Authors:

Camille CHARREAUX
Marc GAUDART
Paul PEYRÉ

Unleashing the power of Open Banking data for affordability assessment

2021/08/02

Transactions Clustering

A new approach to understanding consumers' personal budget configurations & financial habits with unsupervised learning

algoan

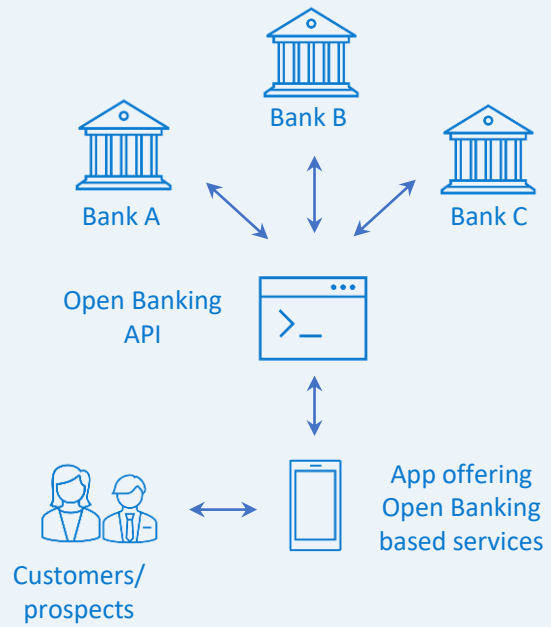
Agenda

1. The new Open Banking paradigm
2. Problem statement:
   Assessing affordability with Open Banking data
3. Our approach:
   Transactions Clustering
4. Results summary

The new Open Banking paradigm

The Open Banking era

Open Banking is designed
to drive greater transparency, security,
innovation and market competition.



87%

of the countries have some form of Open Banking API

Source: The Global Open Banking Report 2019

> 470

third-party providers registered with a National Competent Authority in Europe

Source: Vocalink – Q1 2021 Open Banking tracker

800m

Open Banking API calls each month in the UK

Source: Open Banking Implementation Entity (OBIE)

Open Banking is backed by the latest technological and regulatory developments, e.g. the EU Revised Payment Services Directive (PSD2), or the UK's Open Banking national regulation.

A revolution for credit

A major opportunity for credit risk assessment

- Standard risk analysis methods mostly rely on socio-demographic data and past credit history
- Open Banking channels recent and trustworthy data about the borrower's budget and financial behavior (bank transactions and balances from the last 3 to 12 months)

⇒ Open Banking is the key to enhance risk analysis significantly

⇒ **Credit risk assessment is the top use case for Open Banking in most European countries (Tink)**

For the end-customer

Shorter "time-to-yes"

Improved customer journey

Better alignment between amount granted and needs/affordability

Fairer decision / better access to credit

For the lender

Less costly processes

Higher acceptance rate

More accurate affordability assessment

Better risk management

Several lending use cases emerging

	Consumer loans origination	Buy Now Pay Later	Mortgage loans origination	Debt collection
UX / client interactions improvement	• Less questions • Pay slip replaced		• Less questions • Pay slip replaced • Bank statements replaced	• Bank statements replaced
Processes automation	• Incomes validation simplified • Automatic affordability assessment		• Automated bank data analysis • Automatic affordability calculation	• Automatic affordability calculation
New data / better decisioning	• Regular expenses verification • Credit history • Behavioral scoring	• Affordability assessment for a more responsible lending • Credit history • Behavioral scoring		• Capacity to obtain payment commitments consistent with debtor's financial situation

Concept of affordability

According to the FCA handbook CONC 5.2A (UK)

"Affordability risk: risk to the customer of not being able to make repayments without failing to make any other payment the customer has a contractual or statutory obligation to make or of these having a significant negative effect on their overall financial situation."

The concept of affordability is broadly applicable despite local specifities

⇒ Consequence on income assessment

"Not limited to salary and wages, it can include income from savings, or income from another person (such as where household finances are pooled)"

Note: "It is not generally sufficient to rely solely on a statement of current income made by the customer without independent evidence.''

⇒ Consequence on expenditure assessment

"Non-discretionary expenditure […] includes payments needed to meet priority debts and other essential living expenses […]. It also includes payments the customer has a contractual or statutory obligation to make […].''

Regulators are keen for lenders to have a detailed assessment affordability using a broad range of data

Problem statement: Assessing affordability with Open Banking statement

Business objective

Assessing affordability

Problem

Define all relevant incomes and their regularity, and the same for expenses, to make sure the applicant has the financial capacity to cope with servicing the new debt.

Scope

Rely exclusively on Open Banking data, i.e. the credit applicant bank transactions history.

