NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 3: Recheio – Recommender Systems

Bernardo, Pinto Leite, number: 20230978

Emília, Santos, number: 20230446

Nicolás, Zerené, number: 20230779

Ricardo, Kayseller, number: 20230450

Stepan, Kuznetsov, number: 20231002

Group F

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

May, 2024

INDEX

INDEX OF GRAPHS

INDEX OF TABLES

INDEX OF FIGURES

1. EXECUTIVE SUMMARY

Recheio is embarking on an ambitious digital transformation initiative focused on increasing its customer base and sales through the deployment of an advanced recommendation system, even being a low margin company, meaning that the company cannot explore all technologies and jeopardize the main business (implement digital in favor of the business, not the other way). This project leverages extensive data analysis covering four key datasets—clients, products, transactions, client types—to develop a system that not only understands customer preferences, but also predicts future buying patterns effectively.

The strategic approach to develop this recommendation system involves a thorough analysis of existing market conditions and internal data capabilities. The goal is to provide relevant, actionable recommendations that lead to increased customer retention and higher transaction volumes, focusing only in the HoReCa channel (the one channel that is growing the most, with clients that are not very sensitive to price and have steady shopping habits). This system is expected to significantly improve Recheio's competitive edge in the market, enabling more precise and customer-focused marketing strategies.

Through this initiative, Recheio aims to strengthen its position as a market leader by boosting the relevancy of its offerings to each customer, thereby increasing the percentage of purchases allocated to Recheio over time. The success of this project will not only enhance digital transformation, but also support sustained business growth in an increasingly competitive landscape.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BUSINESS OBJECTIVES:

Market Expansion and Growth: Recheio has consistently pursued a strategy of geographical and market expansion to strengthen its position as a leading wholesale distributor in Portugal. This strategy includes the opening of new stores in underserved regions, enhancing the accessibility of their services to a broader range of business customers. Recheio's approach to expansion includes more than just opening additional stores but choosing locations to better serve and reach their customer base.

Brand and Product Diversification: Recheio operates with multiple brands, each tailored to meet the specific needs of different segments of the market. For example, the MasterChef brand is focused on providing a range of products for the hospitality industry, whereas the Amanhecer brand targets traditional retail trade. This strategic diversification enables Recheio to meet the varying requirements of their clientele, ranging from small businesses to large corporations, securing a comprehensive market presence. The introduction of Gourmês shows Recheio’s commitment to brand diversification, targeting premium segments with specialized products.

Digital Boost: Recheio aims to develop and implement a recommendation system and a user-friendly mobile app to enhance customer engagement and experience. This initiative is designed to personalize shopping, improve service delivery, and streamline interactions, thereby increasing the customer base and boosting sales. Through these technological advancements, Recheio intends to provide tailored product suggestions and convenient shopping options, leading to higher customer satisfaction and loyalty.

Sustainability and Corporate Social Responsibility: The Jerónimo Martins Group, which includes Recheio, has a comprehensive approach to sustainability and corporate social responsibility (CSR): reducing its environmental footprint by implementing many sustainable practices: fighting food waste, promoting sustainable production and consumption, and advancing towards decarbonization in its logistics operations. As a company engages in CSR, it is more likely to receive favorable brand recognition and to have a positive impact on the world through direct benefits to society, nature and the community.

Enhancing Customer and Stakeholder Value: Recheio aims to enhance customer and stakeholder value by proactively anticipating market needs, driving innovation, and fortifying brand loyalty. The company is committed to deepening its engagement with retail partners, notably through the expansion of the Amanhecer chain to 610 locations—an increase of 73 stores from the previous year. Additionally, Recheio seeks to bolster its competitive pricing strategy through the execution of various strategic marketing campaigns.

2.2. PRELIMINARY INFORMATION:

Market Analysis: The retail sector in Portugal, including the Cash & Carry and broader retail markets, showed positive growth in 2023. For the Cash & Carry sector specifically, growth was driven by innovation in store layouts and an emphasis on enhancing the customer shopping experience. This aligns with broader retail trends where digital innovation and technology are crucial in attracting customers. Looking forward, the retail sector in Portugal is expected to continue growing, although consumers may remain price-sensitive in the near term. Retailers are likely to focus on differentiating their offerings and providing specific value to meet consumer demands, leading to further integration of digital strategies to enhance customers shopping experience.

Global Trends: In today's world, staying ahead of market trends is crucial for the success of any business. Cash & Carry model offer several strategic advantages for businesses looking to stay competitive and responsive to market trends. Businesses utilizing cash-and-carry can quickly acquire new products from wholesalers, allowing them to test market and stay ahead of competitors. Unlike traditional wholesale purchasing, Cash & Carry enables businesses to buy in smaller quantities. Cash & Carry often presents lower pricing compared to conventional wholesale. This cost-effectiveness allows businesses to maintain competitive pricing structures, which is particularly appealing in markets sensitive to price fluctuations.

This analysis confirms the global retail and Cash & Carry market's direction, characterized by dynamic growth, and influenced by pivotal trends and critical challenges. Companies operating within this space, including Recheio, can draw on this information to adapt strategies, anticipate market needs, and align their offerings with the forecasted growth and challenges ahead.

2.3. MARKET END USERS:

Commercial End-Users: Market Revenue Share (2023): Commercial users, primarily consisting of small and medium-sized businesses and independent retailers, account for a significant part of the retail market revenue; **Growth Factors:** This dominance is due to the extensive product range that applies specifically to business needs, competitive pricing, and strategic store locations. The offering of bulk products at competitive prices is crucial for businesses looking to manage operational costs effectively; **Key Solutions**: The adoption of digital ordering systems and the creation of user profiles that include their preferences and targeted recommendations.

Residential End-Users: Projected Growth: The residential segment is primed for expansion as the brand broadens its scope to increase the consumer base. **Drivers of Growth:** The growth in the residential sector is often driven by the expansion into consumer-oriented offerings, which bring a company's products and services closer to everyday consumers. **Additional Benefits:** Venturing into the residential market is supported by broader consumer trends that favor convenience and quick access to products and services. Additionally, a company's established reputation for quality and value in the business sector can be a strong selling point in the residential market, appealing to price-sensitive consumers who seek quality without compromise.

