Instacart Shopping Data

Exploratory Data Analysis

Sava Dashev

We investigate the distribution of users according to categories prior order, train order and test order.

The total number of users is 206,209; this is the number of users in prior set.

The number of users in train and test are 131,209 and 75,000 respectively. We see that the number of train and test together is equal to the total number of users.

So, each user is in two sets: prior and train or prior and test. We can make that conclusion by using the formula of total number of elements in an union of two sets. In that formula, we must subtract the number of the intersection of two sets, in our case: train and test. But train and test together add up exactly to total, so the intersection must have zero elements.

When are orders made?

The data has variable for the day in which the order is made. We plot that data. We see that the highest number of orders are in days 0 and 1, followed by 2 and day 5. It looks like 0 and 1 are the weekends.

The other variable is for hour of day of the order. We look at the distribution of that data. For the week, the highest number of orders is 10:00 am and there is secondary pick around 4:00 pm. But is that the case for each day of week?

We break down the week by day and hour and investigate. We see that days 0, 1 and 6 have different type of distribution than days 2 to 5.Days 2 to 5 have more weight and shape the distribution of hours of order for the week. The max of day 0, 1 and 6 come at different time and averaging all hours gives the distribution with two pics.

How many orders per user?

For each user, we have between 4 and 100 orders.

How many items in order?

The mode of items in order is 5. This is right tail distribution.

How many days since prior order?

The distribution of orders grouped by days since prior order has multiple local modes. There are also 0 days since prior order, and the local modes are around 7, 14, 21 and 28 days. The pic at 30 days is artificial. The data is collected in a way such that all orders made more than 30 days since prior are kept at 30. We can make that conclusion by looking at which day of the week that order is made, and it not always corresponds with the day of the prior order. For example, if prior order is at day 2, the 30 day after must be day 4, but in data all days can occur.

Well, for each user, what is max number of days since prior order?

30 days prevail, so we must remove these data points.

When we remove these data points, we see that max of the range of time since prior order varies from 0 to 29 days, with the same multiple of 7 days local modes. Interestingly enough, the 14 days have more users max and 21 even more.

And some users like to shop more than once the same day. We see that more than 30,000 users placed more than one order the same day.

At the other end of the distribution, is the small number, about 3.5% of users that place no more than one order a month.

What about items?

Most popular items: Bananas: more than 490,000 orders. The difference between 1st and 30th order are about 6.5 times less.

Which item is placed first in basket?

The most popular products not always are put in basket first. Whole milk, for example is placed first in basked less than 10k times, and it has more 140,000 orders. The same with limes - about 10k first, and total of 146,600 orders.

What are most popular departments and most popular aisles?

The most popular isles are fresh fruits and fresh vegetables, packaged vegetable fruits. The number 10 aisle is bread.

The most popular department is produce and dairy eggs, the least popular: bulk.

Let take a look at reorders.

The ratio of reordered items. It varies greatly.

The biggest reorder ratio is for Raw Vegetable wrappers: 0.94. The 30 highest reorder ratio is 0.86.

The number of items reordered with ratio more than 0.2 is 38,392.

We can investigate the reorder ratio by day of week, hour of day, aisle and department.

Hypothesis:

We

Hypothesis: The differences between the reordered proportion are statistically significant if we have the difference in reorder proportion no less than 0.2.

We checked lower values of the difference, but in most cases, we failed to reject the null hypothesis.

To check the hypothesis, we wrote a function, that calculates the z-value of the the check and returns that value, error and statement of failing to reject or reject.

Then, we choose random sample of products and choose products with the desired difference between reorder proportion. We used products with small sample size. With product with larger sample size, the statistical significance will increase.