


Marketing Analytics

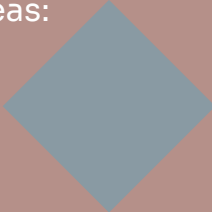
Alex Calabrese
Amelia Acuña
Antonio Sabbatella





Business Objectives & Questions

We aim to leverage data collected from **May 1, 2022, at 07:19:05 to April 30, 2023, at 21:11:41** to uncover actionable insights and develop data-driven marketing strategies across some key areas:



01

Improve Customer Retention

Reduce churn by identifying at-risk customers and implementing targeted strategies.

How can we identify and prevent the churn of high-value customers?

02

Enhance Profit Through Cross-Selling

Utilize MBA to identify product pairs or sets that are frequently purchased together to drive additional sales

How can we increase profit and add value to the customer through product cross-selling?

03

Optimize Customer Experience

Optimize customer experience by effectively managing feedback to minimize negative impacts from detractors and maximize positive engagement from promoters.

How can we reduce the negative impact of detractors and incentivize promoters?

CRISP-DM offers a **structured approach** to data mining, ensuring all steps from understanding the business problem to deploying solutions are systematically covered.

1. Data

Structured Data

The dataset consists of relational CSV tables:

Customer Accounts



Customer Reviews



Labelled Reviews



Orders



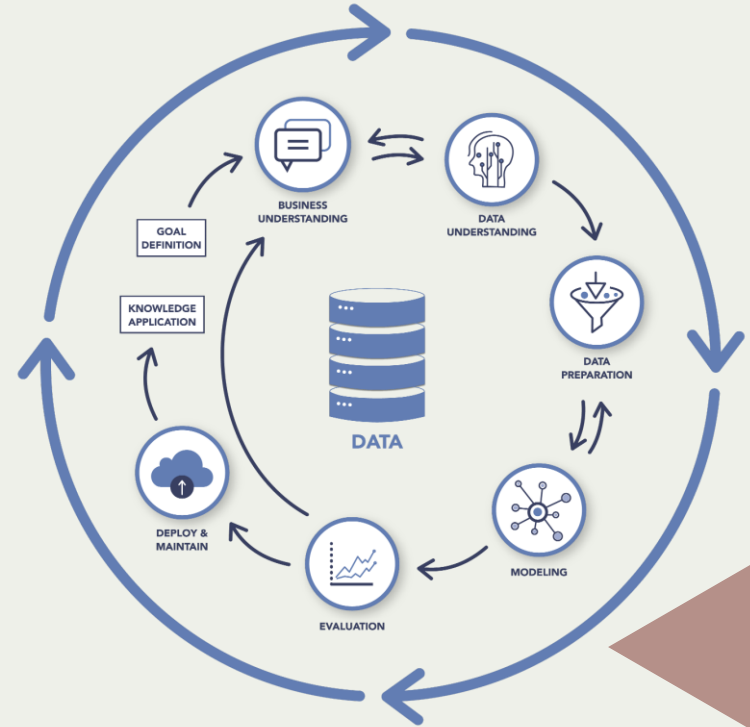
Addresses



Customers



Products



Its iterative nature allows for ongoing improvement and strategy refinement.

2. Data Understanding

Data Relations:

One-to-Many:

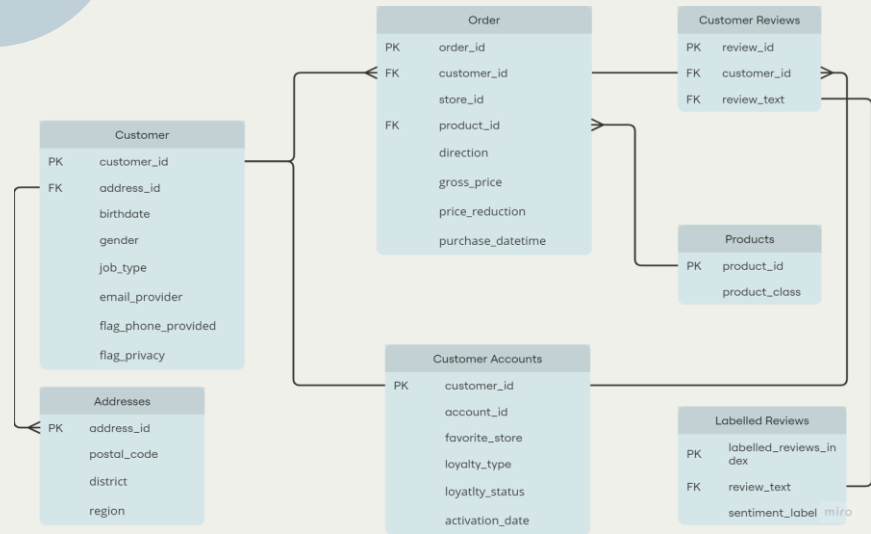
- Multiple orders tied to a single customer
- Each store can process multiple orders
- Address to multi-family homes' customer

Many-to-One:

- Multiple products can be included within a single order
- Each product has a single product class, but each product class can encompass multiple products

Many-to-Many:

- Products are linked to customers through orders. Each product can be ordered multiple times (in different orders)
- A product can appear in many orders



Key Relationships:

Customers ↔ Orders: Purchasing patterns.