Overall data processing

1. Collect ingoing and outgoing transactions
2. Group transactions to highlight recurrences and/or regularities
3. Categorise transactions

   either with a fine category, such as groceries, petrol, loan repayment, etc. or a coarse one such as essential vs. non-essential
4. Calculate residual income (or repayment capacity) based on relevant transactions' groups and categories

Technical challenge

State of the art

The use of bank transaction data to assess affordability is a subject that has been widely discussed in recent years.

This subject is commonly dealt with through categorisation, where revenues and expenses are identified by reading and interpreting the transactions descriptions.

Some related articles

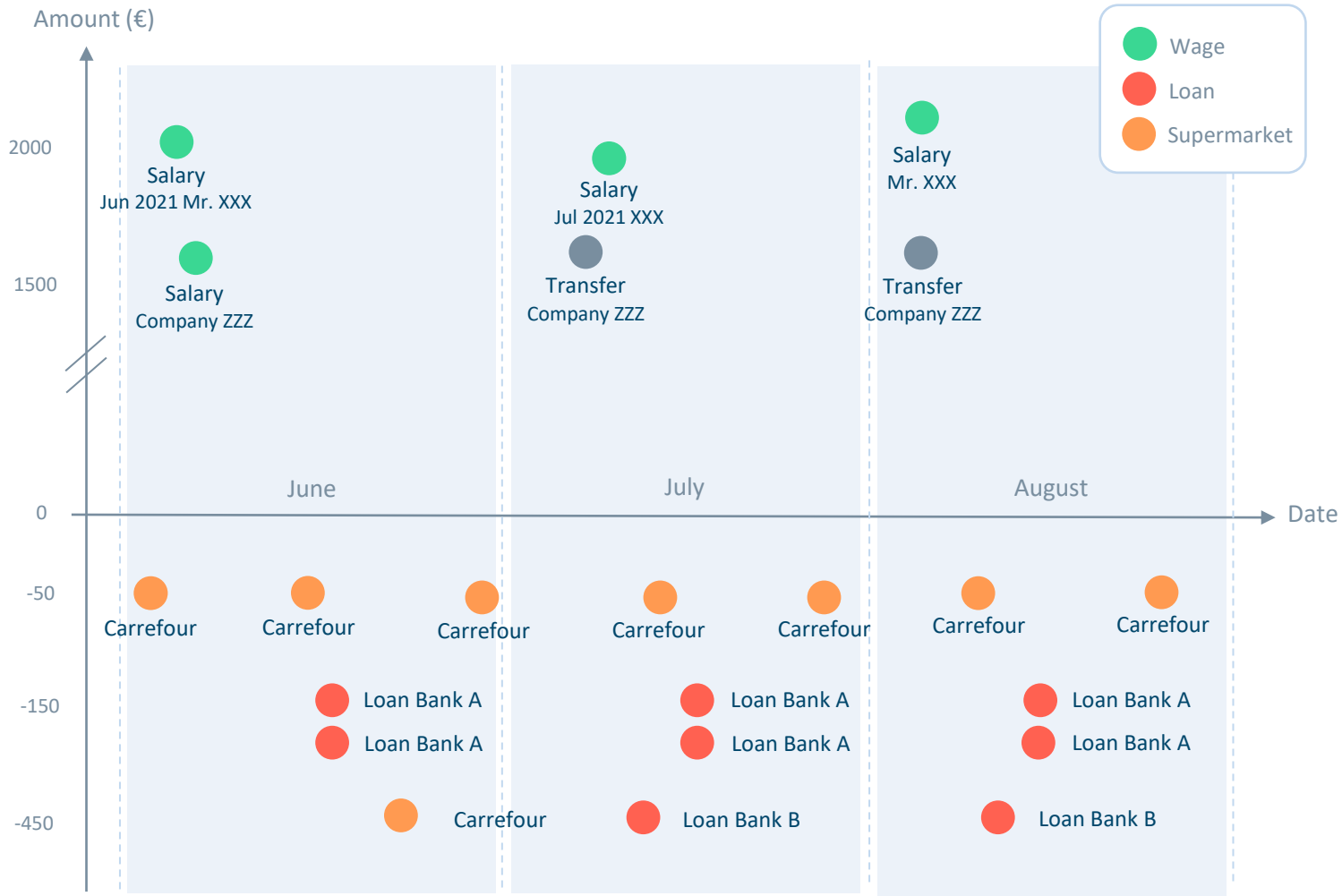
[1] S. GARCÍA-MÉNDEZ et al. Identifying Banking Transaction Descriptions via SVM Short-Text Classification. April 2020.
[2] Olav Eirik Ek Folkestad Erlend, Emil Nøtsund Vollset. Automatic Classification of Bank Transactions. June 2017

Limits of the categorisation approach

- Transactions' descriptions are **not always self-explicit**.
- Categorisation itself generally **does not tell much about the regularity** of a transaction.
- Whilst the above issues can be successfully mitigated by specific approaches (e.g. consensus-based annotation, inclusion of some element of recurrence related to a category, etc.), these can be **manpower, time and/or data hungry**.
- A categorisation engine **treats each transaction separately** and often doesn't use the context, i.e. the other transactions surrounding it.
- The categorisation does not easily cross borders: **a categorisation engine is country and language specific** and cannot be generalised.