2.4. COMPETITIVE AND SWOT FRAMEWORK:

Competitive Analysis: Recheio, part of the Jerónimo Martins Group, faces significant competition in the Cash & Carry sector in Portugal from several major players: Makro, Intermarché and Continente in the Cash & Carry and broader wholesale market in Portugal. The company differentiates itself through the offer of the largest assortment of top-quality products and service, with the best price, the provision of the best Perishables sourced from the best origins, coupled with strict quality control and mutual trust with business partners.

SWOT Analysis: Strengths: strong market position, diverse product range, B2B focus, strong supply chain. **Weaknesses:** limited geographic reach, dependence on Portuguese economy, limited consumer base. **Opportunities:** expansions into new markets, e-commerce and digital transformation, sustainable practices. **Threats:** economic fluctuations, strong competition, regulatory changes.

2.5. STRATEGIC RECOMMENDATIONS FOR RECHEIO

Considering the data and trends observed in the Cash & Carry and retail market, the following strategic recommendations are proposed for Recheio:

1) Modernisation of store network: Continue modernizing store designs to improve navigation and the overall shopping experience.
2) Strengthening of Food Service Offerings: Broaden the range of exclusive food service products, focusing on high-margin and specialty items that are unique to Recheio. Improve logistics to offer more efficient and timely delivery services, ensuring that HoReCa clients can rely on Recheio for quick replenishments.
3) Partnerships and Community Engagement: Continue to build relationships with local producers and suppliers to offer more locally sourced and artisanal products, enhancing the brand's community ties and supporting the local economy.
4) Digital Transformation and Technological Integration: Development of a recommendation system that analyses purchasing behavior and preferences to personalize product suggestions, both in-store and online.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

3.1.1. Data collection and description

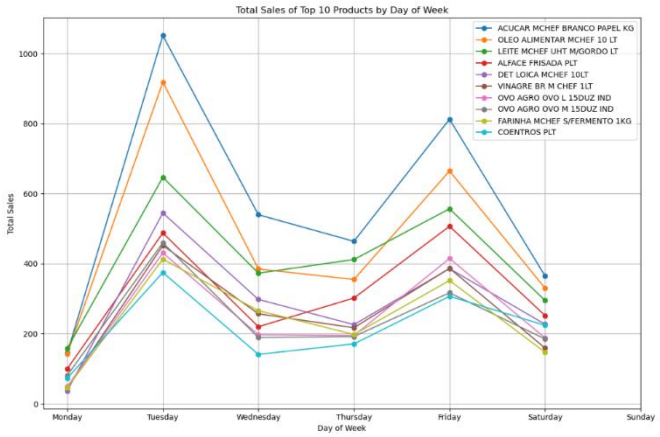
The group received a CSV file from Recheio, composed by 4 sheets, relative to information about clients, products, transactions, and client types.

Regarding the **first sheet, 'Clients'**, it has a total of 930 observations or client ID's in this case and 3 variables: 'Client ID', 'ZIP Code' and 'ID Client Type'. 'ZIP Code' has 116 unique values and the most frequent is 1100, occurring 53 times. "ID Client Type" has 28 unique values, the most frequent is 201 and occurs 220 times. About **'Products' sheet**, there is a total of 2498 observations or Product ID's and, as the last sheet, 3 variables: 'ID Product', 'Product Description' and 'ID Product Category'. There are 33 unique product categories, the most frequent is " ALIMENTAÇAO CORRENTE", occurring 538 times. "Product Description" has 2497 unique values, meaning there might be one duplicate product description. The most frequently occurring product description is "QJ BRIE MG 60% CUNHA PRESIDENT 200G", which occurs twice. Looking at the **third sheet, 'Transactions'**, it is possible to see that we have transactions between 2019-03-01 and 2019-05-31 (total of 234 224 observations) and 3 variables, 'Date', 'Client ID' and 'ID Product'. "Client ID" has 930 unique values (coherence between this sheet and the clients' one), the most frequent is 4426, occurring 2112 times. "ID Product" has 2498 unique values (coherence between this sheet and the products' one), the most frequent product ID is 110110, occurring 3379 times. Lastly, the **'Client Types' sheet**, with 28 observations or client ID types (coherence between this sheet and the clients' one) and 2 variables: 'ID Client Type' and 'Client Type Description', which is unique for each ID client type.

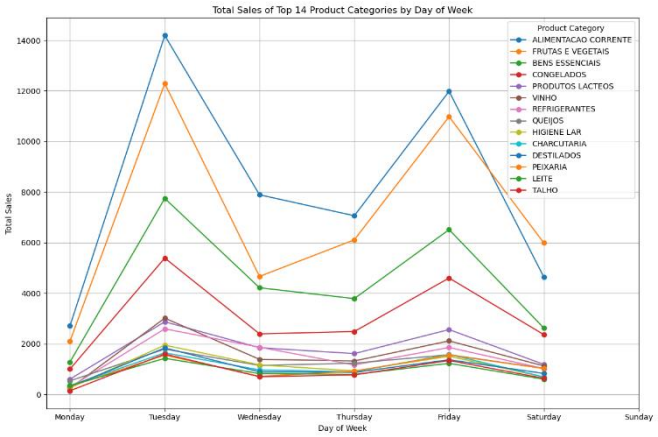
3.1.2. Data Exploration

Before exploring the data, the 4 sheets were merged into one, in order to connect better their information and have more insightful analysis, and some pre-processing was done (for instance, the creation of the variables 'Day of Week' and Weekday Name'). The most important insights we extracted from the analysis were the following.

Regarding the 5 most sold categories, these are 'ALIMENTAÇAO CORRENTE', 'FRUTAS E VEGETAIS', 'BENS ESSENCIAIS', 'CONGELADOS' e 'PRODUTOS LACTEOS', while the least sold are 'ENTREGAS AO DOMICILIO', 'HOMEM', 'QUINQUILHARIA', 'TEXTIL LAR' and 'BACALHAU'. In terms of the products, these were the 5 most frequent ones: 'ACUCAR MCHEF BRANCO PAPEL KG', 'OLEO ALIMENTAR MCHEF 10 LT', 'LEITE MCHEF UHT M/GORDO LT', 'ALFACE FRISADA PLT' and 'DET LOICA MCHEF 10LT'. The transactions of these products and categories that sell the most are mostly done on Tuesday and Friday. The day where the least transactions are made is Monday, considering that on Sunday there are no sales (Graphs 1 and 2).

