Data Story 6: Instacart Customer Segmentation

George I. Hagstrom

Assignment

You have been given a dataset consisting of several files describing customer purchases which took place at Instacart, an online grocery delivery service, during a 365 day period prior to 2020. Your goal on this assignment is to perform a **customer segmentation analysis** to understand the different types of customer behavior exhibited by Instacart customers. You should use dimensionality reduction, cluster analysis, and any other tool you see fit to find and visualize customer segments at Instacart.

Your data consists of a partially processed dataset that Instacart posted to kaggle for a prediction competition. Here we are using the dataset for a different purpose.

Here is a description of the data files provided:

1. user_features.csv This file consists of features generated from the original kaggle dataset.

- `user_id`: this is a number uniquely identifying each user
- `packaged cheese` to `frozen juice`: Columns 2:135 consistent of the name of food categories (called `aisles` in the instacart data). The value in each entry of these columns is the number of items from that category ordered by each customer throughout the entire year. This count is a little imprecise because it does not differentiate the quantity of items per order (which is unavailable in this data).
- `Saturday` to `Friday`: Columns 136:142 correspond to days of the week. The values correspond to the number of orders each user made on a given day of the week.

2. Official Instacart data:

- `aisles.csv`: This file identifies the aisles (food categories) that correspond to each `aisle_id`
- `departments.csv`: This file identifies the departments (a very rough category of product) that correspond to each department id. Departments and Aisles form a hierarchy, each food item is contained within an aisle, and each aisle is contained within a department.

- `products.csv` contains information on each product, including the product name, the aisle, and the department
- `orders.csv` contains high level information about each order, including the `user_id`, the `order_id`, the order number of that user (whether it is the user's 1st, 3rd, or 10th order etc), the day of week and hour of the day the order was made, and the number of days elapsed since their previous order
- `all_order_products.csv` contains detailed information on each specific order, including all the products in that order and the order in which those products were added.

Note on Data Size

There is a lot of data for this assignment. If your computer lacks memory to handle the computational tasks you may subsample- in this case I recommend subsampling the pre-defined feature dataset that I provided. If you do this make sure to note it in your assignment.

More Information on Customer Segmentation

Customer segmentation is an important data analysis task that takes place at probably all e-commerce entities. The goal is to find groupings of customers based on shared characteristics. In typical cases, datasets will include a wealth of demographic information on each customer, as well as data on their behavior on other websites (there is little privacy on the internet), but here we only have data on customer ordering behavior.

A typical framework for segmenting customers by ordering behavior is called **RFM**, which stands for **Recency**, **Frequency**, and **Monetary Value**. The dataset contains direct information on frequency and and indirect information on monetary value. No prices are available for the products, but we do have information on the total number of different products ordered, which is an imperfect proxy for monetary value. There is not enough information to confidently calculate recency. Here you should try to incorporate information on frequency and the total number of items ordered into your analysis in addition to information on the types of behaviors seen in clusters.

There are a number of useful references on customer segmentation- I suggest the wikipedia article on market segmentation and the paper accompanying the UCI Machine Learning online retail dataset. Click here for the UCI Page, and read the pdf posted to the brightspace page for more information.