BC3: Gift Shop Recommender System

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

April 2022

Q Consulting

Laura Isabella Cuna, 20211312

Amelie Florentine Langenstein, 20210637

Tongjiuzhou Liu, 20211012

Nina Urbancic, m20211314

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

Table of Contents

1. Introduction

Online shopping has been growing rapidly over the past few years. In 2021, the number of digital buyers was 2.14 billion, which makes 27.6 percent of the world's population and 900 million more compared to the year before. That's a 4.4 percent year-over-year increase.

There's no real surprise to this as online shopping is becoming increasingly convenient with the speed at which the internet connects the world. Also, online shopping was especially popular during the COVID19 pandemic since many countries closed all physical stores and there were no other options. Following these latest trends in the rise of online shopping, we can see that the potential for e-commerce is enormous (Oberlo, 2022).

Gift-a-Lot, a UK-based company, saw the opportunity and jumped on the bandwagon by selling unique all-occasion gifts. This shift from a traditional store to an online shop has happened in the last couple of years and since then, the company has acquired a healthy number of customers from all parts of the United Kingdom and the world.

The company has now hired us to answer some of the questions they are struggling with. They would like to know which items they should recommend to specific users, based on the data they provided, and to improve user experience. Our work was also supposed to explore which products are frequently bought together and which items should be suggested to new customers which have never made a purchase before.

According to the CRISP-DM methodology, we first elaborated the business objectives and set the data mining goal for this project. Then we moved on to exploring and understanding the dataset and performing data cleaning and transformation necessary for the modeling part.

2. Business Understanding

By operating online for the last two years, Gift-a-Lot has accumulated a huge amount of data about its customers, where many of them are wholesalers. To improve the user experience of the customers and increase sales, the company wanted to better understand the purchasing behavior of their customers.

Offering customers additional guidance on product choices helps to simplify decisions and can increase purchases. This applies especially to online shopping. Therefore, it is crucial to understand the purchasing behavior of customers and identify different product types to get an overview of the customers and the products (Iyengar & Lepper, 2000).

Business Objective

Gift-a-Lot aims to increase sales and improve user experience. Therefore, they want to facilitate user choices by recommending items they might like and suggesting items to new customers.

Data Mining Goal

The goals of this data mining project are to analyze the purchasing behavior of customers, conduct a market basket analysis to identify the relationships between the products (complementary and substitute), create a recommender system based on item/item interaction, and create a solution for the cold start problem.

3. Data Understanding

The dataset we received consisted of 541.909 records that corresponded to 25.900 valid transactions between 01/12/2010 and 09/12/2011. The records were represented by 8 different features and associated with 4.070 unique items and 4.372 customers from 38 different countries.

We observed 1.454 missing values for *Description* and 135.080 missing values for *CustomerID.* The fact that so many values for *CustomerIDs* are missing could be because this company allows its customers to make purchases without logging in. Therefore, we assumed that these samples represent real purchases.

4. Data Preparation

The dataset contained purchases that were canceled after the payment - we separated these samples from the non-canceled ones. This left us with 532.621 non-canceled purchases.

We removed all rows where the *UnitPrice* or the *Quantity* was 0 or less as we identified those as invalid.

We dropped all duplicates for *Description* and *StockCode* as these should be unique and each represents just one product. Consequently, we also dropped all entries for *StockCode* where stock codes represented multiple products. Additionally, we removed any *StockCode* entry that included 'POST' as that represented the postage cost, not a product.

Since the *Descriptions* that represent a product are written with capital letters, we removed all rows where the *Description* contained lower case letters.

After transforming the *InvoiceDate* column to a pandas Datetime object, we created five new

features to be able to assess when the customers complete their purchases: *year, month, day, hour, day_of_week*.

The last feature was created by multiplying the *UnitPrice* by the *Quantity* and showing the *total_price* of the order.

5. Visual Explorations

In this part, we used some visualization techniques and tried to answer the 8 questions below.

Q1: When do people order the most?

Hour of the day: We plotted a bar chart which showed us that people order the most at 10:00 AM and 12:00 PM.

Day of the week: We plotted a bar chart which showed us that people order the most on Tuesdays.

Day of the month: We plotted a bar chart which showed us that people order by far the most on the 12th day of the month. The 12th day of the month sees more than twice the amount of orders as the second most popular day which is the 11th.

