# **Final Project: A Study of Sales through Consumer Behavior on Black Friday**

**G-number and name**:

G01145163 Hung Mao Chen

G01167652 Hsin-Ping Hsu

**Dataset**: <https://www.kaggle.com/mehdidag/black-friday>

**Description**: The dataset collected the transactions made in the retail store. Each record shows the gender, age, occupation and several features of the customers. We decide to make the association rule analysis based on this dataset and expect to create a recommendation system. The retail store can depend on our results to think about the strategy to stimulate the consumption, for example, the customer who is younger than 17 years old buys a product from the store, then we can recommend another product to him.

**Library**: <http://rasbt.github.io/mlxtend/>

The library we used to do the association rule analysis is called mlxtend. The most important parts are creating frequent itemset with Apriori algorithm and generating association rules from frequent itemset.

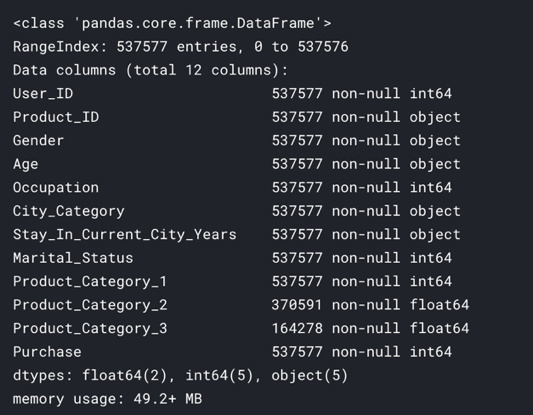
**Challenge from the dataset**:

1. Product category and product, which is better for recommendation? Originally, we tried to make the association rule analysis between product category and consumers’ features because the number of products is much bigger than the number of categories, and we just want to create some representative rules. However, we found that some customers may only bought the products in one category; some may not. Therefore, generating association rules among product categories might become over generalization. We will discuss more about this issue at the discussion section.
2. How to generate association rules based on the transactions and consumers’ features? This kind of association rule which involves not only a single item but also the consumers’ features is referred to multi-dimensional association rule. We will dig deeper on the methodology section.
3. How to choose the thresholds? We need to choose several thresholds for the association rule analysis. If we set the value too high, we would lose a lot of possibilities to recommend a product to a customer; if we set the value too low, there would be so many meaningless rules. We will explain more on choosing the thresholds at the methodology section.
4. Since there are some categorical features in the datasets, classification techniques could be utilized on this dataset. Can we use multiclass classification techniques to predict the categories of good bought based on the consumers features?

**Methodology**:

1. Data Exploration:

Before starting on association analysis, we would do a data exploration first to understand the data types and the dimension of the original dataset. We found that there are 537,577 rows and 12 columns, including 3,623 different items and 5891 consumers.

 一張含有 螢幕擷取畫面 的圖片

自動產生的描述

1. Data preprocessing:

To get the transactions, we need to merge rows of data with the same User\_ID. We drop the columns of Product\_Category\_2 and Product\_Category\_3 because of high percentage of missing values, but we don’t need to worry that it would cause to the loss of information because subcategories are useless in our work. In addition, we also drop the column of purchase because of it is unrelated with the analysis.

1. Association analysis:

Since this dataset contains some consumers’ features, we could generate the multi-dimensional association rules from the relational information and transactions.

Our purpose is to recommend more products to the consumers based on their features. Therefore, the consumers’ features would be included in the antecedent, and more products would be included in the consequence.

Step1: Create multi-dimensional frequent itemset with minimum support of 0.03.

Step2: Create multi-hybrid dimensional association rule from frequent itemset with minimum lift of 2.

Step3: Filter the rules which match the purpose description above (antecedent, consequence).

In the step 1, we observed that setting the value of minimum support is very critical for generating the association rules. The reason that we set it as 0.03 mainly because we want to focus on the top 25% popular products to avoid some outliers. By choosing the third quantile, we found that those top 25% products, approximately 906 different items, were purchased by at least 192 customers; therefore, the minimum support would be 192 / 5891 = 0.03.

