Task 6

Report: Discover Association Between Products

Barcelona, November 2018

Student:
Javier E. Villasmil
Giorgia Felline

ÍNDEX

LIST OF FIGURES

1. SCOPE

Analyze a transaction list from Electronidex, provided by the Blackwell management team, in order to search for customer buying patterns and provide insights about if it is worth it for Blackwell to acquire Electronidex.

2. EXECUTIVE SUMMARY

By analyzing the transactions from Electronidex, we found interesting relationships within product types and categories. For instance, business customers tend to make transactions with a high number of items, and this applies both to high-cost products (i.e. Desktops, Laptops, Monitors) and to low-cost products (accessories, headphones, cords, etc.). On the other hand, for retail customers we found that if a transaction has high-cost products, these are more likely to be purchased together with extra accessories related to said product.

After checking the most frequent items for Electronidex transactions, we found that both for business and retail, Desktops, Laptops and Monitors are the most sold. Blackwell would benefit from these types of customers because these products are the less sold in Blackwell.

By acquiring Electronidex, Blackwell would complement its variety (both in customers and products), increase its volume sales, and, consequently, its profitability.

Also, since the most profitable products for Blackwell are Displays, by adding items that imply the purchase of Displays, you are improving your business.

It is important to notice that, by checking the business of Electronidex, we can conclude that this company handles massive monthly transactions and high-quality products. Therefore, it would be recommended for Blackwell to make some Budget Analysis before considering the idea to acquire Electronidex.

Finally, if the acquisition should be done, it is recommended to make bundles and promotions including the current Blackwell`s products which complement the high-cost products from Electronidex.
It would be interesting for Blackwell to stop focusing on selling many Accessories, and rather invest marketing efforts in pushing products that are more remunerative.

3. METHODOLOGY

The approach used to assess the dataset was to apply basics methods of data mining, descriptive statistics and simple charting to observe the distribution and rules between transactions.

In addition, R / Rstudio helped with the evaluation of histograms, scatter plots and performing the basket market analysis.

On the other hand, we used an R library called ***"Arules"*** which contains functions such as ***"the apriori algorithm"*** that suited perfectly the structure of our data (transaction list).

4. ANALYSIS

4.1 TRANSACTION DATA

The transaction data set is composed of ***9835 transactions*** involving ***125 distinct products***. The frequency of the numbers of items per transaction in the whole dataset is presented in fig.1.

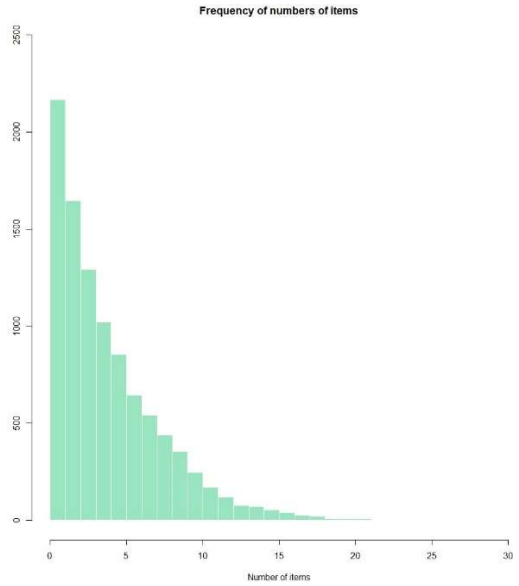

Fig 1. Frequency of number of items per transaction in the dataset.

After analyzing the frequency of the number of items in each transaction, it is important to note that Electronidex has two different kinds of costumers:

- **Retail**: assumed as a natural person or individual buying products in an online store.
- **Business**: assumed as a "legal" person or company buying products in batch from an online store.

To distinguish business transactions from retail transactions a series of rules have been selected.

- Every transaction with more than or equal to three (3) laptops + Pc is considered – Business.
- Every transaction with more than ten (10) Items is considered – Business.
- Every transaction with more than or equal to three (3) monitors is considered – Business
- Every transaction with more than or equal to three (3) printers is considered – Business
- Every transaction with more than or equal to three (3) tablets is considered – Business
- Every transaction with more than or equal to three six (6) headphones (computer+active) is considered – Business
- Every transaction with more than or equal to four (4) mouse + keyboard is considered – Business.
- Every transaction with more than or equal to three (3) mouse and keyboard combo is considered – Business
- Every transaction with more than or equal to four (4) cords is considered – Business
- Every transaction with more than or equal to four (4) accessories is considered – Business
- Every transaction with more than or equal to three three (3) software is considered – Business
- Every transaction with more than or equal to three three (3) speaker is considered – Business
- Every transaction with more than or equal to three (3) printer ink is considered – Business
- Every transaction with more than or equal to three three (3) computer stands is considered – Business
- Every transaction with more than or equal to three three (3) harddrive is considered – Business

After applying all these rules, the ***business*** dataset has ***2.593 transactions*** and the ***retail*** dataset ***has 7.242 transactions.***

In addition, both sets were checked for empty transactions finding that the retail dataset had two empty rows. Both were deleted, therefore the final ***retail dataset*** is composed of ***7.240 rows***.