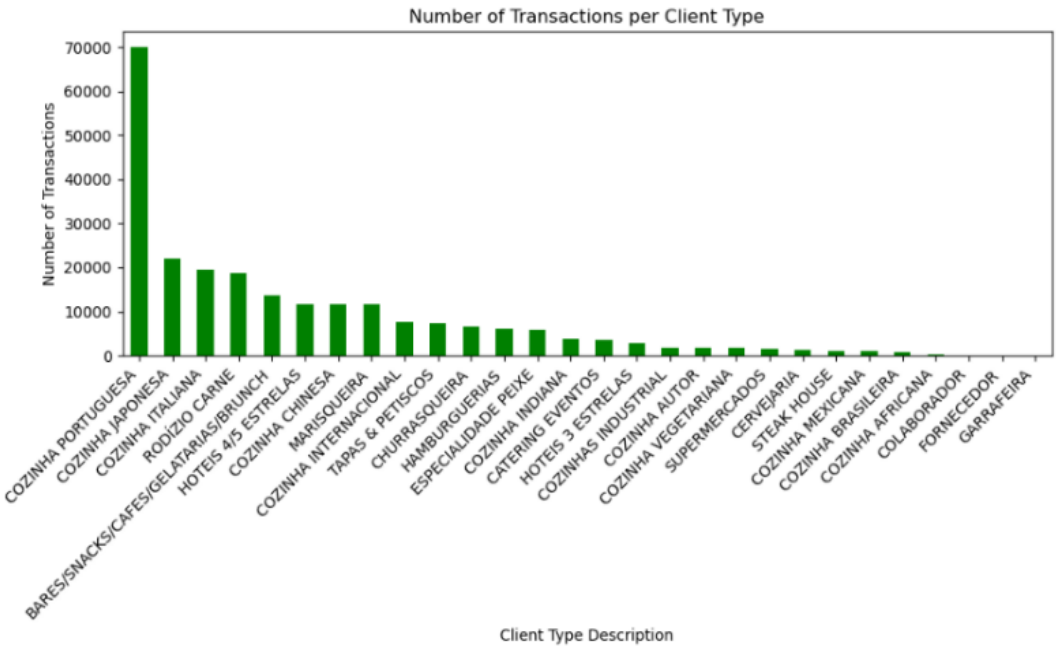


Graph 2 - Top 10 Products sold by Day of the Week



Graph 1 - Top 14 Product Categories sold by Day of Week

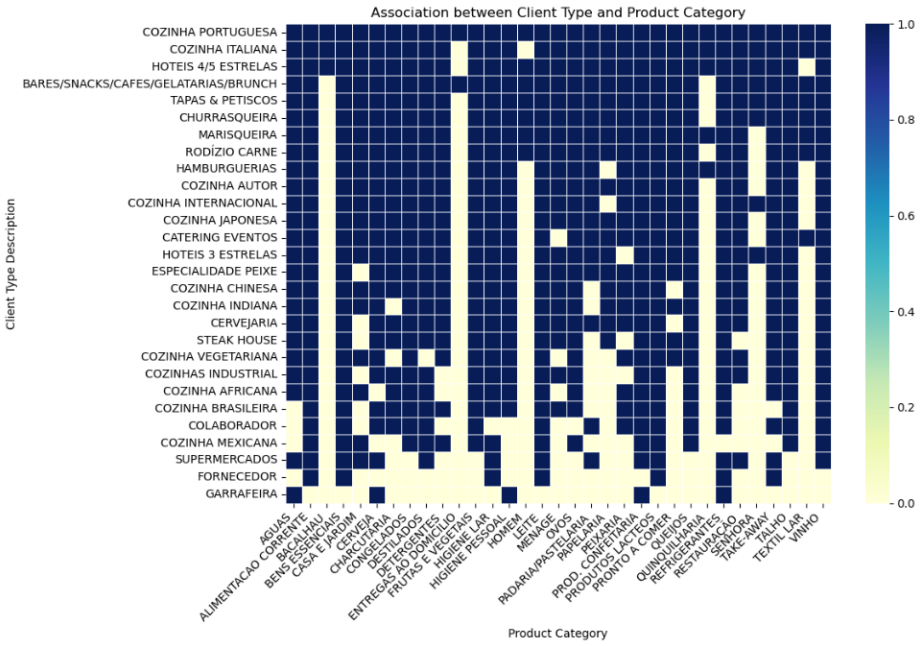
The client types that purchase more from Recheio are the ones related to Portuguese, Japanese and Italian cuisines, 'RODIZIO DE CARNE' and caffes, bars and brunch establishments.



Graph 3 - Transactions per Client Type

When crossing 'Client Type' and 'Product Category' through the following heat map, there are some understandings that can be taken, having in mind that the darker the colour, the more association between the 2 variables there is.

- 'COZINHA PORTUGESA' is the only client type that purchases products from all categories, contrarily to 'GARRAFEIRA';
- The product category that sells to the smallest number of client types is 'ENTREGAS AO DOMICILIO' and the client type that buys the less amount of product categories is 'GARRAFEIRA';
- Each client type usually just buys products from categories which they are commonly associated with. This means that, and can be seen below, for example, 'HAMBURGUERIAS', 'CHURRASQUEIRA' or 'RODIZIO CARNE' don't buy 'BACALHAU'.



Graph 4 - Association between Client Type and Product Category

3.1.3. Data Quality

There are no missing values in the dataset. About the existence of duplicates, it can be said that in 'Client Types' there are none; in 'Transactions' (having the 'Date' as index), 185842 duplicates were identified, however, there is no need to deal with them because they mean that one client bought the same product more than once, in different dates; in 'Products' (having the 'ID Product' as index), there was one duplicated row, which was removed, since the 'Product Description' and 'ID Product Category' were the same; in 'Clients' sheet (having the 'Client ID' as index), there are 419 duplicates, meaning that in the same ZIP Code, there is more than 1 client with the same type of establishment.

