PREDICTION OF CUSTOMER REPURCHASES

Group K: Xuran HUANG
Yun RU
Jingxin FU

CONTENT

Objective of project

❑ **Predict** if a customer is going to repurchase on the next month(09/2020) by using machine learning.

❑ **Make recommendations** to the marketing department for its next promotional email campaign based on prediction result.

Data preparation & Features selection

❑ **Convert** transaction_date & card_subscription to date format

❑ **Delete** Max(multicard) = 21 since multicard is a binary variable.

❑ **Delete** card_subscription = Republique Demogratique".

⚠️ These are the import issues that occurred when the database was created.
→ **Can be avoided.**

Data preparation & Features selection

We also found item_count, gross_amount, discount_amount, basket_value

which are **extremely large or negative**.

→ **Keep** in order to retain more information in train set.

Number: 127571 Clients.

Date: 01/08/2019 – 31/07/2020

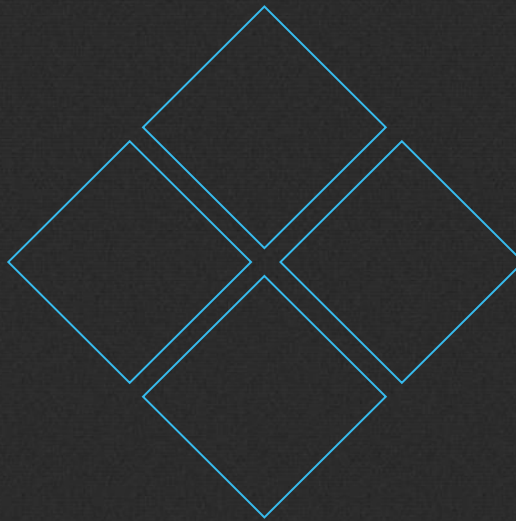# Data preparation & Features selection

Recency

days since the last purchase.

Frequency

- number of purchases over one year

- Frequency by months

(09/2019 – 07/2020)

Monetary value

- cumulative purchase amount

- Cumulative item count

Payment gift

Number of payments with gift cards

Data preparation & Features selection

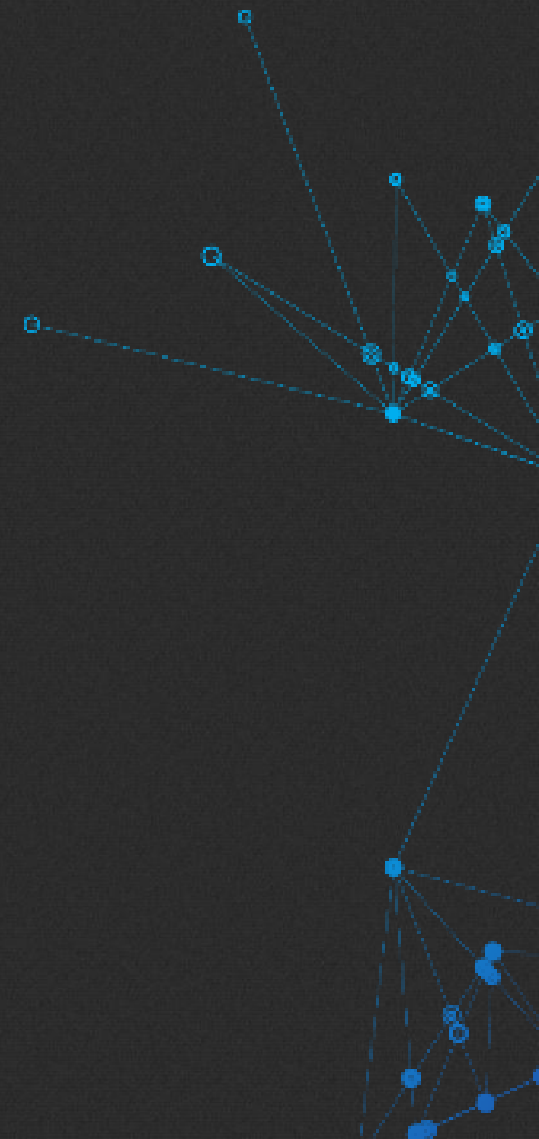Mean values for features according to repurchase

Repurchase	Cumulative amount (euro)	Item count	Payment gift	Number of purchase for whole year	Recency (days)
0	824	276	1	17	138
1	2734	960	5	64	46

- Because we need to predict the repurchase for next month so we also added the purchase frequency by month to have more precise information.

