Recommender Systems

Master Degree Program in Data Science and Advanced Analytics

Business Cases with Data Science

Github repository:

https://github.com/andremforte/BC3_GroupV

Group V:

Anis Tmar (m20211157)

André Forte (m20210590)

Opeyemi Mary Akande (m20211320)

Rafael Nunes (m20210832)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Index

1. Business Understanding

Understanding consumer purchase trends and habits is a critical responsibility for any retail company. Identifying the connections between different types of items, such as complementary, substitute, inferior, and convenience goods, gives to the organization a holistic perspective of its customers and product portfolio.

Gift-a-Lot is a registered non-store online shop situated in the United Kingdom. The company is specialized in selling one-of-a-kind presents for all occasions. The organization decided to recruit a team of external consultants to help them answer some of the concerns they've been pondering.

Using data from transactions between 01/12/2010 and 09/12/2011, Gift-a-Lot hopes to create a recommender system that would help users make better decisions by proposing items they like and improve their shopping experience on its website.

The Data Mining's goals are to present a Market Basket Analysis and to optimize a model that can indicate products that a specific customer could be interested in, suggest customers that could buy a particular product and recommend products that could be bought together.

2. Data Understanding

The retail dataset has 541909 instances and 8 features. During Exploratory Data Analysis, we got more familiar to the data, being able to extract information that was useful for the next steps.

At first, we noticed that the dataset had negative values for 'Quantity' and 'UnitPrice', probably representing cancellations (different purchases of the same products by the same customer with opposite quantity registered in a short period of time) or devolutions. Also, extreme values were seen in boxplots, meaning that we could have here possible outliers.

Additionally, 'CustomerID' and 'Description' registered missing values and the last feature had wrong values like 'barcode problem', 'wrong code?', 'missing', meaning that those rows don't correspond to products that were purchased (most of the descriptions were in lower case).

We found that we had data from 25900 different transactions of 4372 different customers. Besides, we were analysing purchases of 4070 different products. An interesting characteristic that we found was that we registered 8 customers with purchases from two countries, which could mean that they changed their residence during the period that we were considering.

After that, we continue the analysis to get more important patterns by checking:

- Top 10 products that were more purchased (number of products): 'WHITE HANGING HEART T-LIGHT HOLDER', 'REGENCY CAKESTAND 3 TIER','JUMBO BAG RED RETROSPOT','PARTY BUNTING','LUNCH BAG RED RETROSPOT','ASSORTED

COLOUR BIRD ORNAMENT','SET OF 3 CAKE TINS PANTRY DESIGN ','PACK OF 72 RETROSPOT CAKE CASES', 'LUNCH BAG BLACK SKULL.','NATURAL SLATE HEART CHALKBOARD'. (Annex 1)

- Proportion of Sales (money spent) of the Top 5 best countries: United Kingdom, Netherlands, Germany, EIRE, France. (Annex 2)
- Distribution of Customers per Country: most of the customers were from United Kingdom, following by Germany and France. (Annex 3)
- Number of Purchases per month/year: the best month (high number of products sold) was November 2011. (Annex 4)
- Regarding Monthly Analyses, the month with the highest revenue and with the highest quantity sold was November 2011. (Annex 5 and 6)

3. Data Preparation

Using manual filtering, we removed 2.2% of the data as outliers. The reasoning behind the removal was that the observations recorded a greatly high value, almost 80,000 times the 1st and 3rd quartile of the values for that variable. Also, as mentioned above, these items were cancelled a few minutes after placing the order. The other observations that were deleted are those with missing "CustomerID". Though these could be customers who shopped online without prior creation of the account, generating random "CustomerID" for them could bias the analysis.

The "StockCode" have values that could be interpreted as integers or strings. To surpass that problem, we added "SC" as prefix, so it will be considered as strings only. We also removed entries where the description names (mostly written in lower case) and/or the Stock Codes seem strange when compared with the format of other Stock Codes.

In total, we removed 26.9% of the provided dataset. This percentage was acceptable because we still have a large dataset after all the deleted observations.

4. Modeling

4.1. Market Basket Analysis

Market Basket Analysis is a common technique used by companies to understand possible associations among their products. Their goal is to understand the most common combinations of products to create patterns that could be useful to make better suggestions to their customers. With that, it is possible to understand which products are complements and substitutes, using that information to implement better strategies with recommendations when they want to improve the customers' experience.

