Faculty of
Computer Science

Data Science and
Business Analytics

Moscow
2025

Business Analytics

Client Segmentation

Homework #1 / Variant 1

Completed by: Markin Alexander Andreevich | *DSBA-231*

The Data

Like most projects related to data analysis, it all starts with the data.

In this case, the problem revolved around an unknown bank's loan application data which included both existing customers and new applicants

The Problem

The data had to be used in order to conduct customer segmentation and build non-homogenous client profiles which could later be used to pursue specific business goals (target advertising, market positioning, etc.)

Two different segmentation methods had to be applied, and in both cases there had to be at least 5 segments (none of which were allowed to take up more than 50% of the data)

Data Science and
Business Analytics

Client Segmentation

EDA

Exploratory Data Analysis

What's in the data?

Initially, there were 44 variables in the dataset, spanning 10243 rows.

A decision was made to remove "Номер варианта" and "ID" immediately. The former due to its artificial nature (only needed to separate a larger dataset into univariant files) and the latter - due to there being as many unique values of it as rows.

There's a chance that an actual bank could've requested the IDs to be left in the data in order to conduct thorough investigation of some specific outliers, alas - no such request was made in this case

```
Номер варианта
ID
INCOME_BASE_TYPE
CREDIT_PURPOSE
INSURANCE_FLAG
DTI
SEX
FULL_AGE_CHILD_NUMBER
DEPENDANT_NUMBER
EDUCATION
EMPL_TYPE
EMPL_SIZE
BANKACCOUNT_FLAG
Period_at_work
age
EMPL_PROPERTY
EMPL_FORM
FAMILY_STATUS
max90days
max60days
max30days
max21days
max14days
avg_num_delay
if_zalog
num_AccountActive180
num_AccountActive90
num_AccountActive60
Active_to_All_prc
numAccountActiveAll
numAccountClosed
sum_of_paym_months
all_credits
Active_not_cc
own_closed
min_MnthAfterLoan
max_MnthAfterLoan
dlq_exist
thirty_in_a_year
sixty_in_a_year
ninety_in_a_year
thirty_vintage
sixty_vintage
ninety_vintage
```

What's in the data?

A brief look at the data confirmed initial concerns about some variables with numeric types being categorical, therefore a decision was made to separate them into categorical and non-categorical by hand:

Categoricals:

```
['INCOME_BASE_TYPE',
 'CREDIT_PURPOSE',
 'INSURANCE_FLAG',
 'SEX',
 'EDUCATION',
 'EMPL_TYPE',
 'EMPL_SIZE',
 'BANKACCOUNT_FLAG',
 'EMPL_PROPERTY',
 'EMPL_FORM',
 'FAMILY_STATUS',
 'if_zalog',
 'dlq_exist',
 'thirty_in_a_year',
 'sixty_in_a_year',
 'ninety_in_a_year',
 'thirty_vintage',
 'sixty_vintage',
 'ninety_vintage']
```

Non-categoricals:

```
['DTI',
 'FULL_AGE_CHILD_NUMBER',
 'DEPENDANT_NUMBER',
 'Period_at_work',
 'age',
 'max90days',
 'max60days',
 'max30days',
 'max21days',
 'max14days',
 'avg_num_delay',
 'num_AccountActive180',
 'num_AccountActive90',
 'num_AccountActive60',
 'Active_to_All_prc',
 'numAccountActiveAll',
 'numAccountClosed',
 'sum_of_paym_months',
 'all_credits',
 'Active_not_cc',
 'own_closed',
 'min_MnthAfterLoan',
 'max_MnthAfterLoan']
```

Categorical Data: NaN data

With all of the categorical variables, the decision was made to remove any NaNs outright, as they could not be feasibly interpreted based on the existing data without introducing serious biases to the dataset.

NaNs were removed from the Income type, Employment type, Salary size and Bank account flag variables. Additionally, multiple other variables had their NaNs purged, but more on that later.