Illustrating the limits of categorization (1/2)

Date	Amount	Description
2021-06-01	-50,12	Carrefour
2021-06-02	2048,76	Salary Jun 2021 Mr. XXX
2021-06-03	1554,24	Salary Company ZZZ
2021-06-14	-48,26	Carrefour
2021-06-15	-146,78	Loan Bank A
2021-06-15	-189,23	Loan Bank A
2021-06-18	-449,99	Carrefour
2021-06-28	-49,37	Carrefour
2021-07-03	1572,58	Transfer Company ZZZ
2021-07-08	1998,25	Salary Jul2021 XXX
2021-07-10	-445,34	Loan Bank B
2021-07-12	-43,98	Carrefour
2021-07-15	-146,78	Loan Bank A
2021-07-15	-189,23	Loan Bank A
2021-07-27	-55,34	Carrefour
2021-08-02	2048,76	Salary Mr. XXX
2021-08-02	1572,58	Transfer Company ZZZ
2021-08-08	-50,86	Carrefour
2021-08-10	-445,34	Loan Bank B
2021-08-15	-146,78	Loan Bank A
2021-08-15	-189,23	Loan Bank A
2021-08-27	-52,35	Carrefour



Illustrating the limits of categorization (2/2)

Salaries identification

Categorisation on its own does not work well when explicit keywords are not embedded in the description.

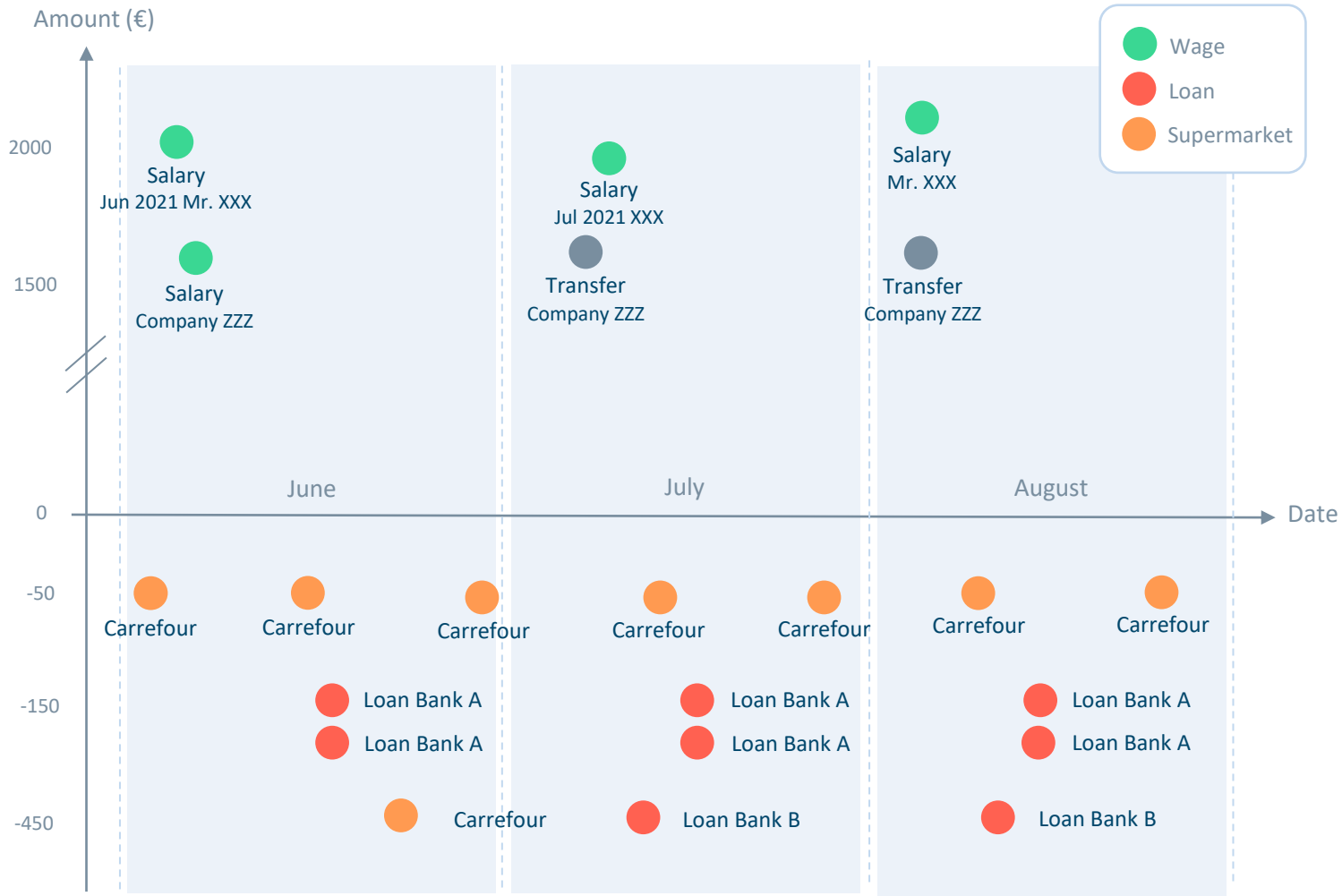
Categorisation on its own cannot separate multiple salaries.