The market basket analysis has been conducted at two different levels: ***product types*** and ***categories***. The first one being more detailed, while the second more general.

The 125 products were divided in 18 categories as follows:

Laptops, desktops pc ,monitors, mouses, keyboards, mouse and keyboards combo , computer headphones , active headphones , cords , software , accessories , speakers, printers, printer ink , computer stands, tablets, hard drives, smart home devices.

4.1.1 BUSINESS SUBSET

CATEGORIES

As you can see from fig. 2, which shows the frequency of categories in the business subset, the most sold categories are: desktops (present in 94% of transactions), laptops (present in 88% of transactions) and monitors (present in 65% of transactions).

We ran the apriori algorithm function with the following parameters:

- ***Support:*** 0.01 – meaning that we are looking for rules explaining at least 25 transactions.
- ***Confidence:*** 0.6 – reasonable value for the accuracy of our rules.
- ***Max Length***: 10 - maximum number of categories in each rule.
- ***Min Length:*** 2 - minimum number of categories in each rule.

This setting led us to 6.766 rules; by removing the redundant rules, we got ***2.136 final rules***. See fig 2.

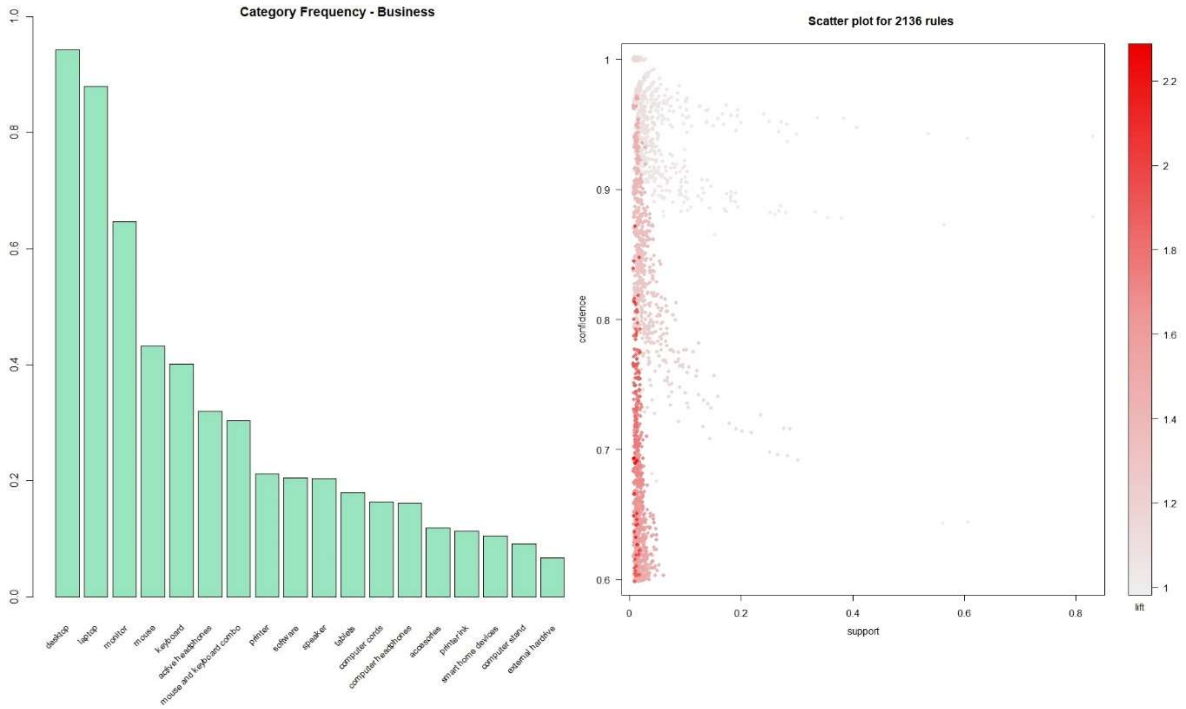

Fig 2. Frequency of categories and rules in the business dataset.

