BUSINESS CASE 3

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Gift-a-Lot Recommender System

Group PN:

Mafalda Garcia, number: 20210763

Simão Pereira, number: 20210250

Tiago Santos, number: 20210548

Rui Ribeiro, number: 20211017

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

INDEX

1. Introduction

Gift-a-Lot is a UK-based and registered non-store online retailer with about 80 members of staff. The company was established in 1981 mainly selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogs, and orders were taken over by phone calls.
Two years ago, the company launched its website and shifted completely to the web. Since then, the company has maintained a steady and healthy number of customers from all parts of the world and has accumulated a huge amount of data about its customers. The company also uses Amazon.co.uk to market and sell its products.
To carry out this project we followed a CRISP-DM methodology. Firstly, we start by reconnaissance of the business, analyzing its objectives, which materials we had access to, and defining the data mining goals for this project. Next, we will perform recognition of the data itself, seeing its composition in detail and we will start the necessary steps for the data preparation, making all the necessary changes so that it is possible to apply the respective prediction models in it.

1. Business Understanding

1.1 Business Objectives

Jane, Gift-a-Lot's newly appointed Chief data officer, has a small team of in-house data scientists, with little time to dedicate to new projects. Consequently, she decided to hire us to address some of the questions they've been struggling with.

With the company's data, they expect us to build a recommender system that can facilitate user choices by recommending items, the user likes, and improve the user experience when making purchases on its website. A particular challenge is the cold start problem.

1.2 Assess Situation

The company provided us a dataset composed of 541909 rows containing information of 4371 customers regarding 7 variables.

The variables included in this dataset are the following: “InvoiceNo”, “StockCode”, “Description”, “Quantity”, “InvoiceDate”, “UnitPrice”, “CustomerID” and “Country”.

1.3 Data Mining Goals

The business objectives already mentioned can also be described in a technical way, more precisely, like Data Mining goals.

Our goals focus on analyzing the given data was to do a Market Basket Analysis, where we should identify the complementary and substitute products and what were the main types of consumer behavior in the business, a recommender system addressing the problem of cold start, and adequate evaluation strategies, select an appropriate quality measure and elaborate on the challenges and recommendations in implementing the recommender system.

3. Data Understanding

3.1 Collect Initial Data

The dataset was presented in a CSV document given by the company and we loaded it into a Jupyter Notebook using the pandas library.

3.2 Description of data

The dataset has a total of 541909 rows each representing a product purchased and 8 related variables.

The dataset had a lot of useful information and it was very complete in terms of information about the company’s customers.

3.3 Exploring data

When exploring our dataset, we used .head() method to see a brief overview of the information, .info() to understand the various types of data we had to deal with and the .describe() to understand some statistics of our data.

3.4 Verify data quality

In the verification of our data quality with found some null values, duplicates, and some inconsistencies in the data set. To discover this, we used .isnull(), .duplicates(), .sum() and .describe() in our variables.

The null values found were in the variables "Description" and "CustomerID" and we also found a total of 5268 duplicates.

In terms of inconsistencies, we found that the dataset had some "Descriptions" that did not make sense for the purpose of this work and also negative quantities and prices.

4. Data Preparation

4.1 Selection and Cleaning of Data

Before uploading the data into the jupyter notebook we used excel to filter some inconsistencies and with this, we ended up erasing some rows with descriptions that did not make sense.

After this, we opened the data with jupyter notebook and proceed to erase the duplicates found using .drop_duplicates().

We used some filters to have only positive quantities and unit prices and to erase 2 records that had very large quantities as this could interfere with our work.

We also changed the data type of the variable "Description" from object to string, using str.strip() to remove any spaces that could be at the beginning or at the end of every description and the type of the variable "InvoiceDate" to DateTime type.

In the variable "InvoiceID" we changed its type to string so that we could remove the rows in which this variable begins with a "C", this meant that the purchase was canceled and it had no meaning for our analysis. In the variable "Description" we erased the rows that had the word "POSTAGE".

4.2 Constructing the data

When constructing data the only thing we've done was to create a new variable called "TotalSpend" by multiplying the unit price by the quantities, this variable gave us the total amount spent with every product for every customer.

4.3 Formatting the data

As we were doing some different analyses like the market basket analysis and the recommendation system, we had to format our data independently for each goal. For this, we mainly used the .groupby() method. For example, in the Market Basket Analysis, we used the "InvoiceNo" as an index and in the Recommendation system, we used "CustomerID" with non-null values instead.

4.4 Data Analysis

As we had to adapt our data to the different data mining goals we had to analyze and understand which variable we should use for each goal. This part was done with analytical thinking and also with the help of some research.