3.2. DATA PREPARATION

In the initial stages of our analysis, a thorough examination of the dataset was conducted to ensure the quality and usability of the data for further analysis. Key steps included:

3.2.1. Data Cleaning

Missing Values: Verified that the dataset contains no missing (NaN) values. This ensures that all variables and observations are complete for each transaction, providing a solid foundation for reliable analysis. **Duplicate Values:** Checked for and found no duplicate records within the dataset. This step is crucial for maintaining the accuracy of transactional data and ensuring that each entry uniquely represents an actual customer purchase.

3.2.2. Outlier Detection and Handling

The Interquartile Range (IQR) was used to identify potential outliers in transaction data per client. Using these bounds, outliers were identified based on their transaction frequency, which significantly deviated from the typical customer behavior. The dataset revealed a small significant proportion of transactions (*approximately 8%*) categorized as outliers. However, these were not to exclude these from the dataset for two main reasons:

1. **Focus on Market Analysis:** The primary aim of this study is to enhance understanding of market dynamics and customer purchasing patterns without losing valuable insights from all customer interactions, including those that are atypically high;
2. **Recommendation Systems Relevance:** In the context of recommendation systems, even outlier transaction data can provide useful insights into niche markets or high-value customer behaviors, which are valuable for personalized marketing strategies.

3.2.4. Feature Engineering

3.2.4.1. Refinement of Geographical Data

To better understand customer purchasing patterns across different regions, a new column was created in the dataset that classifies geographical regions based on postal codes. A focus on areas within and surrounding Lisbon, including Cascais, Oeiras, Sintra, and Setúbal, among others, was made. This classification aids in analyzing regional variations in purchasing behavior.

3.2.4.2. Exclusion of Irrelevant Data

To streamline the dataset of this specific analysis focusing on market dynamics in urban and suburban areas around Lisbon, transactions from districts less relevant to our primary market were removed. These districts include Açores, Beja, Braga, Coimbra, Faro, Leiria, Castelo Branco, and Santarém, which were identified as less central to the current study's scope.

3.3. MODELING – RECOMMENDATION SYSTEMS

All the recommendation systems, apart from MBA, were made based on the Client Type and on the clusters (further explained), due to the existent computational resources and the fact that by doing this the models were easier to interpret.

3.3.1. Market Basket Analysis - Product Category Level

Market basket analysis identifies key purchasing patterns through support, confidence, and lift metrics. We focused on the top ten rules to uncover the most prevalent and impactful associations, highlighting those most likely to influence purchases and demonstrate strong relationships between product categories.

Product Categories	Support	Confidence	Lift
'BENS ESSENCIAIS' -> 'ALIMENTACAO CORRENTE'	49.67%	85.24%	1.233
'ALIMENTACAO CORRENTE' -> 'BENS ESSENCIAIS'	49.67%	71.86%	1.233
'CONGELADOS' -> 'ALIMENTACAO CORRENTE'	36.51%	81.21%	1.175
'ALIMENTACAO CORRENTE' -> 'CONGELADOS'	36.51%	52.82%	1.175
'PRODUTOS LACTEOS' -> 'ALIMENTACAO CORRENTE'	32.78%	88.68%	1.283
'ALIMENTACAO CORRENTE' -> 'PRODUTOS LACTEOS'	32.78%	47.42%	1.283
'CONGELADOS' -> 'BENS ESSENCIAIS'	32.48%	72.26%	1.240
'BENS ESSENCIAIS' -> 'CONGELADOS'	32.48%	55.75%	1.240
'FRUTAS E VEGETAIS' -> 'ALIMENTACAO CORRENTE'	32.04%	75.85%	1.097
'ALIMENTACAO CORRENTE' -> 'FRUTAS E VEGETAIS'	32.04%	46.35%	1.097

Table 1 - MBA per Product Category

Key Observations:

- **Strong Pairing between Essential and Food Items:** Both ‘BENS ESSENCIAIS’ -> ‘ALIMENTACAO CORRENTE’ and the reverse show very high support and significant lift values, indicating that these items are commonly bought together and influence each other's purchase. This mutual relationship suggests that shoppers often combine staple and food purchases, presenting an opportunity for promotions and store layout optimization to capitalize on these patterns;
- **High Purchase Likelihood with Frozen Goods:** The rules involving ‘CONGELADOS’ show strong associations with both ‘ALIMENTACAO CORRENTE’ and ‘BENS ESSENCIAIS’, highlighted by good lift values. This suggests that frozen goods are a central item that pairs well with daily essentials and food items, pointing towards effective cross-selling opportunities;
- **Dairy Products as Strong Predictors of Food Purchases:** The association of ‘PRODUTOS LACTEOS’ -> ‘ALIMENTACAO CORRENTE’ has one of the highest confidence and lift metrics, indicating a very predictable purchase pattern. This can inform inventory and marketing strategies, ensuring that dairy and food items are well-stocked and possibly co-located;
- **Moderate Associations with Fruits and Vegetables:** Although the pairing of fruits and vegetables with current food items has moderate confidence and a lift slightly above 1, it still represents a significant portion of transactions. This suggests a stable but less intense purchasing habit that could be enhanced through targeted promotions or by offering bundled deals.

