Instacart Product Prediction

Step 1: Define the Problem

* Instacart customers make many orders and reorder certain products frequently. In order to better market to their customers, the goal is to recommend to Instacart customers products that they might want to purchase based on past orders.
* Build a recommendation engine based on customer order-products data.

Step 2: Gather the Data

* Orders.csv : gives information on orders
  + order\_id
  + user\_id
  + eval\_set (which data set the order belongs to, the last orders are either in train or test sets)
  + order\_number (which order for the customer is it, i.e. is it the 1st order, 2nd, etc.)
  + order\_dow (day of the week order placed on, Sunday = 0)
  + order\_hour\_of\_day (hour order is placed on)
  + days\_since\_prior (days since previous order)
* products.csv : gives info on products
  + product\_id
  + product\_name
  + aisle\_id (foreign key, what aisle product is in)
  + department\_id (foreign key, what department product is in)
* aisles.csv : gives info on aisle names and aisle\_ids
* departments.csv : gives info on department names and department\_ids
* **order\_products\_prior : dataset to train model**, orders that customers have previously placed
  + order\_id
  + product\_id
  + add\_to\_cart\_order (order in which product was added to cart, ie was it the first product added to the cart)
  + reordered (1 if yes, 0 if no)
  + can add more variables as needed to train model
* **order\_products\_train : dataset to validate model**
  + order\_id
  + product\_id
  + add\_to\_cart\_order
  + reordered
* **order\_products\_test : dataset to test model** and make predictions on products based on order\_ids
  + order\_id
  + product\_id
  + will make new column called predicted\_product\_id that will show predictions, compare this column to product\_id column
* Step 3: Why is business interested in this problem? What other insights are they expecting?
  + Learn customers’ shopping habits
  + What products are worth keeping in constant stock? What products aren’t?
  + Who are the loyal customers?
  + What are the busy times (time of day, week, and month)? 🡪 could help make sure to keep a fresh stock at those times
* Step 4: Possible reasons for customer purchasing behavior that I can explore:
  + Would think customers mostly shop on weekends or after work on weekdays
  + Certain customers probably reorder the same things consistently (is there a correlation between user\_id and reorders or user\_id and product\_id?)
  + Certain products are probably reordered more often (like maybe fruit)
  + Which products are added to cart first. Certain products are probably placed in cart first (is there a correlation between location in cart and product id? Between location in cart and reorder? Between user\_id, location in cart, and product\_id?)
* Observations
  + Organic food, fruits, and veggies are the most popular items ordered, with bananas being the highest
  + most people buy on Sunday afternoon (1-3pm) and Monday morning (9-11am), with a significant decrease during the week
* Model
  + Will be using Random Forest as it is a fairly straightforward ML model to implement, while also getting an accurate and stable prediction
  + Can help with classification of the data (ie what place in the order a product is in) and with regression (predict which products are ordered)