5. Exploratory Data Analysis

5.1 Market Basket Analysis

5.1.1 Association rules

To start building Market Basket Analysis to answer the problems that were posed to us, we used the Association Rules that helped us calculate the probability of existing a relationship between two data items. For this part, we started by encoding the data and removing the 'POSTAGE' column and then to know which the most frequent items were bought, for this we used a priori.

Then, using association_rules() and defining minimum confidence of 50%, we were able to access all association rules of our dataset. We can analyze that the maximum existing support was 0.041696, confidence was 0.904682 and lift was 18.243093.

The first problem that we were able to solve using the Market Basket Analysis model was to be able to identify which types of products can be complementary and substitutes. After trying to identify the complementary and substitute products for the entire dataset, we ended up not finding any relevant results for the sample and so we decided to explore country by country until we found a good example in France. We assumed as complementary products those that had a high lift and simultaneously high confidence, represented in GREEN in the attached table (Table 1.).

Still regarding only transactions carried out in France, in the same table (Table 1) presented in the annexes, we identified in BLUE which was the substitute products (those that are consumed in substitution of another product) and considered as substitutes those that were the same item and the only difference between themselves is the color or the model and also those in terms of support and confidence were relatively lower than in complementary ones.

To help us visualize the complementary and substitute products, we built a network graph (Graph 1.)

5.1.2 Consumer Behavior

The second problem that was solved with the application of the model was being able to verify which were the main types of consumer behavior in the business and for that, we built an RFM Matrix to help us with the analysis. This matrix is a customer segmentation technique that will divide customers into groups according to their past purchase behavior. First, we verified that we didn't have any missing value, no negative value and since the dataset contains transactions until 12/09/2011 we will consider 12/10/2021 for recency. After these verifications, we then started the customer segmentation by building the RFM Matrix, applying quantiles, and defining the score that will allow us to know which are the best customers, being considered as those who verify the lowest recency, highest frequency, and highest monetary value.

Since the best score obtained was = 111, we can see the top 5 of the best customers in the attached table (Table 2) and where the best customer is CustomerID= 14646.0 in which the last purchase was 1 day ago, with a total of 2060 purchases and a total spent of 279138.02. We can also analyze the top 5 customers with a score =311 (Table 3), which are those where the last purchase was a long time ago, but with a nice frequency and had spent a nice amount of money, and for these customers, we suggest further discounts (Black Friday for example) to encourage them to go to the store. Finally, we also have those customers with a score = 444 (Table 4), which we should ignore for analysis because they are considered cheap consumers (super high recency and with a super low frequency and monetary value).

Note: to solve this problem we also started clustering using k-means to cluster the consumer's behavior. However, it wasn't for sure the best approach that's why we changed to RFM Matrix.

5.2 Recommender System

5.2.1 Collaborative Filtering based Product Recommender System

Our work here was to come up with a System to recommend products to a customer. There are three forms of doing so: Collaborative Filtering, Content-Based Filtering, and Hybrid between the last two.

We couldn't use the Content-Based approach given that it requires more detailed information about the products, which we don't have. Therefore, we used Collaborative Filtering.

Given that the recommendations given to the client are going to be found regarding other customers, we disregarded clients who only made one purchase from this site.

This model compares the client, to which it wants to recommend products, to every other client in the DB. We started this by listing all the clients who purchased at least one of the same products that our "wanted client".

The Collaborative Filtering needs a Rating for the product (given by each customer that bought it) and we decided that, given the information we have, the best Rating the client could give to a product was to buy it many times (quantity-wise). Thus, and based on the distribution of the quantity variable, we deployed the following Rate system for each product bought from each similar customer:

- Bough 2 or fewer times by this customer: Rate 2;
- Bought between 3 and 6 times by this customer: Rate 3;
- Bought between 7 and 12 times by this customer: Rate 4;
- Bought more than 13 times by this customer: Rate 5.

Given that this Rate needs to be weighted depending on the similarity of the similar customer who bought the product to our customer, we created a "Similarity Index" that is a ratio between the number of similar purchases and total purchases from each similar client.

Based on this Index, we got a top 30 similar clients and their purchases. Then, we sorted the purchases based on the weighted Rating and grouped this list by product and their mean weighted Rating. By sorting this, we get a list of the better products to recommend to our clients.

5.3 Cold Start Problem

5.3.1 Relevant products to new customers

The cold start problem in recommendation systems means that the current circumstances are not yet 100% favorable for the engine to obtain the best results and for that, the approach that we found most viable to adopt was to suggest to new customers the most popular products in each country (the items that were purchased most frequently by consumers).