4.1.1. Initial Dataset

At first, we decided to analyse the combinations of products considering all transactions. We encoded the variables considering 1 if the "InvoiceNo" included the product and 0 if not. Using mlxtend package (mlxtend documentation, n.d.), we applied Apriori and JP Growth algorithms. Even though Apriori is a simpler algorithm compared to JP Growth, we noticed that the last algorithm is faster than the first after checking the time. Besides, JP Growth is a "tree-based approach that limits data reads to two passes", having the results in linear time complexity, which save computational resources. (Bogart, 2021)

Regarding complementary products, we decided to select a min_support of 0.025. It means that the returned 150 items in both algorithms have occurred in at least 2.5% of the total of purchases. Since the results were the same for both algorithms, we decided to continue with JP Growth due to its advantages mentioned above.

Selecting a lift above 1 (it means that the items are more likely to be bought together) and confidence above 0.7 (it means that we have 70% of possibility of finding the consequent in a purchase given that antecedent is in the purchase), we filtered the 3 most important rules from the given 20. The metrics are presenting in Figure 1 and in Annex 7. The results could be seen in Table 1.

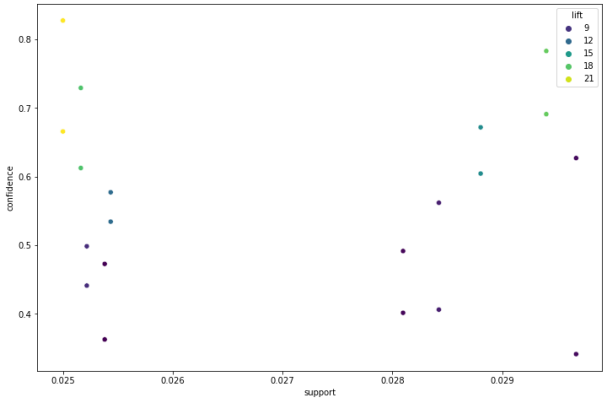


Figure 1. *Scatter plot comparing Confidence, Support and Lift*

Antecedents	Consequents
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
GARDENERS KNEELING PAD CUP OF TEA	GARDENERS KNEELING PAD KEEP CALM

Table 1. *Top 3 Association Rules of Complementary Products (All Dataset)*

Considering substitutes' analysis, we defined a different min_support (0.01) - the returned 987 items using JP Growth have occurred in at least 1% of the total of purchases. In this case, we had 976 rules. However, when we define the criteria to select the substitute rules (lift below 1,

meaning that the products are less likely to be bought together), we haven't found any significant combination.

4.1.2. MBA using the initial dataset split by Quarters

After analysing the general association rules, we decided to understand if splitting the initial dataset by quarters (December2010 and first quarter of 2011, second quarter 2011, third quarter 2011, fourth quarter 2011) we would get different conclusions.

After encoding all sub datasets, we applied JP Growth algorithm using the same criteria mentioned above (min_support of 1% for substitutes and 2.5% for complementary products). The results of each quarter can be seen in Annexes (Annex 8). It is possible to conclude that we have different rules when we analyse transactions by quarter, which could mean that we have different strong combinations of products depending on the period of the year.

4.1.3. MBA based on original dataset split by UK vs Worldwide

We also wanted to have an analysis based on national and non-national customers. We performed the same reasoning and the results can be seen in Annex 9. Since most of the customers were from United Kingdom, the association rules are similar to the rules when we consider the initial dataset. However, analysing non-national customers, it is possible to register association rules with lift below 1 (substitute products).

4.2. Recommender System using LightFM

The goal of Recommender Systems is to suggest items/individuals that could be interested in a specific product, based on interactions between products or individuals. These interactions could be explicit (when it's used a score or a like) and implicit (when it's considered a click, view or purchase). In this case, we performed an implicit analysis based on the purchases of the customers using LightFM algorithm.

LightFM is a "Python implementation of a number of popular recommendation algorithms for both implicit and explicit feedback" (LightFM documentation, n.d.). With this technique, we can analyse product and customer data, having them disposed in matrix factorization models.

4.2.1. Collaborative Filtering

Collaborative Filtering is a method to select products that a particular individual might be interested in based on the behavior of the other customers with similar characteristics.

At first, we started the analysis by creating an interaction matrix based on the following criteria: if a customer has purchased a product anytime, independently of the quantity he/she bought, we attribute 1, if not, the value would be 0. After that, we created dictionaries for customers and products individually since this is the focus of this approach.