Orders ↔ Products: Product affinity and cross-selling opportunities.

Customer Reviews ↔ Orders: Tied customer sentiment to purchasing behavior.

3. Data Preparation

Data Types

Appropriate data types for each attribute, ex. birthdate and purchase_datetime to date formats.



Missing Values

Addressing missing data by imputing defaults (e.g., filling missing job_type and email provider with 'Unknown', for numerical computing with the median, for categorical with “missing”, for binary with mode



Quality Checks

Removed exact duplicates. Consistency by changing flag_phone_provided from float to bool. Identify Outliers (more than 3 standard deviations)

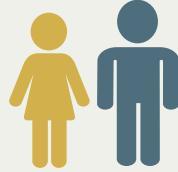
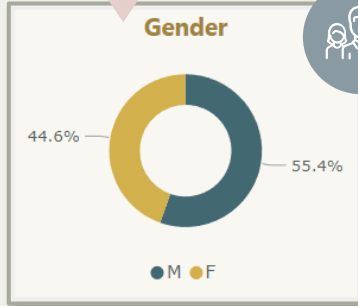


Data integration

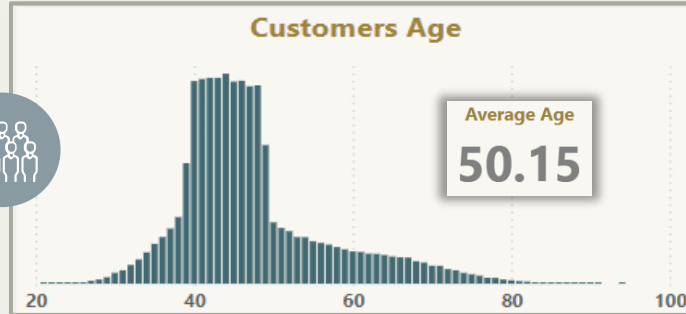
Joins multiple datasets into one comprehensive DataFrame, combining customer, product, order, address, and review data.



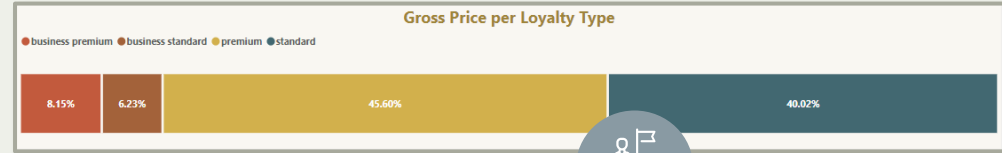
4. Data Exploration



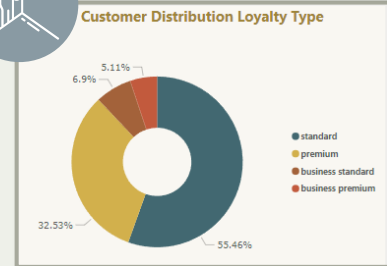
The customer base is 55.4% male, **slightly higher** than Italy's 49%. Not enough difference to targeted this specific segment.



The majority of customers are aged **between 40 and 60**, aligning marketing strategies toward this age group can maximize engagement.



While Standard customers make up the majority, **Premium** customers generate **slightly higher revenue**, indicating potential for targeted campaigns to increase spending among Standard customers.



