Walmart Labs Take Home Data Challenge

Contact Info

- Name: **Ming Jin**
- From: Columbia University, Engineering School, Operations Research
- Email: mj2940@columbia.edu
- Phone: (646)-644-2368

- Date: 02/05/2020
- Applied position: Analyst Intern

1 Business Problem and Approach

1.1 Background

It is obvious that the most successful companies today are the ones that know their customers so well that they can anticipate their needs. Data analysts play a key role in unlocking these in-depth insights, and segmenting the customers to better serve them. Now we were given a retail data containing its transaction between 2016 and 2017, information of the products it sold, and demographics of 100 household. We would like to explore this retail data and provides insightful findings and suggestions so as to get better understand of its customers.

1.2 Business Problem

With the final goal of understand the customers better and boost sales, we would like to resolve the following problems:

- How does the customer engagement change overtime?
- How can we classify these customers based on their engagement?
- What factors influence customer engagement?

1.3 Approach

For this project, I will first use Exploratory data analysis to measure customer engagement from different aspects (time, product categories, etc.). Then I will use cohort analysis to get a graphic understanding of the customer acquisition and lifecycle. Finally, I will use RFM segmentation to classify customers and analyze the factors that affect customer engagement.

2 Assumptions

To simplify our problems, combined with some general analysis of these data, I made following assumptions: (Our analysis may not hold because of the wrong assumptions.)

1. There were not fierce competitors or major events that might change customers' behavior significantly during this period and area.
2. Involuntary churn of customers has been excluded from our discussions (for example, customer move to another country)
3. Current data is accurate and enough for supporting us to draw the conclusion.
4. The ignorance of marking and pricing factors don't affect the result.

3 Customer Engagement Analysis

3.1 General characters of this retail's business

Based on exploration analysis, we find some key insights as below:

- The contribution of different regions and product categories are unbalanced. Store in east region and food product contribute most respectively.
- There is no significant increase or decrease trend of this retail's total revenue overtime.
- Customers behavior are seasonal, especially for food and non-food department

These findings provide us a foundation of analyzing this retail's customer engagement

3.2 Customer Engagement over Time

3.2.1 Customer Acquisition and Retention

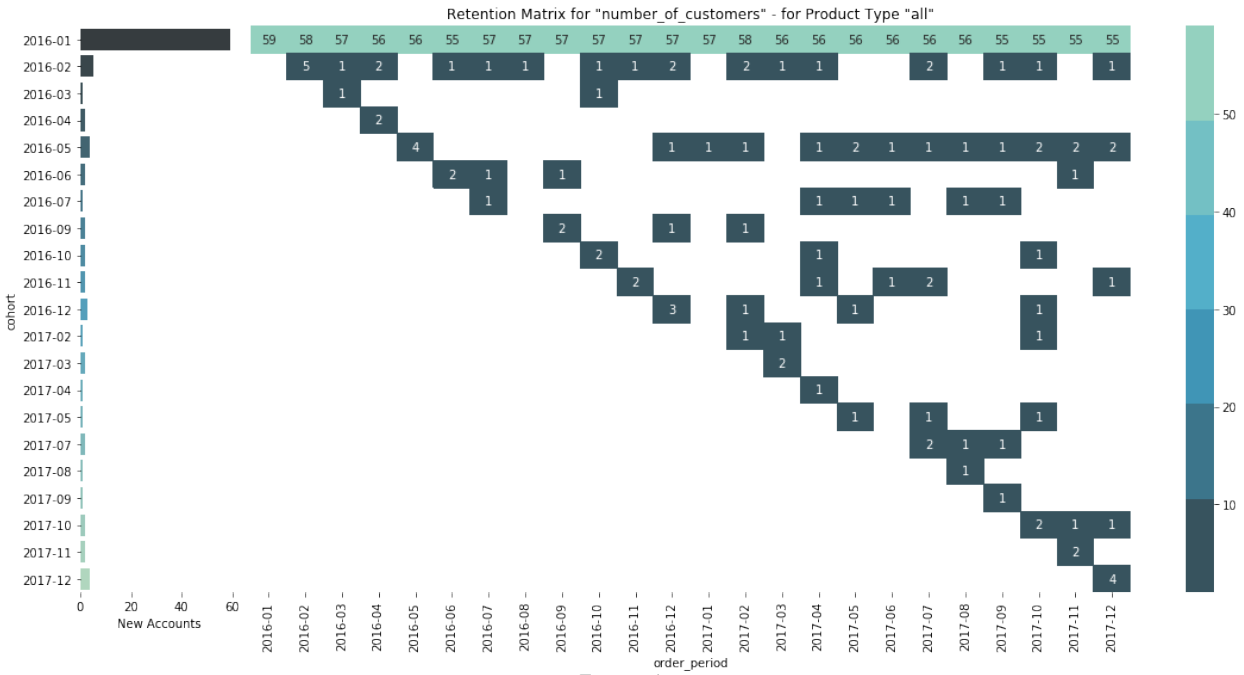

Figure 1

As shown in the above picture,

- The left side represent number of new customers acquired in each month. More than half of these customers are acquired at or before Jan, 2016. Then, this retail gains customer gradually by single digits each month.
- The customers gained at or before Jan, 2016 show loyalty toward this retail, while customers gained afterwards tend to churned