After that, we needed to create the model that we were going to apply. Therefore, we transformed the interaction matrix into a sparse matrix. After that, we performed random train test split to get train and test data, with 80% and 20% respectively, to evaluate the performance of the model. As mentioned above, we applied LightFM and these were the final criteria we used. (LightFM documentation, n.d.)

- N_components = 30 – corresponds to the numbers of embeddings we wanted to create to define products and customers.
- Loss = 'warp' (Weighted Approximate-Rank Pairwise) – this parameter represents the loss function and it is useful when we have only positive interactions, and our goal is optimizing a number of recommendations.
- K = 10 – this value represents "the k-th positive example will be selected from the n positive examples sampled for every user".
- Epoch = 30 – performing 30 epochs to run
- N_jobs = 4 – we used 4 cores for execution

After that, we evaluate the model using AUC and Precision @k (LightFM documentation, n.d.). Regarding AUC, we got 0.959 and 0.875 for train and test data, respectively. These values correspond to how good our model is in predicting the effective recommendations for customers, considering all the products. Analysing the results, it is possible to see that it is model with quality.

The Precision @k analysis allows us to understand the precision of the model when it is generating top k recommendations – it represents the fraction of known positives in the first k positions of the ranked list of results. In our case, the model with k=10 was the one that gave us the highest precision in test data.

After selecting the algorithm and optimizing its parameters, we could start providing recommendations based on different perspectives (examples of the two perspectives can be seen in the notebook):

1. Recommend a product to a specific customer:

In this perspective, we want to recommend a product to a specific customer that could be interested, considering their past behavior. Making use of a personalized function, we got as outputs a list of products previously purchased by a particular customer and a list of 10 recommended products that the same customer could be interested in.

2. Recommend a customer to a specific product

In this approach, we want to provide a list of customers that could be interested in a certain product. Using a personalized function, we selected 15 customers to whom we could suggest a particular product.

4.2.2. Content-based Filtering

Similarly to what we did in Collaborative Filtering analysis, we created a matrix, but, in this case, we used cosine distances, employing products embeddings generated by the previous Matrix Factorization model. This technique helped us to compute similarity between products in order to obtain recommendations that could be similar to the product of interest. After that, we used a customized function to get the top 10 products that could be bought together with a specific product (example can be seen in notebook).

4.3. Cold-Start Problem

The cold start problem is well-known and well-studied in Recommender Systems. Because all of the customers and items had a previous record of purchases, the dataset given by Gift-a-Lot currently had no new users or new products. The business's recommender system may have to cope with two Cold-Start issues in the future.

The first may be introduced via the launch of a new or unpopular item. This problem produces a negative feedback loop in which non-popular things are poorly suggested and have far less publicity than popular items. The second method involves the addition of new users to the database. A freshly enrolled user with no transaction history would be unable to receive customized suggestions.

The techniques available to overcome cold-start difficulties have some characteristics across all categories. The basic technique for dealing with new users is to ask users to enter their demographic data upon enrolling, or to use data that is already accessible, such as in their social network profiles. This information might subsequently be used to create an initial user profile. A balance should be struck between the length of the user registration procedure and the amount of initial data necessary to be submitted, because if the process is too long, too many people will avoid it. Customers may also be requested to offer an after-purchase rating or review for each product they buy.

Another method for addressing the cold-start issue is to use hybrid recommenders. In circumstances when it is impossible to offer a thorough description of the item attributes, this technique balances the drawbacks of one category or model by combining it with additional recommenders.

The methodologies outlined above may be computationally costly, sluggish, or biased. The popular item technique is the cheapest solution Gift-a-Lot might use to remedy the Cold-Start

problem. Popular items in that region are recommended based on their popularity. If no demographic feature is provided or their quality is too low for new users, a popular technique is to offer them non-personalized content.

5. Business Implications

For most business owners, revenue growth is the most visible measure of success. This type of recommendation system may be utilized to improve revenue while also increasing customer satisfaction - key factor to improved company's sales and future purchases by customers. It is used to identify customers who could be interested in new items and to propose preference-tailored products to them (based on their prior behavior and the qualities of the products). With a good recommender system, a consumer feels less worried and more connected to the service, having a better experience. This could trigger them to buy more products because they receive information that is more relevant, increasing the sales of the company.

