Katin Lind
10.3.22
Springboard DSC Capstone 2

What products is an Instacart shopper likely to purchase on this visit?

For this capstone project, I will act as though I am supporting Instacart, analyzing a dataset shared via a Kaggle competition here: https://www.kaggle.com/c/instacart-market-basket-analysis/data.

Problem Statement

How can Instacart lower the ratio of labor:revenue over the course of a quarter by increasing sales per order, with the use of effective recommendations based on predictions of customer purchase behavior?

Context

Instacart uses models to recommend products that customers are likely to buy again, try for the first time, or add to their cart during a shopping visit. Specifically for this competition, the Instacart team asked for support to be able to better predict which products any given user is likely to re-purchase with their next order.

Criteria for Success

- Produce predictions for which products shoppers will buy in the future based on what they have bought in the past. Specifically, develop a model that could repeatedly make this prediction for any given individual user. At least, predict whether a given item will or will not be repurchased by a given user.
- Strive for predictions involving clusters of products that have been purchased together in the past. And predictions that include a temporal element such as a likely time of day, day of week, or duration since last purchase, for any given repurchase.
- Include analysis of how strong predictions are or aren't so that accurate projections of changes in sales can be made, assuming effective recommendation systems within the application interface where repurchase predictions will be applied.

Scope of Solution Space

Although I don't have the opportunity to collaborate with Instacart employees, I can assume at least two reasons why predicting repurchase behavior is a meaningful goal: If recommendations successfully lead to more sales, this means Instacart users are growing trust and increasingly choosing Instacart over other shopping options. Furthermore, given the Instacart model of grocery delivery, the per-sale profit:labor ratio likely increases with larger orders, since larger orders only take a small amount of extra time to prepare and the same amount of time to deliver when compared with small orders. Larger orders likely enable Instacart to charge smaller fees, which may further increase user satisfaction and loyalty.

Constraints

- The problem statement lacks measurability because I lack domain knowledge in the types of quantitative sales goals that make sense in the context of a business like Instacart. I'd like to do additional research in order to better understand what would constitute a realistic target sales outcome based on a fine-tuning of product recommendations such as this.
- The existing dataset does not include information about which products were recommended during prior shopping experiences. This information would allow for more robust model development, stronger predictions at this stage, and the ability to gauge how much a given models' prediction is likely to impact sales in the future. This compelling objective will need to be truncated in the context of this project.

Stakeholders

I imagine that folks within Instacart most interested in this research would be a VP of Sales, Chief Marketing Officer, and VP of Product (app) Design. I learned about Instacart's organizing structure here: https://theorg.com/org/instacart/org-chart

Data Sources

Data is arranged into multiple csv files:

- "Aisles" and "departments" tables organize types of products.
- "Order_products" contains each order, items included therein, and the user who made the order.
- "Orders" indicates whether each order belongs to sets designated as "prior," "train," and "test." These were set up as such for the purposes of the kaggle competition, and it may be wise to reproduce that structure in my work.
- "Products" contains information about each product. This table would link the aisle/department and order_products tables.

Preliminary Plan

I will begin by exploring and cleaning data in case there are order records with aberrations that may lead them to throw off predictions. I'll need to understand well connections among the tables and join them accordingly.

I'll subsequently look to see if there are patterns in products that are often or rarely bought together. This could inform conditional predictions down the road (if x product is in the cart, then y becomes more/less likely to be repurchased).

However, I also want to understand these product associations early on because they could support feature engineering and grouping products for more efficient and effective modeling. For example, one might assume that frozen pizzas can be grouped when making recommendations or that organic frozen prepared foods are connected, but only analysis of how products truly relate will enable groupings. It may be, for instance, that a common pattern is buying craft beer and any interchangeable type of organic frozen pizza but that other types of frozen foods ought not be combined when moving forward with further modeling. I'll need to understand any such nuanced patterns in the data.

Once products within orders are engineered logically, I can work to develop models that will enable the classification of products as those that each user likely will or will not repurchase. A repurchase could be predicted to happen flat out and under varied conditions. This can be followed by multiple-feature classification with an analysis of clusters of co-purchased products or product types. Returning to adjust feature engineering may be useful after analyzing the results of initial modeling.

I'll test models against overfitting and other problems and, once selecting the most effective model, use it to add to existing data with predictions for future re-purchases. These predictions will need to be explained with regard to what they do and don't tell us, with what degree of certainty, about shopping behavior.

All of this will be done in jupyter notebooks that I have pre-arranged using a cookiecuter template and summarized in a project report and summary slides. Everything will be saved here: https://github.com/fractaldatalearning/Capstone2.