Explainable Recommendation System with Association Rule Mining

G6: Zhijia Li, Xiaoyun Zhuang, Liliang Ren

1. Introduction

Existing works in the recommendation system realm have shown impressive performance on various datasets. However, they often lack the ability to explain the reason for their predicted recommendations. In this work, we aim to build an explainable recommendation system that can give recommendations as well as the mined association rules as explanations to justify such predictions. We will first do high dimensional association rule mining, and then give recommendations solely based on such associations. Our model will be trained and evaluated on the Instacart Market Basket Analysis dataset [1].

2. Related Work

Recommendation System: Modern recommendation systems [3] often apply deep neural networks with self-attention mechanisms to infer the item-item relationship from a user's historical interactions and learn the user's transient interests. These lines of work have shown impressive performance to provide recommendation of the next item given a sequence of user interaction history. However, since the next item recommendation for this system is based on the metric learned between the user and the item feature vectors, the explainability for each recommendation decision is still unclear. There are also some non-neural methods [2] using TF-IDF and SVD for neighbor-based item recommendation, but they often ignore the sequential property of the user history.

Recommendation Justification: There have been several works [4] on justifying the recommendation using the filtered review data. However, they need human annotators to filter out the suitable justifications from the review data, while our method based on the association

rule mining does not need such extra annotated data, which means it will be more scalable, and cost efficient.

3. Dataset

The dataset is provided by Instacart, which is one of the largest online grocery ordering and delivery apps in the United States. This dataset contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, this dataset provides between 4 and 100 of their orders with the sequence of products purchased. Besides, the order time will also be included.

4. Exploratory Data Analysis

4.1 Relationship Between Orders and Time

First take a look at the relationship between orders and time. From Figure 1, since 0 represents Saturday and 1 represents Sunday, it is noted that Saturday is the most popular day and Wednesday is the least popular day. Similarly, from Figure 2, we see that hours from 10:00 to 16:00 is the most popular time in a day.

