Sam Rheault January 10, 2019

General Assembly DATR1023 Final Project Proposal

For my final project I will explore Instacart order data. Instacart is an online service that connects customers to personal grocery shoppers, allowing customers to select products through a web application to be collected and delivered by a shopper. Instacart has open-sourced some data from three million Instacart orders, which it uses to make predictions about customer preference and activity.

From the large dataset we have features describing customers’ ordering habits (total number of orders, cart size, days since last order, order hour of day and day of week) and features describing products’ popularity (total number of orders, ratio of reorders to total orders). I will be answering two questions: first, given a user-product pair, what is the likelihood that the user will order the product in the future? Second, given a product, what is the likelihood that the product will be reordered?

The approach is to use K-means clustering to group user-product pairs based on the users’ history with products, and then K-nearest-neighbors classification model to predict what cluster a user-product pair belongs in and thus whether that user will reorder that product.

I predict that users who have a history of ordering a particular product in the past will have a high likelihood of ordering the product again in the future. Similarly, I predict that products that are often reordered will continue to be reordered in the future.

This model is limited because it does not make use of all of the users’ ordering history that is available. If I had the time and expertise, I would like to use a similar approach over all of the order history data to hopefully give a fuller picture of future order predictions. Additionally, I did not use PCA to reduce dimensionality in my Kmeans models, doing so may have made the models more accurate. I also would have liked to have been able to see the value of the cart or price of the order, but that information is not available to us.