GroCart Market Basket Analysis

Which products will a Grocart consumer purchase again?

Whether you shop from meticulously planned grocery lists or let whimsy guide you're grazing, our unique food rituals define who we are. Grocart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favourites and staples when you need them. After selecting products through the Grocart app, personal shoppers review your order and do the in-store shopping and delivery for you.

Grocart's data science team plays a big part in providing this delightful shopping experience. Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session. Recently, Grocart open sourced this data.

In this project, Grocart is challenging you to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order. They're not only looking for the best model, Grocart's also looking for machine learning engineers to grow their team. Be one of them!

Data Description:

The dataset given is a relational set of files describing customers' orders over time. The goal is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Grocart users. For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the week and hour of day the order was placed, and a relative measure of time between orders.

File descriptions

Each entity (customer, product, order, aisle, etc.) has an associated unique id. Most of the files and variable names should be self-explanatory. The task is to predict which products a user will reorder in their next order. The evaluation metric is the F1-score between the set of predicted products and the set of true products.

aisles.csv (134 rows)

- aisle_id : aisle identifier
- aisle : the name of the aisle(passageway between areas of shelves of goods as in stores.

orders (3.4m rows, 206k users):

- order_id: order identifier
- user_id: customer identifier
- eval_set: which evaluation set this order belongs in (see SET described below)
- order_number: the order sequence number for this user (1 = first, n = nth)
- order_dow: the day of the week the order was placed on
- order_hour_of_day: the hour of the day the order was placed on
- days_since_prior: days since the last order, capped at 30 (with NAs for order_number = 1)

products (50k rows):

- product_id: product identifier
- product_name: name of the product
- aisle_id: foreign key
- department_id: foreign key

departments (21 rows):

- department_id: department identifier
- department: the name of the department

order_products__SET (30m+ rows):

- order_id: foreign key
- product_id: foreign key
- add_to_cart_order: order in which each product was added to cart
- reordered: 1 if this product has been ordered by this user in the past, 0 otherwise

where **SET** is one of the four following evaluation sets (eval_set in orders)**:**

- "prior": orders prior to that users most recent order (~3.2m orders)
- "train": training data supplied to participants (~131k orders)
- "test": test data reserved for machine learning competitions (~75k orders)

ALL THE BEST