3.3.1.1. Top 10 Rules with Most Frequent Categories purchased together (Confidence)

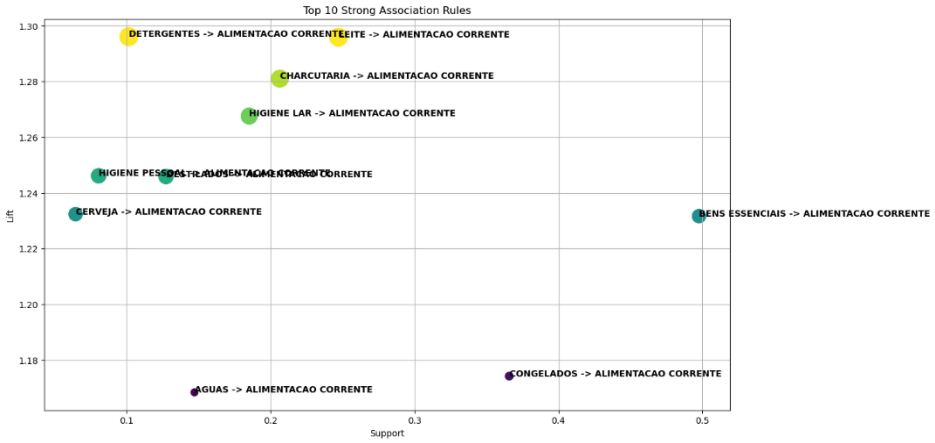
General Observations:

- ‘ALIMENTACAO CORRENTE’ is a highly prevalent category with a support of 0.692119, indicating that it constitutes a significant part of transactions.
- Most categories show a strong association with ‘ALIMENTACAO CORRENTE’, reflected by confidence levels mostly above 80%. This suggests that when items from categories such as ‘AGUAS’ , 'CERVEJA', or ‘LEITE’ are purchased, there is a high likelihood of also purchasing items from ‘ALIMENTACAO CORRENTE’.

Notable Associations:

- The category ‘CHARCUTARIA’ shows one of the highest confidence levels at 88.66% with a lift of 1.280992, indicating a stronger than average association with ‘ALIMENTACAO CORRENTE’. This suggests that deli products are frequently bought in conjunction with routine food items.
- ‘DETERGENTES’ and ‘LEITE’ both exhibit high confidence levels around 90%, with respective lifts of 1.296052 and 1.295859. Despite being non-food items, the purchase of detergents shows a significant co-purchase pattern with food items, which could indicate common shopping patterns or promotions influencing these purchases.

Overall, the analysis suggests optimizing retail strategies around more frequently occurring item combinations to better leverage the frequent purchase behavior associated with ‘ALIMENTACAO CORRENTE’ (Graph 5).



Graph 5 - Top 10 Strong Association Categories

3.3.2. Market Basket Analysis - Product Description Level

The analysis conducted here focuses on identifying the strongest association rules within a dataset of product transactions. Market Basket Analysis (MBA) is used extensively in retail to understand customer purchase behaviors by discovering products that frequently co-occur in transactions. This insight allows retailers to make informed decisions about marketing strategies such as product placements, promotions, and inventory management. The specific criteria for filtering these rules are based on a lift greater than 1, indicating a positive association between item sets, and a confidence level of at least 0.7, ensuring the reliability of these associations. The top 15 rules, sorted by descending lift, reveal the most significant relationships.

Antecedents	Consequents	Support	Confidence	Lift
(GEL.CAT.GOURMES BAUNILHA 4,5LT, GEL.CAT.GOURM...)	(GEL.CAT.GOURMES MORANGO 4,5LT)	0.011302	0.853448	43.21260
(GEL.CAT.GOURMES CHOC.4,5LT)	(GEL.CAT.GOURMES BAUNILHA 4,5LT, GEL.CAT...)	0.011302	0.704626	42.71401
(GEL.CAT.GOURMES CHOC.4,5LT)	(GEL.CAT.GOURMES MORANGO 4,5LT)	0.012558	0.782918	39.64145
(ICE TEA LIPTON MANGA LATA 33CL, ICE TEA LIPTO...)	(ICE TEA LIPTON LIMÃO LATA 33CL)	0.012158	0.803774	32.97731
(ICE TEA LIPTON MANGA LATA 33CL, ICE TEA LIPTO...)	(ICE TEA LIPTON PÊSSEGO LATA 33CL)	0.012158	0.938326	32.42314
(ICE TEA LIPTON MANGA LATA 33CL)	(ICE TEA LIPTON LIMÃO LATA 33CL)	0.012957	0.725240	29.75521
(ICE TEA LIPTON MANGA LATA 33CL)	(ICE TEA LIPTON PÊSSEGO LATA 33CL)	0.015126	0.846645	29.25519
(ICE TEA LIPTON LIMÃO LATA 33CL)	(ICE TEA LIPTON PÊSSEGO LATA 33CL)	0.018209	0.747073	25.81453
(GEL.CAT.GOURMES MORANGO 4,5LT, GEL.CAT.GOURME...)	(GEL.CAT.GOURMES BAUNILHA 4,5LT)	0.011302	0.900000	25.55446
(GEL.CAT.GOURMES MORANGO 4,5LT)	(GEL.CAT.GOURMES BAUNILHA 4,5LT)	0.016496	0.835260	23.71624
GEL.CAT.GOURMES CHOC.4,5LT	GEL.CAT.GOURMES BAUNILHA 4,5LT	0.013243	0.825623	23.44260
ALHO FRANCÊS CORTADO 750G 4G MCHEF RCH	PEPINO RCH	0.010217	0.848341	14.38731
REF COCA COLA ORIGINAL LATA 33CL, ICE TEA LIPTON PÊSSEGO LATA 33CL	REF COCA COLA ZERO LATA 33CL	0.010503	0.773109	13.37029
ICE TEA LIPTON PÊSSEGO LATA 33CL, REF COCA COLA ORIGINAL LATA 33CL	REF COCA COLA ZERO LATA 33CL	0.010503	0.851852	12.74432
MAÇÃ GOLDEN CAL65/70 50*30 2CAM RCH	LARANJA CAL7 (67/76) RCH	0.010160	0.812785	12.58991

Table 2 - Top 15 Association Rules - MBA

The market basket analysis reveals significant associations across a variety of product categories, particularly within beverages and desserts. This analysis provides a robust foundation for targeted marketing strategies, optimized store layouts, and innovative product offerings.

1. **Gourmet Ice Creams and Lipton Ice Teas**

- The strong intra-category associations among different flavors of gourmet ice creams and Lipton Ice Teas indicate that customers who purchase one flavor are highly likely to purchase others. This suggests a strong base for promotions such as bundle deals or flavor variety packs which can drive sales across these categories.
- Marketing Action: In-store promotions and cross-merchandising strategies can be employed to capitalize on these associations. Placing related flavours together and offering discounts on the purchase of multiple flavours can enhance sales.

