February 2024

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Marketing Analytics

PROJECT

Enrico Mannarino

Giorgia Prina

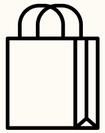
Master's Degree in Data Science

Main approaches

We conducted a data-driven analysis leading to the generation of insights to support/guide strategic decisions.

Customer focus

Prevent high-value customer churn by a marketing campaign for customer retention using **Churn** models and **RFM** analysis (and also estimate the repurchase probability of one-shooter customers via **Repurchase** models).



Product focus

Increase profit by a marketing campaign for product cross-selling, using **Market Basket Analysis**.

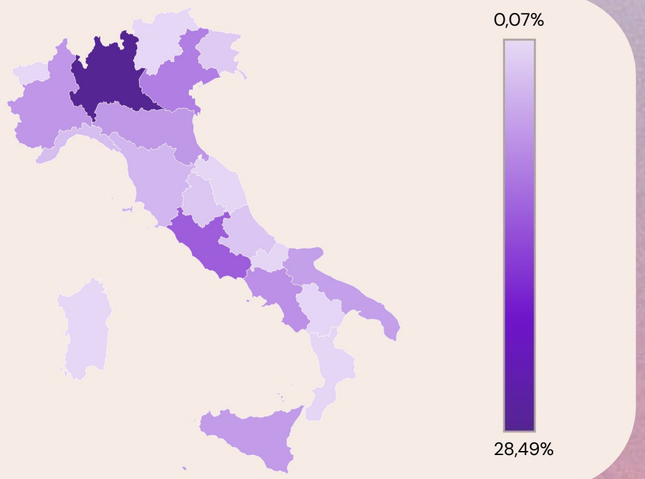
Feedback focus

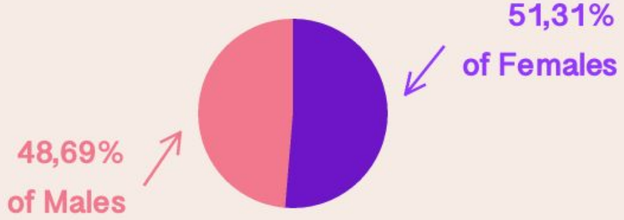
Address detractor and promoter customers with a loyal engagement marketing campaign to reduce the negative impact of detractors and to incentive the positive effect of promoters through a **Sentiment Analysis**.

Data exploration

How is the **customer base** composed?

Check out our dashboards on Tableau





REGION DISTRIBUTION

Lombardia is the region with the highest number of customers (28.5%), followed by Lazio (14%) and Veneto (9%).
On the other hand, Sardegna (0.14%), Molise (0.14%) and **Valle d'Aosta** (0.07%) have the fewest customers.

GENDER DISTRIBUTION

The majority of customers are **female**.

AGE DISTRIBUTION

The vast majority of customers (80%) are in the age range of **35 to 54 years old**.

Data exploration

We calculated and then not considered **inactive customers** in the following analysis, so as to focus on the actually active ones.
We then identify and distinguish between **one-shooters** and **repeaters**.

Those who have NOT PLACED any order from **1st May 2022 to 30th April 2023**

Those who have placed AT LEAST one order in the last year

Those who have placed MORE THAN 1 order in the last year

Customer focus:

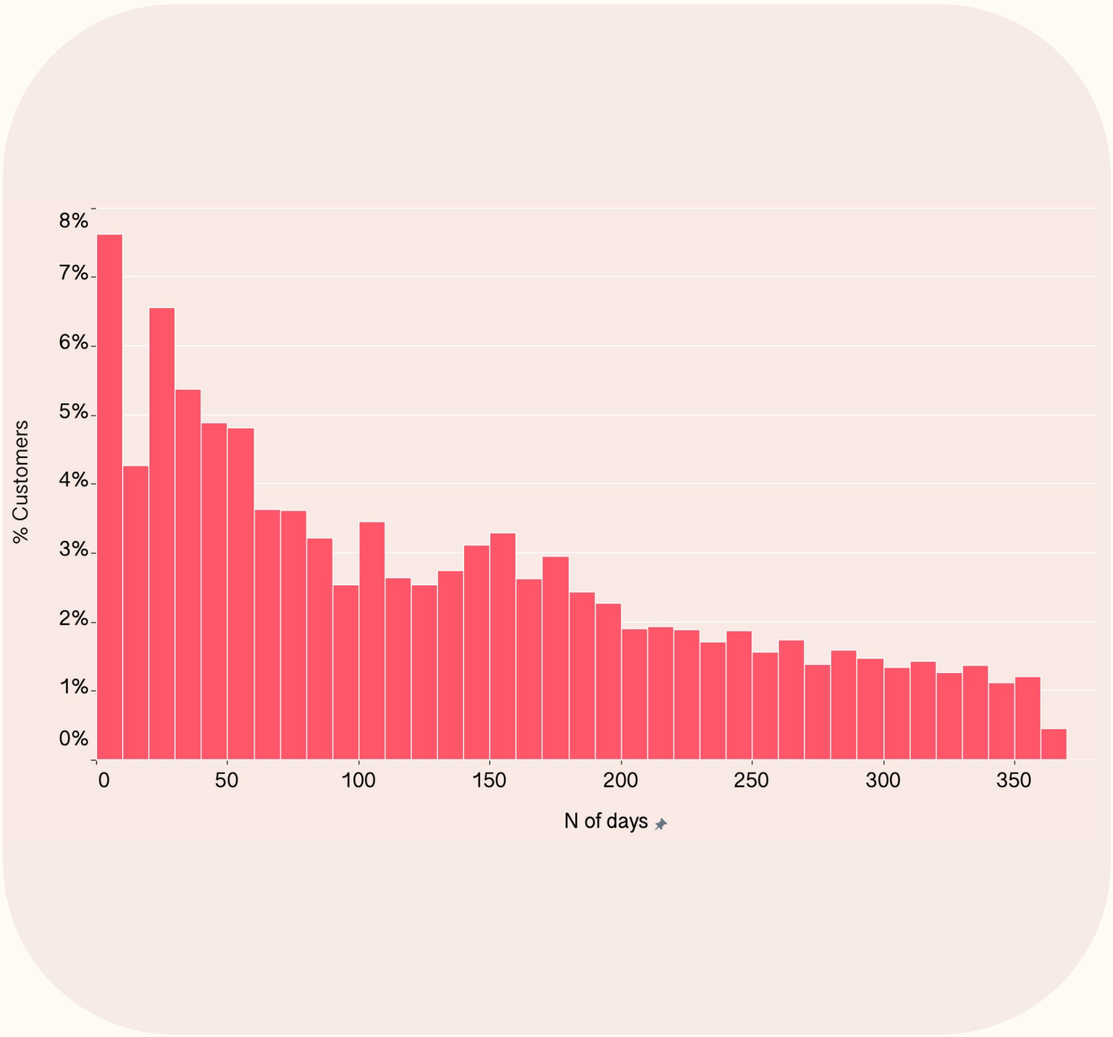
RFM analysis

RFM is a customer behaviour analysis model to segment customers according to their historical behaviour.
It is based on three main dimensions: **Recency**, **Frequency** and **Monetary**.

RFM analysis is usually performed by assigning each of the three dimensions a numerical score based on customer behaviour. These scores are then combined to obtain distinct customer segments, e.g. **identifying the most valuable customers** (e.g. those with high purchase frequency, large recent spending and frequent interactions).

Goals

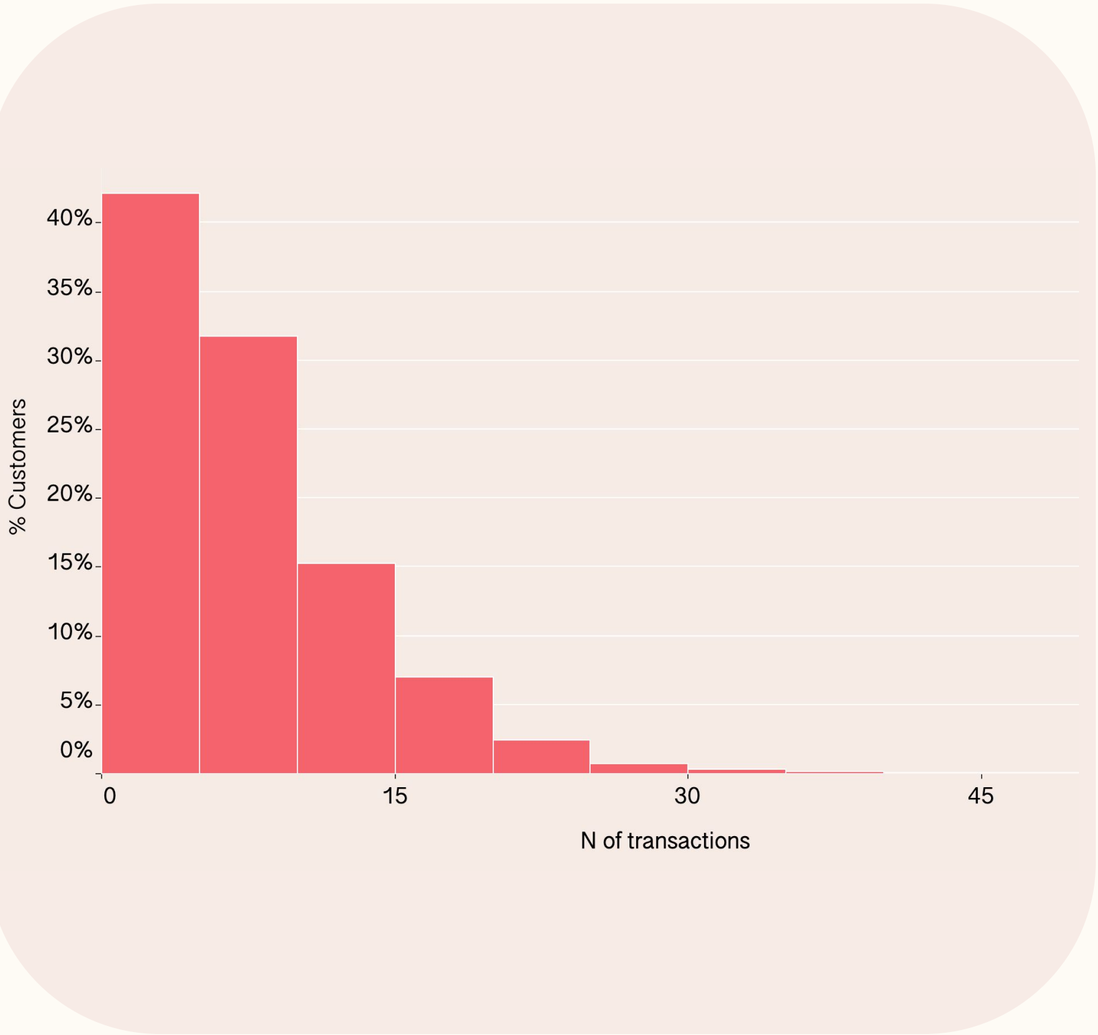
- How to segment the customer base? → **RFM**
- Which customers are at risk of dropping out? → **Churn models**
- What is the likelihood of a one-shooter customer returning? → **Repurchase models**



Recency

Amount of time elapsed since the customer's last transaction.