- We found that there was an obvious difference in every month frequency.

Algorithm of models

❑Logistic regression (Lasso,Ridge)

❑Bagging & Random forest
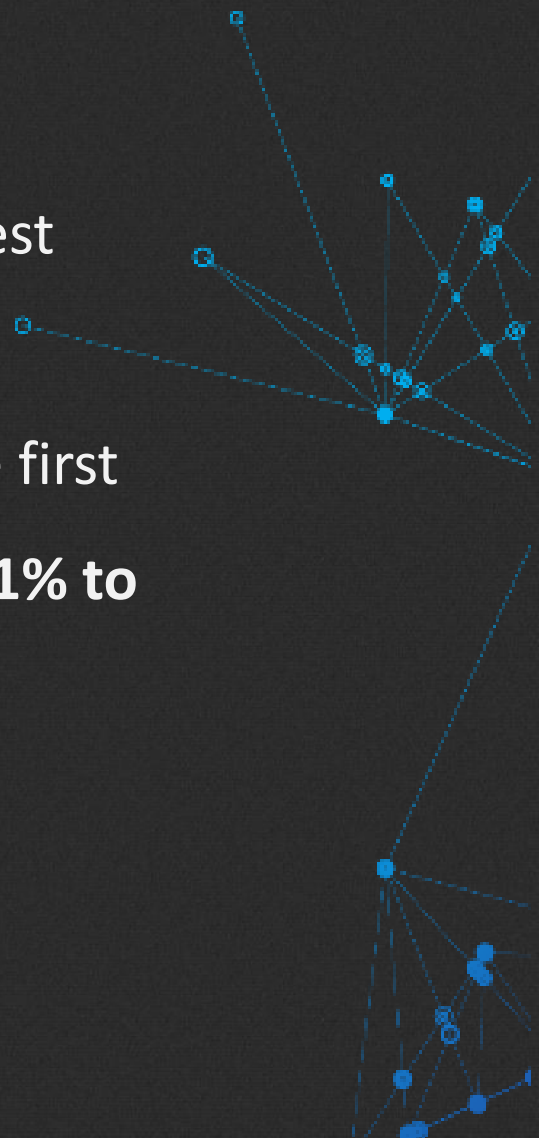
❑Boosting (Gradient boost,Lightgboost,**Xgboost**)

Algorithm of models

❏ "Xgboost" is one of the most powerful machine learning tools available for tabulated data.

❏ Regularization: XGBoost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting.

❏ Handling Missing Values: XGBoost has an in-built capability to handle missing values.

❏ Based on an article named *Completed Guide to Parameter Tuning in XGBoost with codes in Python*, we tuned the hyperparameters using GridsearchCV().

❏ 'max_depth': 6, 'min_child_weight': 6, 'n_estimators': 100

Logloss: 0.19376

Analysis & Recommendations

❑ We choose the first **10%** (4010) clients who have the highest repurchase probabilities to analyze their features.

❑ There are **2549 clients** who have email address among the first 10% clients and their repurchase probabilities vary from **21% to 93%**.

❑ These 2549 clients are our target clients for the next email marketing campaign.