Regarding company's strategies, market basket analyses could be important to understand the more important and valuable combinations of products for their customers. This technique could be used to define promotional campaigns, offering bundles with those combinations. Moreover, it could be important to control stocking essentials quantities (products that are more valuable) or to solve problems with products in stock during too much time (making use of recommender system to suggest them to a specific group of customers that could be interested in).

Finally, with well-structured and planned business strategies, it is possible to reduce the staffs' workload and overhead, surpassing the challenges that a company faces when it doesn't have those techniques (e.g., when a company doesn't have a recommender system, it is difficult to surpass problems with stock quantities or promotional campaigns, leading to a less effective service).

6. Deployment[1]

The deployment of this model should be in an embedded Web Application. When a customer selects an item to his/her cart, similar products will be displayed based on the selected item or based on the preferences indicated on the demographic profile of the customer.

Deploying the recommender system could be hindered by factors such as cold start problem, as we discussed above, and wrong reviews or customers giving biased details about their personal information. Besides, it is important to consider that this step requires a big investment to deploy.

[1] (appier, 2021)

Funding is required to hire and train technical staffs who are responsible for analysing, monitoring and refining the recommendation engine in order to ensure that the system is working properly.

Finally, another challenge is related to privacy concerns. Having more information about customers will make the algorithm provide better recommendations. However, due to recent events regarding data leaks, individuals could feel reluctant to provide personal information. Therefore, it is necessary to make the customers have confidence that their information is secured.

References

appier. (2021, February 02). *7 Critical Challenges of Recommendation Engines*. Retrieved from https://www.appier.com/blog/7-critical-challenges-of-recommendation-engines

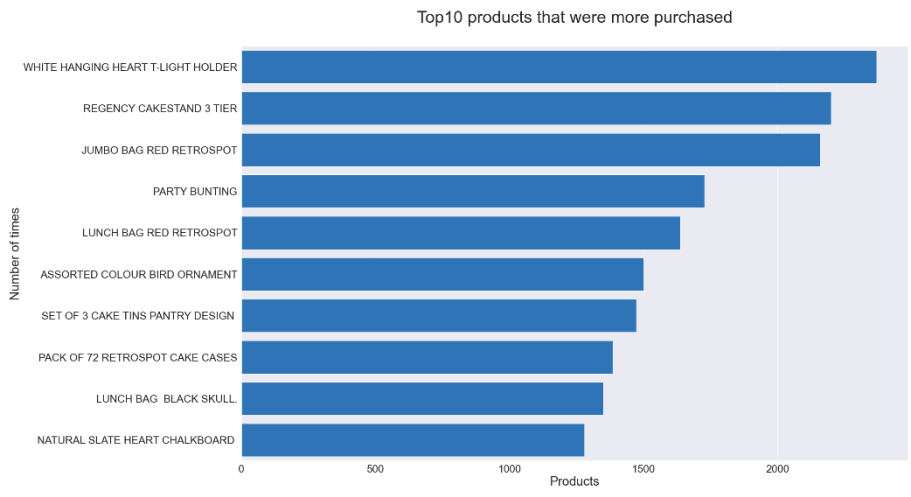
Bogart, B. (2021, October 5). *Towards Data Science*. Retrieved from The "Frequently Bought Together" Recommendation System: https://towardsdatascience.com/the-frequently-bought-together-recommendation-system-b4ed076b24e5

LightFM documentation. (n.d.). *LightFM*. Retrieved from https://making.lyst.com/lightfm/docs/index.html

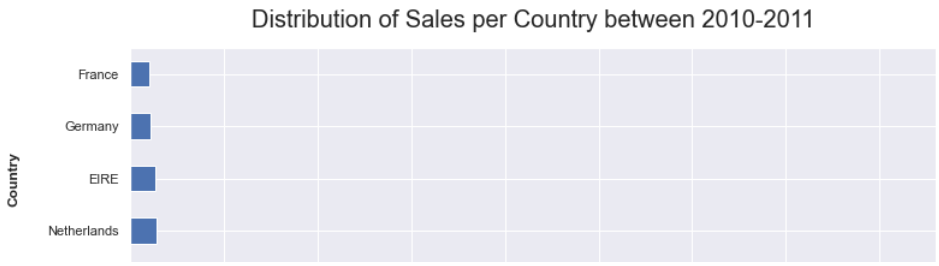
LightFM documentation. (n.d.). *LightFM*. Retrieved from Model evaluation: https://making.lyst.com/lightfm/docs/lightfm.evaluation.html

mlxtend documentation. (n.d.). *mlxtend*. Retrieved from https://rasbt.github.io/mlxtend/#examples

