Luca Galli

Marketing Analytics

University Project

Master's Degree in Data Science

University of Milano-Bicocca

Goals

Engage promoters and high-value customers, convert detractors and potential churners

01

Exploration & Preparation

02

Cluster customers with RFM

03

Predict churners to retain

04

Extract reviews sentiment and insights

01. Exploration & Preparation

Brief overview of the preprocessing and exploration phase

01. Preparation

Customer and products entities

Basic cleaning steps: NULL VALUES · WRONG TYPES · DUPLICATES

***Main focus:* Addresses cleaning and enrichment process:**

1. Regions/districts abbreviations extended with *dictionary* lookup + *manual fixes*
2. Missing/abbreviated districts inferred by postal code (*dictionary* lookup)
3. Residual uncleaned regions first inferred from enriched districts, then by postal code
4. Manual replacements for the most common null regions (Campania and Lazio)

Regions improvement: 6530 to 679 null values
Districts improvement: 15621 to 738 null values

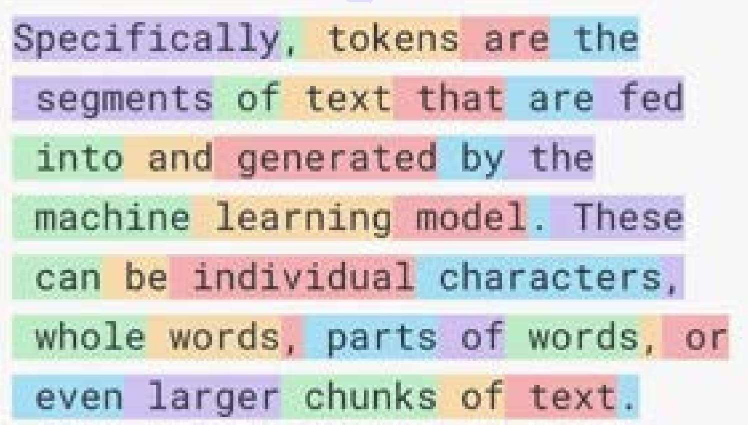
Note: lookup dictionary was created by joining trusted dictionaries available for free on the Internet, containing italian postal codes with their corresponding district and region, and their respective abbreviations.

01. Preparation

Review entities

Review text pre-processing

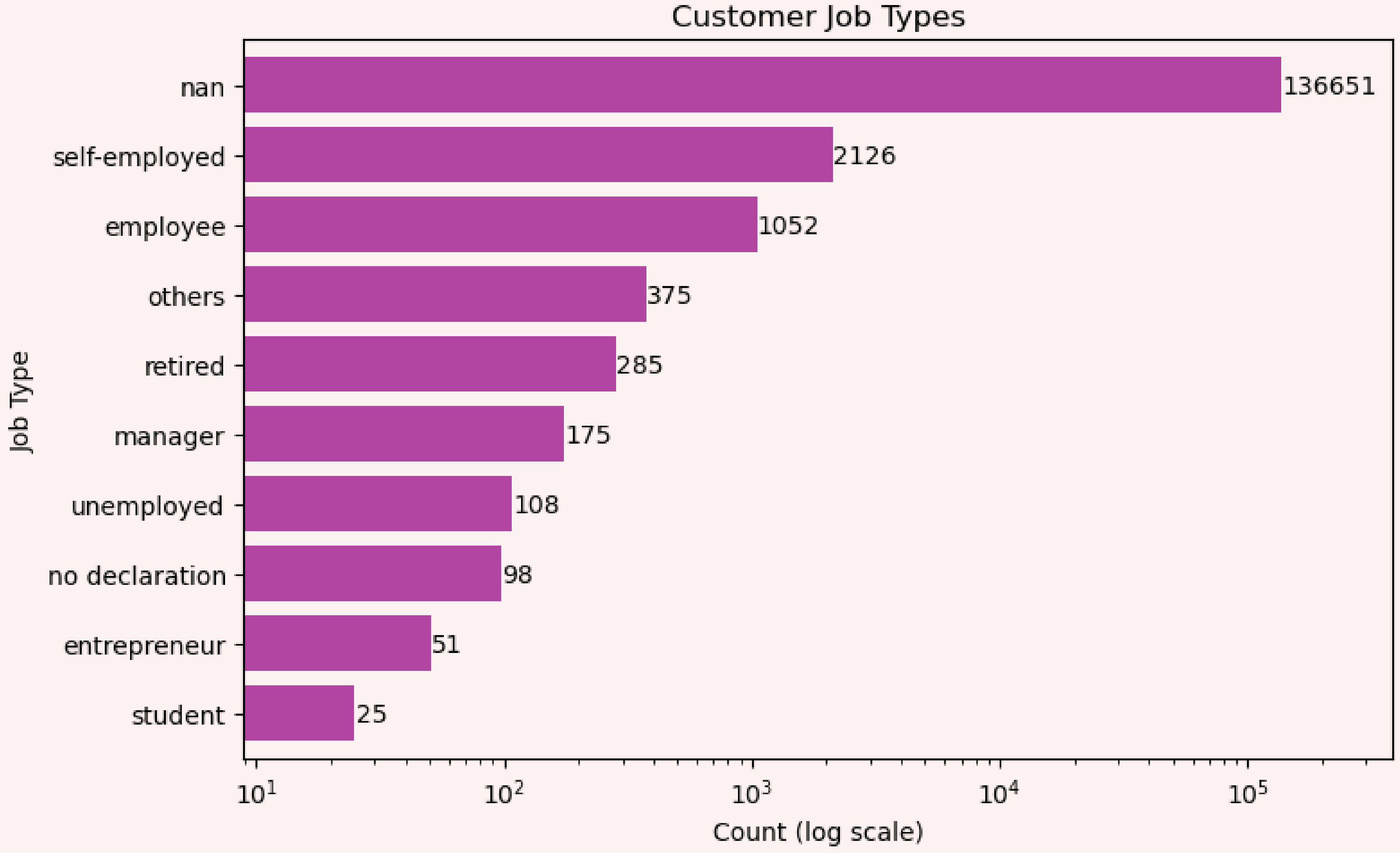
- **URLs** and **tags** removal
- **Numbers** removal
- **Punctuation** removal
- Character **repetitions** removal
- **Stop-words** removal
- **Concatenated whitespaces** removal
- Word **tokenization**
- Tokens **lemmatization**