Multiple loans identification

Categorisation on its own cannot distinguish between multiple loans.

Regular grocery expenses estimation

Categorisation on its own cannot assess whether a grocery expense is regular or not.



Our approach:
Transactions Clustering

Methodology: Transactions Clustering (TC)

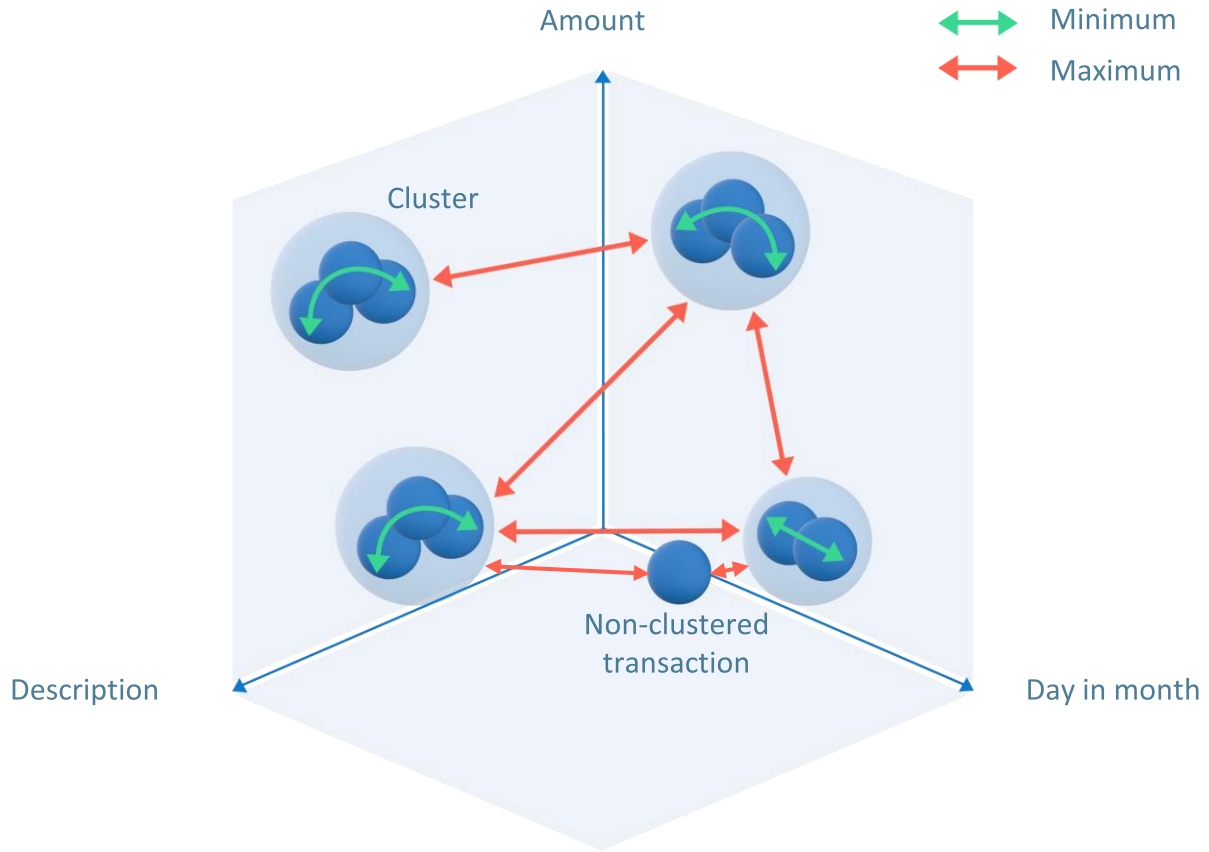
‘’Transactions Clustering’’ (TC)

Rather than looking at bank transaction data as a sequential series of payments, we project the transactions into a multi-dimensional space with the following dimensions:

(amount, day in month, description)

… and use clustering methods to highlight relationships between transactions.