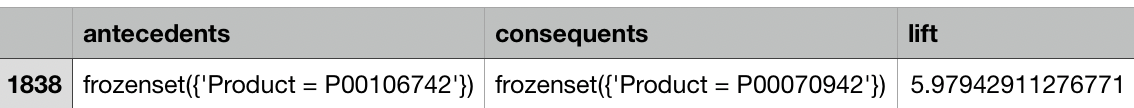
In addition, we use the Apriori’s algorithm to generate the frequent itemset. To create the association rules, we prune the rules whose lift is lower than 1 (at least positively correlated) and extract the rules whose lift is greater than 2.

In the step3, the antecedent of our multi-hybrid dimensional association rule only includes single attribute of the user. We found that most of our rules only show one group of the single attribute. Take association rules of city for example, the city included in the antecedents are all city ‘B’. We consider it caused by the distribution of the dataset.

[The reference of multi-dimensional association rule](https://www.vskills.in/certification/tutorial/data-mining-and-warehousing/single-and-multidimensional-association-rules/)

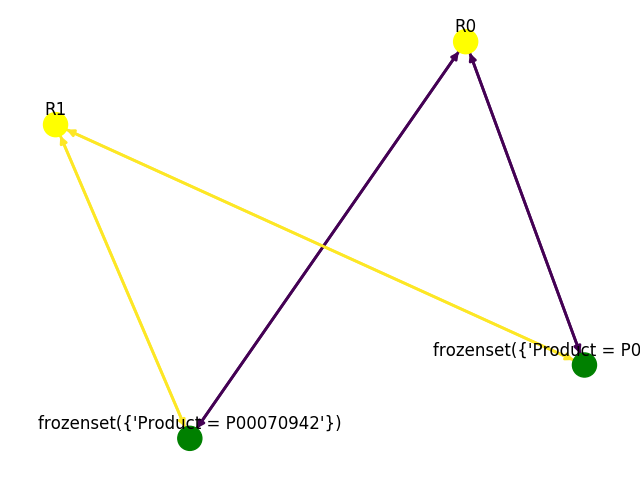
**Results**:

1. Rules upon products



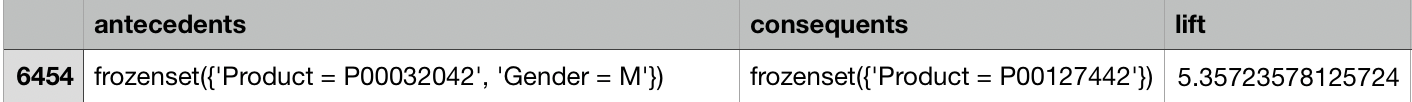
The result above shows the association between products. Based on the result, we can recommend the product 70942 to any customers who have already bought the product 106742. The number of rules here is near to 7500, so we would avoid showing results with the visualization. In the association rules we found, the highest lift can be near to 5.98!

Rule: Buy (Product = 106742) 🡪 Buy (Product =70942)



The above is the visualization of the association rules between products which have the highest lift. There are nodes and edges in the picture, the nodes have two kinds of colors: yellow and green. Yellow nodes represent the index of rule while green nodes represent the itemset. The direction of edge determines the nodes are antecedents or consequences. We would use the same way of visualization in the following part.

1. Rules upon products and gender



Based on the association with age just like above, we can recommend the product 127442 to the male consumer who already bought the product 32042. In the association rules we found, the highest lift can be near to 5.36!

Rule: Buy (Product = 32042) ^ Gender (’M”) 🡪 Buy (Product = 127442)