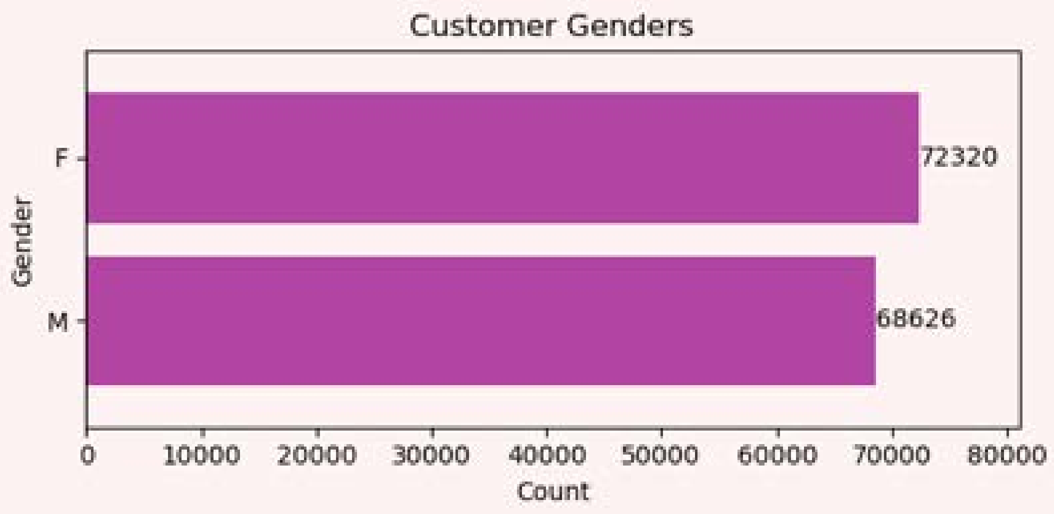


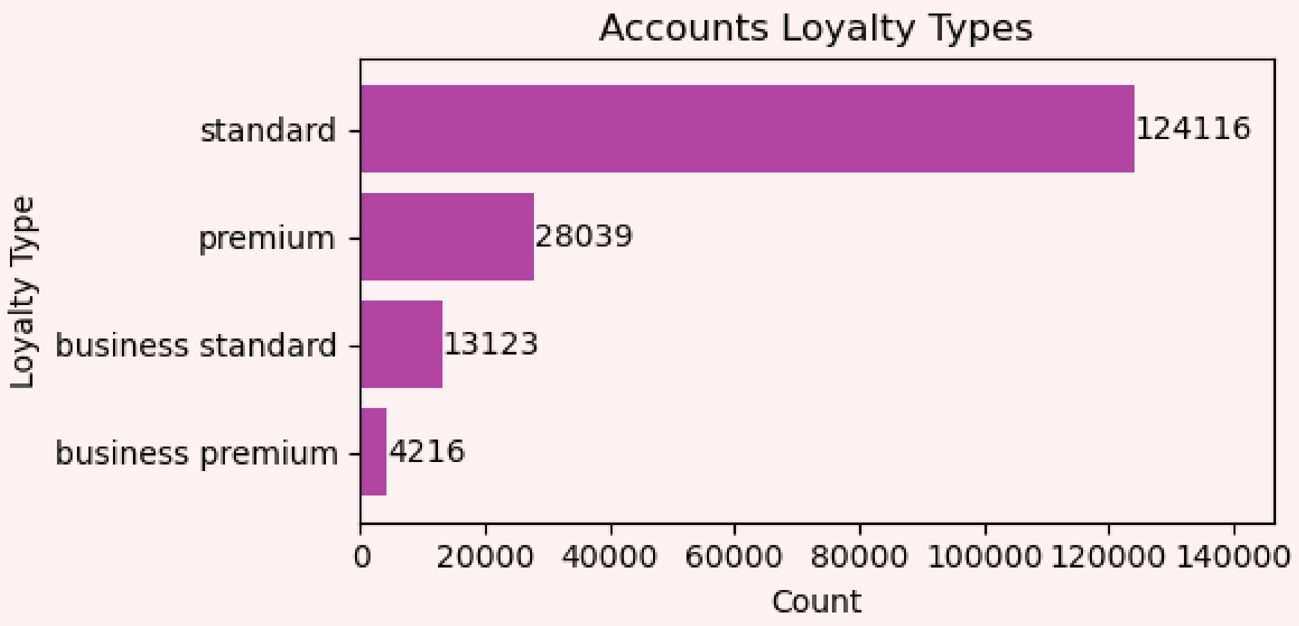
01. Exploration

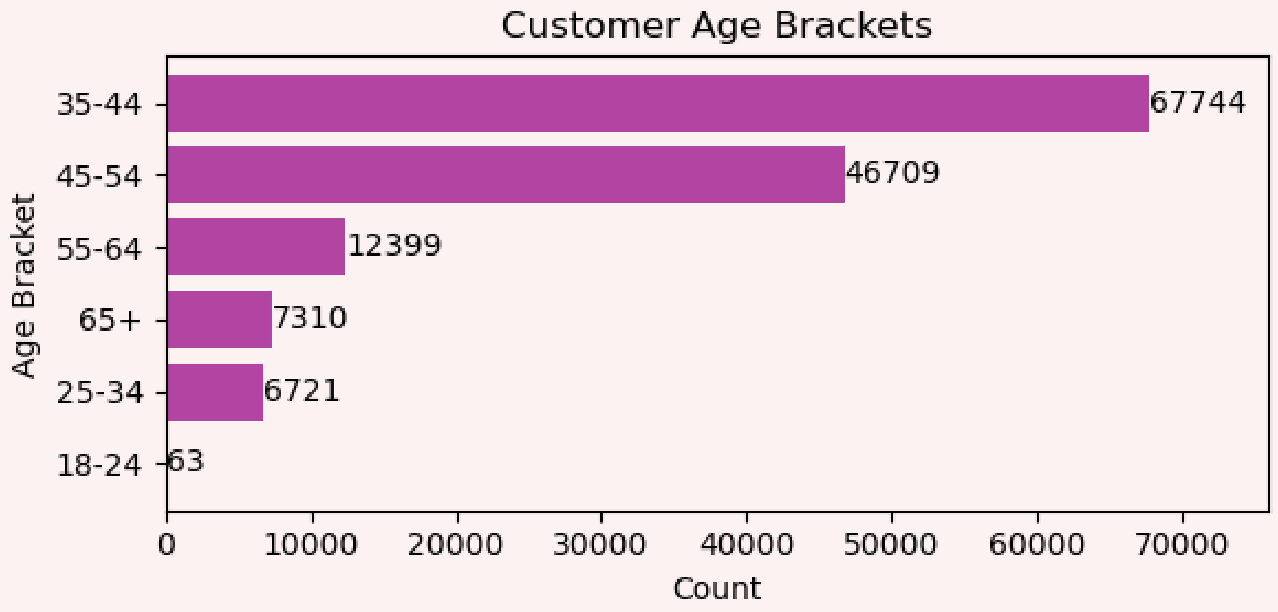
Customer entities

Total: 140,946 customers









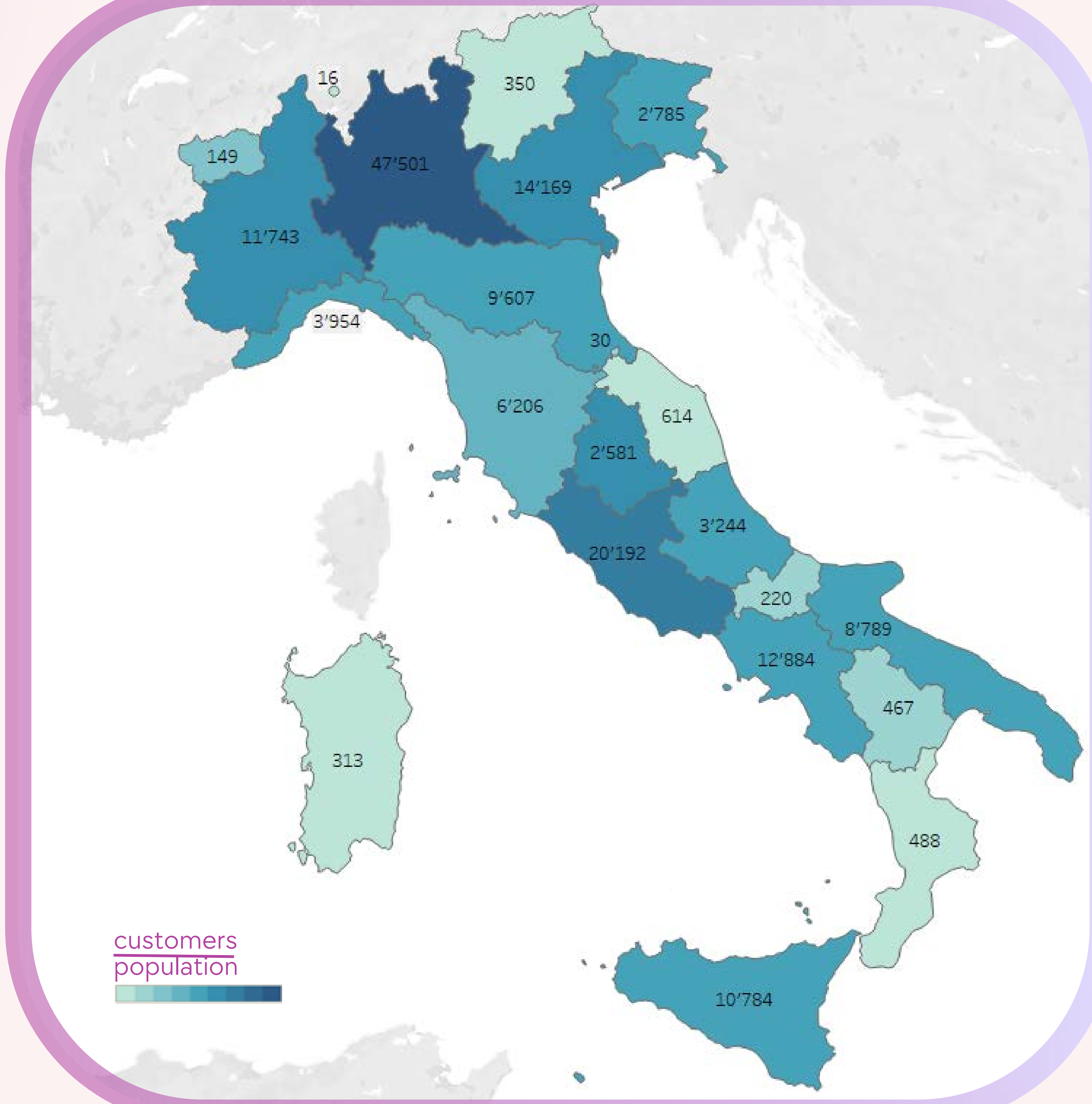
01. Exploration

Customer entities

Districts

TOP 10

1.	Roma	18058
2.	Milano	15800
3.	Napoli	8621
4.	Torino	8134
5.	Monza-Brianza	6163
6.	Palermo	6061
7.	Bergamo	5807
8.	Bari	4360
9.	Bologna	4231
10.	Venezia	3235

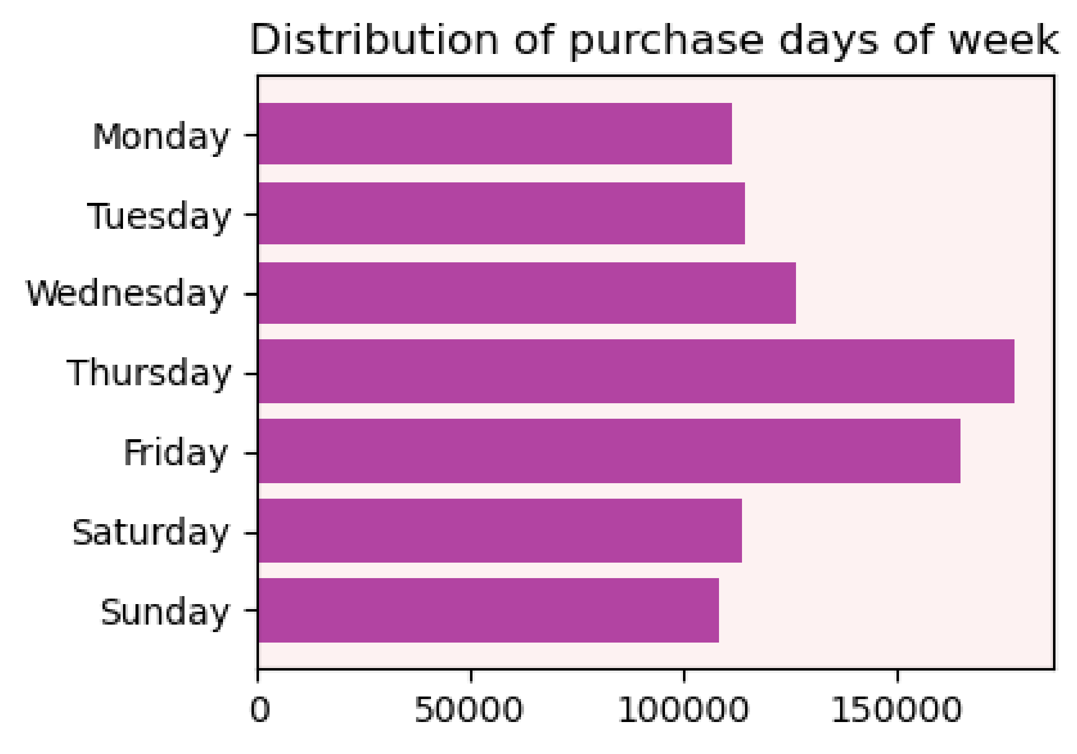


Buyer persona: standard user from Lombardy,
around 40 years old,
account activated in mid-August 2022.

01. Exploration

Product entities

Orders are mainly placed in the evening or late morning.
Most common day of the week to place an order is Thursday.
Most common expense for a single order is around €13.



Average order:

- **Gross price**
  - mean: €48.50
  - median: €13.30
- **Gross price** + *reductions*
  - mean: €45.58
  - median: €12.94

Purchases **overview:**

From the 1st of May 2022 to the 30th of April 2023 (~1 year):

- **Gross total income**: €18,031,477
  - with *reductions* applied: €16,945,528
- 917,000 **purchases**
- 371,804 **orders** (~2.5 *purchases per order*)

01. Exploration

Review entities

Labelled Reviews

- 297008 positive
- 123386 neutral
- 42350 negative

UNBALANCED
Required **oversampling**
and weighted metrics

Customer Reviews

Strongly focused on food taste and ingredients.
Similar to labelled reviews most common words
=> Good train-test split