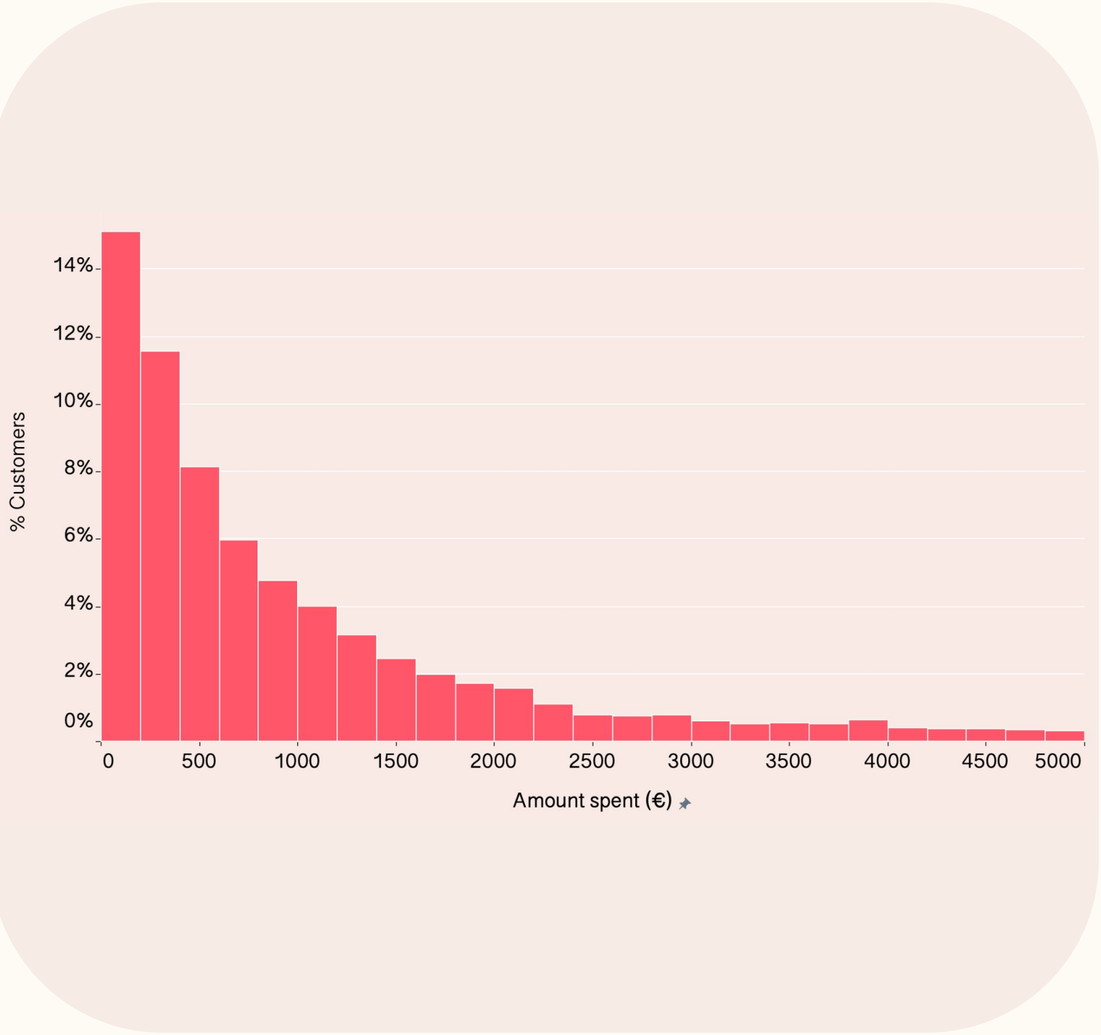
- Average days since last purchase: **130 days**
- 50% of customers haven't made a purchase in at least **109 days**.



Frequency

Total number of transactions made by each customer between 1st May 2022 and 30th April 2023