Month: We plotted a bar chart which showed us that people order the most toward the end of the year, more specifically in September and November.

Q2: When do they order again?

The histogram and the boxplot we plotted here showed us that by far most customers order again within the first 4 days of the first purchase. It's quite common for customers to order up until the first 44 days, but after that, it's not so common for customers to order again.

Q3: What is the hottest product?

With the help of a treemap, we identified that the most frequently bought product - and therefore the hottest product - was Paper Craft, Little Birdie which was sold more than 80.000 times and represented 2% of all products sold.

In terms of total sales, Regency Cakestand 3 Tier generated the most revenue at around 174.484,74$ and represented 2% of all sales. The only thing that generated more revenue was the Dotcom Postage, but we did not consider that as it was not a product of the company but the postage cost.

Q4: Which products are most often sold together?

We discovered that the following products were bought together most frequently:

Most frequently bought together products: 4 Purple Flock Dinner Candles, Assorted Colour T-Light Holder, Multi Colour Silver T-Light Holder, Victorian Glass Hanging T-Light, 3 Tier Cake Tin Green And Cream, Round Cake Tin Vintage Green, Belle Jardiniere Cushion Cover, Poste France Cushion Cover, Gardeners Kneeling Pad Cup Of Tea, Union Stripe With Fringe Hammock, and Recycled Acapulco Mat Pink.

Q5: What is the largest order?

The largest order both per quantity sold and the total price was from the same order. A customer from the United Kingdom purchased 80.995 products with a total price of $168.469. The *InvoiceNo* was 581483.

Q6: What is the number of products sold and the total sales for countries?

When looking at the number of products sold and the total sales, we divided the world into two parts: the United Kingdom and the rest of the world (International). The UK customers bought 4.047.031 products and generated 7.704.739$ in revenue.

The rest of the world's sales are shown in the Choropleth Map. We can see that the country with the second-highest sales is the Netherlands, followed by Germany, France, and Australia.

Q7: What are the most popular products in different countries?

The three most popular products in the UK that were sold the most were: Paper Craft, Little Birdie; Medium Ceramic Top Storage Jar, and World War 2 Gliders Asstd Designs.

The three most popular products in the Netherlands that were sold the most were: Rabbit Night Light, Spaceboy Lunch Box, and Dolly Girl Lunch Box.

The three most popular products in Germany that were sold the most were: Round Snack Boxes Set Of4 Woodland, Assorted Colours Silk Fan, and Woodland Charlotte Bag.

The three most popular products in France that were sold the most were: Rabbit Night Light, Mini Paint Set Vintage, and Red Toadstool Led Night Light.

The three most popular products in Australia that were sold the most were: Mini Paint Set Vintage, Rabbit Night Light, and Red Harmonica In Box.

Q8: What are the most frequently bought products?

The top three most frequently bought products were Regency Cakestand 3 Tier which was bought 1723 times, Jumbo Bag Red Retrospot which was bought 1618 times, and Assorted Colour Bird Ornament which was bought 1408 times.

6. Modeling

Clustering

For Modeling, we started with Clustering to analyze the customers based on four characteristics: *Product_distinct*, *Order_times*, *Avg_product_price* and *total_price*. We used the k-Prototype algorithm, which works with numeric and categorical data. We normalized the four numeric features we created.

The Elbow Method helped us to derive the optimal number of Clusters. The customers are represented in 4 different Clusters: In Cluster 1, the customers have the lowest values in every single characteristic. The customers of Cluster 2 behave similarly. Clusters 2 and 3 consist of only a few customers. The customers of Cluster 2 order more often and also order different products, whereas the customers of Cluster 3 order products with a higher average product price. We used t-SNE to visualize our Clusters.

Market Basket Analysis

Association Rules

Association rules help to establish the relationships between different purchases and products. We built two association rules for cluster 0 and cluster 1, as they represented most of the customers.

Based on the cluster means, clusters 2 and 3 indicate that these customers are probably resellers and premium customers, because of a high order quantity for cluster 2 and high average price for cluster 3. However, clusters 2 and 3 had small quantities and were therefore not considered. It would be difficult to build general association rules based on these clusters.

The values for *product_distinct*, *order_times,* and *total_price* of cluster 0 are slightly higher than the ones of cluster 1, except for the average price of goods. Cluster 0 contains fewer customers, so we assumed that 0 is the priority group and 1 is the general group.

