# Clustering Customers & Predicting Purchasing Behavior with InstaCart Data

Team 24: Marion Cassim, Nicole Flowerhill, Eugene Huang, Dongwon Jang, Pruek Vanna-iampikul

## Intro

The emergence of COVID-19 and its subsequent impact on consumer behavior created a shockwave through the global retail industry, creating winners of those who prioritized digital and omnichannel funnels, and losers of those that prioritized a physical retail footprint.

To succeed in this new environment, the use of data analytics is necessary to better understand customers while ensuring that they have the right product mix, sufficient inventory, and appropriate levels of staffing to guarantee an experience that will help generate customer loyalty.

## Problem Definition

This project seeks to determine the possibility of segmenting grocery customers based on the items that they purchase and repurchase, the order that they purchase items, the length of time between purchases, and the time and day that a customer makes a purchase. We also seek to predict whether a customer will repurchase an item or similar items in the future.

## Dataset

We plan to use Instacart's Online Grocery Shopping Dataset from 2017. It contains anonymized data of over 3 million orders from over 200,000 users, covering between 4 and 100 orders per user. The dataset has order information, product information, department and aisle information per product, the order in which products were added to cart and whether the product was ordered by a user previously. Given the dataset size and limitations of our personal computers, we plan on using a randomized subset of the data.

## Methods:

For unsupervised methods, we can use clustering to segment customers, where data points that have similar features are grouped [Kashwan]. Our study can cluster based on the features mentioned in the problem definition. We will apply the k-means algorithm, a time-efficient method which starts with an initial k amount of clusters, finds centroids of the clusters, and reiterates until convergence [Kashwan], [Hwang]. The results of k-means can initialize a Gaussian Mixture Model, a soft clustering method with the additional parameter covariance to cluster the model more flexibly [Cikovic]. GMM is computationally expensive and may take issue with high-dimensional data, so it may require preprocessing for dimensionality reduction [Cikovic].

For supervised approaches, we could use a logistic regression and/or a decision trees approach. Logistic regression would output binary data that we can use to predict customer satisfaction [Lawson] or to make product recommendations for a customer based on their previous transaction information [Peng], [Huang]. Logistic regression can recommend future purchases by fitting the engineering features from the dataset into the binomial curve. Since the purchase information will only contain the product information, we plan on combining and engineering new features out of existing information to provide a more meaningful product prediction. With the decision trees

method, we could measure splitting performance through Information Gain and other metrics. [Khalili-Damghani].

Potential Results:

For clustering, we will use the elbow method and distortion score to determine the best number of clusters. To evaluate cluster separation, we will use TSNE plots. To evaluate the prediction models, we will generate confusion matrices and compare accuracy, sensitivity, and specificity between the models.

Gantt Chart:

https://github.com/instakartproject/cs7641/blob/main/CS7641_Fall2022_Group24_GanttChart.xlsx

Contribution table:

| TASK TITLE | TASK OWNER |
|---|---|
| Project Proposal | All |
| Introduction & Background | EH |
| Problem Definition | EH |
| Methods | NF, MC, DJ, PV |
| Potential Results & Discussion | EH |
| Video Recording | All |
| GitHub Page | EH |

Citations:

Briedis, H., Kronschnabl, A., Rodriguez, A., & Ungerman, K. (2020). *Adapting to the next normal in retail: The Customer Experience Imperative*. McKinsey & Company. Retrieved October 4, 2022, from https://www.mckinsey.com/industries/retail/our-insights/adapting-to-the-next-normal-in-retail-the-customer-experience-imperative

Cikovic, K. F. (2020). Customer Profiles in the Antiques and Collectibles Industry in Croatia Using Gaussian Mixture Model Clustering: An Empirical Study. *Economic and Social Development: Book of Proceedings*, 148-156.

*Covid-driven recession impact on retail industry*. Deloitte United States. (2020). Retrieved October 4, 2022, from https://www2.deloitte.com/us/en/pages/consumer-business/articles/retail-recession.html

Huang, B. C. L., Xiang, Y., & Huang, Z. H. (2014). Use Logistic Regression to predict user'behaviors. In Applied Mechanics and Materials (Vol. 651, pp. 1695-1698). Trans Tech Publications Ltd.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)* (pp. 21-34).

Jiang, P., Zhu, Y., Zhang, Y., & Yuan, Q. (2015). Life-stage prediction for product recommendation in e-commerce. In Proceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1879-1888).

Kashwan, K. R., & Velu, C. M. (2013). Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering*, 5(6), 856.

Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing*, *73*, 816-828

Lawson, C., & Montgomery, D. C. (2006). Logistic regression analysis of customer satisfaction data. Quality and reliability engineering international, 22(8), 971-984.

Dataset:

"The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on <date>