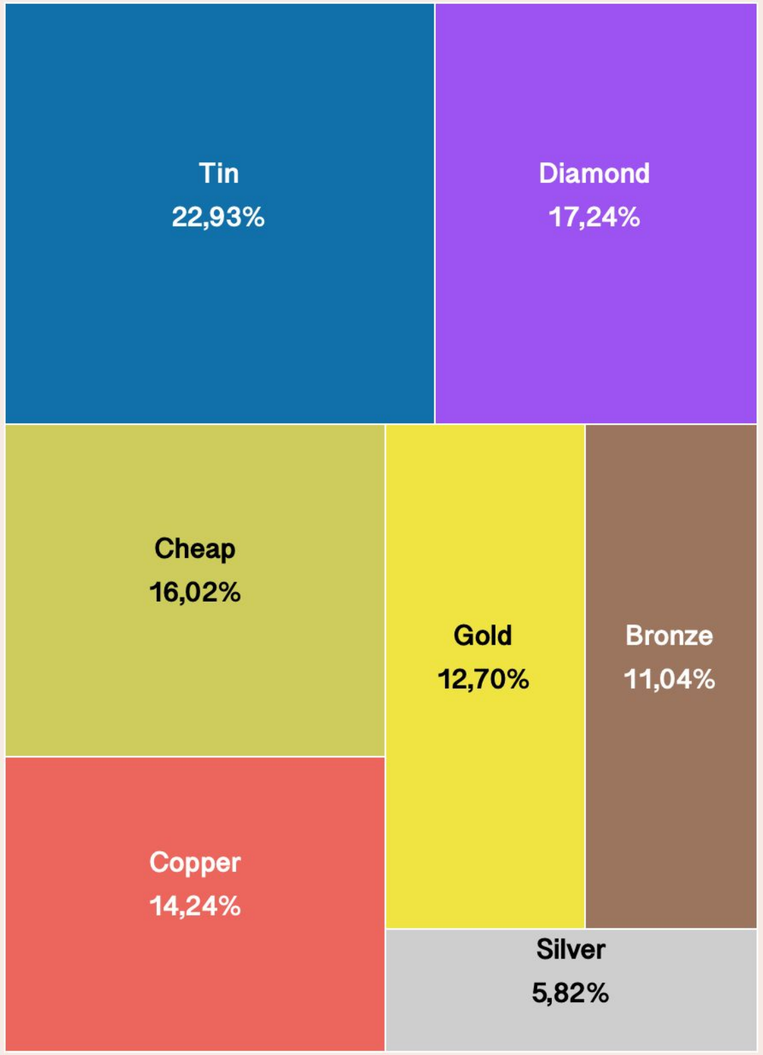
- Average of purchases: **3.5**
- 75% of customers have purchased **up to 4 times** in the time period considered



Monetary

Total amount each customer spent on all transactions between 1st May 2022 and 30th April 2023

- Average expenditure: **265.8€**
- 50% of the population spends 56.7€ per order, however one particularly large order reached a maximum amount of **331.303,56€**



RFM classes

- The active customers are divided in 3 groups (Low, Medium and High) using the tertiles of each one of the RFM measures.
- Carry out the cartesian product of the categories to identify possible RFM segments.
- From the combinations we obtain the RFM classes: **Cheap**, **Tin**, **Copper**, **Bronze**, **Silver**, **Gold**, **Diamond**.

The majority of customers fall into the **Tin** category (23%), not the most active in our business along with the Cheap category (16%) and towards which it is advisable to activate targeted marketing policies such as **sending emails with personalised discounts** to encourage them to purchase.
On the other hand, 17% of customers fall into the **Diamond** category with the highest value: these are customers who tend to spend a lot and frequently, with a decisive impact on turnover, and whom it is worthwhile to maintain and retain in order to maximise their contribution to the business.

Customer focus:

Churn models

Churn model is a predictive model that, at an individual level for each customer, assesses the propensity (or susceptibility) to disengage. For each customer at any given moment, it provides an indication of the associated risk level of potential loss in the future.

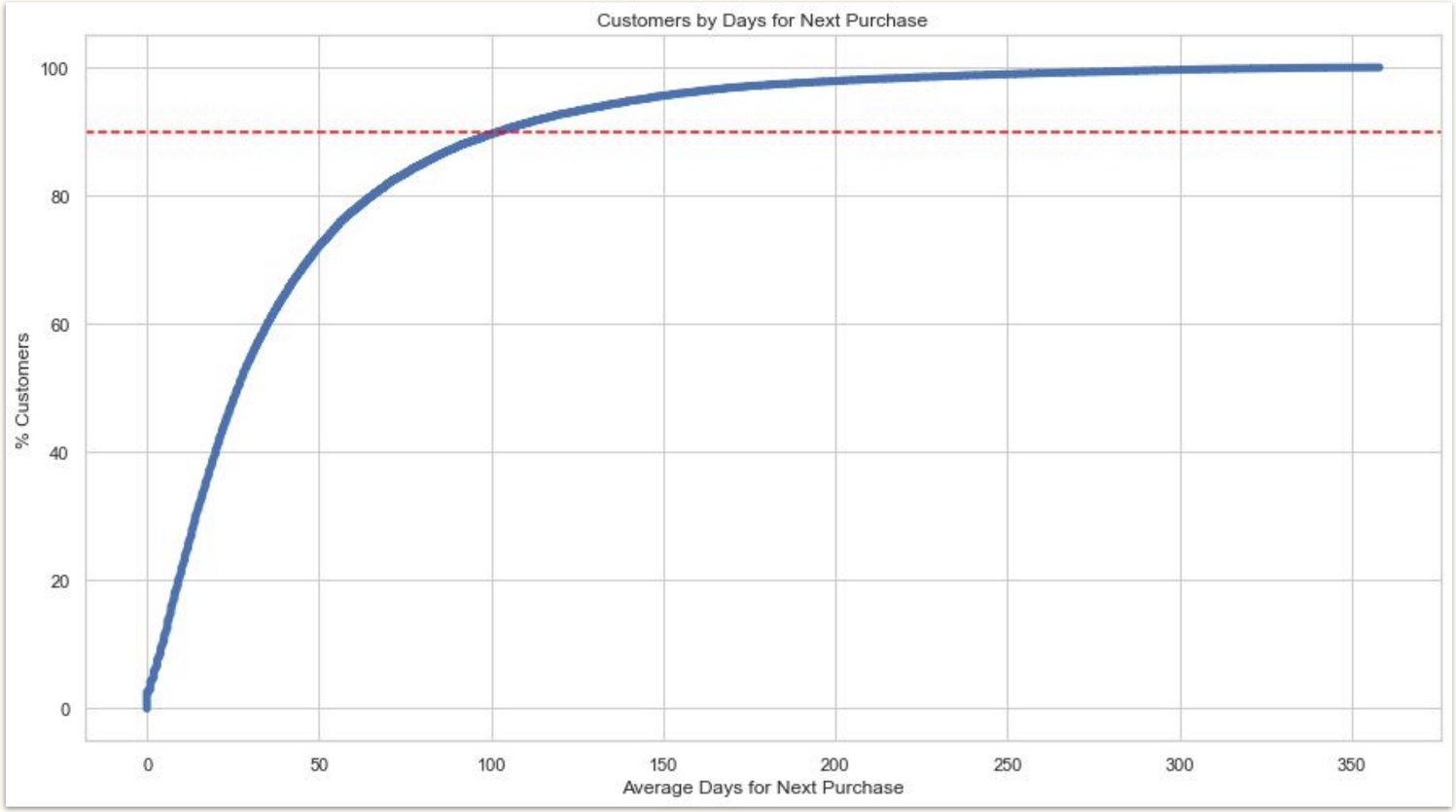
It's a system that categorizes customers into two groups: those who leaves and those who stay. Alongside assigning them to one of these groups, it also provides the likelihood of a customer belonging to a particular group.