There were also records labeled as *n.a.* for the education and salary size variables. Those were not removed as they convey the client not having an education/salary.

```
['INCOME_BASE_TYPE',
 'EDUCATION'
 'EMPL_TYPE',
 'EMPL_SIZE',
 'BANKACCOUNT_FLAG']
```

CD: additional cleanup

- The EDUCATION variable featured separate values for “Higher education”, “Secondary higher education” and “PhD”, as well as a catch-all value combining the three. A decision was made to combine all of these under just the catch-all variable in order to minimise bias
- The EMPL_SIZE (Salary size) variable had both a “>100” and a “>=100” value. This was the only number to have a “>” in the value and the distinction didn’t provide much information, so these were combined as well. It also seems as if the values starting with “>=” all signified bins of [X; X+50) for salary size.

```
['EDUCATION',
 'EMPL_SIZE']
```

CD: additional cleanup

- From the list attached to the problem, it's known that the BANKACCOUNT_FLAG variable only holds three meanings depending on the values: 0, 1 or >=2. Therefore, any value equal to or higher than 2 was merged under just "2".

```
['BANKACCOUNT_FLAG']
```

CD: The problem with EMPL_FORM and FAMILY_STATUS

- Both the EMPL_FORM and the FAMILY_STATUS variables have >60% NaN values.
- Moreover, whenever one of them lacks a value, the other does too!
- And so do all the variables below them in the table to the right:

EMPL_FORM
FAMILY_STATUS
max90days
max60days
max30days
max21days
max14days
avg_num_delay
if_zalog
num_AccountActive180
num_AccountActive90
num_AccountActive60
Active_to_All_prc
numAccountActiveAll
numAccountClosed
sum_of_paym_months
all_credits
Active_not_cc
own_closed
min_MnthAfterLoan
max_MnthAfterLoan
dlq_exist
thirty_in_a_year
sixty_in_a_year
ninety_in_a_year
thirty_vintage
sixty_vintage
ninety_vintage

CD: The problem with EMPL_FORM and FAMILY_STATUS

- And when neither is missing, the values below still have ~200-300 missing values.
- Out of these variables, there are eight remaining categoricals: if_zalog (is there any property to facilitate the loan), dlq_exist (is there an overdue payment right now), thirty/sixty/ninety_in_a_year/vintage (was there an overdue payment g.t. 30/60/90 days this year/ever), all of them booleans.

=> Therefore, the NaNs will be removed from the data in both cases in order to make sure that the final dataset ONLY contains bank clients.

Numerical data:

After the initial categorical cleanup (which left us with a total of 3730 records), most non-categorical variables barely have any NaNs left.

Out of those that have missing values:
FULL_AGE_*CHILD*_... has 1 missing value,
Period_at_work has 1 missing value as vell, max90days-max14days all have 2 missing values in the same row and avg_num_delay has 16 missing values.

All of them will be replaced with means over existing data, in contrast to what had to be done with categoricals.