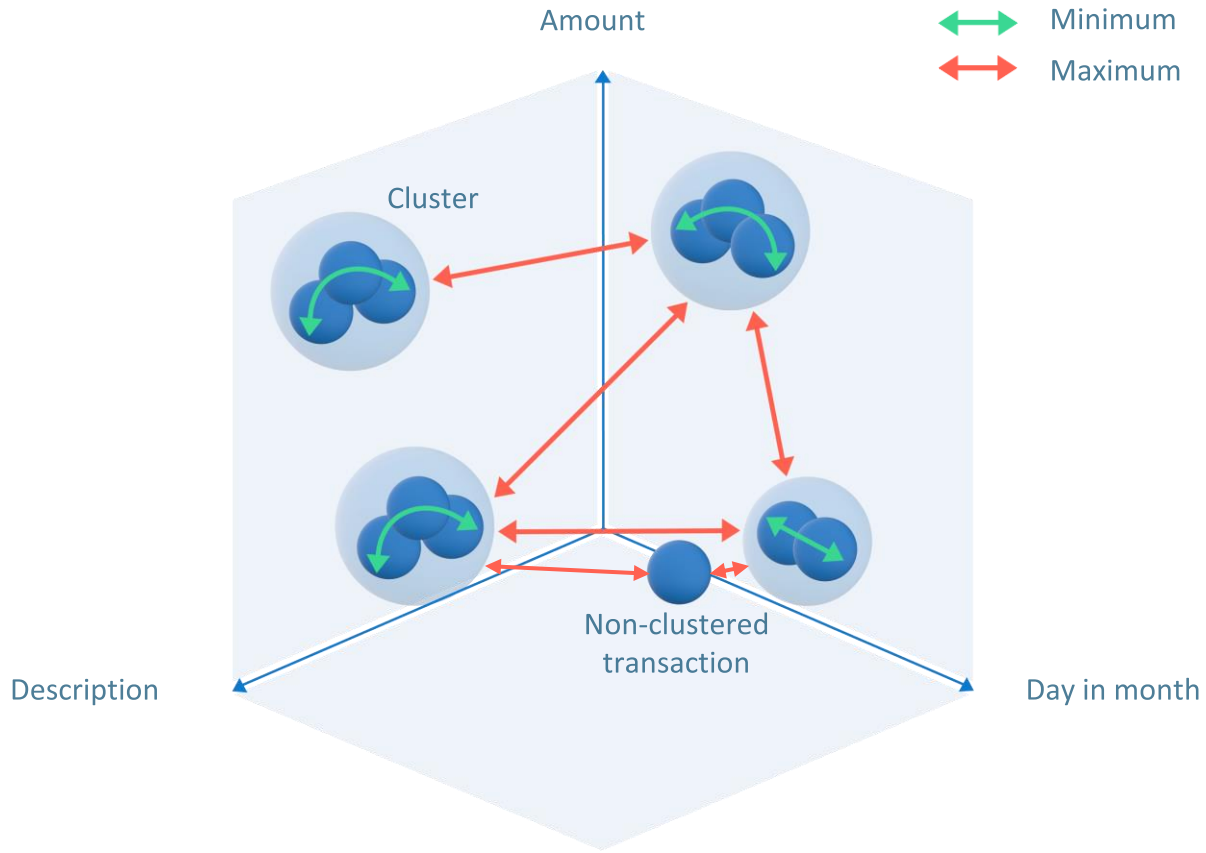
Projection of transactions for TC



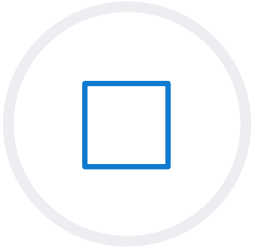
Critical issues we encountered

1. Combining heterogeneous data
2. Choosing the right distances

Projection of transactions for TC



Algorithmic variations on TC

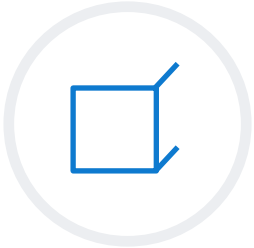


2D TC

Leveraging only 2 dimensions: amounts and dates.

Does not use transactions description.