After inspecting the most meaningful rules in the business category subset, we can draw the following conclusion:

1. Active headphones (RHS) are usually bought together with tablets and smart home devices (LHS)
2. Low-cost equipment (LHS) are usually bought in batch and they often imply the purchase of a mouse or headphones (RHS).

The rules were selected after inspecting the relevance of support, confidence and lift (in that order). This meant that we looked for rules relevant to plenty of transactions, high accuracy and lift greater or equal than two (2).

PRODUCT TYPE

As you can see from fig. 3, which shows the frequency of the ***top 20*** product types in the business subset, the most sold products are: iMac (present in 54% of transactions), Hp Laptops (present in 46% of transactions) and Lenovo Desktop Computer (present in 36% of transactions).

We ran the apriori algorithm function with the following parameters:

- ***Support:*** 0.01 – meaning that we are looking for rules explaining at least 25 transactions.
- ***Confidence:*** 0.6 – reasonable value for the accuracy of our rules.
- ***Max Length***: 10 - maximum number of categories in each rule.
- ***Min Length:*** 2 - minimum number of categories in each rule.

This setting led us to 492 rules; by removing the redundant rules, we got ***425 final rules***. See fig 3.

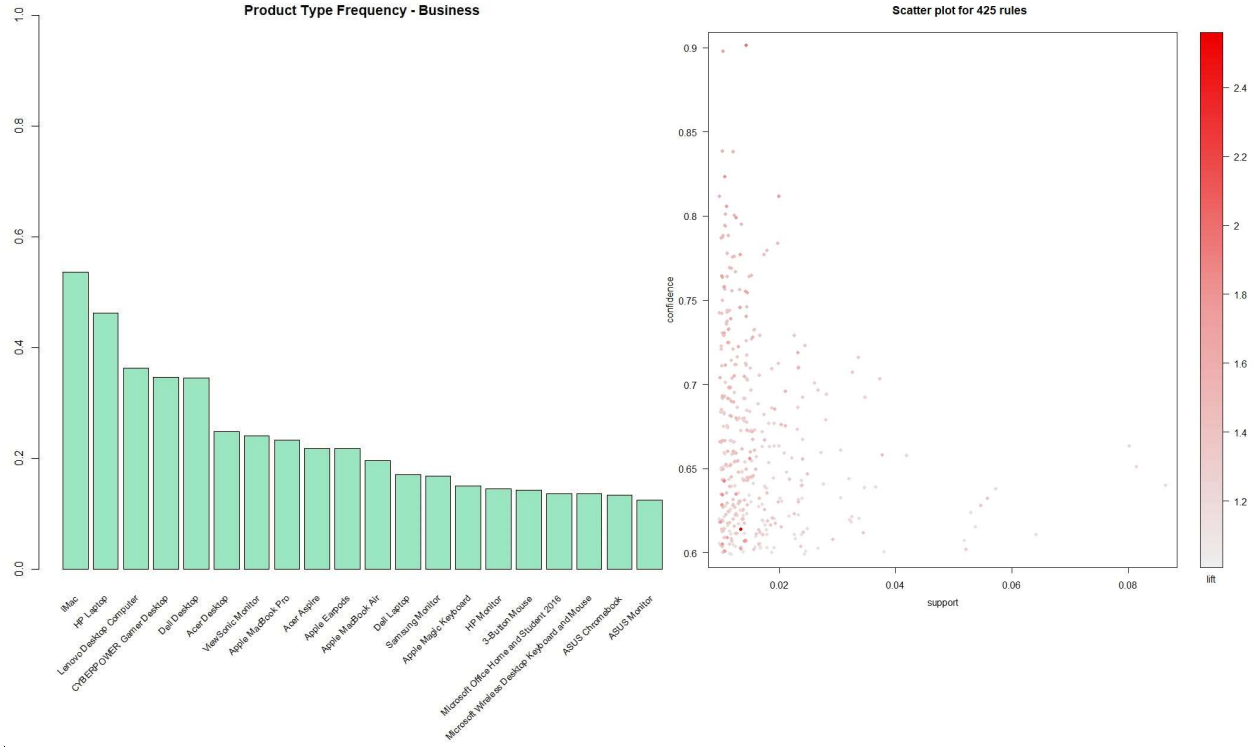

Fig 3. Frequency of Product Types and rules in the business dataset.