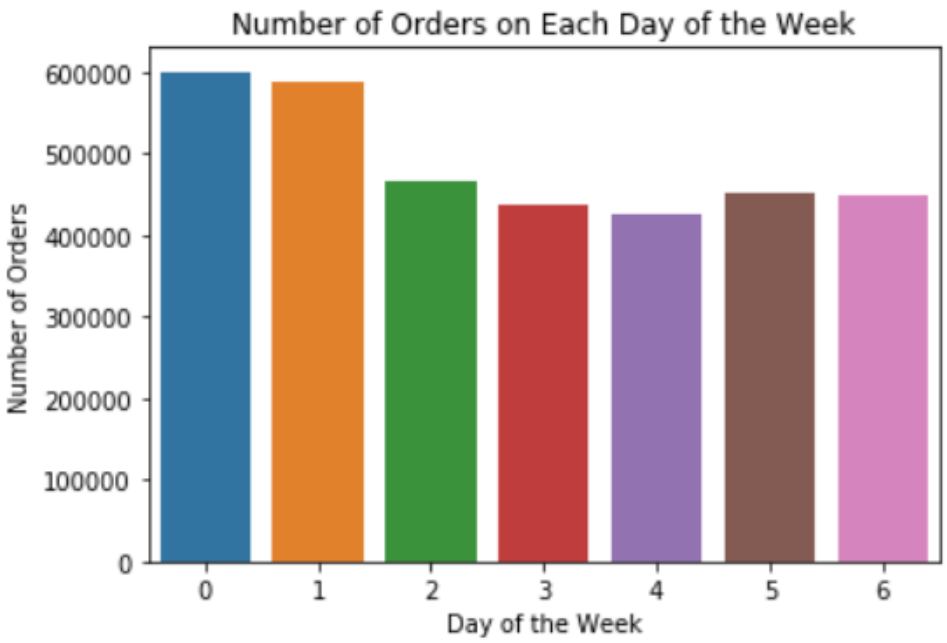

Figure 1: Number of Orders Every Day

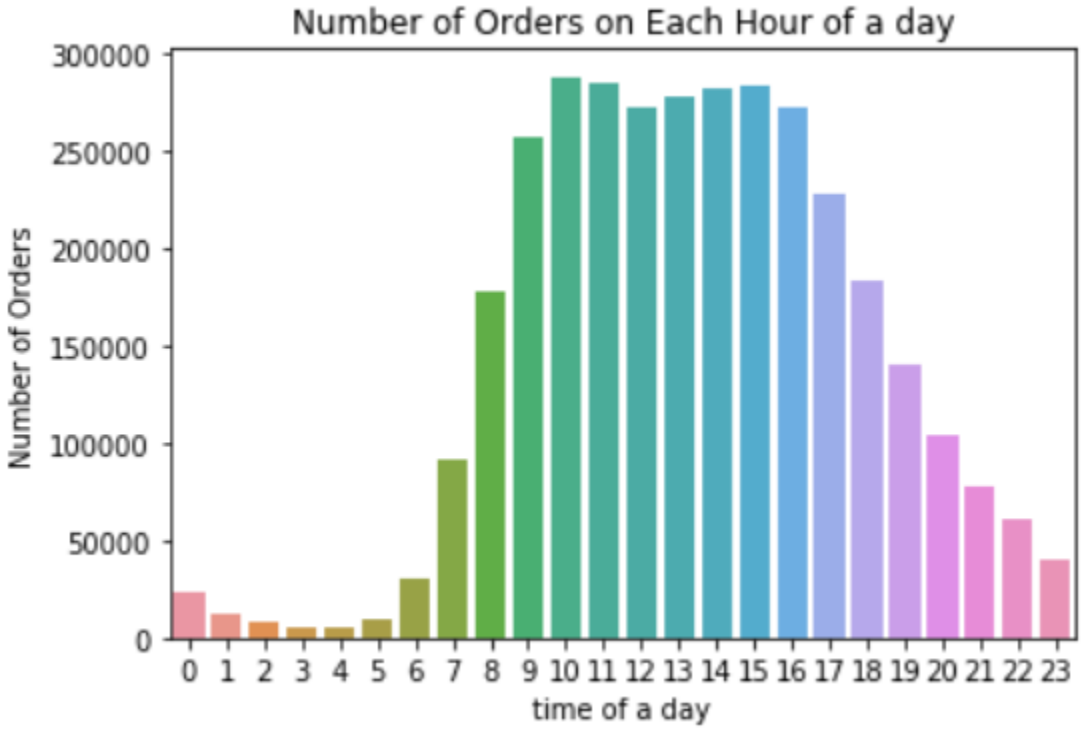

Figure 2: Number of Orders Every Hour

If we look at the busy hour and busy days as a whole (Figure 3), we can see that saturday 12-16 and sunday 8-10 is the most popular time.