02. RFM Analysis

Identification and retention campaign for high value customers

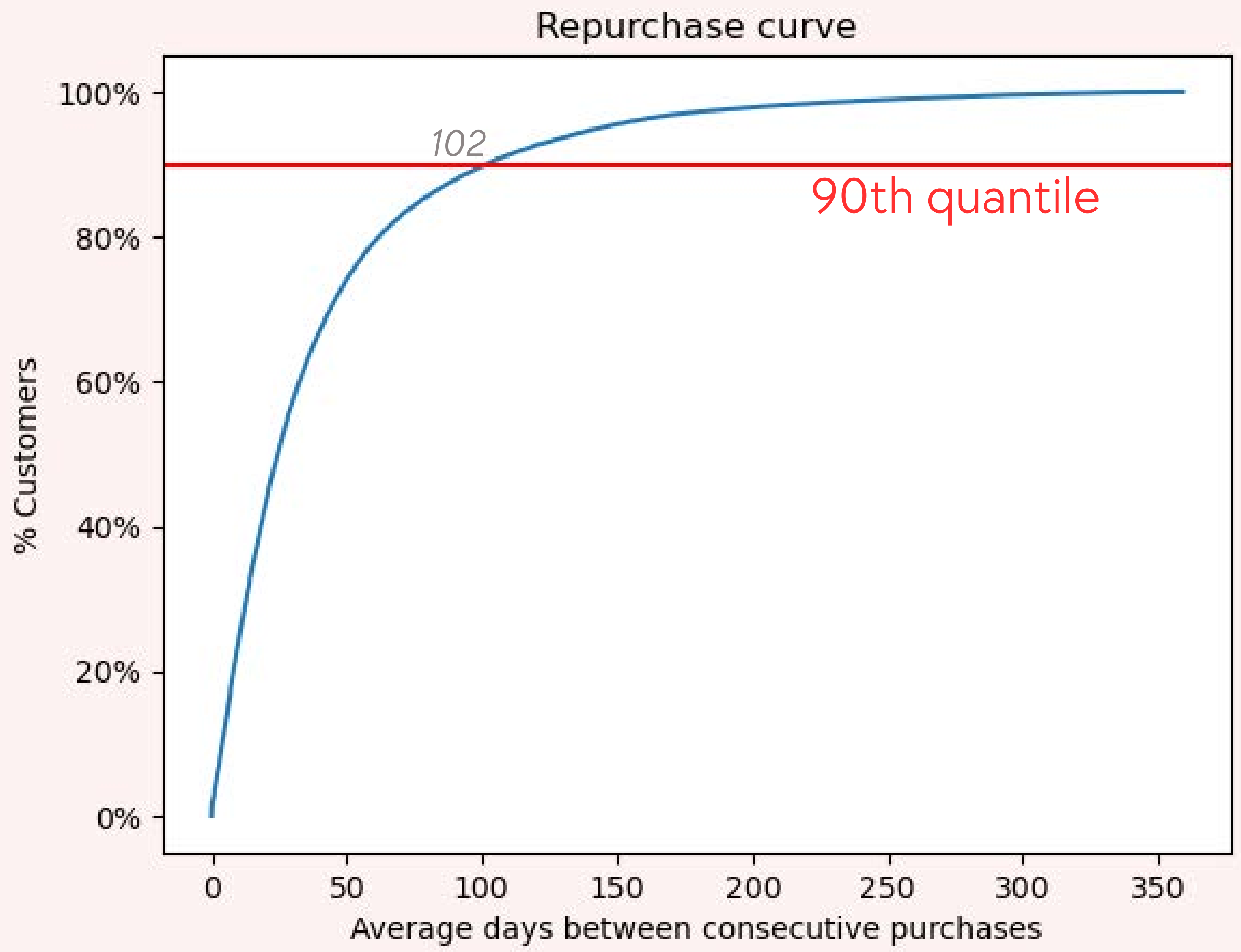
02. RFM

Repurchase curve

End date: 1st May 2023

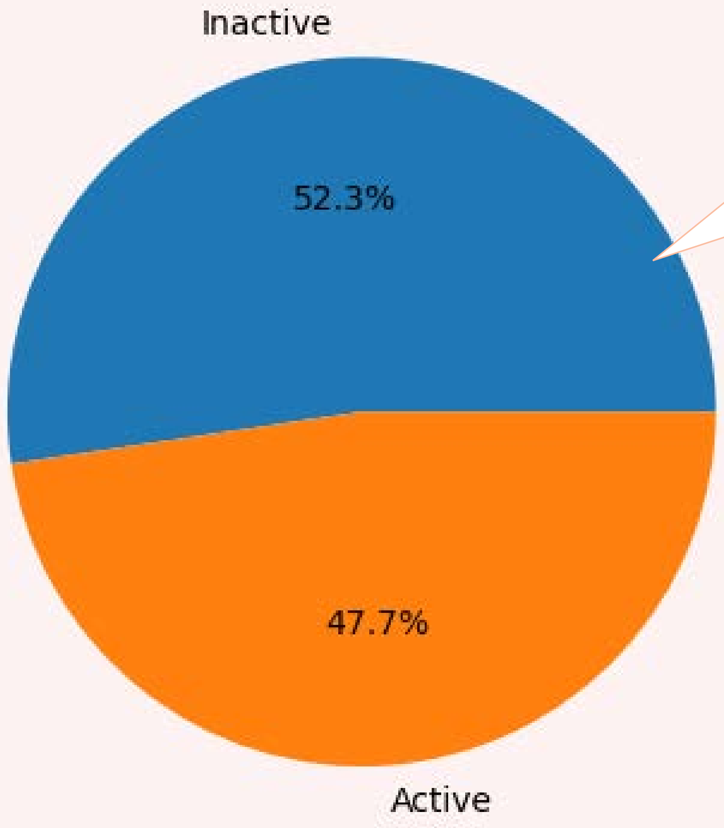
Total: 140946 customers

68.7% repeaters



90% of the repeater customers repurchases within **102 days** on average.

Therefore, we define **inactive** a customer not repurchasing within 102 days before the end date