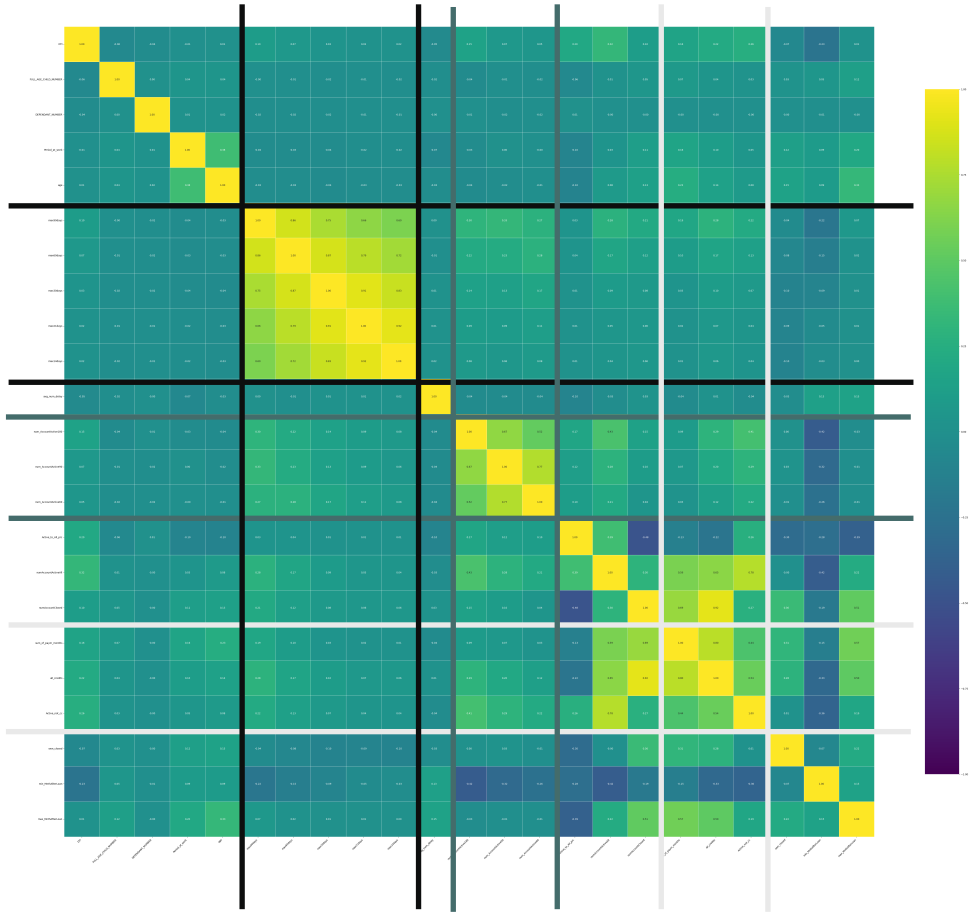
Non-categoricals:	Count NaNs
'DTI',	0
'FULL_AGE_CHILD_NUMBER',	1
'DEPENDANT_NUMBER',	0
'Period_at_work',	1
'age',	0
'max90days',	2
'max60days',	2
'max30days',	2
'max21days',	2
'max14days',	2
'avg_num_delay',	16
'num_AccountActive180',	0
'num_AccountActive90',	0
'num_AccountActive60',	0
'Active_to_All_prc',	0
'numAccountActiveAll',	0
'numAccountClosed',	0
'sum_of_paym_months',	0
'all_credits',	0
'Active_not_cc',	0
'own_closed',	0
'min_MnthAfterLoan',	0
'max_MnthAfterLoan']	0

Data after cleanup

Column	Count Unique	Count Zeros	% of Zeros	Count NaNs	% of NaNs	data type
INCOME_BASE_TYPE	4	0	0	0	0	object
CREDIT_PURPOSE	10	0	0	0	0	object
INSURANCE_FLAG	2	1460	39.1421	0	0	int64
DTI	58	0	0	0	0	float64
SEX	2	1897	50.8579	0	0	int64
FULL_AGE_CHILD_NUMBER	7	2269	60.8311	0	0	float64
DEPENDANT_NUMBER	3	3719	99.7051	0	0	int64
EDUCATION	6	0	0	0	0	object
EMPL_TYPE	8	0	0	0	0	object
EMPL_SIZE	3	0	0	0	0	object
BANKACCOUNT_FLAG	3	3061	82.0643	0	0	Int64
Period_at_work	258	0	0	0	0	float64
age	39	0	0	0	0	float64
EMPL_PROPERTY	5	0	0	0	0	object
EMPL_FORM	6	0	0	0	0	object
FAMILY_STATUS	6	0	0	0	0	object
max90days	20	1048	28.0965	0	0	float64
max60days	18	1567	42.0107	0	0	float64
max30days	14	1994	53.4584	0	0	float64
max21days	14	2381	63.8338	0	0	float64
max14days	12	2602	69.7587	0	0	float64
avg_num_delay	1148	1587	42.5469	0	0	float64
if_zalog	2	2510	67.2922	0	0	float64
num_AccountActive180	7	2642	70.8311	0	0	float64
num_AccountActive90	4	3202	85.8445	0	0	float64
num_AccountActive60	4	3407	91.3405	0	0	float64
Active_to_All_prc	96	499	13.378	0	0	float64
numAccountActiveAll	13	471	12.6273	0	0	float64
numAccountClosed	26	447	11.9839	0	0	float64
sum_of_paym_months	325	16	0.428954	0	0	float64
all_credits	30	0	0	0	0	float64
Active_not_cc	7	1266	33.941	0	0	float64
own_closed	9	2161	57.9357	0	0	float64
min_MnthAfterLoan	101	138	3.69973	0	0	float64
max_MnthAfterLoan	134	9	0.241287	0	0	float64
dlq_exist	2	1603	42.9759	0	0	float64
thirty_in_a_year	2	3194	85.63	0	0	float64
sixty_in_a_year	2	3417	91.6086	0	0	float64
ninety_in_a_year	2	3483	93.378	0	0	float64
thirty_vintage	2	3612	96.8365	0	0	float64
sixty_vintage	2	3679	98.6327	0	0	float64
ninety_vintage	2	3677	98.5791	0	0	float64

