Instacart Market Basket Analysis

Introduction:

Nowadays, the retail & e-commerce sectors are experiencing dramatic changes as more and more Americans prefer online shopping. Common business problems that online retailers face including: how to implement more data-driven customer retention strategy, how to understand the customer preferences by figuring out who they are, what they want. At the same time, these online businesses are also obtaining millions of transactions data. Data scientists thus could leverage this big data to help business gain more values.

Problem Definition:

Instacart is a grocery ordering and delivery app, which allows you to select products through their app, and then personal shoppers review your order and do in the in-store shopping and delivery for you. In other words, Instacart delivers groceries from your favorite stores to your door. The company is expanding its platform to cover 90 millions US household in 2018. With millions of transactions in real time, Instacart's problem is a representative of a problem I would like to work on as a data scientist: predict customer behaviors with large amount of data. This project will focus on:

- Which products a user would buy again, try for the first time, or add to their cart next during a session?

Data Sources:

Instacart open-sourced 3 Million of their Instacart Orders. This data is also available on Kaggle: https://www.kaggle.com/c/instacart-market-basket-analysis/data

This anonymized dataset contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, Instacart provide between 4 and 100 of their orders, with the sequence of products purchased in each order. Data also provide the week and hour of day the order was placed, and a relative measure of time between orders.

File descriptions

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

aisles (134 rows):

- aisle_id: aisle identifier
- aisle: the name of the aisle

deptartments (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)

Methods:

1. Load, store, clean data
2. Exploratory Data Analysis: visualize customer's order patterns using matplotlib
3. Product Recommendation, possibly use Pyspark to load such a large amount of data.
4. Feature Engineering/Predictive Modeling using apriori algorithm, XGBoost, LightGBM

Deliverables:

1. Code for the project on github.
2. A final paper explaining the problem, approach and findings in complete technical detail. Include ideas for further research, as well as up to 3 concrete recommendations for your client on how to use the findings.
3. A slide deck or a blog post which presents my analysis to clients (e.g. non-technical and business teams) in an easy to understand, but compelling way.