After inspecting the most meaningful rules in the business product type subset, we can draw the following conclusion:

1. A strong rule that appears is that the combination {ASUS Chromebook, Dell Desktop, HP Laptop} often implies the purchase of a ViewSonic Monitor;
2. In a lot of transactions HP Laptop is at the RHS, as a consequence of purchasing combination of Desktops and Monitors.

The rules were selected after inspecting the relevance of support, confidence and lift (in that order). This meant that we looked for rules relevant to plenty of transactions, high accuracy and lift greater or equal than two (2).

Please notice that we are dealing with business customers, therefore it is normal to observe transactions with a high number of different items, meaning that maybe these rules are general. Moreover, it is recommended in future analysis to lower the max length of items in the apriori algorithm from ten (10) to three (3) in order to get meaningful insights and specific patterns across the most frequent products.

4.1.2 RETAIL SUBSET

CATEGORIES

As you can see from fig. 2, which shows the frequency of categories in the business subset, the most sold categories are: desktops (present in 45% of transactions), laptops (present in 41% of transactions) and monitors (present in 30% of transactions).

We ran the apriori algorithm function with the following parameters:

- ***Support:*** 0.001 – meaning that we are looking for rules explaining at least seven (7) transactions.
- ***Confidence:*** 0.3 – reasonable value for the accuracy of our rules.
- ***Max Length***: 10 - maximum number of categories in each rule.
- ***Min Length:*** 2 - minimum number of categories in each rule.

This setting led us to 3.284 rules; by removing the redundant rules, we got ***1.947 final rules***. See fig 4.

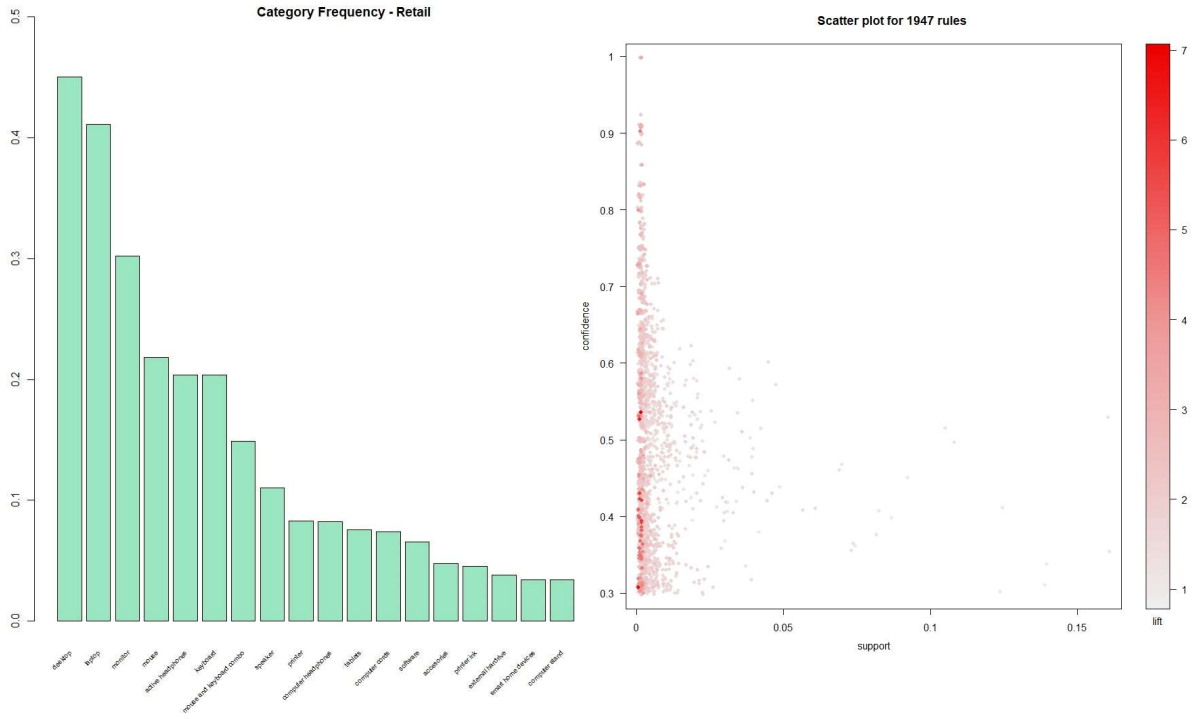

Fig 4. Frequency of Categories and rules in the retail dataset.