→ **Useful when dealing with anonymous data** (i.e. when transactions description have been anonymised)

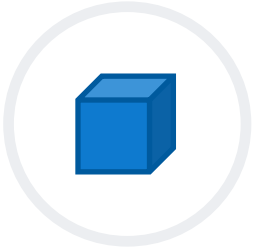


Combined TC (live)

(1) 2D TC and description-based clustering are run separately.

(2) Clusters are merged using expert rules taking into account category and regularity considerations

→ **Allows for more accuracy**, but expert rules may lead to robustness issues



3D TC (in progress)

Leveraging all 3 dimensions (amounts, dates and descriptions) with one single embedding and distance.

→ **Optimal in terms of accuracy and robustness,** but difficult to calibrate

Illustrating Combined TC

Salaries identification

We have been able to consistently group transactions even when descriptions, amounts and dates vary.

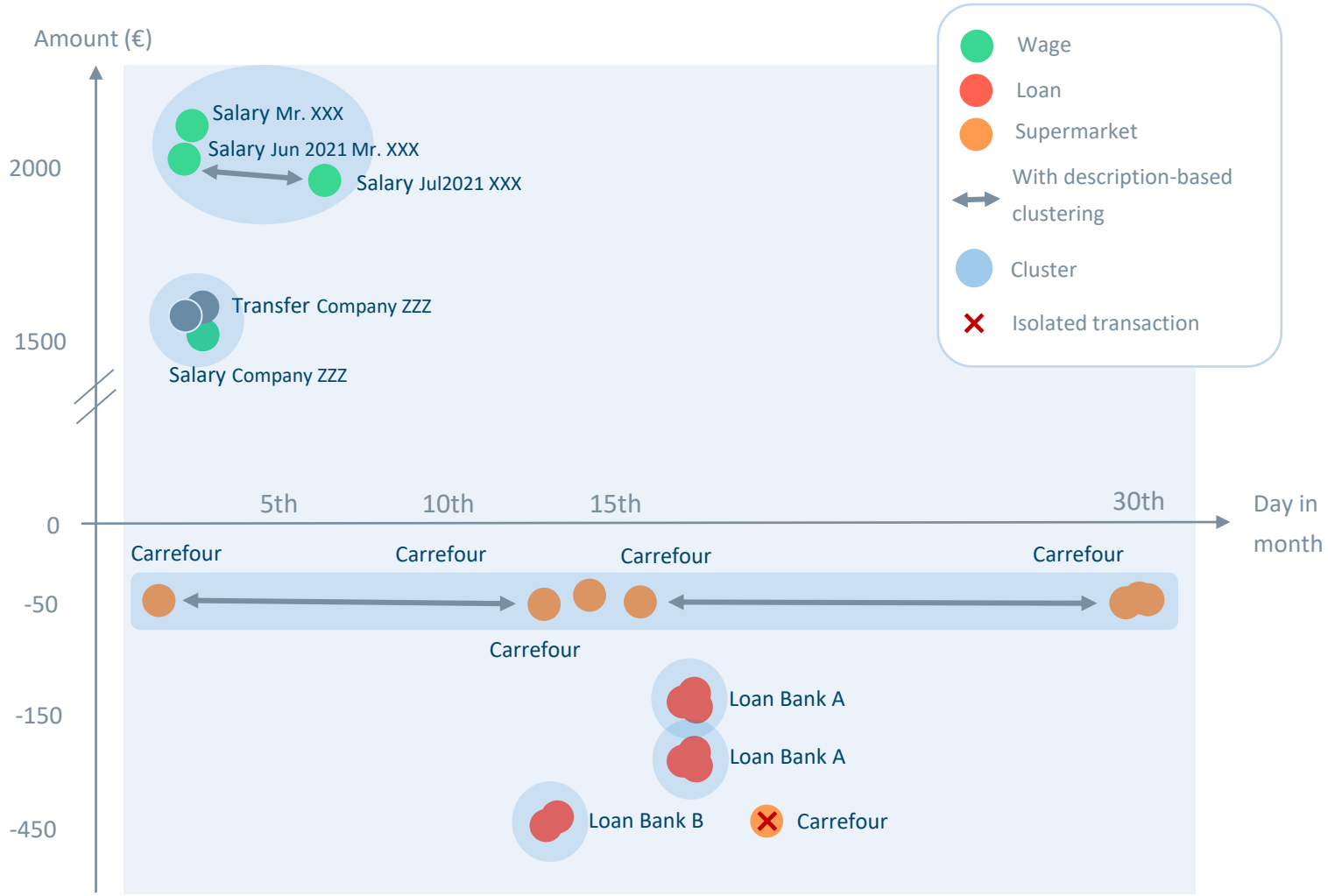
We have been able to detect 2 salaries.