Total Gross Price

18,031,477.05



Total Price Reduction

1,085,948.57

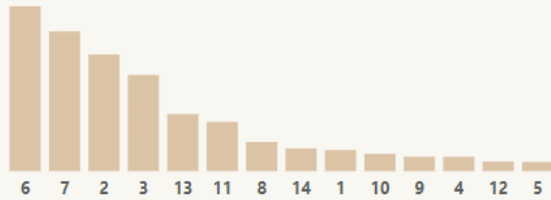


Devolutions

30,719

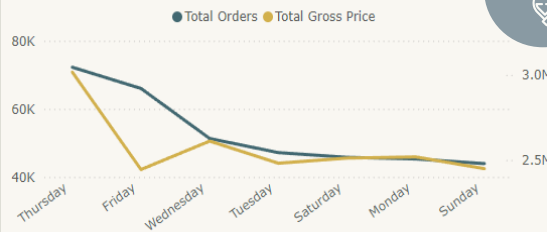
4. Data Exploration

Total Gross Price - Product Class



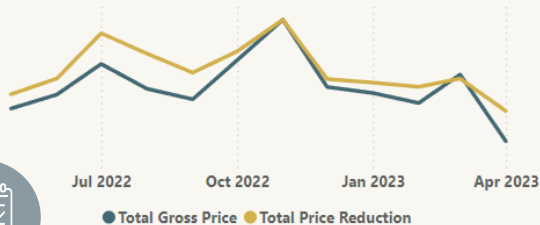
Product class **6** generates the highest total gross price, followed by **7** and **2**.

Gross Price and Order per Day



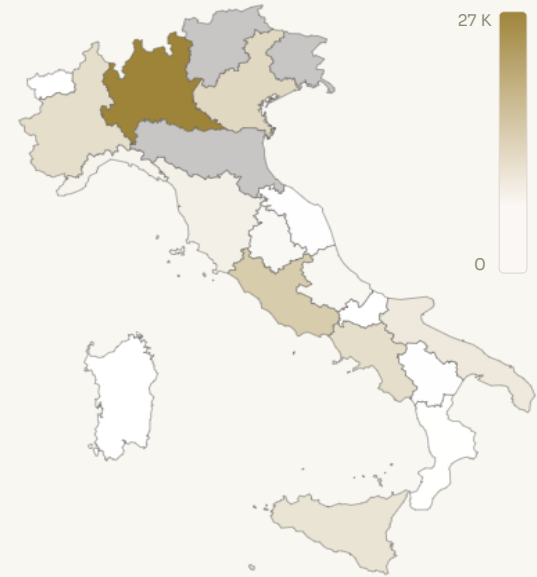
Thursday sees the highest volume of orders and gross price, with activity gradually decreasing throughout the week, reaching the lowest on Monday.

Total Gross Price - Total Price Reduction



Total gross price fluctuates over time, with notable peaks, while price reductions closely, indicating periods of high activity and discounting (July 2022, Nov 2022, March 2023)

Count of Customers per Region



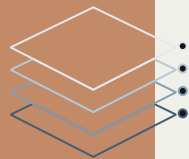
Northern Italy, particularly the Lombardy region, has the highest customer concentration, providing opportunities for regionally targeted marketing strategies.

5. Model

Churn Analysis

Churning refers to the moment a client stop buying a company's products or using its services

Retaining existing customers is often more profitable than acquiring new ones.



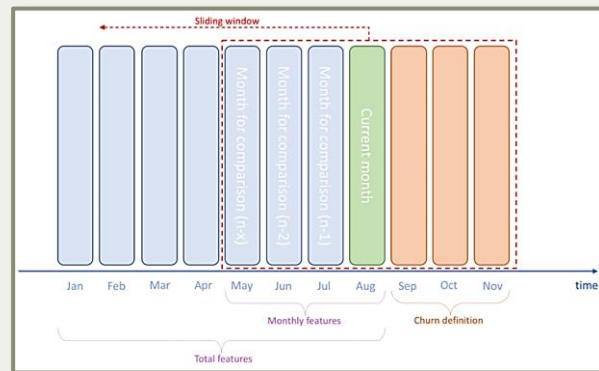
Observation Period (90 days): Tracks customer purchases to mark them as active.



Control Period (30 days): If no purchases follow the observation, customers are labeled churned



Step Size (30 days): Windows shift every 30 days, ensuring overlap to capture varying patterns.



The target variable was defined as a **binary indicator** of churn based on the sliding window approach.

Churn Analysis

We benchmarked four algorithms for the task:

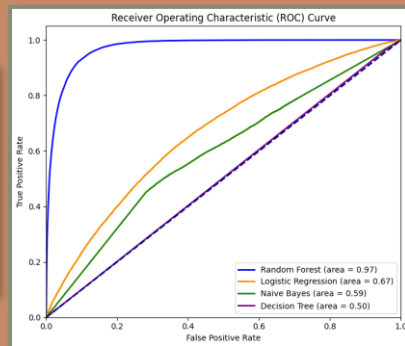
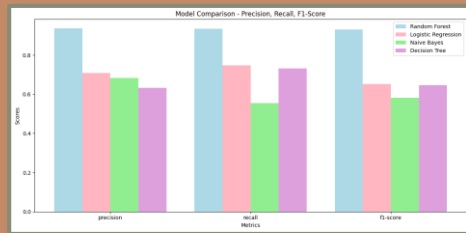
Random Forest