Goals

- How to segment the customer base? → **RFM**
- Which customers are at risk of dropping out? → **Churn models**
- What is the likelihood of a one-shooter customer returning? → **Repurchase models**

Repurchase curve

How many days after the last purchase can we say that a customer has stopped buying?



- On average, **90%** of customers repurchase within **102 days**.
- In our case, a customer is considered a **churner** if on average it takes more than 102 days to repurchase.

Based on the insights from the curve, we have determined the threshold value for segmenting the customer base for all the **models** we will be exploring.

Churn models

We considered the following set of explanatory variables, which were not found to be strongly correlated, for the churn prediction models:

- **Gender** (binary)
- **Age**
- **Region** (one-hot encoding)
- **Loyalty type** (one-hot encoding)
- **Days from activation of the account**
- **Recency**
- **Monetary**
- **Whether the customer has ever made a return** (binary)
- **Whether the customer has ever made a review** (binary)

The binary target variable was obtained on the basis of the repurchase time scale by assigning 1 to customers who on average took more than 102 days to repurchase (**churners**), 0 otherwise.

After dividing the customers into training (80%) and test set (20%), we solved the **class imbalance** problem of the target variable on the training set (only 10% of churners) by **oversampling** - through **ADASYN** algorithm - and applied the following classification models: **Logistic model, Decision Tree, Random Forest, XGBoost** and **Multilayer Perceptron**. Finally, we evaluated the results on the test set according to the metrics of **Accuracy, Precision, Recall, F1-score** and **AUC**.

ADASYN note: it involves computing k-nearest neighbors for minority class instances, assessing the imbalance ratio, and creating a density distribution for each minority instance based on neighbor counts in majority and minority classes. Subsequently, synthetic samples are generated for each minority instance according to the determined desired number.

Models

	Accuracy	Precision	Recall	F1-score	AUC
Logistic	0.74	0.25	0.78	0.38	0.83
Decision Tree	0.84	0.29	0.44	0.35	0.65
Random Forest	0.84	0.34	0.61	0.44	0.87
XGBoost	**0.86**	**0.37**	**0.54**	**0.44**	**0.88**
Multilayer Perceptron	0.70	0.24	0.91	0.38	0.86

The best performing model is the **XGBoost** which, with an AUC of 0.88, is the most useful in identifying customers with the highest probability to churn.

Furthermore, the variables that had the greatest impact on classification were **Recency**, **Age**, **whether or not at least one return was made** and **days since account activation**.

Customer focus:

Repurchase models

Repurchase models explains the likelihood that a customer will make a repeat purchase or buy a product or service again. It is a key metric used to assess the probability of customer retention or loyalty.

The process is similar to the one for churn prediction except for the target variable which in this case is the binary **one-shooter** which takes the value 1 if the customer has made only one purchase in the reference period, 0 otherwise - so if it is a **repeater**.

Goals

- How to segment the customer base? → **RFM**
- Which customers are at risk of dropping out? → **Churn models**
- What is the likelihood of a one-shooter customer returning? → **Repurchase models**

Models

If the model indicates that a customer might be inclined to make another purchase, we can send them a **special offer or a limited discount!**

	Accuracy	Precision	Recall	F1-score	AUC
Logistic	0.75	0.63	0.50	0.56	0.80
Decision Tree	0.79	0.66	0.66	0.66	0.75
Random Forest	0.84	0.77	0.68	0.72	0.90
XGBoost	**0.85**	**0.79**	**0.72**	**0.75**	**0.91**
Multilayer Perceptron	0.79	0.73	0.51	0.60	0.84

Considering the same procedure of dividing into training and test set and also taking into account the previous set of explanatory variables, the **XGBoost** model also turns out to be the best in this case, identifying customers with the highest probability of repurchase.