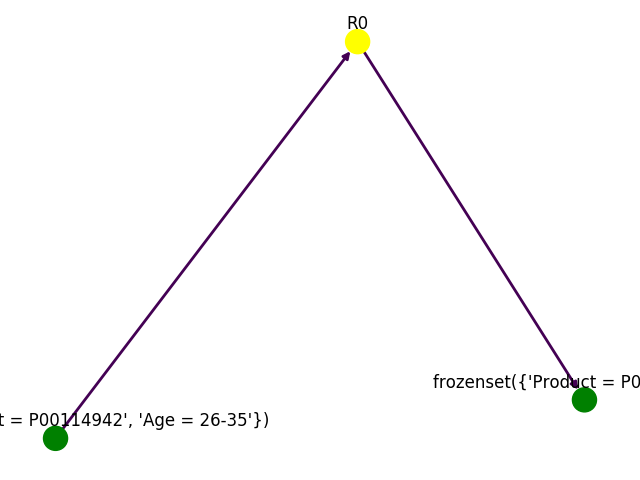
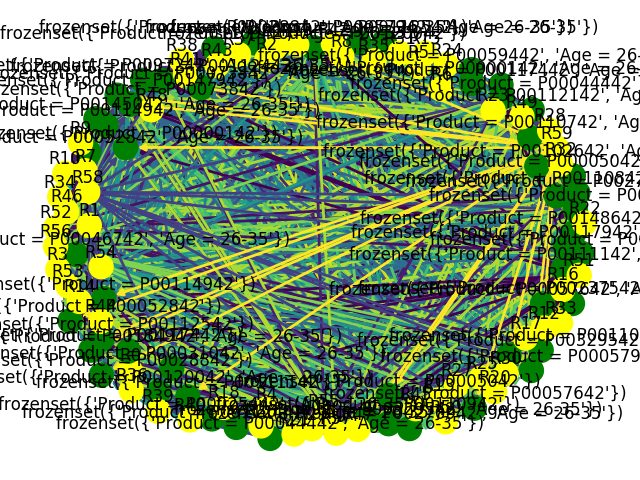
1. Rules upon products and ages

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

Based on the association with age just like above, we can recommend the product 329542 to the customer between 26 to 35 years old who already buys the product 114942. In the association rules we found, the highest lift can be 3.05!

Rule: Buy (Product = 114942) ^ Age (’26-35’) 🡪 Buy (Product = 329542)

  
The picture above is the visualization of the association rule between age and product.

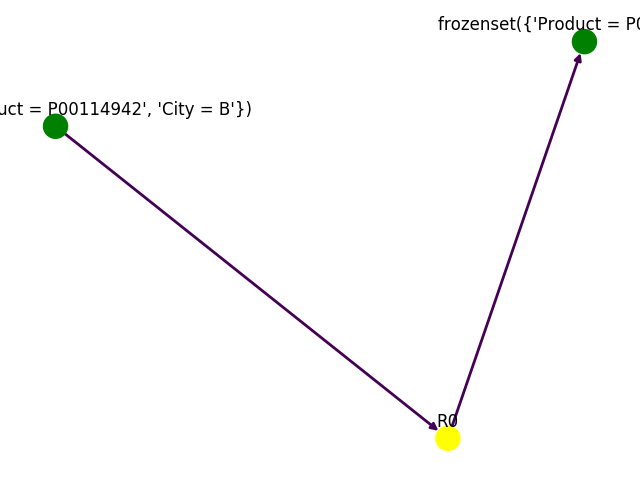
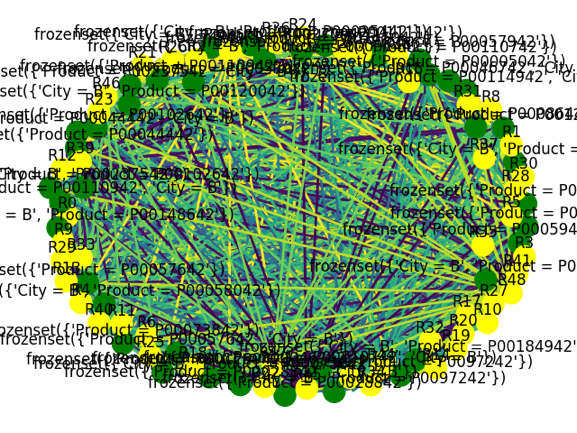
1. Rules upon products and cities

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

Based on the association with city just like above, we can recommend the product 97242 to the customer living in B city who already buys the product 114942. In the association rules we found, the highest lift can be near to 2.6!

Rule: Buy (Product = 114942) ^ City (’B’) 🡪 Buy (Product = 97242)



The picture above is the visualization of the association rule between city and products.