Logistic Regression

Naïve Bayes

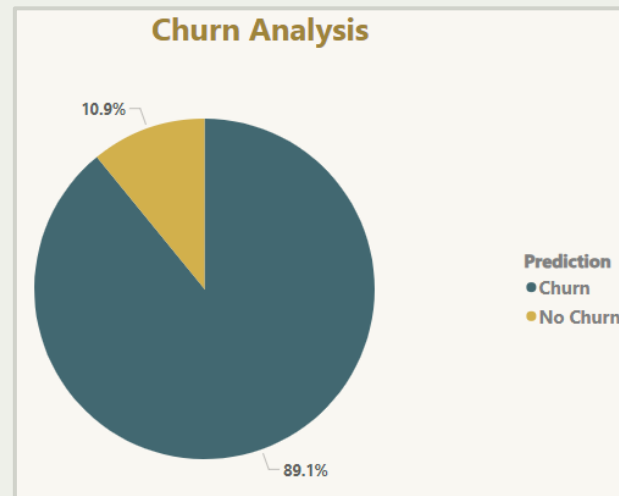
Decision Tree

The Random Forest model performed the best, **precision** of 0.95, **recall** of 0.88, and an **F1-score** of 0.91, demonstrating its strong ability to manage the complexity and non-linear patterns in churn prediction.



From 104134 customers, 97925 made at least one purchase in the 3 months of observation of time window, therefore taken into account for the analysis.

Time frame of the graph: Most recent time window





Primary campaign:

High-Value Customer at Risk Identification

Identified the **top 25%** of customers by total monthly spend as high-value

Filter **churn probability > 70%**, as per results in the probability distribution

Segment them into “Low,” “Medium,” and “High” based on their RFM scores

Select higher profiles

High-value customers at risk (**11.620**) predominantly fall within the **40–50 age range**, a demographic known for its purchasing power and brand loyalty. Customers are heavily engaged in our **standard** and **premium loyalty** programs, with **nearly all** having **provided phone** numbers and agreed to **privacy** terms.

Theme of campaign:

Taking into consideration their age we want to showcase: convenience, value for money, family-oriented products, health and wellness services.



Loyalty Program Upgrade

Instant and free upgrade to the next loyalty program level.

Additionally, we will provide **double points** on their next purchase.



Historical Spend-Based Discounts:

Limited-time offer, encouraging immediate action

20% off
Top Spenders

15% off
Medium Spenders

10 % off
Low Spenders



Direct phone calls for the highest-value customers (**the top 5%**) and for the second peak of customers between **60–70 years old**, this age group might appreciate **more direct, personal contact**.



SMS and personalized emails



5. Model

RFM Analysis

Identifies an organization's best and worst customer segments



Recency - How recently did the customer make a purchase? If the customer has made a purchase more recently, then they are likely to be more responsive to outreach and promotions



Frequency - How often does the customer place an order? If a customer purchases more frequently, then they are more engaged and satisfied



Monetary Value - How much do your customers spend per order? Every purchase is valuable, but combining this factor with how recently or frequently a customer purchases could indicate if they are a brand loyalist

RFM Segments



Champions - bought recently, buy often and spend the most

Loyal Customers - spend good money and often, responsive to promotions

Potential Loyalist - recent customers, but spent a good amount and bought more than once

New Customers - bought most recently, but not often

Promising - recent shoppers, but haven't spent much

Needing Attention - above average recency, frequency and monetary values; have not bought very recently