The variables that had the greatest impact on classification were **Monetary**, **Age**, **days since account activation** and **Recency.**

Product focus:

Market Basket Analysis

The Market Basket Analysis is an unsupervised model used to identify and measure relationships among products within commercial transactions. This approach enables businesses to pinpoint items that are frequently purchased together, providing strategic insights to optimize product assortments and implement targeted promotional tactics.

Goals

- Which products can be suggested as complementary purchases to optimize cross-selling strategies?
- In what ways can promotions be personalized based on basket analysis to increase the effectiveness of promotional campaigns?

Products overview

product classes
14

There are 88,538 products divided into 14 classes.

The distribution within the various categories is shown in the graph on the right. We note how **category 12** contains about 16% of the total of all products.

We also noticed how out of 88,538 total products, only **2000 (2.3%) are sold at least once**, the remainder (97.7%) are in the product database but never appear in the purchase or return transactions.

It would be interesting to conduct an in-depth analysis of why this happens. Offering a variety of products can be an advantage, however, if this advantage is not exploited it represents a **loss** for the company.

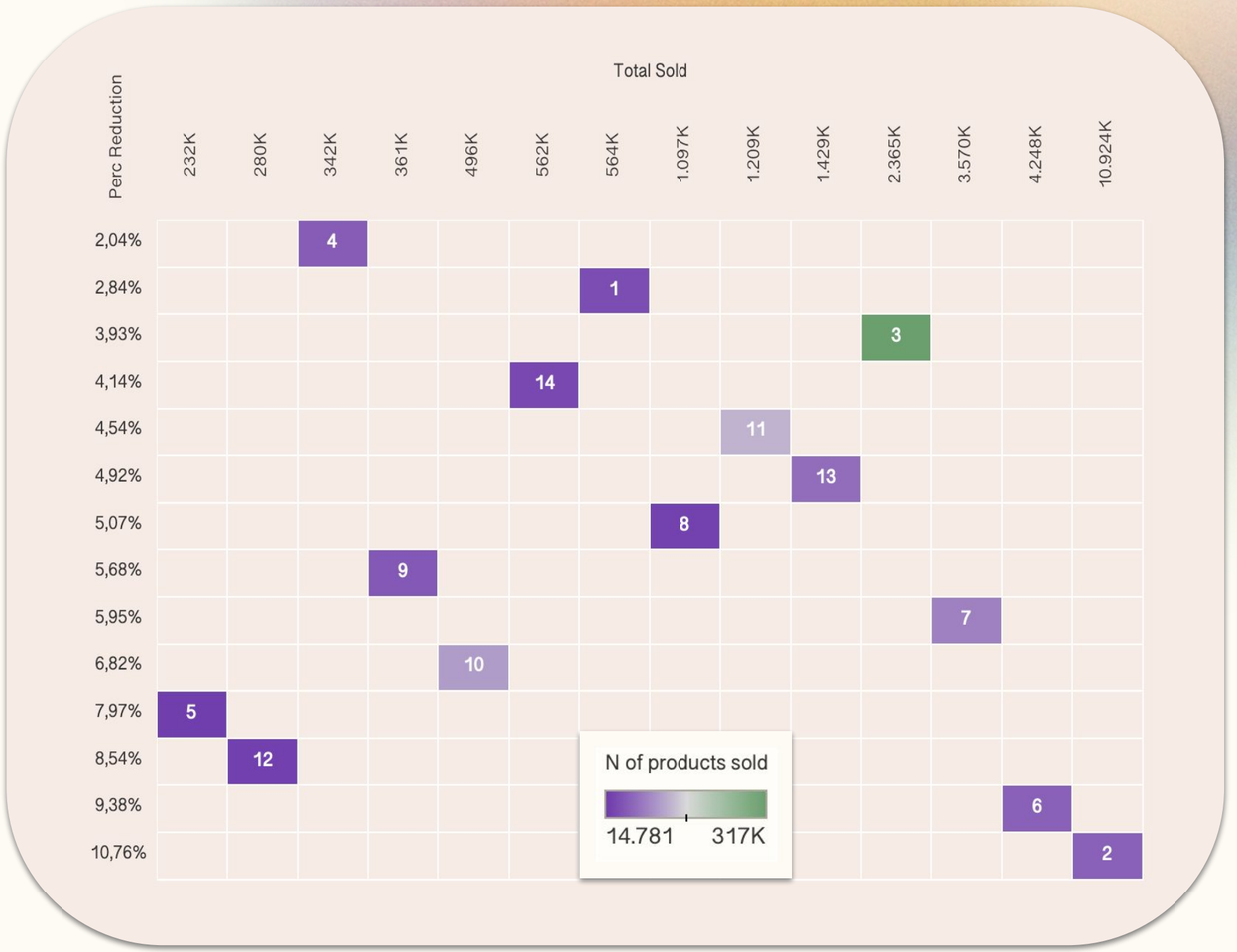
Products overview

There are three crucial dimensions for the company regarding the product:

1. Analyzing the **quantities sold** in terms of volume
2. Calculating the **profit** generated by each product
3. Evaluating the extent of **applied discounts**, measured through the average percentage of discount granted

The image on the right illustrates this three dimensions.

- Top three product categories with the highest sales in terms of **quantity**: **3**, **11**, and **10**.
- Top three product categories with the highest amount in terms of **expenditure**: **2**, **6**, and **7**.
- Top three product categories with the highest amount in terms of percent of **discount** applied (on average): **2**, **6**, and **12**.