Multiple loans identification

We have been able to detect 3 different loans, 2 distinct loans from Bank A and 1 loan from Bank B.

Regular grocery expenses estimation

We have been able to distinguish a one-off expenditure from regular essential grocery expenses.



Distances examples

For amounts and dates

- Euclideean
- Mahalanobis

For transactions descriptions

We tested the following distances:

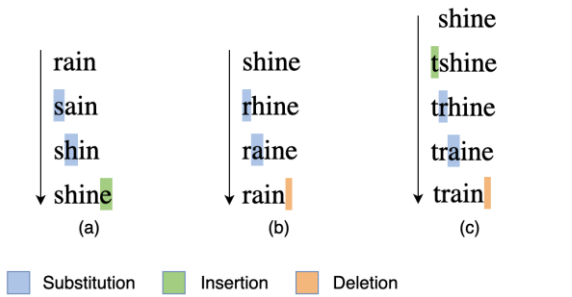
- Hamming distance
- Levenstein distance
- Jaro-Winkler distance
- Cosine similarity
- Ratcliff-Obershelp similarity

Most relevant distances

We selected the Levenstein distance and the Jaro-Winkler distance for descriptions.

Levenstein similarity

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}(\operatorname{tail}(a), b) \\ \operatorname{lev}(a, \operatorname{tail}(b)) \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$



Jaro-Winkler similarity

The Jaro distance d_j of two given strings s_1 and s_2 is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

where:

- m is the number of *matching characters* (see below);
- t is half the number of *transpositions* (see below).

"martha" and "marhta".

```
dw = 0,944 + ( (0,1*3)(1-0,944)) = 0,944 + 0,3*0,056 = 0,961

Jaro-Winkler distance = 96,1%
```

Results and conclusion

How to measure performance

Target

We have chosen a clear business target, measuring the accuracy* for the following classification problem:

Is the credit applicant residual income** above a minimum threshold***?

Household composition	Threshold
Single	550€
Single with one child	826€
Single with 2 children	986€
Single with n > 2 children	986€ + (n-2) * 220€
Couple	826€
Couple with one child	991€
Couple with 2 children	1156€
Couple with n > 2 children	1156€ + (n-2) * 220€

Minimum residual income for basic expenditures (French Central Bank)

* Based on comparison with the result obtained by a human underwriter
** Calculated as the difference between all regular inflows and wages and all regular outflows (including contractual obligations and priority debts)
*** Minimum budget for basic expenditures as established by the French Central Bank (see table on the right)

Data

We have used a dataset of Open Banking data related to 7m transactions related to French credit applications collected between 2019 and 2021