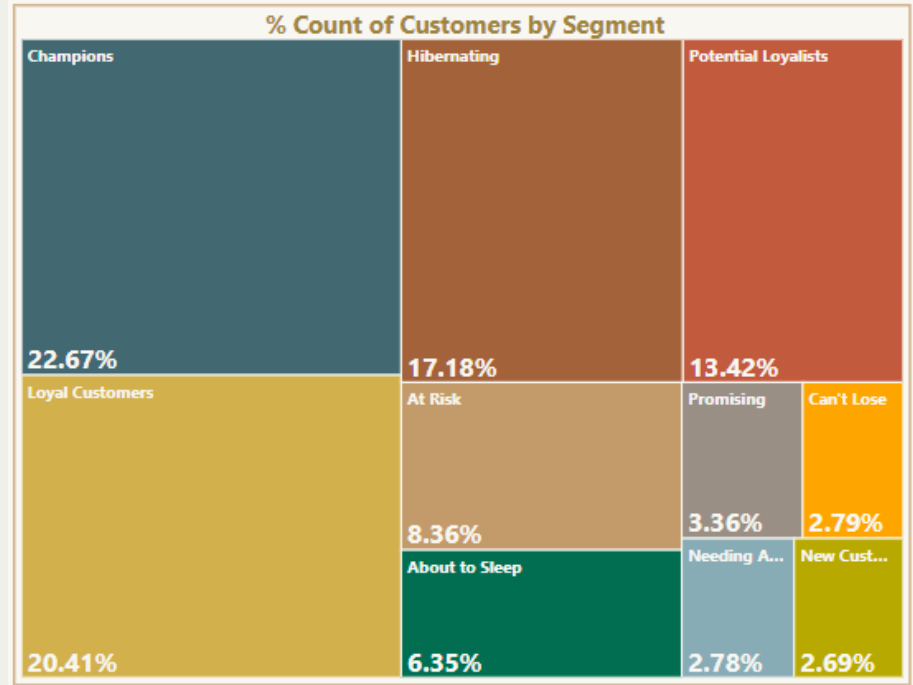
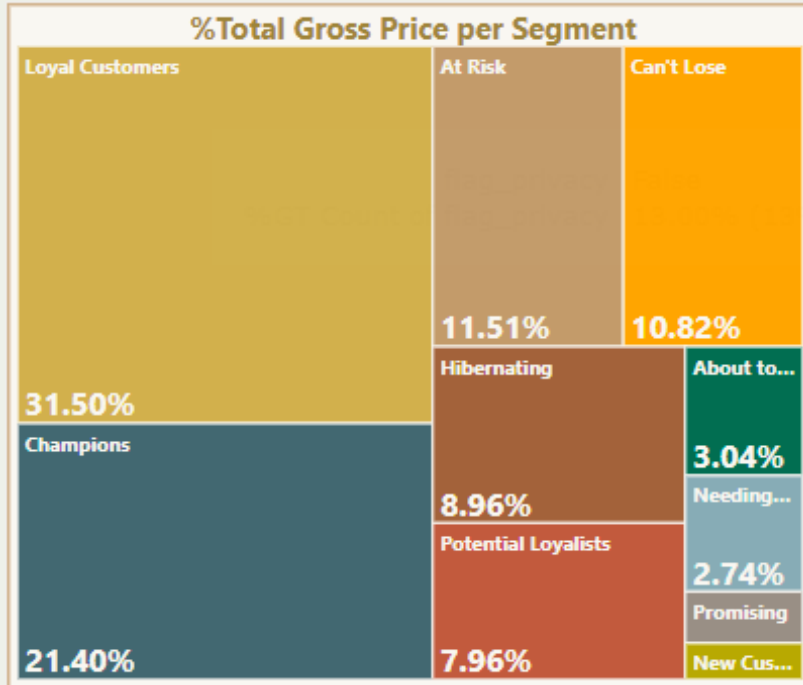
About To Sleep - below average recency, frequency and monetary values; will lose them if not reactivated

At Risk - spent big money and purchased often but long time ago; need to bring them back

Can't Lose Them - made biggest purchases, and often but haven't returned for a long time

Hibernating - last purchase was long back, low spenders and low number of orders

RFM Analysis



These visuals highlight the customer segments that drive **the most revenue (left)** versus **their overall count (right)**. This helps **prioritize** high-value segments like Loyal Customers and Champions while focusing retention efforts on At Risk and Hibernating customers.

Secondary Campaigns

Loyal Customers (20.4% of customers, 31.5% of gross price)

Engage & Reward:

- Double **loyalty points** on the next purchase within a specific timeframe.
- Invite them to **exclusive events** and give them early access to products in their **favorite categories**. This keeps them engaged and feeling valued.

Channels: Targeted Email, SMS reminders, VIP Event Invitations.



At Risk (8.36% of customers, 11.5% of gross price)

Re-Engage & Recover:

- A **72/48/24/12** hour **exclusive sale** with a 30% discount that decreases over time to create urgency.
- Survey to understand why they've disengaged, offering a chance to win a 100€ **gift card**.
- Highlight free returns to reduce purchasing hesitation.

Customer Profile: Youngest segment, average age 44.65 years.

Channels: Email with Countdown Timer, social media, Direct Mail Surveys.

Champions (22.6% of customers, 21.4% of gross price)

Leverage Loyalty:

- **Referral program** with a 20€ credit, (slightly above their average order value).
- 20€ **voucher surprise** on their birthday or holidays.
- Direct customer **service line**, tailored for their older demographic (average age 56.83 years) who appreciate personal contact.