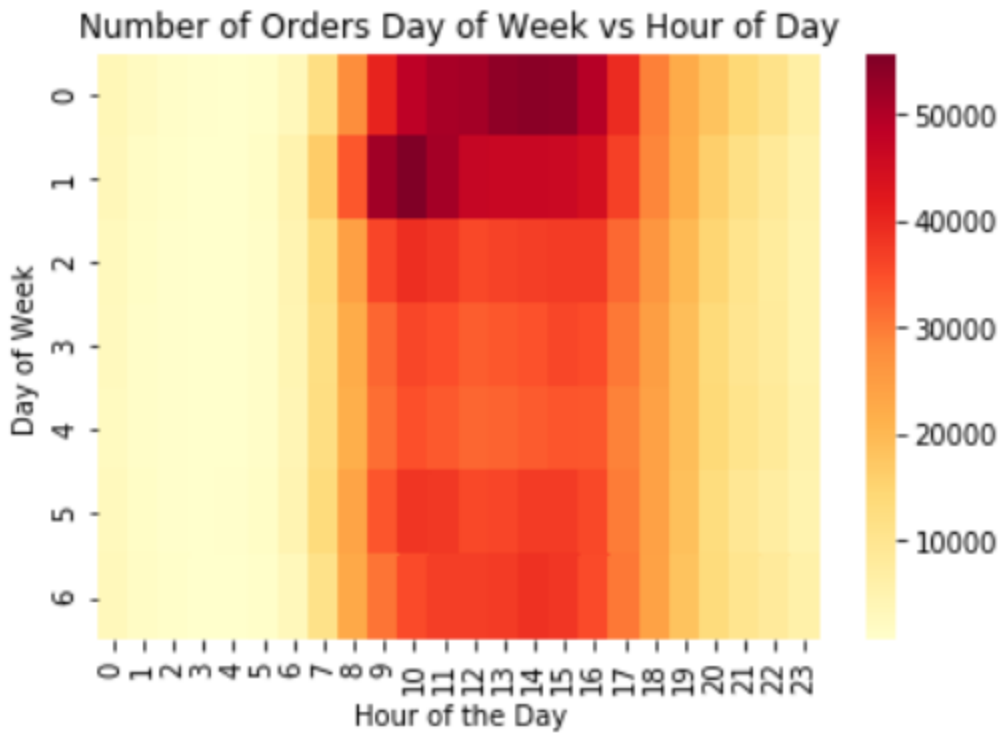

Figure 3: Matrix of Busy Day amd Busy Hours

4.2 Order Information

Then, we are interested in the number of products in each order. By laying out the distribution of number of items of each order, we find out that people tend to but 3 to 9 items together each time.