Regarding the Gift-a-Lot dataset, we applied this strategy to the first 11 countries where customers made the most purchases, and through the total amount, we were able to check the top of most frequently purchased items in each one (Table 5).

This strategy is super simple, easy to implement, and an efficient way for Gift-a-Lot to deal with the cold start problem.

6. Evaluation

6.1 Evaluate Results

Concerning Recommender Systems, using an original method we got the output we wanted which is the top products that would be more interesting for a given customer, based on their similarity to other customers and their "evaluations" of the products they bought.

When it comes to the Market Basket Analysis, we used the RFM method to determine the consumer behavior and it was possible to successfully analyze Gift-a-Lot's top 10 best customers. Also, inside the Market Basket Analysis, we used association rules and the apriori algorithm to help us understand which types of products are substitutes and complementary. We could not identify complementary and substitute products for the whole dataset because there were mainly substitute products however when we filtered the information by the country, we managed to find good examples for this matter in France.

We consider that we have fulfilled our goals and outputs which can be an important part of their website in terms of recommendations and give great insight into a very important part of the business, through which the management can make thoughtful and confident key decisions.

6.2 Review Process

We believe that we achieved our goals in logical and simple ways but after analyzing our algorithm we conclude that there could be other ways of approaching the problems:

- Regarding RS we could have done the similarity index in a different way, using a better understanding of the business and other libraries/methodologies in python.
- Regarding Market Basket Analysis, we could have used a more effective/efficient method, with a dataset with more product and customer characteristics.

Despite this, we believe we did a good job.

6.3 The next steps

When it comes to the improvement of our work, we conclude that with some extra time we could either simplify our method or study to discover a better one. Considering the deadline, we have and the fact that we consider our job to give the wanted result and that the algorithm doesn't take much time to run, we disregarded these next steps.

7. Deployment

Regarding this topic, our only recommendation is about the Recommender System. We suggest that it is only used when the customer has made 2 or more purchases on the website. From then on, the algorithm should present thoughtful Recommended Products based on their previous ones.

This system should be implemented on the main page of the website after the client logs in and should be presented at the top of the page under the title "Recommended products for you". If the customer does not log in or they are still making their first purchase the webpage should present the products based on the cold-start problem algorithm under the same title, given the country of the IP address from the customer's computer.

When it comes to market basket analysis, we don't find any limitations in terms of purchases given that it depends on the selected product. When the client adds to the basket a product it should be suggested to them a complementary one or if the product that the client is trying to buy is out of stock the website should automatically present a substitute product based on our market basket algorithm.

8. Conclusion

Using our recommendation system and our market basket analysis we believe that Gift-a-Lot can increase the sales by recommending the right products to frequent customers that will make them spend more money in the store and also by putting together complementary products and replacing some products with others that will give similar value to the clients.

Although this method will increase the sales it is clear that there is a lot of room for improvement in the recommender system algorithm and market basket analysis.

So to better deploy these strategies we advise the company to collect more data about the customers' interaction on the website and also gather more information about the characteristics of the products.