Channels: Personalized Emails, Phone Calls, Loyalty Program Portal.

Secondary Campaigns

Hibernating (17.1% of customers, 8.9% of gross price)

Selective Re-Engagement:

- Target top 5% past spenders with a 10% discount.
- For the rest, connected with a brand video that highlights our mission. Avoid over-communication to maintain email reputation.

Channels: General Email, YouTube



Potential Loyalists (13.4% of customers, 7.9% of gross price)

Upsell & Engage:

- A 20€ discount on additional products to encourage bigger purchases.
- We will get them onboard with our newsletter by offering a 5€ signup bonus, keeping them engaged over time.

Channels: Email, In-App Notifications, Newsletter Signup Forms.

Can't Lose Them (10,8% of gross price)

Personal Touch:

- Personalized communication, acknowledging their value.
- We will offer immediate loyalty tier upgrades and customized discounts on products they love.

Channels: Direct Mail, Phone Calls, Personalized Emails.



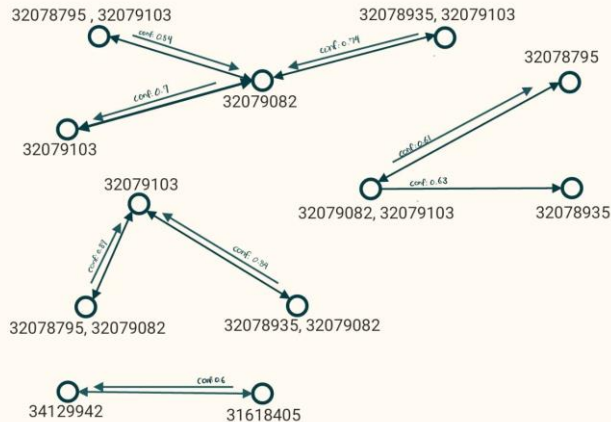
5. Model

Market Basket Analysis

We grouped products in each order to find frequent combinations for cross-selling, focusing on pairs that **appear in at least 1%** of orders. We then filter for strong relationships, where the **likelihood of buying** one product with another (measured by lift) is **0.7 or higher**.

Market Basket Analysis

Top 15 lift – Cross Selling Opportunities



If a customer buys product (32078795, 32079103), there's a 84% chance they'll also buy (32079082), with a strong association indicated by a lift of 111.9.



*Second row indicates the highest confidence when relation is bidirectional

Targeted Cross-Selling:

Leverage purchase data to recommend **complementary products**.

Example: CustomerA bought 32079082 in the last 4 weeks, we will promote 32078795 and 32079103.

Channels: “Frequently Bought Together” sections on product pages. *In-store:* Link customer history to loyalty accounts, offering targeted suggestions at checkout through POS systems.

Loyalty-Driven Cross-Selling

For Premium loyalty customers, an offer of **exclusive bundles**.

Example: 32079082 + 32079103 with a 15% discount for high-spending customers and 10% for the rest.

Loyalty points to drive repeat purchases **1 euro spent = 1 point**.
100 points = 5 euros off.

Channels: *In-store* Promotions in screens, printed coupons. Loyalty Program customers also see these offers on their loyalty dashboards.

Personalized emails, app notifications, SMS

5. Model

Sentiment Analysis

Categorize customers into **Detractors**, **Neutral**, and **Promoters** based on their predicted sentiment scores from their reviews.

We used machine learning techniques and Natural Language Processing:

- Traditional Naive Bayes, LR, DT
- Advanced (BERT transformer).

We used a labeled dataset, which was then applied to predict the sentiment of an unlabeled dataset.

Score ≤ 0.3
Detractors



$0.3 < \text{Score} \leq 0.7$
Neutral



Score > 0.7
Promoters

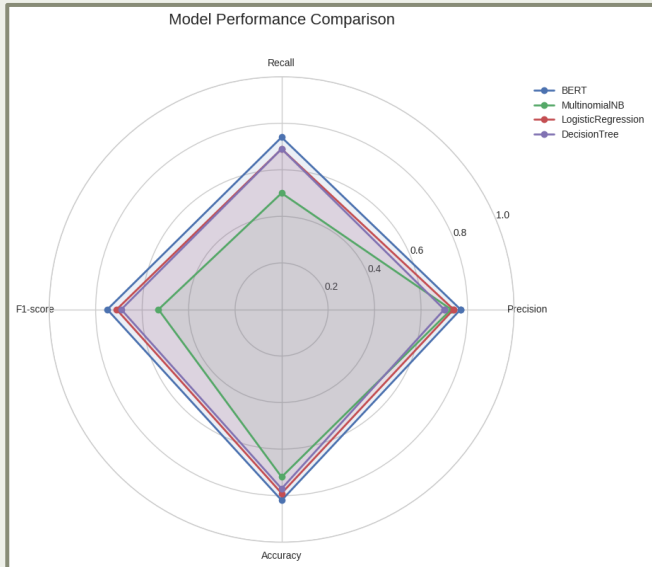


Consistency: We verified that both the labeled and review datasets were in the same language and described customer reviews.