For clusters 0 and 1 we built association rules for complementary and substitute products. The association rules for complementary products were created based on which products were bought together and the ones for substitute products were based on which products could replace each other.

To create the Associated rules we created a new dataset only containing customers with IDs. After that, we used the customer data to create the new data frame only with unique *InvoiceNo* and *StockCode*. This dataframe was then transformed into a matrix to apply the function of the TransactionEncoder. This was used to apply the *Apriori algorithm*.

Complementary Products

From the Apriori algorithm, we got the product combinations and the corresponding support value. This was used to generate the association rules. We set the lift greater than 2 to get the list of complementary products. Usually, the lift should be 1, we just set it to 2 to get more plausible results. After that, we sorted our table by the lift score in descending order. This shows the association rules starting with the ones with the highest confidence to the lowest confidence.

Substitute Products

We assumed that when the lift is lower than two, the products are substitutes. Therefore, we built a substitution system based on the substitute product list. In cluster 0 we found 30 combinations of substitute products. In cluster 1 the products with the stock codes 22726, 22727, and 22728 are substitutes.

Recommender System

We built a function for the recommender system. The function has three inputs: the StockCode of the product, the rules based on the association rule of cluster 0, and the number of results. The default number of the results is 5. The function will use the StockCode to locate the consequents first. This gave us the first 5 consequent StockCodes. In the end, we generated the product name by the StockCode provided.

Cold Start Function

The last thing we were asked to explore was which products should be offered to new customers who have never visited our website before. For this, we generated a new function based on their IP address which gave us their location (country), and the login time which is computer generated and completely automatic.

We split this recommendation into two categories:

Countries where we have previously recorded data

First, we listed the top two most popular products of all time. We continued by listing the two most popular products of that day of the week (for example, if users view the website on Monday, they will see the products that were most popular on other Mondays), the two most popular products of the hour (for example, if users view the website at noon, they will see the products that were most popular on other days at noon), and the two most popular products for that month (for example, if users view the website in April, they will see the products that were most popular in April).

Countries where we do not have any recorded data

Similarly, as for the category above, we first listed the top two most popular products, but here we took into account the country they were visiting the page from. Meaning, that if we have any past data for that specific country, we will display the top two products bought by that country's users. The same goes for the rest of the recommendations; two most popular products of that day of the week, two most popular products of the hour, and the two most popular products for that month - all for that specific country.

We tested our Cold Start function manually by entering a few values for different days, time, months and countries. The test was successful and gave us accurate results.

7. Evaluation

We used the elbow method to find the optimal number of clusters. While using the elbow rule, we also focused on the *cost* resulting from the different numbers of clusters. For the clustering we used a clustering algorithm that can take into account both - categorical and numerical features.

Another evaluation strategy that helped us to measure the quality of our model was building two association rules instead of one. The first rule is based on Cluster 0 which is represented by the second highest number of customers and also the customers with higher value. The second rule is based on Cluster 1 which has the highest number of customers. Having different rules for different customer segments helps to formulate our recommendations more precisely.

Because of the limitation of features, our rules only stand on the basis of transactional data and not on customer information. If we could get more information about our customers, we could

develop our rules more precisely. The same applies to the information about the product category. If we had more information on that, we could enhance our association rules.

8. Deployment

Our Recommendation and Cold Start systems can be used in the future to help Gift-a-Lot recommend relevant products to their existing and new customers.

The deployment of both systems differs, so we provided some instructions below.

Recommendation System

For registered users (CustomerID is known)

Firstly, the model selects the cluster in which the user belongs, and chooses the association rule of this cluster. The model then reads their purchase history and makes recommendations based on their previous purchases by looking at the StockCode and the list of recommended products for that specific product.

For unregistered users (CustomerID not known)

The recommendation can be made based on the cookie's purchase history or browsing history, and the product information contained in the cookie will be put into the system to get the recommended product list.

Cold Start Problem

For new users, the date and country can be determined based on their login time and their IP address. The model will recognize that this is a completely new user and use our Cold Start system to recommend products.

9. References

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

Oberlo. (2022). *How Many People Shop Online in 2022? [Feb 2022 Update]*. https://www.oberlo.com/statistics/how-many-people-shop-online