Consider reactivation

02. RFM

RFM customer segments

RFM scores for **active** customers

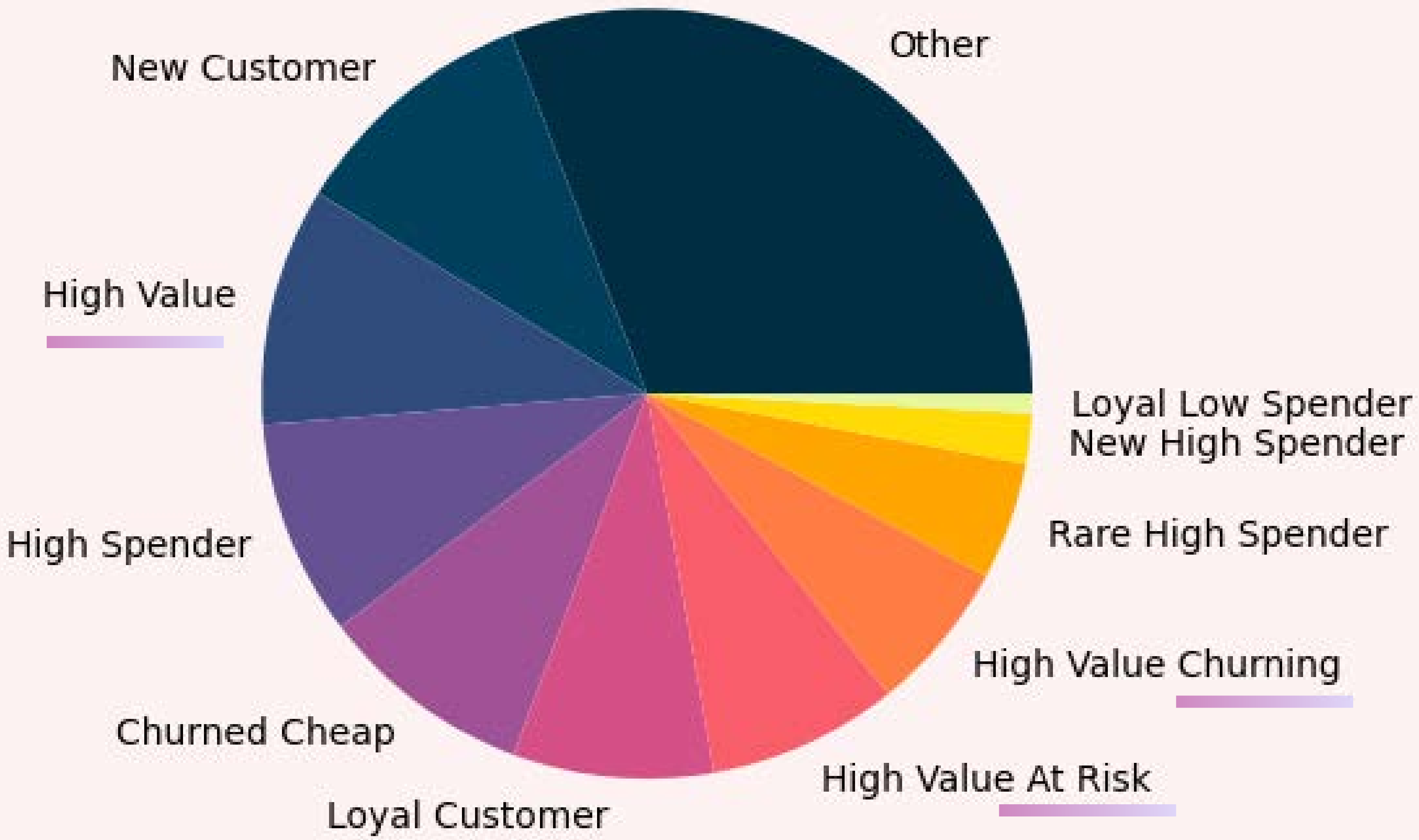
(equal-buckets quantile approach, excluding outliers)

Recency *(in days)*	**3**	< 25 ≤	**2**	< 56 ≤	**1**
Frequency *(in days)*	**1**	< 2 ≤	**2**	< 4 ≤	**3**
Monetary *(in euros)*	**1**	< 29.39 ≤	**2**	< 108.15 ≤	**3**

RFM	MAIN SEGMENTS
333	High Value
233	High Value at Risk
133	High Value Churning
313	New High Spender
213	Rare High Spender
113	RARE High Spender

02. RFM

Segments proportions



Here we see how many customers we have for each segment and what are the segment RFM values averages

segment	count	recency mean	frequency mean	monetary mean
Inactive	54509	210.2	2.5	143.7
Other	15259	47.2	2.4	40.2
New Customer	5233	12.2	1.5	26.1
High Value	4964	10.8	10.1	514.9
High Spender	4489	44.5	3.5	377.9
Churned Cheap	4478	78.1	1.3	11.0
Loyal Customer	4176	38.8	6.6	68.9
High Value At Risk	3968	39.9	9.1	498.0
High Value Churning	3157	77.5	8.6	520.6
Rare High Spender	2457	61.1	1.6	356.3
New High Spender	1037	11.9	1.7	345.1
Loyal Low Spender	407	41.1	5.5	21.1

02. RFM

Actionables

LATEST NEWS

- Recency is improving fast over time
    - Active **customer base is growing!**

- Prevalence of **small purchases**
    - ... and few big spenders!

What can we do with our segments?

- High value... *DELIGHT!*
    - No price incentives
    - Loyalty programs and updates
- High value at risk... *RETAIN!*
    - More competitive pricing *(e.g. 25%)*
    - Personalized offers and mails
- High value churning... ***RETAIN!***
    - Aggressive price incentives *(e.g. 50%)*
    - Personalized offers and mails/SMS

...in the meanwhile...

- New/Rare high spenders... *ACTIVATE!*
    - Up-selling + Cross-selling
    - Observe when they buy and be there!

(e.g. seasonal behaviours)

we want to increase their frequency to change segment!

03. Churn Prediction

Predict churners and retain them

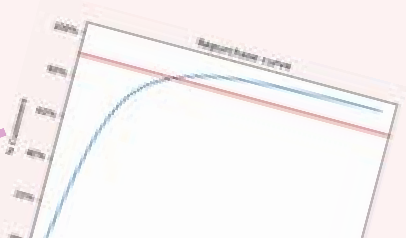
03. Churn Prediction *pipeline*

We want to be able to predict a customer that will become a churner from the reference date (19th Jan 2023)

Define a churner

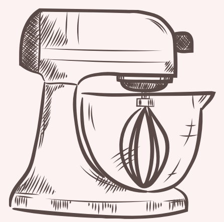
“A churner is a customer not purchasing any product within 102 days from reference date”

(see repurchase curve)