We fine-tuned the BERT pre-trained model on label data to perform sentiment analysis on customers reviews.

Sentiment Analysis

We compared different machine learning models' performance using **text vectorizer embedding**, and the **pre-trained BERT model**. The metrics used for comparing the models are F1, accuracy, precision, and recall.



DETRACTORS WordCloud



We suggest further quality analysis in product categories such as **food** (with characteristics 'flavor' and 'taste') **hygiene** ('shampoos', 'conditioner')

Our company might offer some products through **Amazon**. Here "taste," "flavor," "coffee," "food," are positive features.

PROMOTERS WordCloud



Sentiment Analysis



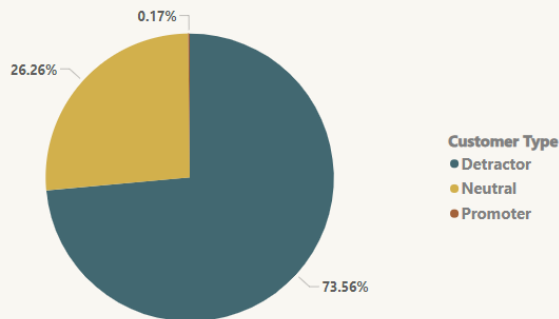
"We hear you"

We will target detractors who purchased last month. and has **highest gross price** and number of **orders**, we'll use a **dedicated support line** to resolve issues and offer tailored incentives (**free returns, free repairs**), aiming to convert them into neutral.

Detractors: Personalized Problem-Solving

Top 10% detractors with the highest gross price and segmented based on specific complaints (e.g., product quality, pricing, service). We will address them with **direct communication** and **quick-fix**.

Number of customer per Sentiment



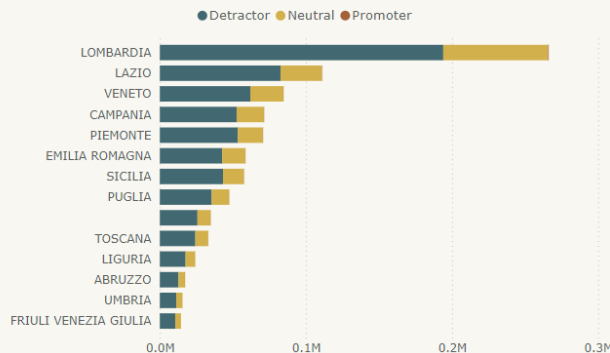
From the total of customers (104.134), 73,5% are identified as detractors, 26,32% as neutrals, and only 0.187% are promoters.

Exclusive Experiences:

Promoters receive **early access** to new products or services through **private shopping events** in the favorite stores with an **accompanying guest**.

Neutral that purchased in the **last month** and had a score **of 0.5 or above** could also be included for faster conversion.

Distribution of Sentiment by Region



(EXTRA 1): Marketing campaign with Sentiment analysis + MBA

Sentiment-Guided Product Recommendations for Neutral Customers

Target Audience: Neutral customers with mid-range sentiment and **PREMIUM** loyalty plan.

$0.3 < \text{Score} \leq 0.7$
Neutral



MBA criteria: **have purchased one** or a set of products that have high **cross-sell potential**
 $(32079103, 32078795) > (32079082)$
 $(32079082, 32078795) > (32079103)$

We will encourage these customers to **explore complementary products** that can enhance their overall experience.



We will offer the neutral customers a in-store "**try-before-you-buy**" option at the customer's preferred store location or a **trial period** for the complementary product.

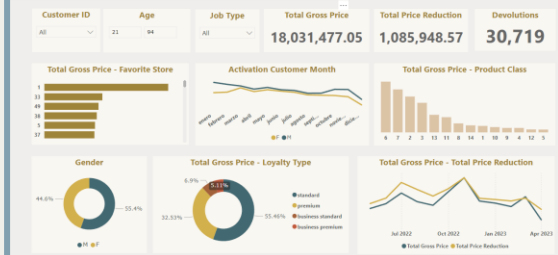


After the trial, **encourage feedback** through a simple review process with **loyalty points** or **25 points*** to redeem in the next purchase.