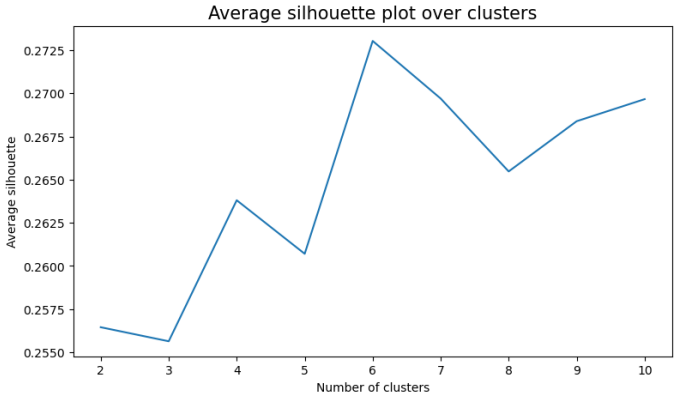
2. Coca Cola Original and Coca Cola Zero

- The association between Coca Cola Original and Coca Cola Zero indicates that consumers often purchase both, suggesting that households might cater to different dietary preferences or sugar intakes. This provides an opportunity to promote Coca Cola Zero alongside Coca Cola Original, appealing to both traditional cola lovers and those seeking sugar-free options.
- Marketing Action: Implement cross-promotions and dual placements in store layouts. Advertising campaigns could highlight the choice between “Classic Enjoyment” and “Health-Conscious Delight,” catering to a diverse consumer base.

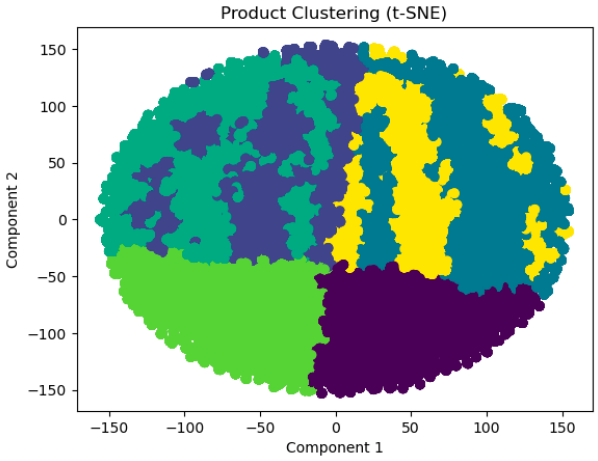
3.3.3. Clustering (not a recommendation model)

Although clustering is not a recommendation model, it was used to support the analysis. Also, given the importance of segments in this project, as part of a process to better determine recommendations, it was decided to perform this algorithm. This was done to improve or compare whether client types are a good representation, as well as enrich the analysis.

K-Means was the chosen algorithm, for it all the numeric features were considered ('ID Product', 'Client ID', 'ID Client Type', 'Day of Week', 'Season', 'Week Value', 'Count'). We utilized the Silhouette score method, alongside the elbow method, to determine the optimal number of segments for our analysis. The results indicate that the best number of segments is 6. These 6 segments showed all the differences between the categories (image below). It is to note that the t-SNE visualization shown below displays the clusters in the same dimensional plane, making it seem like two clusters are mixed with other 2. However, in practice, the 6 clusters are well defined. It is important to remember that the t-SNE is a low dimension visualization.



Graph 6 – Silhouette plot

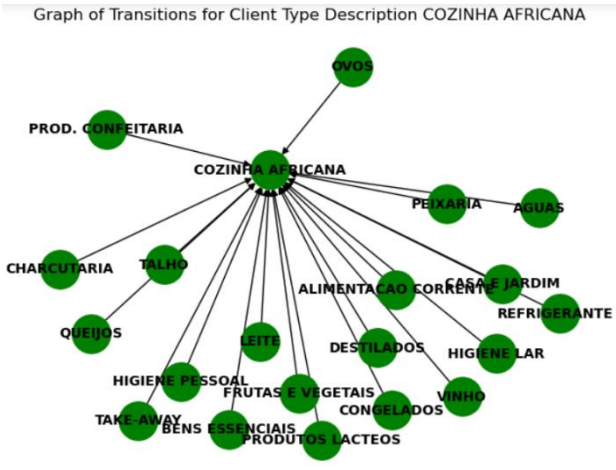


Graph 7 – t-SNE Product Clustering plot

3.3.4. Page Rank

Page Rank algorithm is an algorithm specially used in graph databases. Created by Google, the objective was to get a quick way to rank the websites and it measures the importance of each node based on the relationships (5.2.1. PageRank - 5.2. Centrality Algorithms, n.d.). The algorithm, therefore, calculates a probability distribution to obtain the likelihood that a person will get to a particular page by randomly clicking links (Bisht, 2017), in our case, the likelihood of purchasing a product.

For this project, the data was converted to a representation of a graph database so the links between clients and products could be shown (the image below shows an example of this). In this scenario, the probability a client would purchase specific products was calculated and the higher the PageRank score, the more likely would that product be purchased by that client.



Graph 8 – PageRank graph of transitions for the Client Type ‘COZINHA AFRICANA’

3.3.5. Collaborative Filtering

Collaborative filtering is a technique widely used as a recommender system. The basic idea of CF-based algorithms is to provide item recommendations or predictions based on the opinions of other like-minded users (Sarwar et al., 2001).

In terms of this project, the use of a collaborative filtering algorithm is to suggest new items or to predict the utility of a certain item for a particular client based on their previous purchases and the purchases of other like-minded clients. There are different types of collaborative systems, however, in the scope of this project, it was decided to use the item-based Cosine Similarity (described below), the Singular Value Decomposition and a Correlation-based Similarity.

3.3.5.1. Item-Based Cosine Similarity

This approach to produce recommendations is based on the interactions of the clients. The item-based Cosine Similarity looks at what items a client has purchased, finds items similar to those, and then recommends those similar items (Sarwar et al., 2001). The way this approach works is by considering products as vectors and calculates the cosine angle between each of them. This gives the similarity between those products.