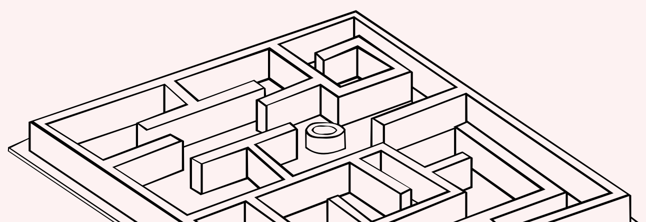
Prepare data

- Enrichment and cleaning of joined customer data
- Feature selection



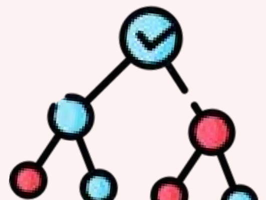
Build a model

- Use and tune different models
- Identify the most accurate



Explain the model

- Build a surrogate simpler model that explains the complex one

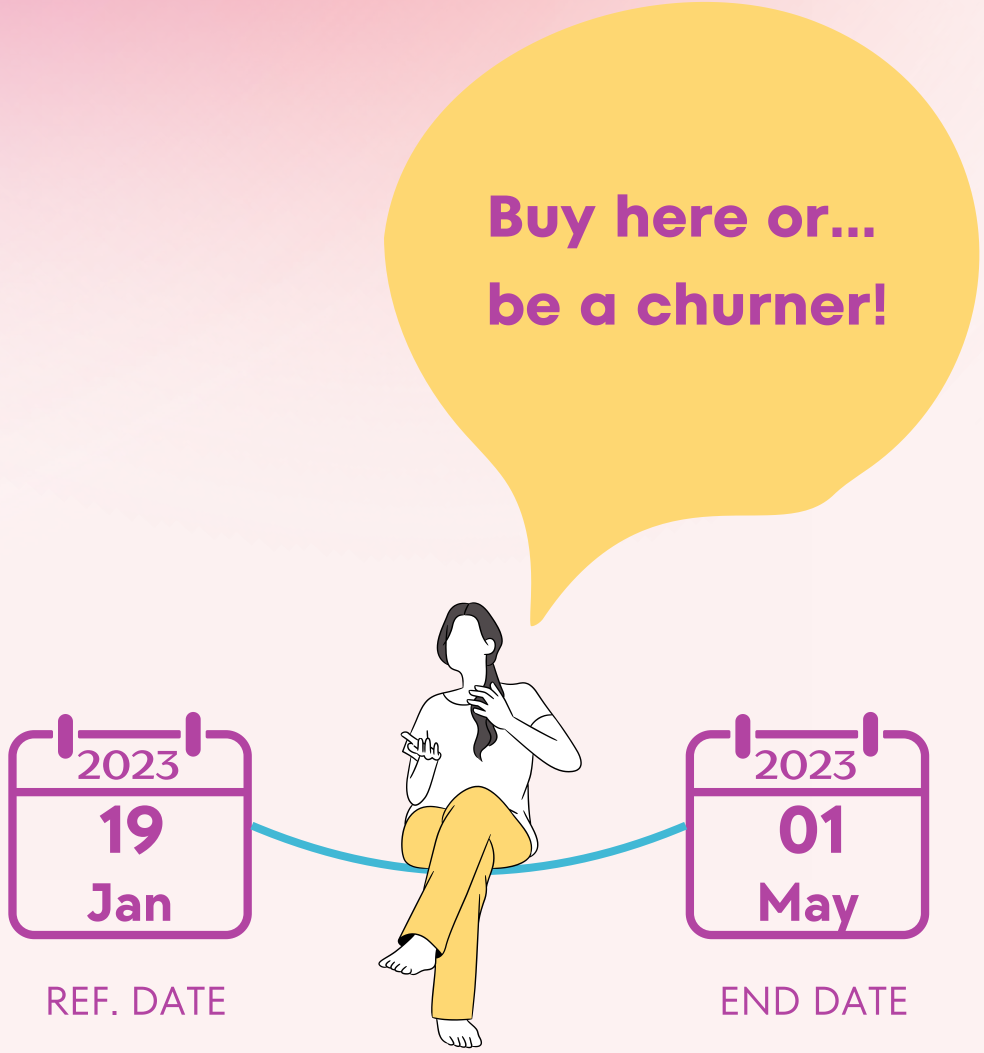


03. Churn Prediction

Joined unique customer dataset

Customer (102822 unique)

- Age, gender, job type
- Address (district, region)
- Privacy and provided phone flags
- Loyalty type and status *(standard, premium...)*
- Days from account activation to ref. date
- Most common product class and ID
- Favourite store and most common store
- Average order expense and reduction
- RFM scores and segment
- Average repurchase days
- **Days from last purchase to end date**




```
if < 120:   NO CHURN (0)
    else:   CHURN (1)
```

RESPONSE VARIABLE!

03. Churn Prediction

Feature selection

Customer (102822 unique)

• **Age**, **gender**, **job type**	“no declaration” job type removed for collinearity
• ~~Address (district, region)~~	collinearity and redundancy with common store
• **Privacy** and provided **phone** flags	
• **Loyalty type** and **status**	“standard” type removed for collinearity
• **Days from account activation** to ref. date	
• **Most common product** class ~~and ID~~	too sparse + correlated w/ product class
• ~~Favourite store and~~ **most common store**	less representative than most common store
• ~~Average order expense and reduction~~	gross expense – reduction = **net expense** NEW
• ~~RFM scores and segment~~	too high correlation w/ each other and w/ churn
• ~~Average repurchase days~~	
• ~~Days from last purchase to ref. date~~	used to compute response variable **churn**

03. Churn Prediction

Feature selection

Final Customer entity

- *Churn (response)*
- Gender
- Age
- Job Type
- Provided phone
- Privacy
- Loyalty type
- Loyalty status
- Account activation days
- Common store
- Class of most common product
- Average Net Expense

How do we choose the best model?

"in churn prediction, the cost of losing a customer that could have been retained is generally much higher than the cost of an unnecessary retention activity"

So we'll give *recall* more importance than *precision* *by computing* ***F2-Score****, a trade-off metric between precision and recall that gives more weight to the latter*

Predicted

Actual	0	1
0	TN	FP
1	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Main metrics: F2-Score, Accuracy

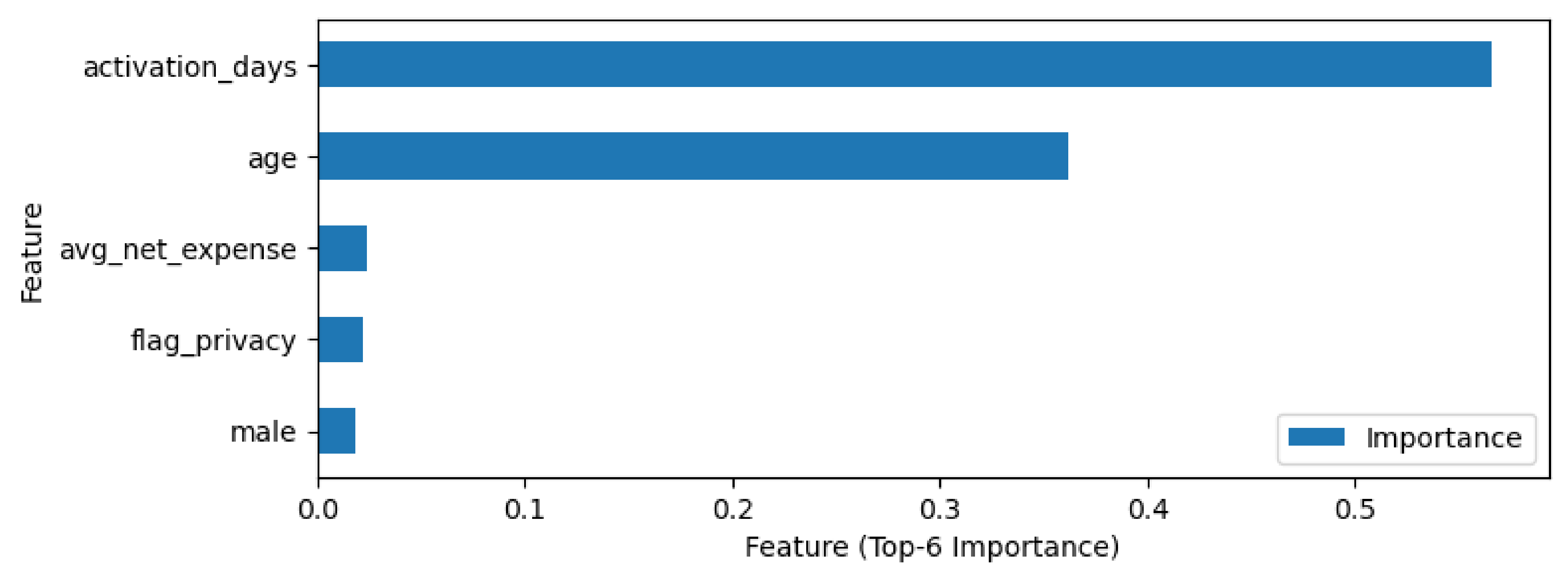
03. Models performances

Validation set metrics

Classifier	Accuracy	F2-Score	Precision	Recall	F1-Score	AUC
Logistic Regression	*70%*	*75%*	*69%*	*77%*	*73%*	*70%*
Decision Tree *(pruned)*	74%	87%	68%	93%	79%	73%
Random Forest *max depth 7*	**74%**	**87%**	**68%**	**93%**	**79%**	**73%**
MLP	73%	84%	69%	89%	78%	73%

03. Model explainability

Random Forest is the best model! But it's a black box hard to explain...

