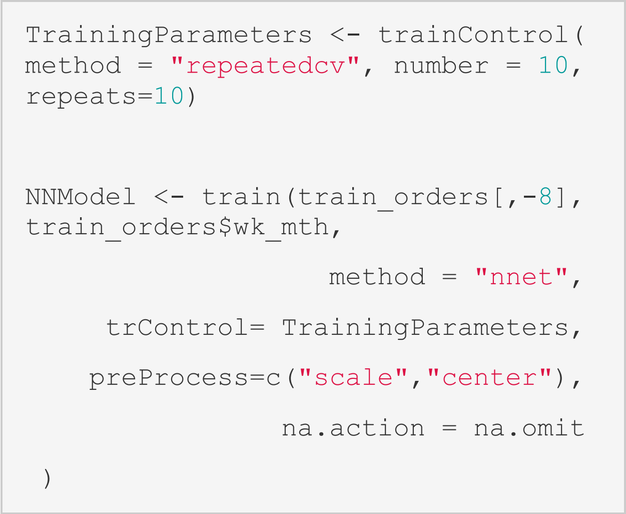
The top ten products purchased are all produce which is interesting given that Instacart is an online grocery store. The top selling item by far is bananas, followed closely by organic bananas. It is important to note that over half of the top selling items are also organic. Instacart actually partners with local grocery stores to fulfill the orders. The company prides itself on quality products with safe fast delivery of fresh ripe produce; even eggs. [1]

Customer Reorder Frequency

Placing emphasis on the frequency of individual customer reorders. How often an individual customer chooses to reorder provides a different perspective of the shopping habits of the dataset.



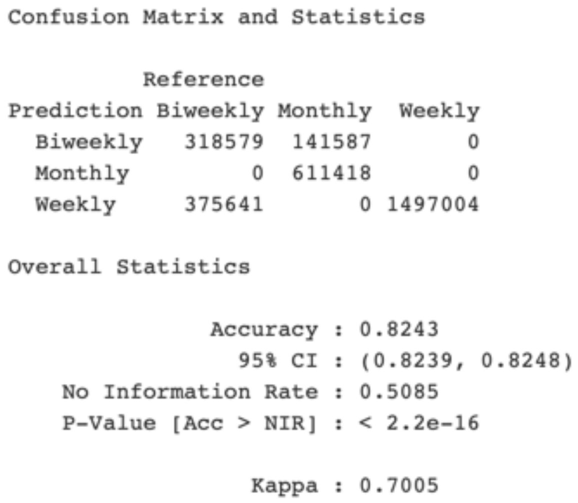
This view shows a clear pattern among the frequency between purchases of individual customers. The spike at 30 days between orders is undeniable as well at seven and fourteen days. These distinct trends led to a desire for us to categorize the customers into, weekly, bi-weekly, and monthly shoppers. Using the data provided to train using a neural network model was not the first choice, but was found to be the best choice. It took over two hours to run the training, but it yielded 85% accuracy. A neural network has the ability to train in supervised and unsupervised situations and the model is designed to attempt a recreation of a human neural network by making connections that a human mind would make. The algorithm for learning using a neural network was introduced in the late 1950s, and since then algorithms have been modified and refined, and in this case, we are using the one integrated into the R studio programming language. The decision to use supervised learning derived from the prior knowledge of the customers and the business problem to solve. The goal of the use neural net was to train and learn a function that given a set of sample data and the desired outputs produced a model to apply future transactions with observable outcomes. The sample data revealed a strong case to classify customers into distinct groups, thus simplifying the structure of future marketing campaigns. The model gives the company the ability to predict future media applications without the need to guess on periodicity and save from materials and resources.



```
TrainingParameters <- trainControl(
method = "repeatedcv", number = 10,
repeats=10)


NNModel <- train(train_orders[,-8],
train_orders$wk_mth,
                                 method = "nnet",
                       trControl= TrainingParameters,
                     preProcess=c("scale","center"),
                               na.action = na.omit
 )
```

Accuracy was used to select the optimal model using the largest value. The final values used for the model were size =1 and decay = 0.1.



```
Confusion Matrix and Statistics

          Reference
Prediction Biweekly Monthly  Weekly
  Biweekly   318579  141587       0
  Monthly         0  611418       0
  Weekly     375641       0 1497004

Overall Statistics

               Accuracy : 0.8243
                 95% CI : (0.8239, 0.8248)
    No Information Rate : 0.5085
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7005
```

This trained model makes it possible for Instacart to identify which of their shoppers fall into each of the categories. The largest category is monthly, so it would be best to put most of their marketing dollars into a campaign for monthly shoppers. This could be in the form of a direct mailer to all monthly shoppers or perhaps an incentive program to encourage customers to shop more often which could increase revenue.

Transactional Data Preprocessing

In order to look at the data from the perspective of individual transactions the data needed to be joined into two data sets the prior purchases and the current purchases. Each of these data sets are joined to include all of the transactional data about the products being purchased, from which department and the names of the individual items. This preprocessing took quite a while, but allowed the transactions to be analyzed in several new ways.