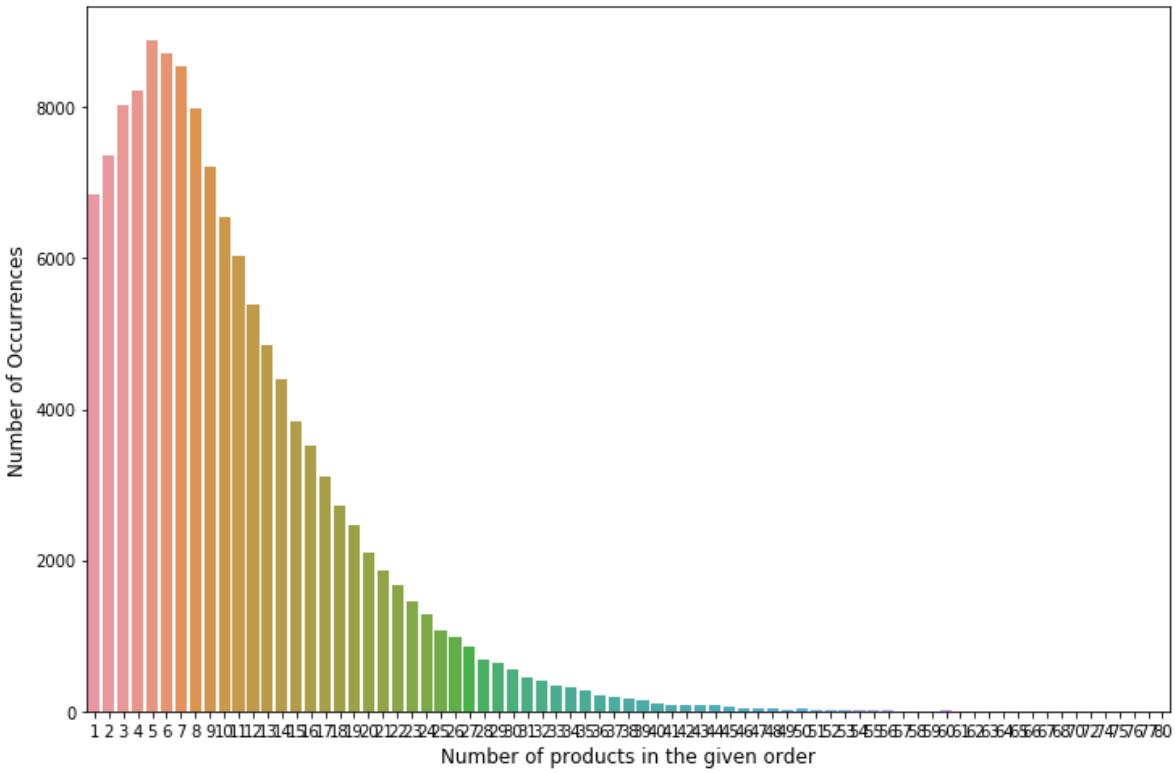

Figure 4: Number of Products in Each Order

Furthermore, Banana, Bag of Organic Bananas, Organic Strawberries, Organic Baby Spinach, Organic Hass Avocado, and Organic Avocado are the top 5 popular products across the orders. And Raw Veggie Wrappers, Serenity Ultimate Extrema Overnight Pads, Orange Energy Shots, Chocolate Love Bar, Soy Powder Infant Formula, and Simply Sleep Nighttime Sleep Aid are the top 5 reordered items across the orders. It is interesting to see that the most popular items are from Produce aisle while the most popular reordered items vary from household to beverage. And it is counter-intuitive that nothing from the top 5 reordered items and top 5 ordered items is overlapping.

5. Methodology

- **Data Cleaning and Merging:** Merge the multiple tables to have a dataset where each row represents a transaction and each column represents corresponding information (user_id, time, etc) of this transaction.
- **PCA and K-Means Clustering:** Principal components analysis (PCA) is applied to reduce the number of features for the K-means clustering. In order to do the PCA, we merge the chart again to have a dataset at customer level which each row (customer) has 134 features (aisles) representing whether the customer buys items from this aisle. The principle of PCA is to search for k n-dimensional orthogonal vectors that can best be used to represent the data, where k <=n. The original data are thus projected onto a much smaller space, resulting in dimensionality reduction. After the PCA, the number of features was reduced from 134 to 6. And the first three features explain the majority of the variance in our data. After plotting all pairs in these three components, we choose column 0 and column 2 pairs to do the k-means clustering.

  k-means is a centroid-based technique to do clustering. The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. It proceeds as follows. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean. The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration. All the objects are then reassigned using the updated means as the new cluster centers. The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round. After several times of cluster, we use the Elbow Method to confirm the ideal k value, since 3 is a bend of the elbow, we choose 3 as the k value.
- **Association rule mining:**

  With the clustering ready, we expect to use frequent association rule mining so as to explore the relationship between products in terms of how likely they would be purchased together in an order. For each of the clustered user sets, we split the dataset into train and test samples. The Apriori algorithm is used to mine the association rules amongst the ordered items. First, 15 products are found when the support is set at 0.03,

which means these items are found in 3% of the total transactions. Then, we set the support to 0.003 to find the combination of 2 and 3 items.

6. Results

After tuning the parameters with the minimum relative support of 0.001 and the minimum confidence at 0.25, around 30 rules are generated in each cluster. After inspecting the top 10 rules each, here are some interesting results and possible interpretation:

(1) There is some reinforced relationship between the type of yoghurt and the type of sparkling water, with lifts as high as 80, and the confidence around 0.6; it implies the preferred flavor of the customer and might suggest the potential of bundle sales promotion campaign

(2) There are associations between items with similar names. It implies that customer tends to buy multiple items from the same brand in a single transaction;

(3) Organic produce is highly associated with other organic products, which implies the customers with such lifestyles tend to purchase more organic products.

7. Conclusion

In this work, we proposed a pipeline model that mines frequent association rules for next-item recommendation. Specifically, we first do PCA and clustering on the dataset to group the users and the products, then we do frequent association rules to mine the high level associations between these groups. These association rules can indicate the meta relationships that can either justify specific recommendations or provide fuzzy recommendations for the user. Our model can be applied to shopping companies to increase their sales revenue and improve their customer experience.

Reference

[1] "The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.kaggle.com/c/instacart-market-basket-analysis/data on Mar. 5, 2021

[2] http://cs229.stanford.edu/proj2020spr/report/Qian_Xu_You.pdf

[3] Zhang, Shuai, et al. "Next item recommendation with self-attention." arXiv preprint arXiv:1808.06414 (2018).

[4] Ni, Jianmo, eT al. "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects" EMNLP 2019