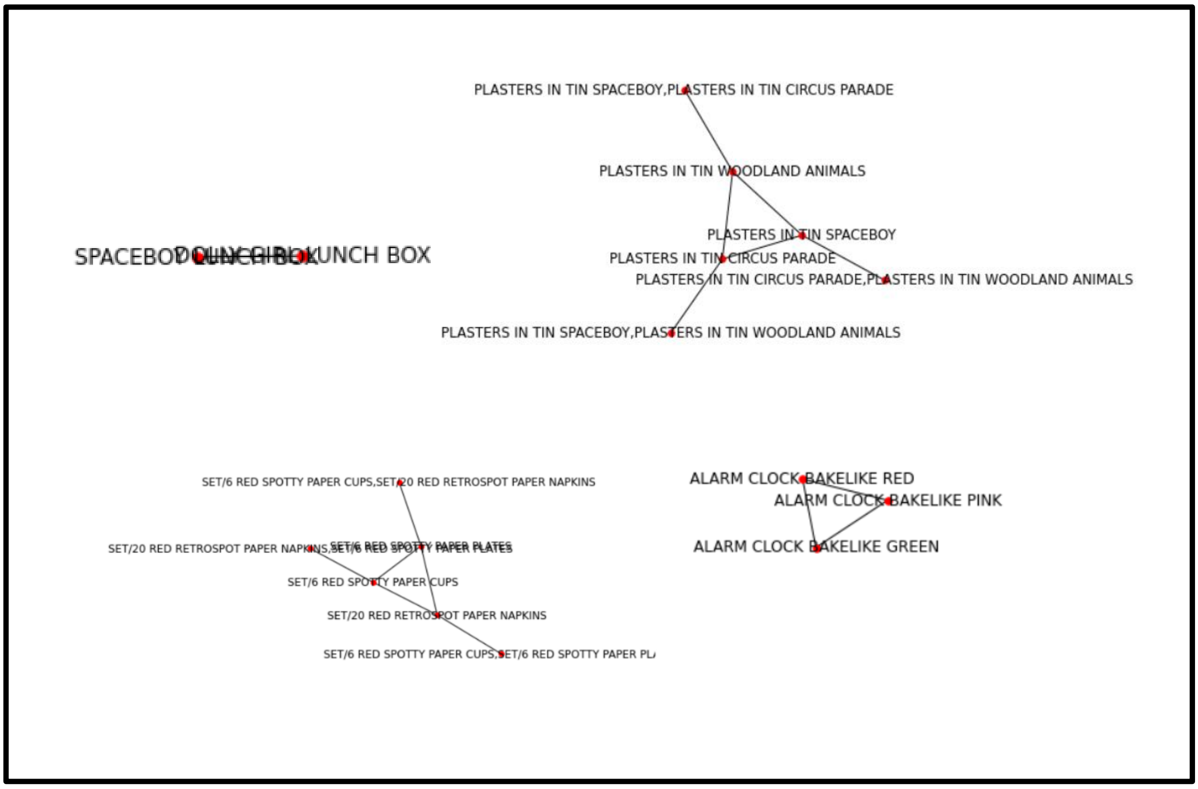
This will also help Gift-a-Lot to better understand what kind of customers they have and what kind of products are being bought so that they can better address the cold-start problem and recommend more precisely products for new customers.

Appendix

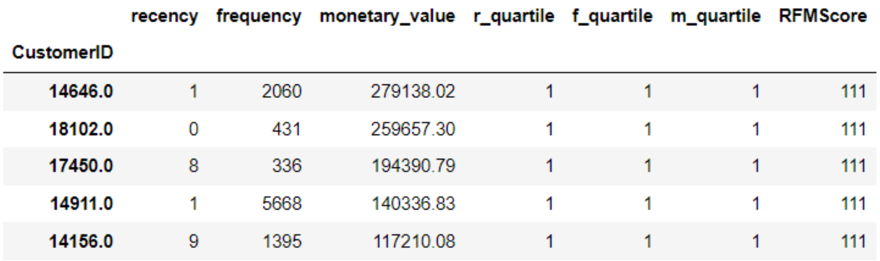
Antecedents	Consequentes	Support	Confidence	Lift
ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	0.080940	0.815789	8.444523
ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	0.080940	0.837838	8.444523
SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	0.125326	0.888889	6.808889
SET/6 RED SPOTTY PAPER PLATES	SET/6 RED SPOTTY PAPER CUPS	0.125326	0.960000	6.808889
SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS	SET/6 RED SPOTTY PAPER PLATES	0.101828	0.975000	7.468500
SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER CUPS	0.101828	0.975000	6.915278

Table 1. – France: Substitutes & Complementary products

(Blue: substitutes; Green: complementary)

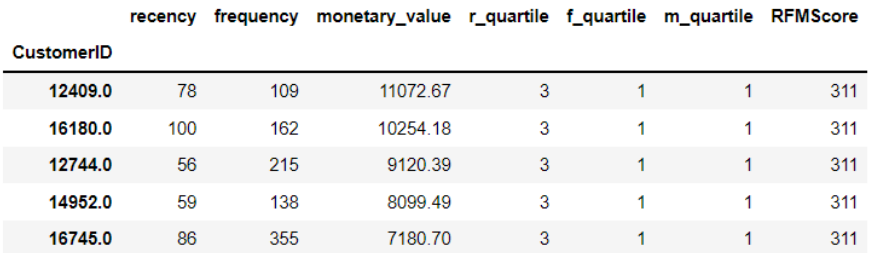

PLASTERS IN TIN SPACEBOY,PLASTERS IN TIN CIRCUS PARADE
PLASTERS IN TIN WOODLAND ANIMALS
PLASTERS IN TIN SPACEBOY
SPACEBOY LUNCH BOX
DOLLY GIRL LUNCH BOX
PLASTERS IN TIN CIRCUS PARADE
PLASTERS IN TIN CIRCUS PARADE,PLASTERS IN TIN WOODLAND ANIMALS
PLASTERS IN TIN SPACEBOY,PLASTERS IN TIN WOODLAND ANIMALS
SET/6 RED SPOTTY PAPER CUPS,SET/20 RED RETROSPOT PAPER NAPKINS
ALARM CLOCK BAKELIKE RED
ALARM CLOCK BAKELIKE PINK
SET/20 RED RETROSPOT PAPER NAPKINS,SET/6 RED SPOTTY PAPER PLATES
SET/6 RED SPOTTY PAPER PLATES
ALARM CLOCK BAKELIKE GREEN
SET/6 RED SPOTTY PAPER CUPS
SET/20 RED RETROSPOT PAPER NAPKINS
SET/6 RED SPOTTY PAPER CUPS,SET/6 RED SPOTTY PAPER PL