Product Associations

Market basket analysis is a sector of data mining in itself. Identifying which items are purchased together provides retailers with the ability to make suggestions for additional items the customer might be interested in as well as tailor marketing to individuals who have previously purchased items that might be currently on sale.

Market Basket Analysis/Association Rules

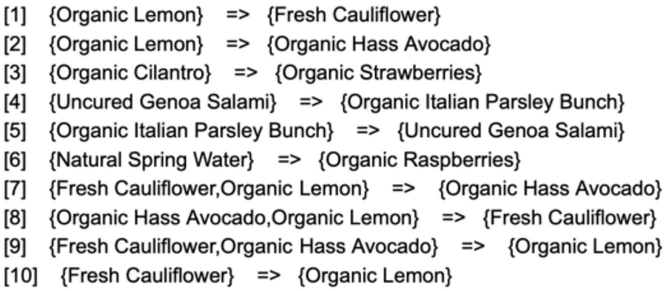
Market basket analysis, also known as association rules in machine learning literature, is used widely in market research to uncover interactions between products. We used the following metrics:

Support of A, P(A), is the probability of a product A appearing in a user's cart. As described above, the vast majority of items for sale by Instacart are purchased very rarely; a marketer could do worse than to recommend the most popular items, especially when lacking user information.

Confidence A→B, P(A∩B)/P(A), measures the probability of a shopper purchasing Item B given that Item A is included in the order.

Lift A→B, P(A∩B)/(P(A)∗P(B)), is confidence A→B scaled by the support of B.

This is designed to find interactions including lower-probability items missed by the confidence rating. We generated support scores for all products, along with confidence and lift scores for all pairs of items appearing in the same cart at least once, partitioned into user clustered and non-clustered data sets (clustering described in more detail below). Parameter tuning included setting a minimum number of orders in order for a product to appear in a recommendation, using different scaling factors (taking P(B) to powers between 0 and 1), and testing whether user clustering improved recommendations.



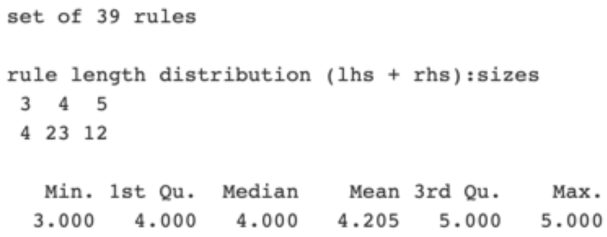
```
[1]  {Organic Lemon}    =>  {Fresh Cauliflower}
[2]  {Organic Lemon}    =>  {Organic Hass Avocado}
[3]  {Organic Cilantro}    =>  {Organic Strawberries}
[4]  {Uncured Genoa Salami}    =>  {Organic Italian Parsley Bunch}
[5]  {Organic Italian Parsley Bunch}    =>  {Uncured Genoa Salami}
[6]  {Natural Spring Water}    =>  {Organic Raspberries}
[7]  {Fresh Cauliflower,Organic Lemon}    =>    {Organic Hass Avocado}
[8]  {Organic Hass Avocado,Organic Lemon}    =>  {Fresh Cauliflower}
[9]  {Fresh Cauliflower,Organic Hass Avocado}    =>    {Organic Lemon}
[10]  {Fresh Cauliflower}    =>  {Organic Lemon}
```

```
#Build Rules with at least confidence of 70% and support of 0.0001555 (500 Transactions).

rules <- apriori(transactions, parameter = list(supp=0.0001555, conf=0.7, minlen=2, maxtime=120))

rules <- sort(rules, by='lift', decreasing = TRUE)

summary(rules)
```

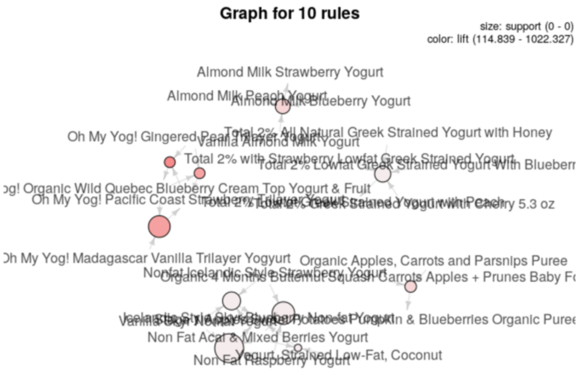


```
set of 39 rules

rule length distribution (lhs + rhs):sizes
 3  4  5
 4 23 12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   4.205   5.000   5.000
```

The Apriori model yielded 39 association rules for the products in the nearly 200,000 transactions provided in the data. Several of the rules included very similar product types. The strongest association was between different types of organic yogurt. If a customer orders pear and strawberry they are likely to also order blueberry. There were several variations of yogurt flavor association rules, but the next strongest association is baby food flavors, and finally flavors of sparkling water. These associations are clearly predictable because these items are purchased as individual servings of items that come in many flavors and a customer who is purchasing for a week or a month would likely purchase several varieties.