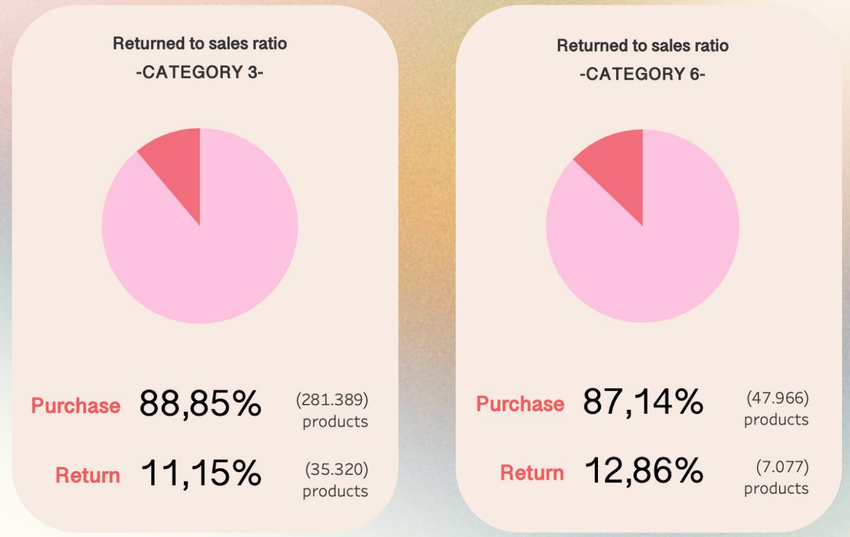


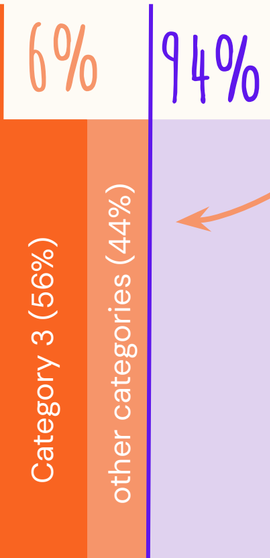
Products overview

In this context, it's noteworthy that approximately **6%** of all sold products are **returned**.

What's intriguing is that **Category 3**, despite being the top-selling category in terms of quantity (representing 30.5% of the total), is also the category with the **highest rate of returns (56% of returned products are part of Category 3)**.

However, when looking at the **return-to-sales ratio in percentage terms**, **Category 6** takes the lead. Specifically, approximately **13% of products in Category 6 are returned**, as depicted in the image on the right.





It is crucial to delve into a comprehensive analysis to understand the reasons behind product returns. This analysis should not only consider material issues, such as defects, but also explore the temporal aspect—whether these problems have recently emerged or have been persistent over the considered timeframe.

One method could be to propose a **post-purchase questionnaire** to users and evaluate their responses, in particular those of customers who have made a return.