3.2.2 Does customer spend more or less?

We choose **average customer spend value per month** as metric to evaluate the customer engagement at first glance. Specifically, to better represent customers' behavior over time, we set number of months since they were acquired as variable. According to figure "'the retention matrix of 'avg_customer_spend_value' "for all product type) and supplement figure "the retention matrix of 'avg_customer_bought_items'" for all product type), after exclude some outliers, customer seems to spend more over time. In detail, customers spend more on food as they become more settled, especially for meat, beverage (Figure 2,3). While customer spend less

on produce as time goes by (Figure 3). This is reasonable because they need produce products when they first moved to some place. (However, as the actually acquisition date of these customer are not guaranteed, these deductions might be wrong.)

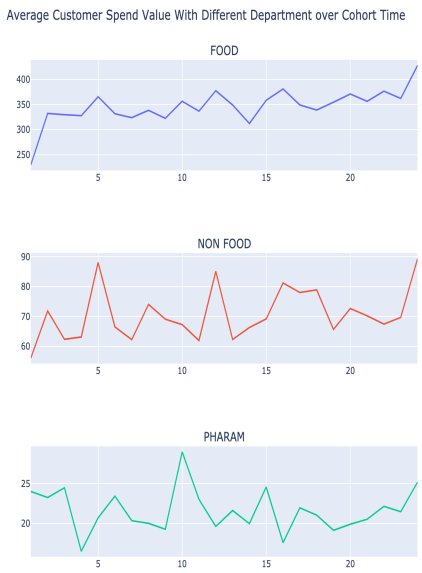

Figure 2

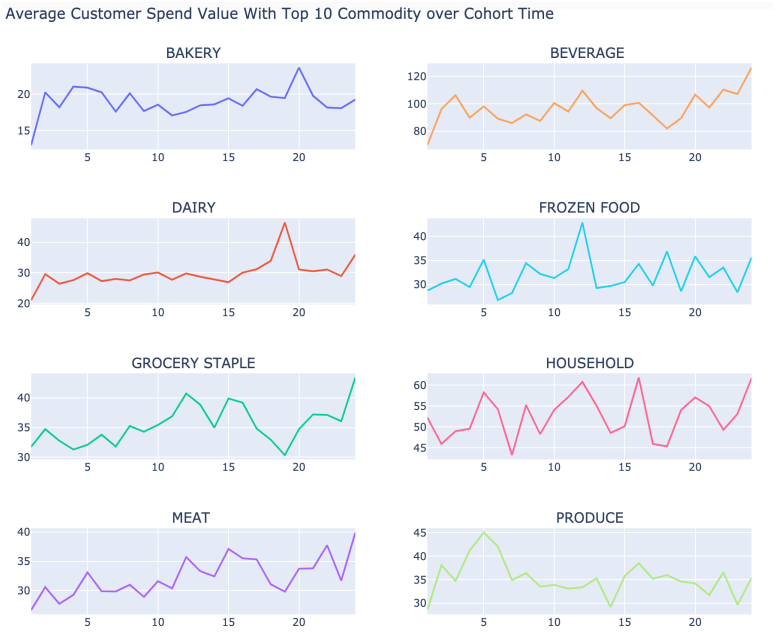

Figure 3

3.3 Customer Segmentation

3.3.1 RFM Score

RFM Segmentation is classical method to analyze customer value. RFM stands for recency, frequency, and monetary. These three metrics provide a well description of customer engagement. **Recency** measures the time between when customer last ordered to today. **Frequency** measures how many total orders the customers had, and **Monetary** is the average amount they spent from those orders.

In this analysis, we select the maximum date of transaction overall as 'today', and calculate recency in days. Also, we only analyze the department of food since they contribute to majority transactions.

The results are as follows:

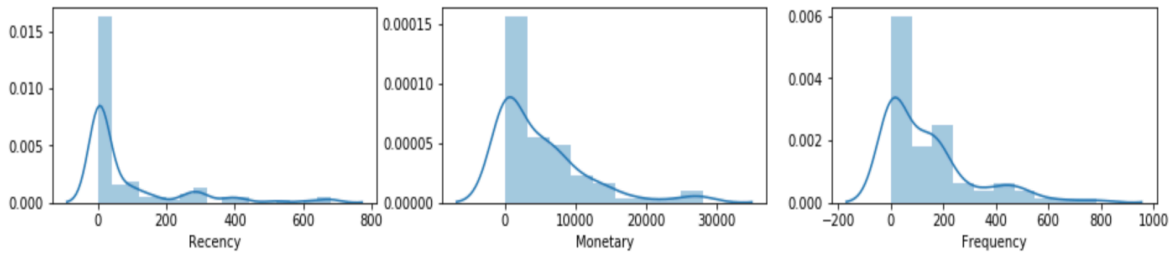

Figure 4 The Distribution of customers' RFM Score

We can figure out that a good number of customers come to the retail very often, spend a lot and are still active recently, in other words, they are very engaged. While some other customers just come across and never step in this retail again, these are so-called lazy or even lost customer.