Graph 1. – Network graph: France



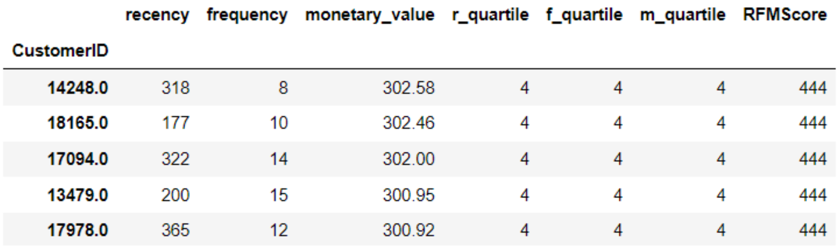
	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
CustomerID							
14646.0	1	2060	279138.02	1	1	1	111
18102.0	0	431	259657.30	1	1	1	111
17450.0	8	336	194390.79	1	1	1	111
14911.0	1	5668	140336.83	1	1	1	111
14156.0	9	1395	117210.08	1	1	1	111

Table 2 – Consumer Behavior: Best customer



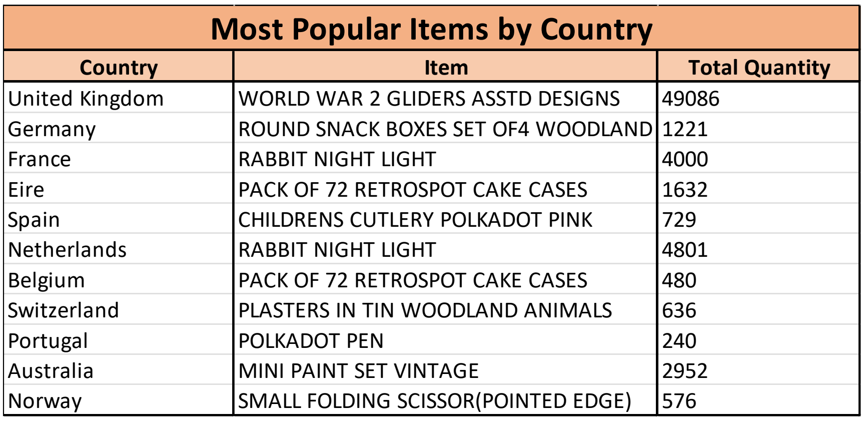
	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
CustomerID							
12409.0	78	109	11072.67	3	1	1	311
16180.0	100	162	10254.18	3	1	1	311
12744.0	56	215	9120.39	3	1	1	311
14952.0	59	138	8099.49	3	1	1	311
16745.0	86	355	7180.70	3	1	1	311

Table 3 - Consumer Behavior: Medium customer



	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
CustomerID							
14248.0	318	8	302.58	4	4	4	444
18165.0	177	10	302.46	4	4	4	444
17094.0	322	14	302.00	4	4	4	444
13479.0	200	15	300.95	4	4	4	444
17978.0	365	12	300.92	4	4	4	444

Table 4 - Consumer Behavior: Worst customer



Most Popular Items by Country		
Country	**Item**	**Total Quantity**
United Kingdom	WORLD WAR 2 GLIDERS ASSTD DESIGNS	49086
Germany	ROUND SNACK BOXES SET OF4 WOODLAND	1221
France	RABBIT NIGHT LIGHT	4000
Eire	PACK OF 72 RETROSPOT CAKE CASES	1632
Spain	CHILDRENS CUTLERY POLKADOT PINK	729
Netherlands	RABBIT NIGHT LIGHT	4801
Belgium	PACK OF 72 RETROSPOT CAKE CASES	480
Switzerland	PLASTERS IN TIN WOODLAND ANIMALS	636
Portugal	POLKADOT PEN	240
Australia	MINI PAINT SET VINTAGE	2952
Norway	SMALL FOLDING SCISSOR(POINTED EDGE)	576

Table 5 – Cold Start