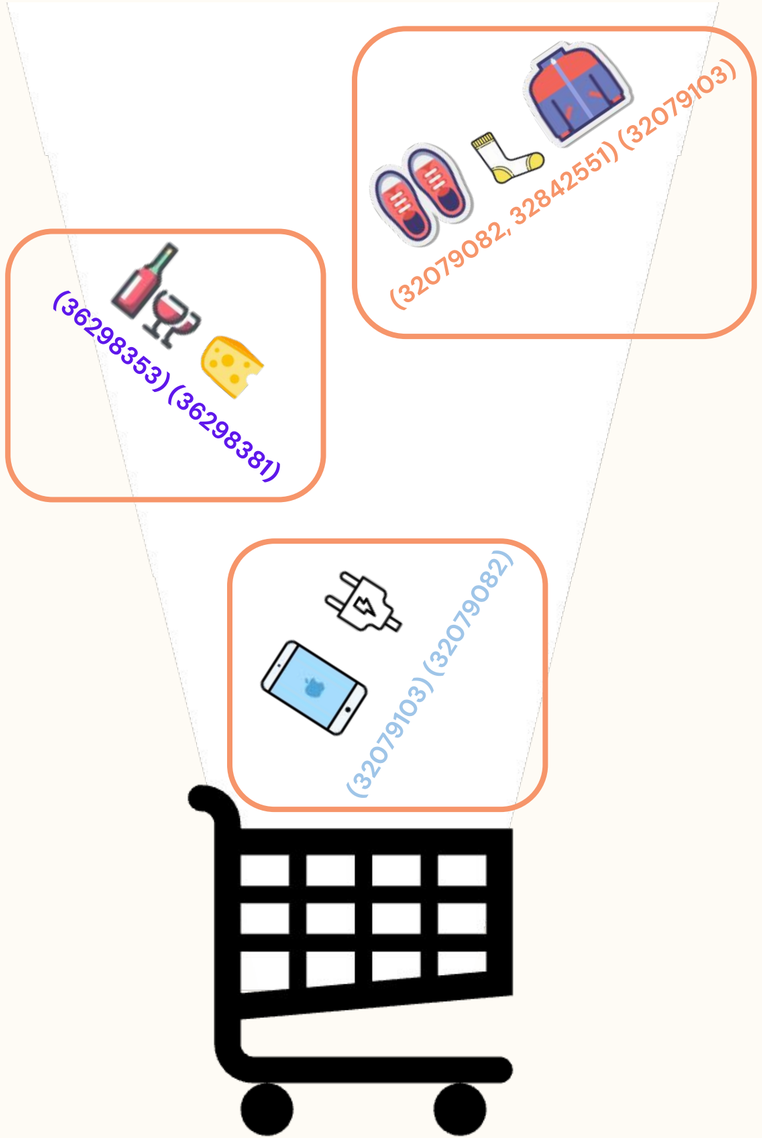
MBA description

- We **excluded** returns and transaction with only one product.
- We generate a **preliminary table** where Order ID is reiterated multiple times along the rows, corresponding to the number of products in the transaction. This initial matrix consists of a single column containing the Product IDs.
- With **TransactionEncoder** from mlxtend.preprocessing we simulate pivoting.
  The resulting transactional matrix has 'order_id' on the rows, and each distinct 'product_id' allocated to its own column.
- We applied the Apriori algorithm with a **minimum support of 0.5%** and a **maximum relation length set to 3**. Our emphasis was directed towards rules exhibiting a **confidence level of ≥ 0.5**.
- We **sorted by support, confidence and lift** to give importance to the products that are most often purchased and their complementaries.

  Note: two products could in fact be purchased 99% of the time together but have very little impact in terms of sales on the total sold.

We note that the majority of products involved are in **Category 3**. It seems natural as it's the top-selling category in terms of quantity.

MBA description



How to leverage the results of the Market Basket Analysis to our advantage?

Identifying top-selling products is certainly an initial strategy to enhance sales, especially if they represent a significant portion of our revenue.

However, MBA provides more detailed information, paving the way for a more targeted strategy on products sold together. Exploring these relationships allows us to develop focused **cross-selling** or **up-selling strategies**.

In this scenario, we aim to suggest complementary products to guide the customer towards a more comprehensive purchase, using targeted advertising, discounts, or promotional packages designed for the purchase of related products.

Feedback focus:

Sentiment Analysis

Sentiment analysis is a natural language processing (NLP) technique that involves determining and extracting sentiments or opinions expressed in text data.

The primary goal is to analyze the subjective information within the text to understand the sentiments, emotions, and attitudes of the writer or speaker.

Goals

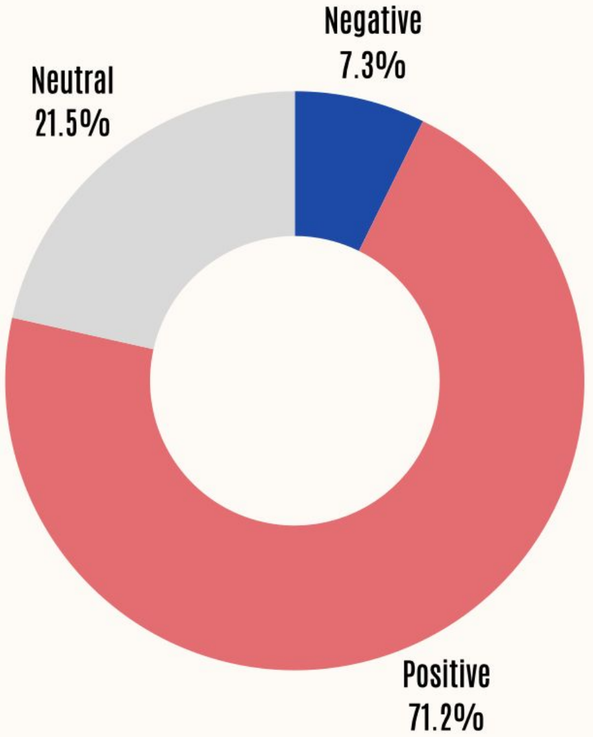
- What do customers think about us?
- What do our customers like/don’t like about our products and services?
- What topics or themes are associated with different sentiments?

We started analysing the text of the 462,744 reviews with their respective ratings – positive, negative or neutral – by performing a preprocessing phase involving the following operations:

- lowercase and remove links, html, http, emoji, tabs, punctuation, numbers and extra white space.
- **Tokenization** with stopwords removal
- **Lemmatization**, a process which involves the reduction of words to their basic form or root.

We then divided the texts into training (80%) and test set (20%) solving the **class imbalance problem** in the training set (64% of reviews are positive, 27% neutral and 9% negative) by **undersampling** the majority classes.

After transforming the structure into a vector representation using the **TF-IDF** approach, we evaluated the performance of the **Logistic model**, **Random Forest** and **SVM**.
The Logistic one was the best with an **accuracy of 72%** on the test set and was then used to classify the sentiment on the set of new customer reviews.



There are few expected negative reviews, only **7%**.

It would be useful to understand whether they refer to a particular product or store!

We analyzed reviews using text mining techniques, creating two Word Clouds to examine the frequency of words in both positive and negative reviews. The aim was to understand, based on the language used, which strengths are praised in positive reviews and which weaknesses are criticized in negative ones.

Note: in positive feedbacks, the recurrent words 'flavor' and 'taste' are accompanied by words which describes the **experience** itself such as 'good,' 'great,' or 'love.' Conversely, in negative feedback, 'taste' prevails, often coupled with 'product'. In this case the words seem to refer to more specific comments about what was **disliked** in the product or experience.

Top 5 most frequent positive words:
taste, **flavor**, **love**, **great**, **good**

Positive feedbacks

Top 5 most frequent negative words:
taste, **product**, **food**, **one**, **buy**

Negative feedbacks