3.3.2 Customer Segmentation

We then clustered these customers into 3 class based on the RFM score by kmeans algorithm. The result shows as below:

Cluster	Recency Mean	Frequency Mean	Monetary Mean	Count
0	4.0	240.0	8773.0	56
1	250.0	6.0	131.0	33
2	15.0	6.0	117.0	9

Figure 5 The Result of RFM Segmentation

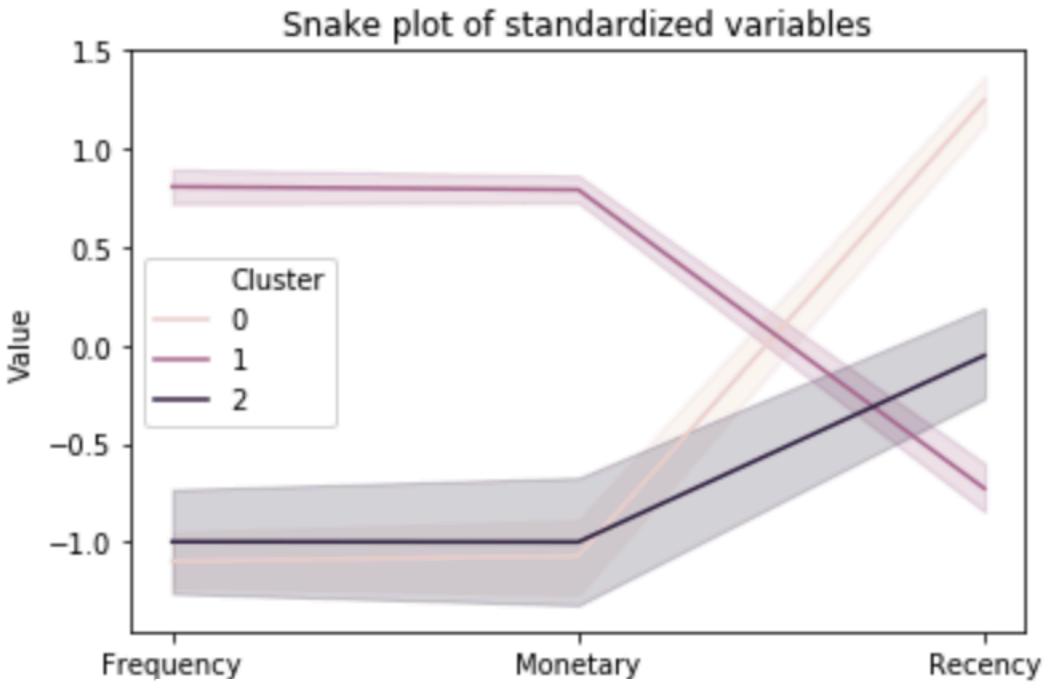

Figure 6

According to the result, we can label our customer into 3 type:

Active/Loyal	Lazy	Lost
recently has been to the retail	it has been sometime but not very long since its last visit time	it has been a long while since its last visit time
goes very often	visit few times	visit few times
spend very less	spend very less	spend very less

The active customers are our most important customers. They keep making contributions to the retail' revenue and generate more profits within one purchase. The lazy group are customers we should focus on currently to transfer into active/loyal customers. While the customers nearly never visit this retail again can be treated as lost, and there are few profits to put efforts to save them.

3.4 Factors Affect Customer Engagement

I then analyze the influence of household demographic factors on customer engagement based on the label obtained at 3.3. While the number of customers in lazy group are too small, I at this time only analyze the most and least engaged customers, namely active group and lazy group.

Based on the analysis, we find that:

- people with loyalty flag tends to be more engaged
- there is no considerable difference of marriages
- single males are more engaged
- people with lower income, especially under 35k are more engaged

Due to time and data limited, these findings are rather constraint.

4 Conclusion

From above analysis, we can conclude that:

1. The business of this retail is stable during 2016-2017
2. Customers tend to spend more on food and spend less on produce products as time goes by.
3. Customer engagements are influenced by loyalty program, salary, etc.

Also, we provide a cluster segmentation model for this retail which could be used for marketing campaign , customer management and other business strategies.

5 Next Step

5.1 Data Aspect

- Have more amount of household data
- acquire some macroeconomic data
- have pricing and marketing data
- have customer service data

5.2 Model and Analysis Aspect

- train a predict model based on customer segmentation label
- further analyze factors importance and correlation based on models