3.3.5.2. Singular Value Decomposition (SVD)

SVD is a matrix factorization technique applied to improve the performance of the recommendations. It is used to decompose the user-item interaction matrix into lower-dimensional matrices. This decomposition can reveal latent factors that explain the observed interactions between users and products. It is possible to give a personalized and unique recommendation to the customer, which is part of the main goal.

3.3.5.3. Correlation-based Similarity

This approach is very similar to Item-Based Cosine Similarity, however it does not consider the products as vectors. In this case, the similarity is calculated with Pearson's correlation. Clients' purchases of products give information about their needs and the rate of purchase of products for certain clients, give the similarity. The recommendation is done from that.

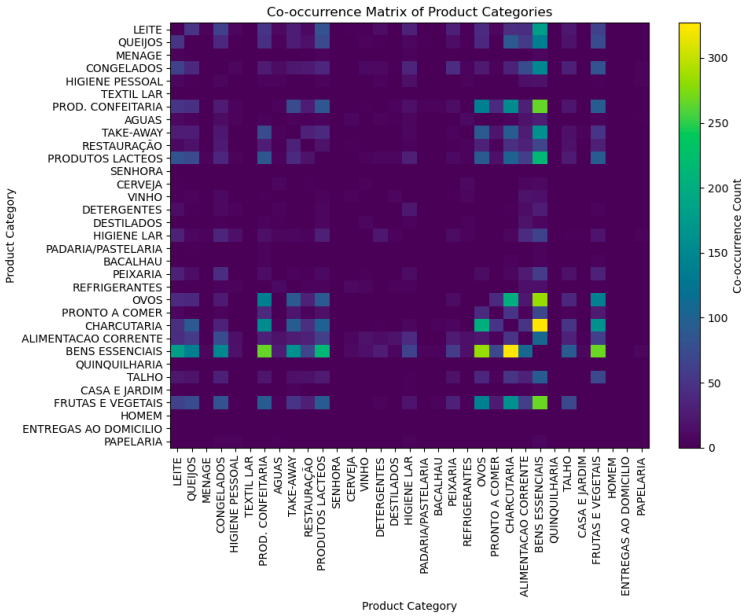
3.4. EVALUATION

For the evaluation process, it is needed to understand the process of modelling, because every algorithm gives a different output, and also a different way to support this recommended system.

In the case of page rank, it will consider a group of all the items recommended for each Type of Client, so the way to evaluate was look for each iteration in the Matrix M, and then calculate a score for each nod, considering the maximal real eigenvector. This way, it is possible to align a Client ID with a Product Category, showing the best scenario for the PageRank Score (Float). Also, were plotted several iterations with the relationships for each transition for Product Category and also for Client Type, showing all the nods that could be related to the nucleus of the diagram. To finalize the evaluation of the model, there were two important steps. First a new data frame, which counts how many times a product category is recommended with other categories for each Customer Type, so

after that it was possible to see a correlation between the products. Columns and rows will be the same categories with a null diagonal (graph bellow).

After this, the process was repeated for 'Product Description' and 'Product Categories' that haven't been consumed by a client, and were also made recommendations for each cluster.



Graph 9 - Co-occurence Matrix of Product Categories

On the other hand, for Collaborative Filtering the main idea was to look for the vector with the information recommended. This was applied for Products, Clients and also Clusters.

To do that, it was needed to count the number of purchases for each user and product combination, then convert the purchase counts to a sparse matrix, and compute the cosine similarity matrix between the products.

Then it was possible to recommend items for a user based on their purchase history. The process was for Product Category and for each Product. Also there was included each cluster, to see what would be a general vision for that segment of the clients.

At the end, the important measure to understand the performance of CF is the cosine similarity. Because each vector, starting from the same point, will have an angle. The idea is to look for each similarity, with all the iterations, and when a product/client/cluster has a similarity, it could be a recommendation. A good practice was creating a data frame with all the measures and differences between the angles of the vectors.

In the case of Singular Value Decomposition it was needed to compare the actual values with the predicted with RMSE. For this it was need to define SVD as al algorithm, and then apply a Gridsearch. This way, looking for data just considering the features to predict, it was a path to compare the actual data, with the predicted. So after doing that, the main Idea, was to locate if the product, were inside the top 10 products for each Client ID, Client Type, and Cluster. But if they were not there, just stop. At the end, this way could prioritize the best predictions and recommendations, so the process could finalize with the best product to recommend.

4. RESULTS EVALUATION

The different methods and approaches were used to obtain a recommender system. The scope of this project was to recommend 10 products to a client. The clients in our data, as mentioned above, correspond to HoReCa businesses. Using each client's transactions, it was possible to infer the interactions and similarities of each company and product purchased. After accounting for the metrics of novelty and coverage, a good recommender system should provide high values. For this case, novelty refers to how diverse or unique the recommendations are, and coverage measures the extent to which a recommender system is able to recommend items from the entire dataset. Based on this, the model that showed the best values was item-based with Cosine Similarity for Recommendations of products to Client Types with novelty score of 906141.80 and coverage of 0.06.

For example, using this approach, for clients with category of "CATERING EVENTOS" the products to recommended are 'PALITOS GOURMES EMB.IND.850UN', 'REF GINGER ALE SCHWEPPES TP 25CL', 'BEB ESPIRITUOSA MACIEIRA 1LT', 'BOL RECH OREO TUBO ORIGIN 154G', 'LEITE UHT MGORDO PRADO VERDE 1LT'. For the case of clients in "CERVEJARIA" business, the products recommended are 'REF COCA COLA ZERO LATA 33CL', 'PALITOS GOURMES EMB.IND.850UN', 'REF 7UP ORIGINAL LATA 33CL', 'CARAMELO LÍQUIDO MCHEF 1,3KG', 'ÁGUA TÓNICA SCHWEPPES ORIGINAL TP 25CL'. This makes sense as their main products are drinks and beverages.