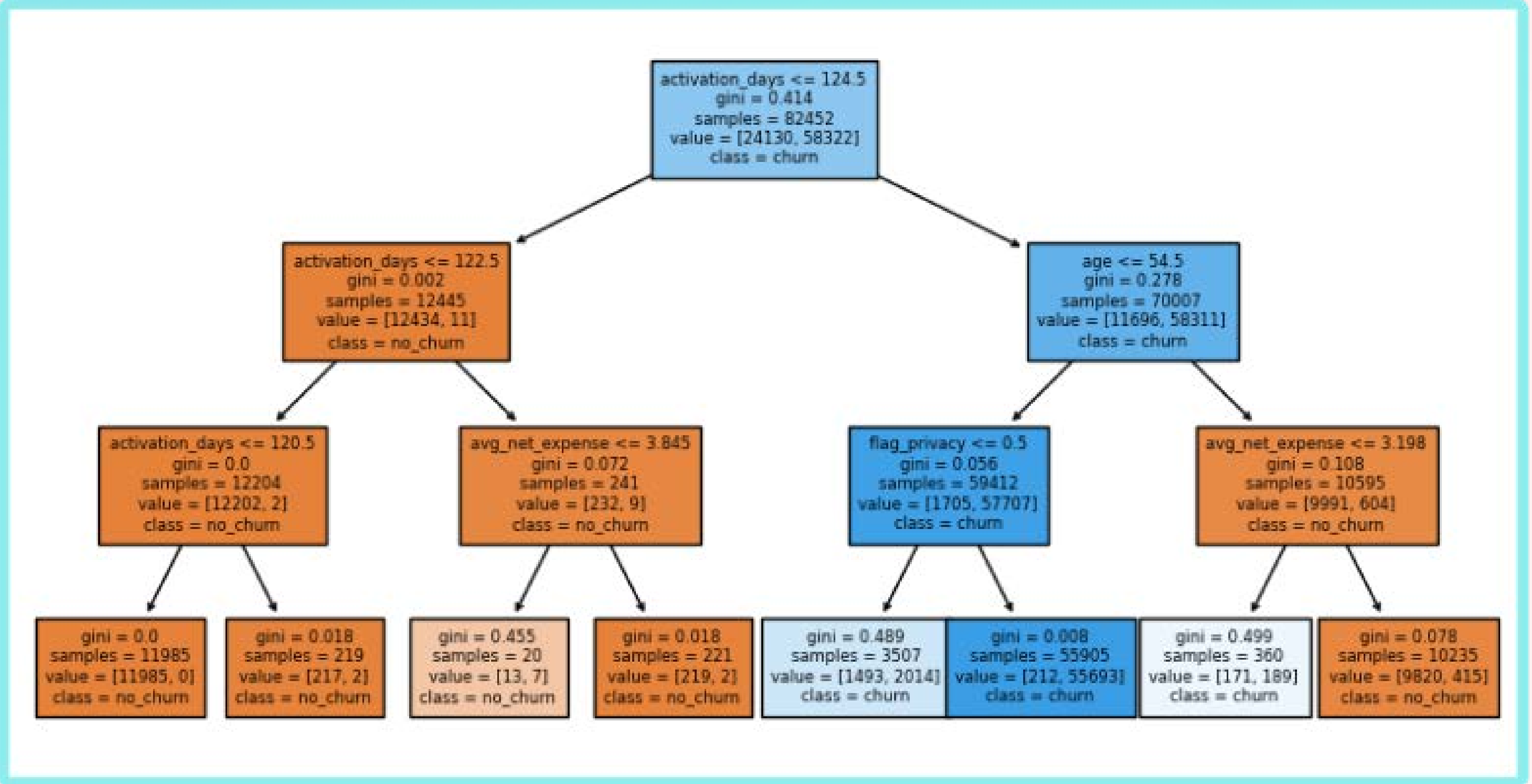


We only know these are the most important variables for prediction

Let's build a **surrogate** simple **decision tree** that fits the Random Forest predictions.

03. Surrogate Decision Tree

This 4-layers pruned decision tree is much easier to interpret...
... but also describes 97.2% (R^2) of the Random Forest predictions variance!



We can say this Random Forest instance tends to predict as churners customers that are younger than 54 y.o. and have activated their accounts more than 124 days before reference date

Now we understand the predictions better and have a potential churner persona to retain!

04. Sentiment Analysis

Find insights in reviews, identify promoters and detractors

04. Sentiment Analysis

We want to find what works and what can be improved for users.
We want to delight **promoters** and reach out for **detractors**.

Pre-Processing

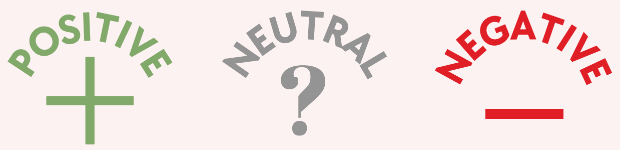
Convert text into pre-processed tokens *(already seen in Sec. 01)*

Text Representation

Make text machine-readable by using a numerical representation

Build a model

Find the best model and representation to classify reviews into:



Extract insights

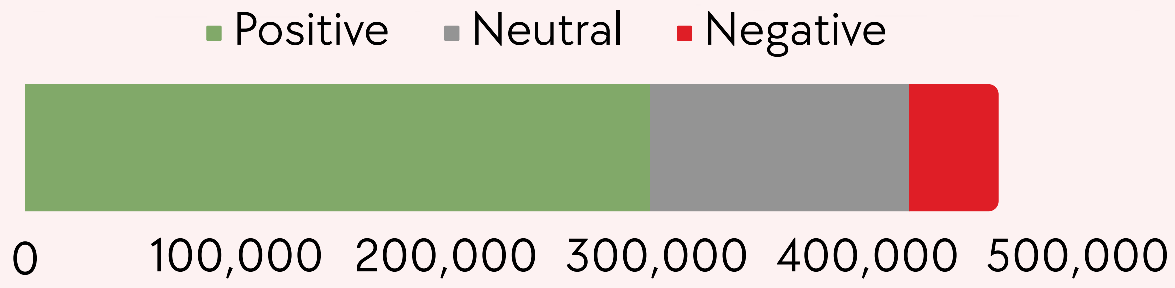
- Find hot positive or negative topics by common words
- Identify promoters and detractors

04. Sentiment Analysis

Text Representations and Models

Train set:

Preprocessed **Labelled Reviews**



- We balance classes with random oversampling of the under-represented classes
- We do a 80-20 train-validation split with stratified sampling

Test set:

Real **Customer Reviews**

Text Representations

max 5000 features

Bag of Words
Each word in a document is represented by its frequency

TF-IDF
Each word frequency is weighted by its uniqueness in the corpus

BoW and TF-IDF with Bigrams
Mantain positional information

Models used

Logistic Regression
Non-linear regression

Linear SVC
Support Vector Classifier

Random Forest
Bagging on decision trees

XGBoost
Boosting on decision trees

Multinomial NB
Naive Bayes classifier

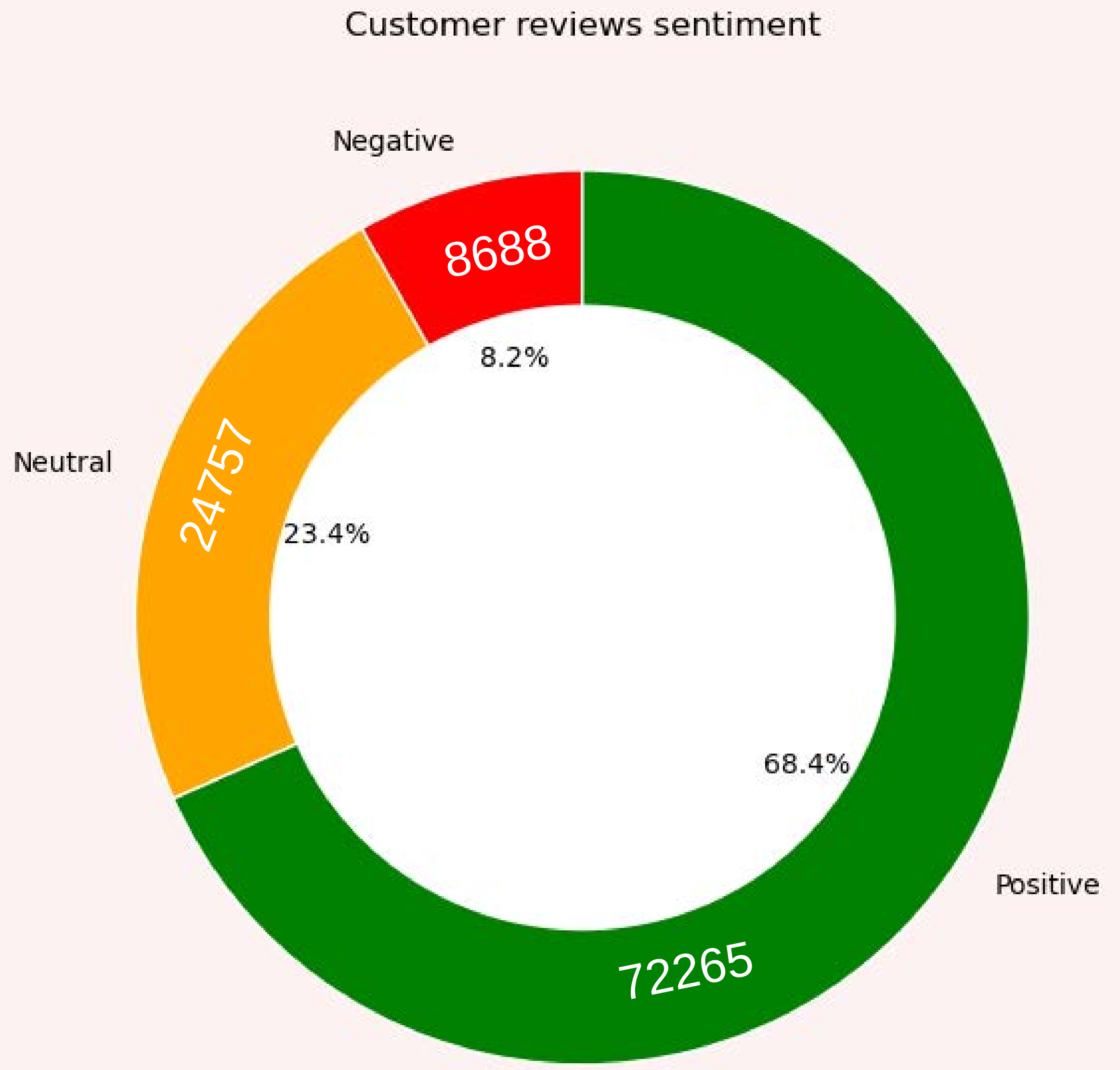
04. Models performances

Validation set metrics (averages weighted on class size)

	Bag of Words		TFIDF		BoW w/ Bigrams		TFIDF w/ Bigrams	
Classifier	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
Logistic Regression	*71%*	*72%*	*71%*	*72%*	*71%*	*72%*	*71%*	*72%*
Linear SVC	72%	72%	71%	72%	72%	72%	71%	72%
Random Forest	83%	82%	83%	82%	83%	82%	**83%**	**82%**
XGBoost	69%	70%	69%	71%	69%	70%	69%	71%
Multinom. NB	68%	69%	68%	69%	67%	69%	67%	69%

04. Best model results

Random Forest on TFIDF with Bigrams has a 83% accuracy!



POSITIVE CLASS
promoters

NEGATIVE CLASS
detractors

Most common words
(excluding too common and generic words shared by both classes)

Most common bigrams

04. Some detected insights

<u>From the most common words</u>

- **Positive reviews** seem more **focused on the product**. Beverages are particularly mentioned
- **Negative reviews** focus more on the **company** and customer **experience** (shipping, quality...)

<u>From the most common bigrams</u>

- **Positive reviews** appreciate products **taste** and the fact they are **gluten-free**, given that they are **cheaper** with respect to their local stores
- **Negative reviews** complain about **unhealthy**, **expired** and **"made in China"** products.
  - Peanut butter is particularly mentioned mainly due to *poor artificial taste and texture*, and receiving a different product from the advertised one (*wrong flavour or product, damaged package*)

<u>From promoters and detractors customer profiles</u>

- Higher % of **self-employed** customers in **promoters** with respect to detractors
- Higher % of **retired** customers in **detractors** with respect to promoters
- Products 33120913 and 48020504 more **commonly purchased by promoters**
- Products 35209202 and 35761670 more **commonly purchased by detractors**

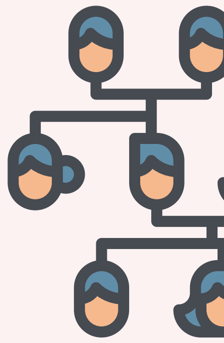
To sum up...

Let's get some actionables from our analysis

My campaign

We used RFM analysis to cluster customers into segments

- We have 4964 **high value customers** to delight and involve in **loyalty programs**
  - "You've been with us for X years!": fidelty card, VIP exclusive products, try new products previews!
- We have 3157 high value customers that are churning (and 3968 at risk)
  - **At risk**: "We haven't seen you for a while", abandoned cart mail reminder, 25% coupon
  - **Churning**: "Something wrong? Give us your feedback!", limited time 50% discount (also SMS)
- We have 2457 rare **high spenders** and 1037 new **high spenders** to activate with **up-selling** and **cross-selling**
  - **New**: "Bring 3 friends for a coupon", "Reach 200 points and we'll send you a gold card!", products updates
  - **Rare**: "Check out our new product!", fidelty points, be there when seasonal buyers usually activate

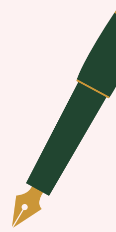


We used Churn prediction to predict potential churners

- We can use the model to **predict churn** on new users: **retaining** customers is easier than getting new ones!
- In the chosen period, attention should be paid for customers younger than 54 y.o. that have activated their accounts more than 124 days before reference date: the model seems to mainly classify them as churners

We used Sentiment analysis to find hot topics in reviews

- Detractors suggest improving customer service, shipping, ingredients quality and provenance
  - We could reach out to understand what's wrong, hoping to make users change their review in better
- We could invite promoters to write reviews, since they have a great impact on new customers



Questions? Reach out

Email

l.galli40@campus.unimib.it

Student ID

905236