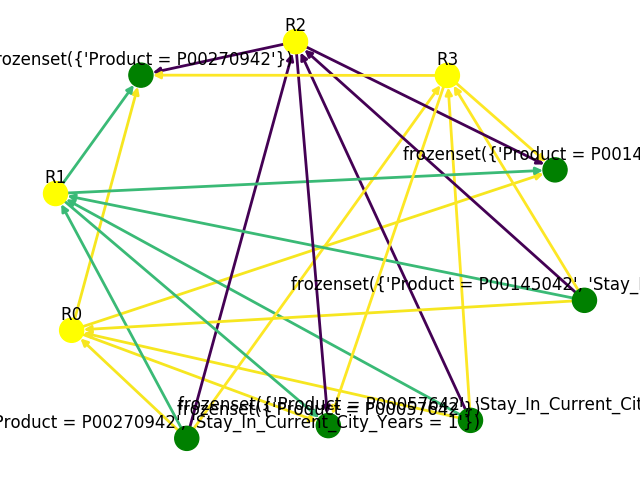
1. Rules upon products and stay in current city

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

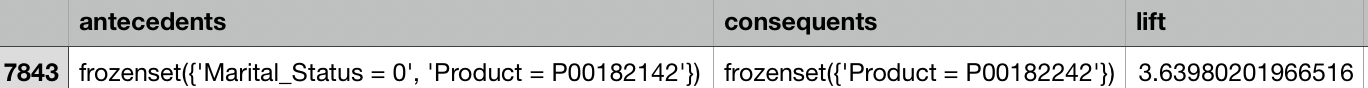
Based on the association with the time staying in the current city just like above, we can recommend the product 270942 to the customer staying for one year who already buys the product 57642. In the association rules we found, the highest lift is near to 2.09!

Rule: Buy (Product = 57642) ^ Stay in current city year (1) 🡪 Buy (Product = 270942)



The visualization for this time is clearer than the previous one because there are only 4 rules matched our threshold (Therefore, there are only R0, R1, R2, and R3). Take a look at the yellow node R2, there are several purple arrows pointing to it which means antecedents, and there is an arrow pointing to the green node of product 270942 from the node R2 which means consequence.

1. Rules upon products and marital status

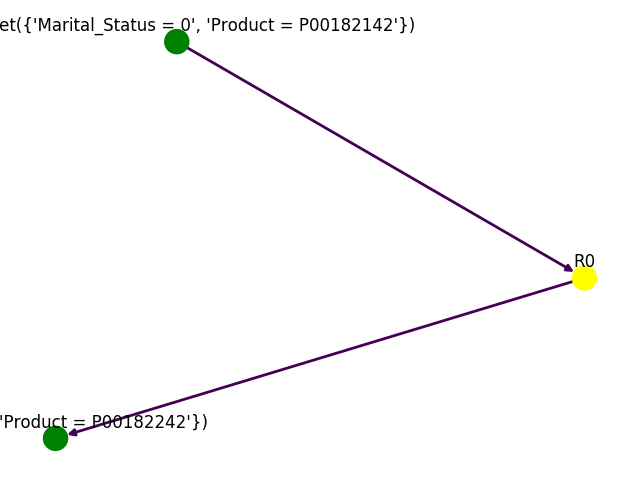
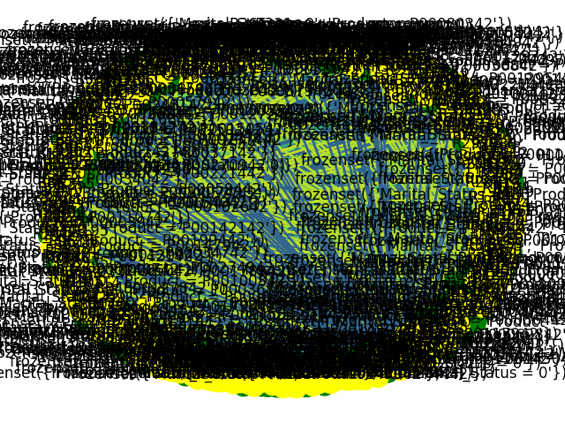




Based on the association with the marital status just like above, we can recommend the product 270942 to the customer still not marrying and already buys the product 57642 and 145042. In the association rules we found, the highest lift can be near to 3.64!