After inspecting the most meaningful rules in the retail categories subset, we can draw the following conclusion:

There were not relevant rules in this subset because:

- If we look for a high lift, we find ad hoc rules for few transactions (low support less than 10 transactions) and with unacceptable values for confidence (less than 0.5).
- If we ask for high support (above 30 transactions) and high confidence (above 0.6), then the lift is unacceptable (approximately 1)

It is recommended in future analysis to lower the max length of items in the apriori algorithm from ten (10) to three (3) in order to get meaningful insights and specific patterns across the most frequent categories.

PRODUCT TYPE

As you can see from fig. 5, which shows the frequency of the top 20 product type in the retail subset, the most sold product types are: Apple Earpods (present in 15.8% of transactions), iMac (present in 15.6% of transactions) and Apple MacBook Air (present in 14.1% of transactions).

We ran the apriori algorithm function with the following parameters:

- ***Support:*** 0.001 – meaning that we are looking for rules explaining at least seven (7) transactions.
- ***Confidence:*** 0.3 – reasonable value for the accuracy of our rules.
- ***Max Length***: 10 - maximum number of categories in each rule.
- ***Min Length:*** 2 - minimum number of categories in each rule.

This setting led us to ***149 rules***; there were not found redundant rules. See fig 5.

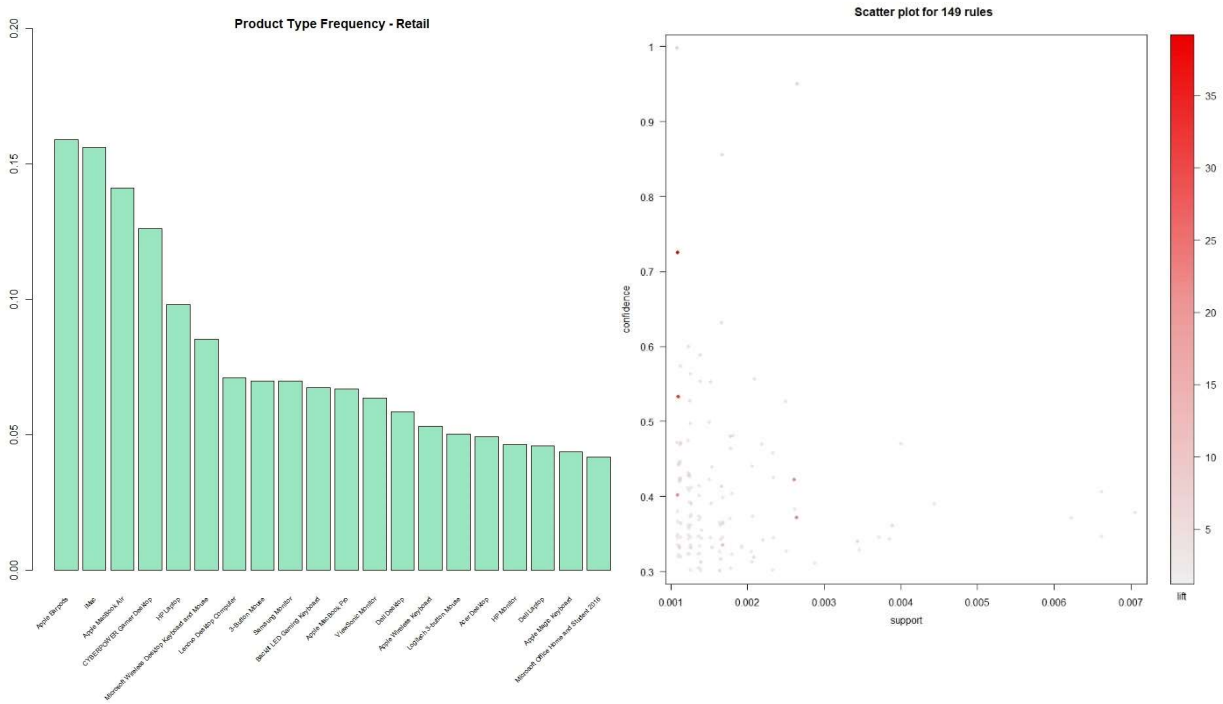

Fig 5. Frequency of Product types and rules in the retail dataset.

After inspecting the most meaningful rules in the business category subset, we observed the following rules:

1. {iPhone charger, DELL Wireless Keyboard and Mouse} -> Apple Macbook Air

2. {Large MousePad, Backlit LED Gaming Keyboard} -> Apple Macbook Air

Both rules had very high lifts, around 6. This means that we should consider reversing the LHS and the RHS. By doing this, the Apple Macbook Air is usually being bought with different kinds of accessories, such as the ones that you can read in the rules above.