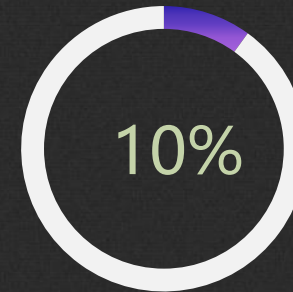# Analysis & Recommendations

63%
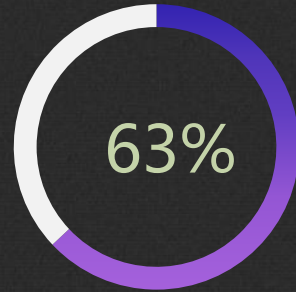
Woman

63% of these clients are women, we can use colorful background in the email and promote some new products which are suitable for women to attract female clients.
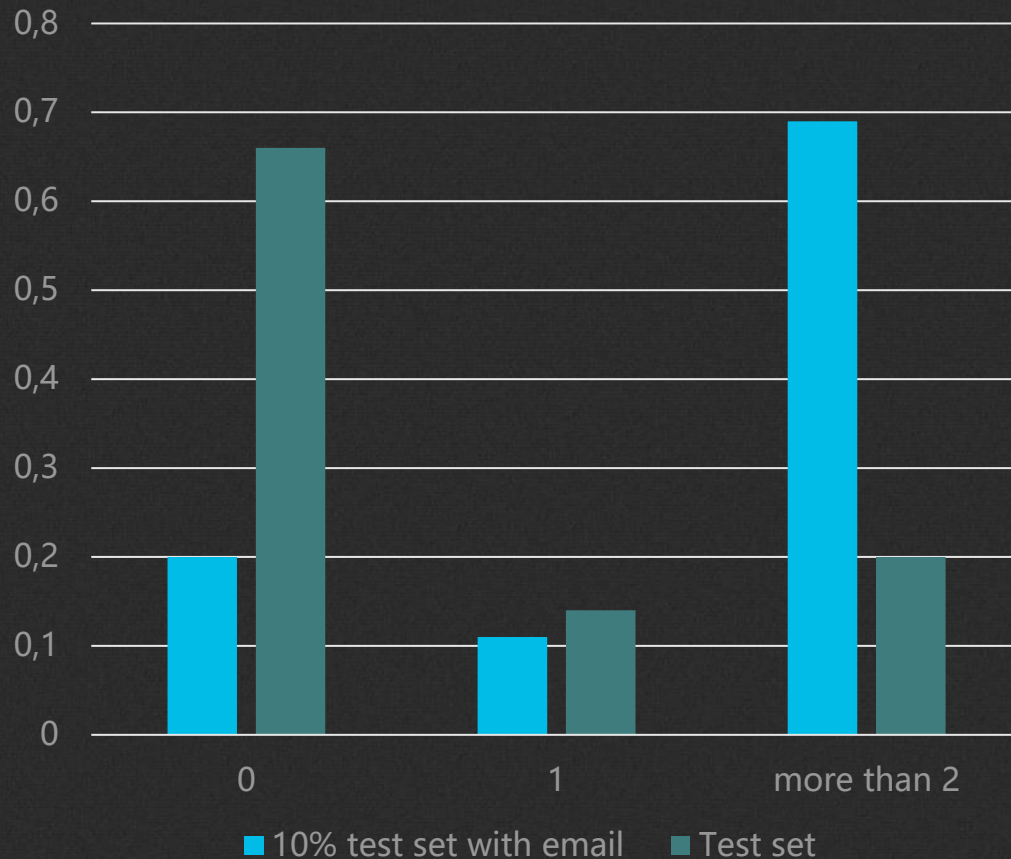
Only 10% of clients have subscribed the card among our target clients for email campaign. We can include membership card introduction and the benefits of possessing a membership card in the email.

10%

Possession of Card

Analysis & Recommendations

Distribution of number of payment by gift card

Target clients have more payments by gift card then others do. We can deduce that gift card is one of their repurchase reasons. We may offer them for example a 10 euro gift card for 55 euro. (On average they spend 47 euro each time)

Potential improvement

Neural network

Try to use neural network **with variable frequency by months.**

Hyper-parameter optimization

- **GridsearchCV** costs a long time.
- **Bayesian Optimization.**

Thank you!