Rule: Buy (Product = 182142) ^ Marital (0) 🡪 Buy (Product = 182242)

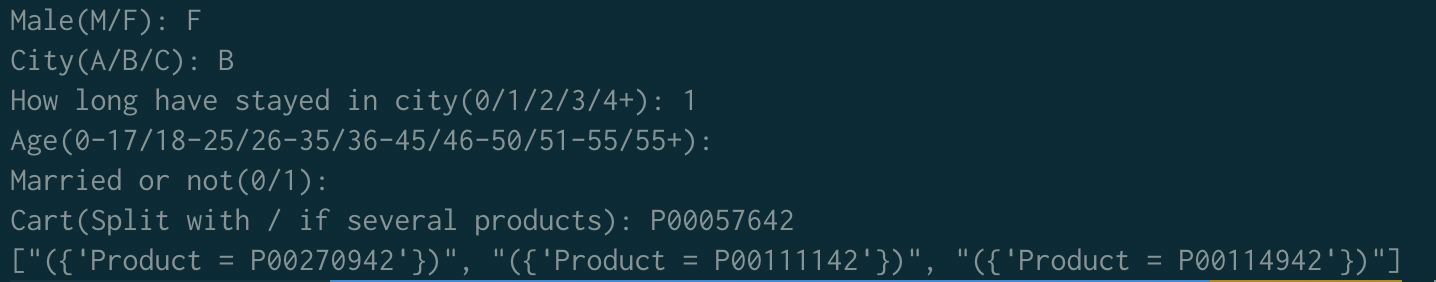
Rule: Buy (Product = 57642) ^ Buy (Product = 145042) ^Marital (0) 🡪 Buy (Product = 270942)



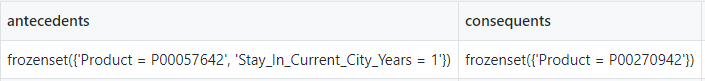
For the attribute occupation, there is even no any rules match our threshold. In next part, we would show the results for multiple attributes.

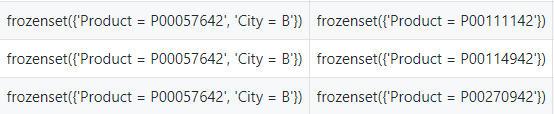
**Conclusions**:

1. Based on the result of association rule analysis, we also create a program (<recommend.py>) of recommendation system for the retail store. Following is the screenshot of the program:



The program would ask for the information of customers. In the real world, stores always cannot get all the information of the customers’, so it’s ok to leave the answer blank if they are the missing values. The program would union the consequents of the association rules on single attribute, and would also sort it by the lift value, which can make it convenient for the manager to choose what to be recommended. The example in the screenshot is the union of city, the time to stay in the city and the products, and there are no any matched rule for female to pass the threshold of lift 2.0.





The rules above are the matched ones found by the program, the lift values from the first one to the fourth are 2.08, 2.14, 2.11, 2.33. After sorting by the lift and removing the duplicate ones, the recommended list would be P00270942, P00111142, and P00114942. This program tends to keep all the data instead of removing the ones with lower lift because false positive and true positive are both important for the recommendation system.

1. Comparisons of different models based on product categories:

* Association rules: the number of product categories is only 18, we thought we could generate some simple association rules upon based on it; however, we observed that the lifts of most rules are only around 2.0, which are lower than the lifts of rules based on products and even some rules containing consumers’ feature.
* Multiclass classification: we tried to build different classification models for multiclass classifications such as SVM, decision tree, random forest, and KNN; however, none of these provide a good result, the accuracy is only about 0.35. Although the dataset contains more than 537577 rows, only 5891 consumers’ features in it; moreover, most of them tended to purchase products from multiple categories. As a result, predicting the product of goods bought based on the consumers’ features via classification techniques is not a good approach for this data set.

Through the association rules we generated, the store can promote their products based on the consumers’ purchase behavior and features. We believe that it would be more useful and interesting to analyze deeper on the relationship among multiple consumers features and purchase behaviors in the future.

**References:**

1. <http://rasbt.github.io/mlxtend/>
2. <https://www.kaggle.com/dabate/black-friday-examined-eda-apriori>
3. <http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-6.html>
4. <http://dataminingzone.weebly.com/uploads/6/5/9/4/6594749/ch15multilevel_association_rules.pdf>
5. <https://www.vskills.in/certification/tutorial/data-mining-and-warehousing/single-and-multidimensional-association-rules/>
6. <http://intelligentonlinetools.com/blog/2018/02/10/how-to-create-data-visualization-for-association-rules-in-data-mining/>
7. <https://www.kaggle.com/mehdidag/black-friday>