Instacart Market Basket Analysis

Group 10:
Tiantian Zhang, Shengshi Yuan, Fan Zhang, Xinyi Tan, Siyu Shen

Background and Goals

- Instacart is a growing e-commerce company that provides grocery delivery and pick-up services.

- With the massive amount of customer data, the goal of this project is to predict which products will be in a user's next order so that the company can **adapt communication to different customer groups** and further boost revenue.

- In this project, we would perform customer grouping, examine different predictive models, and optimize the performance of our final model.

Photo Source: https://www.supermarketnews.com/online-retail/instacart-sees-2020-year-grocery-pickup

instacart

Dataset

- The dataset from the Kaggle competition [1] is a relational set of files describing customers’ orders over time.

Order_products_prior.csv	Order_products_train.csv	Orders.csv	Products.csv	Aisles.csv	Departments.csv
Order_id	**Order_id**	**Order_id**	**Product_id**	**Aisle_id**	**Department_id**
Product_id	**Product_id**	Buyer	Product_name	aisle_name	Department_name
Add_to_cart_order	Add_to_cart_order	The day of week	**Aisle_id**		
Reordered	Reordered	Days since prior order, etc.	**department_id**		

Table 1. Dataset Descriptions

Dataset

- The anonymized dataset includes over 3 million grocery orders samples from more than 200,000 users. For each user, between 4 and 100 of their orders are given, with the sequence of products purchased in each order.

- Prior dataset gives the past behaviors of a user, while the train and test dataset give the future behaviors that we would like to predict.

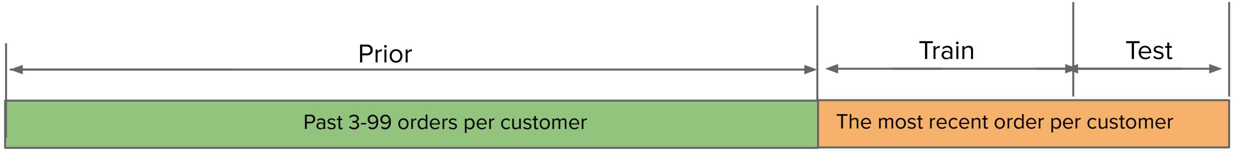


Figure 1. Train / Test Split

Methods - Preprocessing and Feature Engineering

User Features

How often the user reorder items

Time between orders

How much does the user reorder products

Product Features

How often the product is purchased

Probability being reordered

User x Product Features

Number of days since the user last purchased the product

Number of orders in which the user purchased the product

Datetime Features

Counts by day of week

Counts by hour

Method - Modeling

1. **Baseline: Popularity-based bias model** [4]

Basically we model each interaction as a combination of global, item, and user terms.

- Global bias

$$\mu = \frac{\sum_{u,i} R[u,\ i]}{|R| + \beta_g}$$

where $|R|$ is the number of all purchasing counts records and β is the damping value

- Item bias

$$b[i] = \frac{\sum_u R[u,\ i] - \mu}{|R[:,\ i]| - \beta_i}$$

where $|R[:,\ i]|$ is the number of purchasing counts records associated with item i

- User bias

$$b[u] = \frac{\sum_i R[u,\ i] - \mu - b[i]}{|R[u,\ :]| - \beta_u}$$

where $|R[u,\ :]|$ is the number of purchasing counts records associated with user u

The interaction between item i and user u is modelled as $R[u, i] \fallingdotseq \mu + b[i] + u[i]$

instacart

Method - Modeling

2. Xgboost

- Prepare a train data frame that merge the future orders (train & test) with prior orders (prior).
- Generate a label ‘Reordered’ for each (user_id, product_id) pairs to indicate whether the pair was reordered or not (1/0).
- Train xgboost for binary classification.

3. Word2Vec [5]

- Interpret every order as a sentence and every product in an order as a word.
- Find products that are usually bought together from the order history of all users.
- Recommend the product with similar learnt vector representations to the user.

instacart

Method - Modeling

Method - Evaluation

- In this project, we adopt F1 Score (1), the harmonic mean of precision and recall, as our evaluation metric.

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{1}$$

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

(2)

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

- To calculate F1 Score, we need to convert the predicted probabilities into binary labels using a threshold value.

- We plan to use the grid search method to find the optimal threshold.

Method - Optimization

In this project, we plan to explore a range of performance optimization techniques covered in class.

Runtime Optimization:

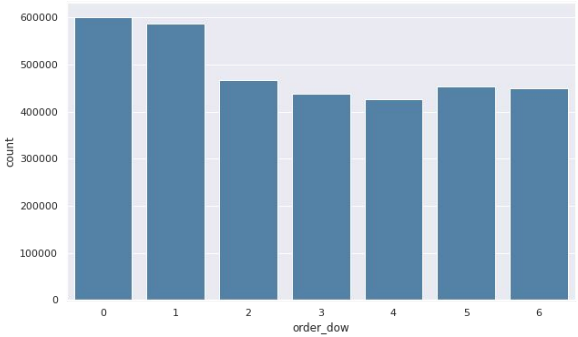
- Avoid function call overheads
- Vectorization
- Code profiling: line profiler
- Improve pandas performance
  - Set_index on merging column
  - sort_index
- ‘n_jobs’
- Multiprocessing: pool