Another example shows that for "COZINHA PORTUGUESA" the recommended products are 'V.ALENTEJO MARQUÊS DE BORBA BCO 75CL', 'SANGRIA CASAL GARCIA TTO 75CL', 'LEITE EM PÓ NIDO GORDO 700G', 'SIDRA SOMERSBY BLACKBERRY 33CL' and 'CANELA MOIDA DIAMIR PET 710G'. These products do make sense as Portuguese cuisine is based on wine (people love to drink alcoholic drinks for lunch) and like to have coffee (with milk) and some cinnamon to give flavour.

Thus, this recommender system manages to recommend better to the client base when compared to the other approaches.

5. DEPLOYMENT AND MAINTENANCE PLANS

Figure 1 - Release Managment Cycle

1. **Request for changes**

The main requests and business challenge asked by Recheio were to develop a general recommender system, generating relevant recommendations. By doing so, the main objectives were to increase Recheio's relevancy to each customer and try to increase at least 1 product, in each purchase.

2. Release planning and designing

The official release of the recommendations made would be done after all tests have been completed and the approval of the individuals responsible for the process. The design of the recommendations should follow an adaptable system, with no more than 10 recommendations suggested per purchase.

3. Software build

In order to meet the business challenge, we resorted to Recommender System models, after doing some data preparation and pre-processing, using Python programming language. Each one of these models was built with the purpose of recommending a maximum of 10 products or product categories (meeting the criteria above), since this would be the maximum number presented to customers, being that they usually order quickly and don't have much time to waste.

4. Review

The change of parameters and adjustments on pre-processing were made multiple times, with the objective of meeting the business challenge, improving the existent models and having a system that can be flexible, i.e., can be replicated to other customers, products, among others.

5. Test

In order to test and measure the performance of these recommendations, Recheio can compare them with what was really bought by the clients and if they bought more (diversity and quantity) than what they usually did. An A/B testing can also be done, i.e., randomly assign users to the recommendations generated by the different recommender systems built by the team (Item-Based, SVD, Page Rank) and its different versions (product level, product category level and by cluster). Analyse each user interactions and behaviour through metrics such as CTR (Click Through Rate), Conversion Rates and Retention Rates, or even through satisfaction surveys. With the insights obtained from these and other metrics, Recheio can refine, optimize, or choose the best, most accurate and effective system.

6. Deployment

After applying the recommendations of products that a certain client had already bought, an analysis on recommendations of products or product categories that were never bought by the client was developed. One example of this would be recommending the purchase of fresh fruits and vegetables to a 'HAMBURGUERIA'. Having this is mind and knowing that these are suggestions of items that possibly are not part of the client's menus, Recheio could offer alternative menu options that would require these new items. Recheio could for instance, if the purchase is made online, include below each product recommended a recipe or meal that uses that product, easing the creative process of implementation this new item. These implementations would lead to a broader client target, due

to the attraction of new clients to their establishment, while maintaining the current ones. A consequence of this for Recheio would be the reinforcement of their relationship with the businesses (Recheio's clients), which could potentially lead to the acquirement of other non-bought products, and, of course, an increase in profit. To correctly apply all these recommendations and have the business aligned with them, some changes or maintenance actions would have to be made, such as customization of the CRM model and coordination with sales force.

7. **Support, issue reporting and collection**

Recheio's team will address and resolve any issue that might arise, leading to adjustments on the models and its involving processes. Finally, the company would monitor the performance of the deployed recommendations, gathering feedback and performance data. Issues and suggestions for improvements would be documented to inform future development cycles, potentially leading to further improvements on the recommendation systems.

6. CONCLUSIONS

The Recheio recommender system project represents a significant advancement in improving both the shopping experience and operational efficiency for the company. Throughout the project, we utilized sophisticated data analytics and machine learning techniques, including Market Basket Analysis, association rules, clustering, PageRank, collaborative filtering, and Singular Value Decomposition, to develop a system that provides personalized product recommendations to customers.

Our recommender system for Recheio uses sophisticated similarity algorithms to enhance customer interactions by delivering tailored product recommendations, which should significantly boost sales and strengthen customer loyalty. Implementing our recommender system will lead to a comprehensive reevaluation and transformation of Recheio inventory management practices, ensuring that high-demand products are strategically stocked and promotions are targeted with precision. Overall, this system aims to improve Recheio's competitive edge by increasing operational efficiency and delivering insights for strategic decision-making, positioning the company for continued success in the dynamic retail landscape.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Neural Collaborative Filtering: Use deep learning to model the nonlinear interactions between users and items, potentially capturing complex patterns missed by traditional methods.

Ensemble Techniques: Combine predictions from different models (e.g., clustering, collaborative filtering) using ensemble techniques like stacking or blending to enhance overall performance.

Context-Aware Recommendations: Since traditional models may not consider context, integrating contextual information (location, device, time of day) can significantly tailor recommendations.

User Intent Detection: Employ machine learning models to infer user intent in real time, which can dynamically influence recommendation strategies.

Feedback Systems: Implement systems that can leverage real-time feedback from users to immediately refine the recommendations.

Exploration Strategies: Use exploration strategies like greedy algorithms to regularly suggest new or lesser-known items to users, helping to prevent them from only seeing the same types of recommendations repeatedly.

Transaction history data: The more data there is, the more confident the recommendations are. With this logic, data needs to be gathered about if the clients are buying the recommended products and longer timespans (every 6 months, update the recommender system to account for different seasons)

7. REFERENCES

- *Recheio*. (n.d.). https://www.recheio.pt/portal/pt-PT/category/latic%C3%ADnios/queijos/queijo-estrangeiro/0ZGQD00000001HD4AY
- 5.2.1. PageRank - 5.2. Centrality algorithms. (n.d.). Neo4j.com. https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/
- Bisht, J. (2017, August 30). Page Rank Algorithm and Implementation - GeeksforGeeks. GeeksforGeeks. https://www.geeksforgeeks.org/page-rank-algorithm-implementation/
- Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the Tenth International Conference on World Wide Web - WWW '01. https://doi.org/10.1145/371920.372071
- Towards Data Science (2020). T-distributed Stochastic Neighbor Embedding(t-SNE) Learn the basics of t-SNE, how it is different than PCA and how to apply t- SNE on MNIST dataset https://towardsdatascience.com/t-distributed-stochastic-neighbor-embedding-t-sne-bb60ff109561