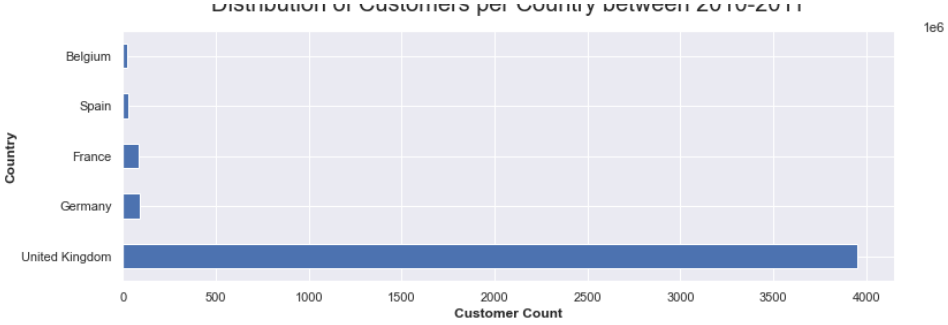
Annexes



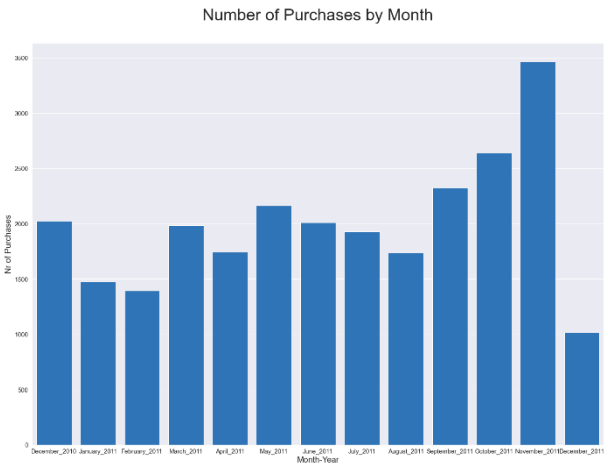
Annex 1. *Top 10 products that were more purchased*



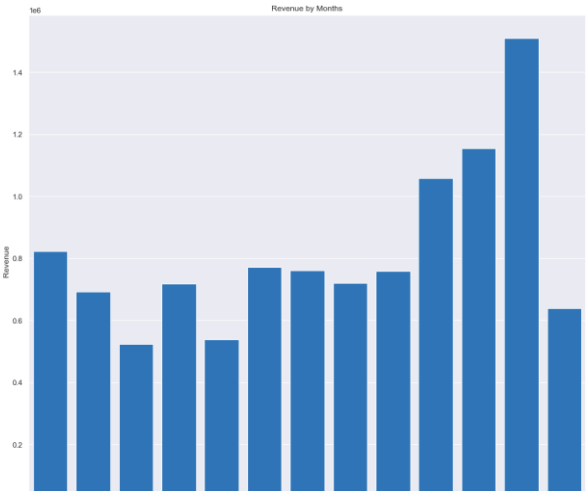
Annex 2. Distribution of Sales per country



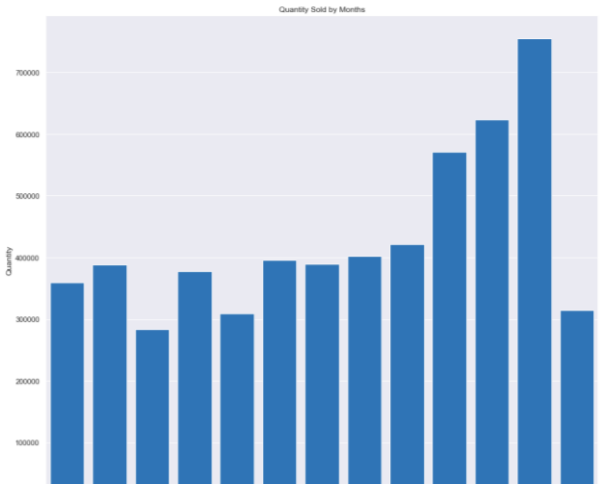
Annex 3. *Distribution of Customers per country*