*100 points = 5 euro off

(EXTRA 2) PowerBI Dashboard

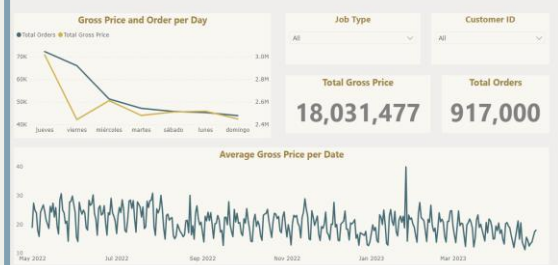
Marketing Analytics - EDA



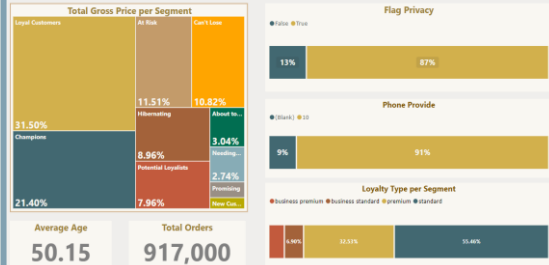
Marketing Analytics - EDA



Marketing Analytics- EDA



RFM Analysis



CLICK to access Dashboard:



*Access granted to members of UNIMIB

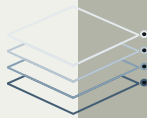
(EXTRA 2) New Strategy

Propensity modelling and multi-category predictions for products

The model incorporates **key customer features** like loyalty status, past interactions, and how they've responded to promotions.

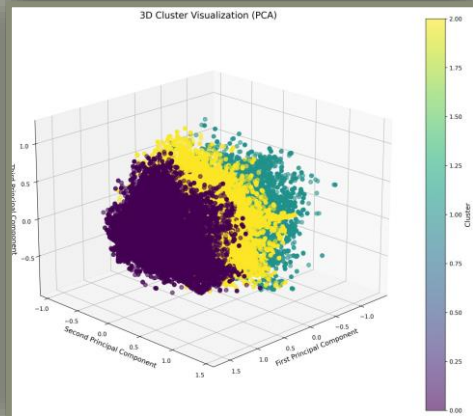
Sliding window approach to create our target variables, which has given us a **more dynamic view** of customer behaviour.

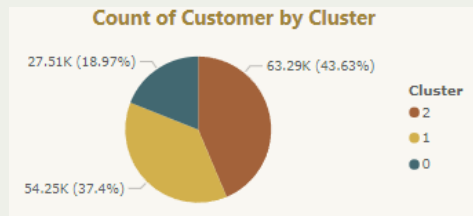
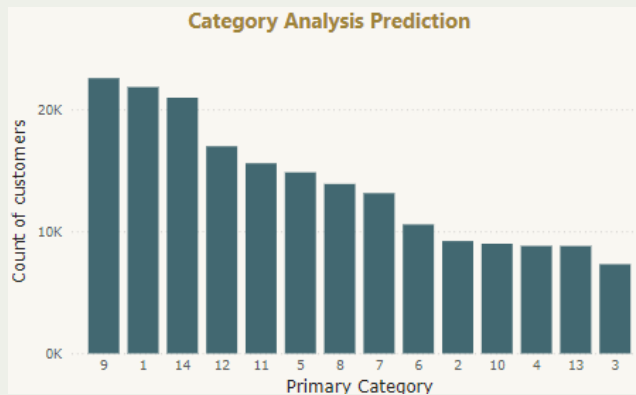
The model's performance looks promising, with well-balanced class distributions and strong ROC AUC scores.



We trained a **Random Forest classifier** to predict a customer's propensity to purchase products in multiple categories (multi-label classifier)

After calculating propensities, the data can be segmented using K-Means clustering. This helps group customers with similar behavioural patterns, allowing for better-targeted marketing strategies.





Head of Output for Category Analysis

Customer ID	Primary Category	Secondary Categories
476735	13	['8']
649064	4	['5', '3', '12', '9']
643161	11	['3', '14', '9', '5', '12', '1', '4', '10', '1...']
398150	2	['12', '11', '6', '10', '1', '9', '3', '13', '1...']
493049	9	['13']



Personalize every campaign based on the strongest interests of each customer.

Generalize our campaign to each cluster.

Smart cross-selling & upselling based on predicted interest

Create product bundles that align with their preferences, driving higher value per transaction

Helps optimize marketing spend (we know which category to focus on)



Thanks!