Graph for 10 rules
size: support (0 - 0)
color: lift (114.839 - 1022.327)
Almond Milk Strawberry Yogurt
Almond Milk Peach Yogurt
Almond Milk Blueberry Yogurt
Oh My Yog! Gingered Pear Trilayer Yogurt
Total 2% All Natural Greek Strained Yogurt with Honey
Vanilla Almond Milk Yogurt
Total 2% with Strawberry Lowfat Greek Strained Yogurt
Total 2% Lowfat Greek Strained Yogurt With Blueberr
og! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit
Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt
Total 2% Greek Strained Yogurt with Cherry 5.3 oz
Oh My Yog! Madagascar Vanilla Trilayer Yogyurt
Organic Apples, Carrots and Parsnips Puree
Nonfat Icelandic Style Strawberry Yogurt
Organic 4 Months Butternut Squash Carrots Apples + Prunes Baby Fo
Icelandic Style Skyr Blueberry Non-fat Yogurt
Pumpkin & Blueberries Organic Puree
Non Fat Acai & Mixed Berries Yogurt
Yogurt, Strained Low-Fat, Coconut
Non Fat Raspberry Yogurt

These associations would suggest that Instacart should maintain a large variety of flavors for yogurt, and baby food and that it would be a beneficial to advertise new flavors and perhaps to provides some sort of bulk discount or incentive to increase the number of these items purchased in each transaction.

Add to Cart Order

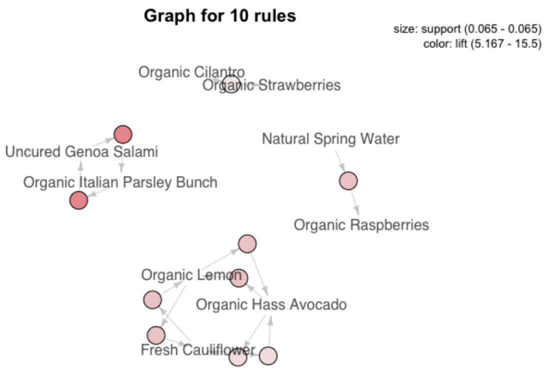
The final market basket analysis completed was to observe the order in which items are added to the cart by customers. This analysis was particularly resource intensive so to produce results in a reasonable amount of time the number of transactions used for creating the rules had to be reduced significantly.

```
cart_rules <- apriori(trainTransactions,
        parameter = list(supp=0.05, conf=0.5, minlen=2, maxlen=5))

cart_rules <- sort(cart_rules, decreasing=TRUE, by="confidence")


inspect(head(cart_rules, 10))
```

Because of the size of the data being processed each time the apriori is run the results are different. In order to get a more accurate picture of the market basket order of purchase association we would need much more computing power than we have in our personal laptops. In one run of the data the following results were generated.



As you can see in this subset of transactions the purchases include produce as well as other grocery items. When you look at the graph you can see that some of the items lead to the purchase of others in a continuum while others have a reciprocal relationship to the order of purchasing. This information makes it possible for Instacart to make suggestions about what a customer should also add to their cart based on the items already there. This idea of add on marketing could easily increase the revenue of the site.

Conclusion

Instacart provided a huge amount of data, and using a few of the datamining techniques learned in class our group is able to make several suggestions for marketing with the goal of increasing revenue to the site. Online shopping allows customers to literally shop at their convenience making weekends a more popular shopping time, so it would be prudent for Instacart to have plenty of staff on the weekends to fulfill the larger order volume. Customers shop between 8a and 6p making it important for Instacart to have plenty of employees for pulling orders as well as be sure that the site is working well during these times and able to handle the larger order volume. This fact also makes it important that marketing materials be delivered electronically prior to 8am, and suggests that site updates should happen between 3a and 4a, to cause the least disruption to customers.

Repeat customers are an important part of every retail business, the neural network makes it possible for Instacart to categorize its customers by frequency of purchase. This allows them to appropriately spend marketing budget dollars. A monthly mailer to infrequent customers would remind them of the convenience and benefits of shopping with Instacart. The large number of weekly shoppers suggests that changing promotions and sale items weekly would be the most beneficial to increasing their total revenue.

Apriori is a powerful data analysis model to identify trends in shopping. This model was able to predict that customers who buy a few flavors of yogurt or baby food would likely buy additional flavors as well. This result suggests that Instacart expand their flavor offerings or institute a bulk purchase discount to increase the number of items of the type in each purchase. This method of analysis also creates order of addition association rules that could make it possible for the site to make suggestions of items for customers to add to purchases. Upselling is an important part of all retail shopping so providing suggestions to the customer increases the likelihood of last-minute additions. Though we were not able to make associations based on the large amount of data provided due to our lack of computing power, it is still an important idea that should be considered by Instacart to increase their total revenue in the future.

References

[1] “Groceries Delivered from Local Stores.” *Instacart*, 2019, www.instacart.com/.

[2] “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on May 1, 2019.