Annex 4. *Number of Purchases by Month*



Annex 5. *Revenue by month*



Annex 6. *Quantity sold by month*

***Annex 7**. Complementary products of initial dataset by lift*

***Annex 8.** Association Rules by quarters*

Antecedents	Consequents
CANDLEHOLDER PINK HANGING HEART	WHITE HANGING HEART T-LIGHT HOLDER
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER

***Table 2**. December and First Quarter Top Complementary products*

Antecedents	Consequents
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER
PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER
PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER
GARDENERS KNEELING PAD CUP OF TEA	GARDENERS KNEELING PAD KEEP CALM

***Table 3**. Second Quarter Top Complementary products*

Antecedents	Consequents
JUMBO BAG PEARS	JUMBO BAG APPLES
ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE GREEN
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
DOLLY GIRL LUNCH BOX	SPACEBOY LUNCH BOX
GARDENERS KNEELING PAD CUP OF TEA	GARDENERS KNEELING PAD KEEP CALM
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER
PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
SET OF 12 MINI LOAF BAKING CASES	SET OF 12 FAIRY CAKE BAKING CASES
SET OF 6 TEA TIME BAKING CASES	SET OF 12 FAIRY CAKE BAKING CASES
SET OF 6 SNACK LOAF BAKING CASES	SET OF 12 FAIRY CAKE BAKING CASES

***Table 4**. Third Quarter Top Complementary products*

Antecedents	Consequents
ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED
ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN
SET OF 3 WOODEN TREE DECORATIONS	SET OF 3 WOODEN STOCKING DECORATION
SET OF 3 WOODEN STOCKING DECORATION	SET OF 3 WOODEN TREE DECORATIONS
WOODEN STAR CHRISTMAS SCANDINAVIAN	WOODEN HEART CHRISTMAS SCANDINAVIAN
WOODEN HEART CHRISTMAS SCANDINAVIAN	WOODEN STAR CHRISTMAS SCANDINAVIAN
WOODEN TREE CHRISTMAS SCANDINAVIAN	WOODEN STAR CHRISTMAS SCANDINAVIAN
WOODEN TREE CHRISTMAS SCANDINAVIAN	WOODEN HEART CHRISTMAS SCANDINAVIAN
WOODEN STAR CHRISTMAS SCANDINAVIAN, WOODEN TREE CHRISTMAS SCANDINAVIAN	WOODEN HEART CHRISTMAS SCANDINAVIAN
WOODEN TREE CHRISTMAS SCANDINAVIAN, WOODEN HEART CHRISTMAS SCANDINAVIAN	WOODEN STAR CHRISTMAS SCANDINAVIAN

***Table 5**. Fourth Quarter Top Complementary products*

***Annex 9.** Association Rules by national and non-national customers*

Antecedents	Consequents
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER
GARDENERS KNEELING PAD CUP OF TEA	GARDENERS KNEELING PAD KEEP CALM

***Table 6**. UK top complementary products*

Antecedents	Consequents
DOLLY GIRL CHILDRENS CUP	DOLLY GIRL CHILDRENS BOWL
DOLLY GIRL CHILDRENS BOWL	DOLLY GIRL CHILDRENS CUP
DOLLY GIRL CHILDRENS BOWL	SPACEBOY CHILDRENS BOWL
SPACEBOY CHILDRENS BOWL	SPACEBOY CHILDRENS CUP
SPACEBOY CHILDRENS CUP	SPACEBOY CHILDRENS BOWL

***Table 7**. Worldwide top complementary products*

Antecedents	Consequents
JUMBO BAG WOODLAND ANIMALS	REGENCY CAKESTAND 3 TIER
JAM MAKING SET PRINTED	ROUND SNACK BOXES SET OF4 WOODLAND
RED RETROSPOT MINI CASES	REGENCY CAKESTAND 3 TIER
ROUND SNACK BOXES SET OF 4 FRUITS	REGENCY CAKESTAND 3 TIER
ROUND SNACK BOXES SET OF 4 FRUITS , ROUND SNACK BOXES SET OF4 WOODLAND	REGENCY CAKESTAND 3 TIER
RED RETROSPOT CHARLOTTE BAG	REGENCY CAKESTAND 3 TIER
RABBIT NIGHT LIGHT	REGENCY CAKESTAND 3 TIER

***Table 8**. Worldwide top substitute products*