Course task: Shopping cart analysis

Retail Analytics (DAT-4182)

Beau Giannini and Pavel Paramonov

Hult International Business School, San Francisco

IDENTIFYING PRODUCTS FREQUENTLY BOUGHT TOGETHER is commonly done based on the data collected from individual shopping carts.[1] In this task, you will be analyzing real-world point-of-sale transaction data from a representative grocery store. The objective is to identify most reliable product association rules of potential interest for cross-selling and promotion of specific product combinations.

Note that association rules may involve more than two products.[2] Depending on the tools you choose for this task, you are welcome to explore just product pairs and/or more complex product association subsets.

Key metrics

Support: for a product A occurring N_A times among the total number of

transaction records N,

$$Support(A) = \frac{N_A}{N}.$$

Confidence: for the association rule $A \rightarrow B$ between the products A and B,

$$Confidence(A \rightarrow B) = \frac{N_{A,B}}{N_A},$$

where $N_{A,B}$ is the number of transaction records where both products occur together. Intuitively, support quantifies how predictable the association rule is, i.e. the likelihood that B is also bought if A is bought.

Lift measures the increase in the ratio of sale of B when A is sold:

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} = \frac{N_{A,B}N}{N_A N_B}.$$

Intuitively, the lift metric quantifies the departure from independent sales of the products A and B.

Note that many rules with high lift typically have low support. Numerically, the most interesting rules are those where support is at least 0.001, confidence exceeds 0.4 (40%), and lift values are as high as possible (above 2, and can reach 10 or more)[3].

Course task description

For this task you are provided with the transaction dataset containing one month of real-world point-of-sale grocery store data, with 9835 transactions and 169 products.[4] The files is available on Github: https://github.com/multidis/hult-retail-analytics/tree/main/shopping_cart

Different files representing the same dataset in different formats are available:

- Binary matrix representation with products as columns and transactions as rows; 1-entries correspond to the products included in a given transaction (shopping cart).
- List of products representation with rows of strings representing product names included in each transaction.

Please note that **you do not have to use both files**, they are just two different representations of the same data.[5]

Step 1

Identify top-30 most frequently bought products during the time period of the dataset.

Step 2

Identify at least 5 most promising product association rules that involve top-N^{th} most frequently bought product, with `N` being your team number.[6] Support your conclusions with relevant metrics (support, confidence, lift).

Please make sure to **include a full, explicit listing of top product names and most promising association rules** in your presentation and assignment submission.