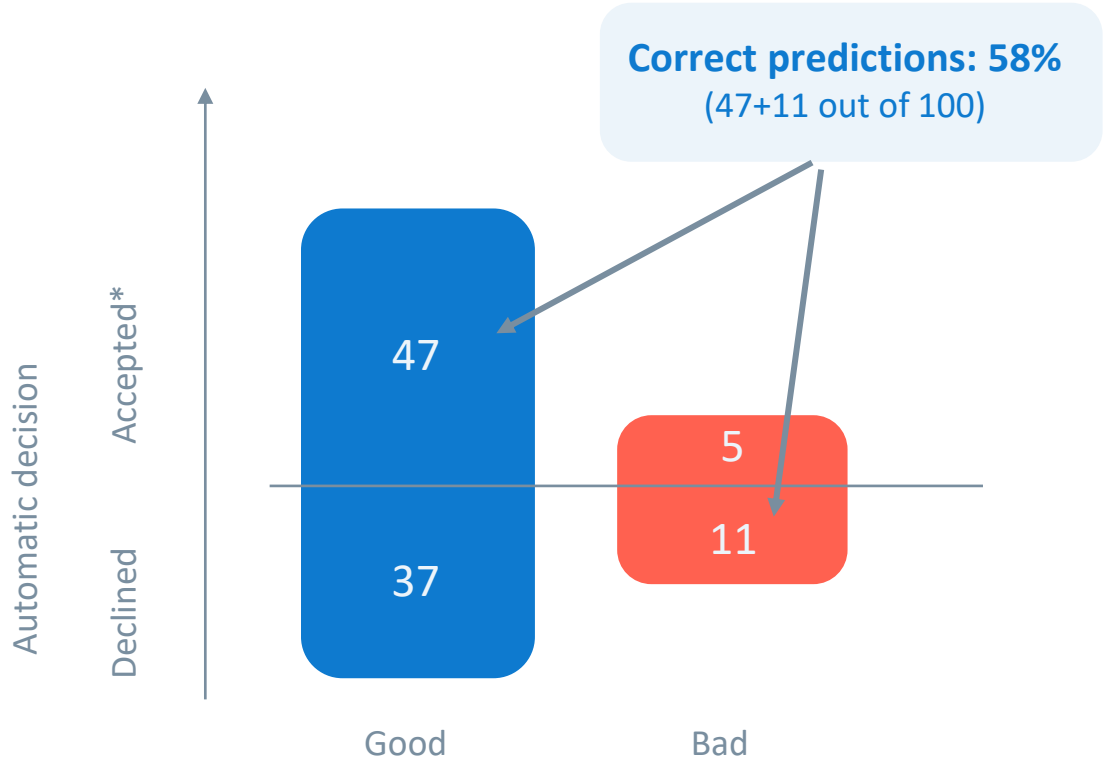
Champion challenger

We have compared the performance between
1. a categorisation only based approach
2. our combined TC approach (including categorisation)

Results

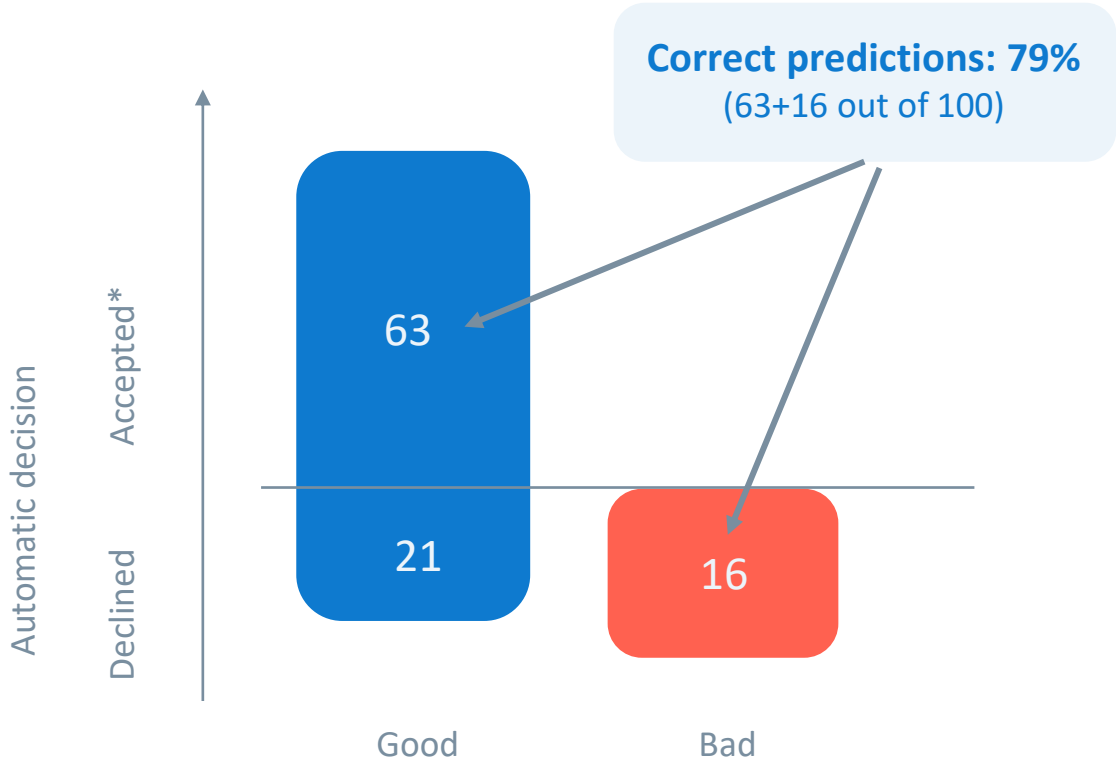
Categorisation only

% of total applications



Categorisation + Combined TC

% of total applications



*Accepted if residual income above minimum threshold

Conclusion

Currently live **in France** with **10 leading financial institutions** & under deployment with lenders **in Spain, Portugal & Belgium**

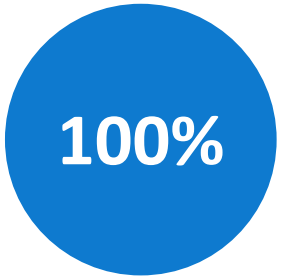
More accuracy with TC

Accuracy increased by…

To 79% with Combinated TC

from 58% with categorisation only

Prediction errors always on the conservative side



… of « Bad » are declined

… of accuracy expected with 3D TC

Conclusion

Implications for the credit industry
of providing an automatic and accurate affordability assessment


Mortgage loans


Consumer loans


BNPL

Shorter “time-to-yes”

U/W process automation

More responsible lending

Comply with regulation

Fight over-indebtedness

www.algoan.com

20, rue Drouot
75009 Paris - France

contact@algoan.com