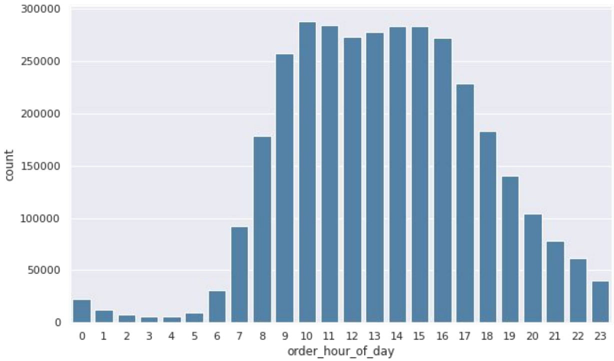
Memory Optimization:

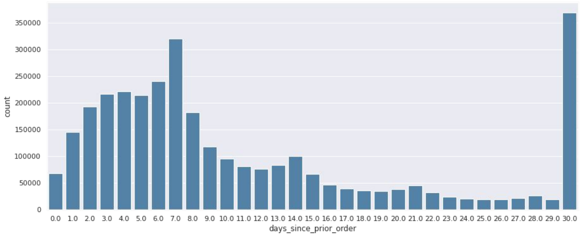
- Code profiling: Memory profiler
- Process large dataset in smaller chunks
- Data type optimization

instacart

Exploratory Data Analysis



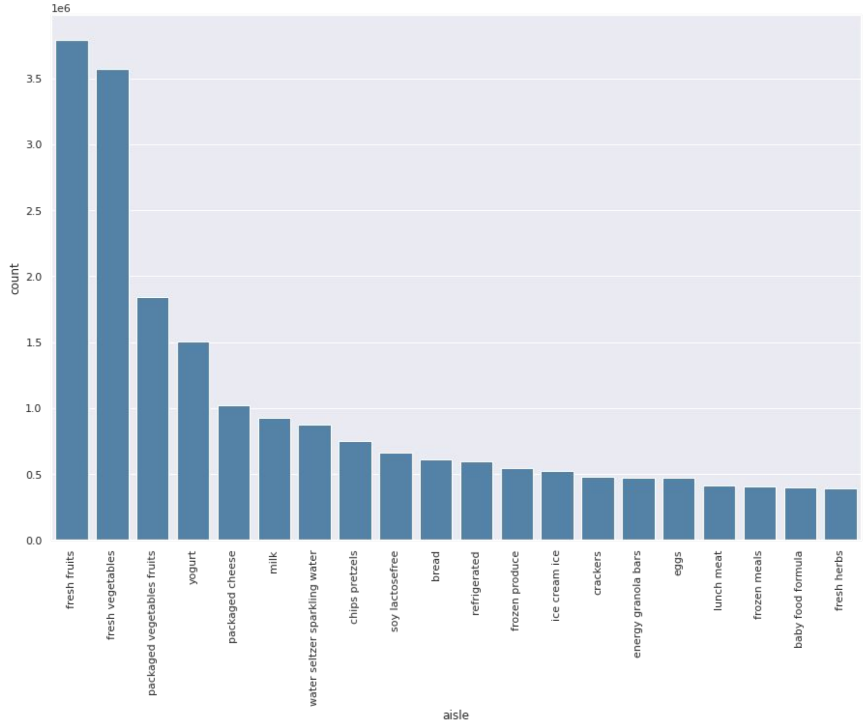




- Customers tend to order on weekends and during day time.
- More customers order once a week or once a month.

instacart

Exploratory Data Analysis



	aisle	reordered
0	milk	0.781812
1	water seltzer sparkling water	0.729930
2	fresh fruits	0.718823
3	eggs	0.706359
4	soy lactosefree	0.692361
5	packaged produce	0.691977
6	yogurt	0.686501
7	cream	0.685184
8	bread	0.670552
9	refrigerated	0.663006
10	breakfast bakery	0.651302
11	energy sports drinks	0.649473
12	soft drinks	0.639301
13	packaged vegetables fruits	0.639275
14	white wines	0.631928

- The **most popular product categories** are fresh fruits / vegetables, packaged vegetables / fruits / cheese, and yogurt.
- Milk, water / seltzer / sparkling water, fresh fruits, eggs, and soy lactose free have the **highest reordered rate**.

instacart

Timeline

Week 7

Proposal presentation

Select the topic of interest and research various techniques to improve the performance

Week 8-9

Modeling

Preprocess the dataset and build the machine learning models

Week 10-13

Optimization

- Examine techniques for improving performance and finalize the codes
- Start working on the report

Week 14

Final Project Presentation

Finish the final report and present the results to the class

instacart

References

[1] https://www.kaggle.com/c/instacart-market-basket-analysis/overview

[2] https://towardsdatascience.com/10-tips-tricks-for-working-with-large-datasets-in-machine-learning-7065f1d6a802

[3] https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

[4] https://newclasses.nyu.edu/access/content/group/8bce5f3c-c1f9-48a9-9f2a-7abb59e6b9a5/Slides/Week%2010.1%20Recommender%20systems.pdf

[5] https://arxiv.org/pdf/1301.3781.pdf

[6] https://link.springer.com/article/10.1057/dbm.2012.17

instacart