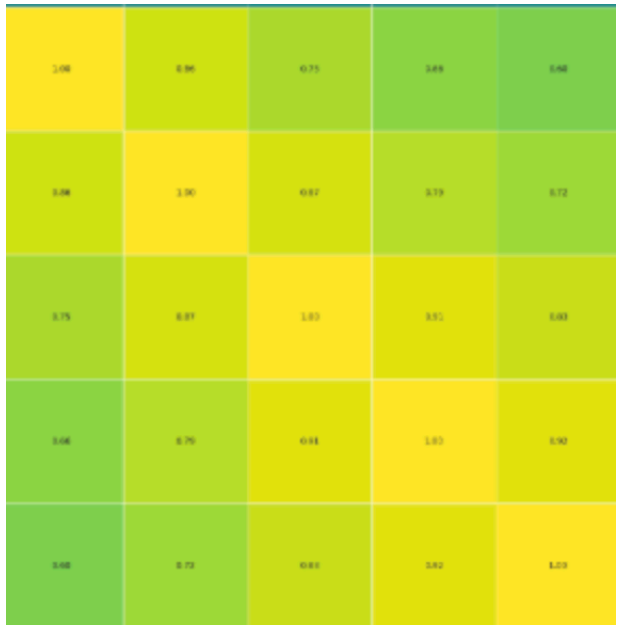
Correlation matrix

First, it was decided to look for correlation between numerical variables. With ~30% as the threshold for correlation, four correlated classes were found:



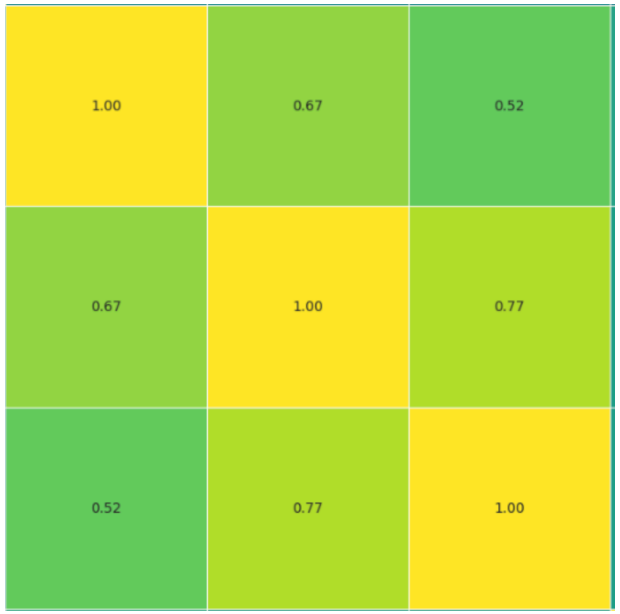
Correlation matrix

maxNdays [data requests about a client's credit history] (N $\in 14, 21, 30, 60, 90$) - highly positive correlation (even higher for N close to each other). Makes sense, as the value for N = 90 cannot be lower than one for N = 60 and so on.



Correlation matrix

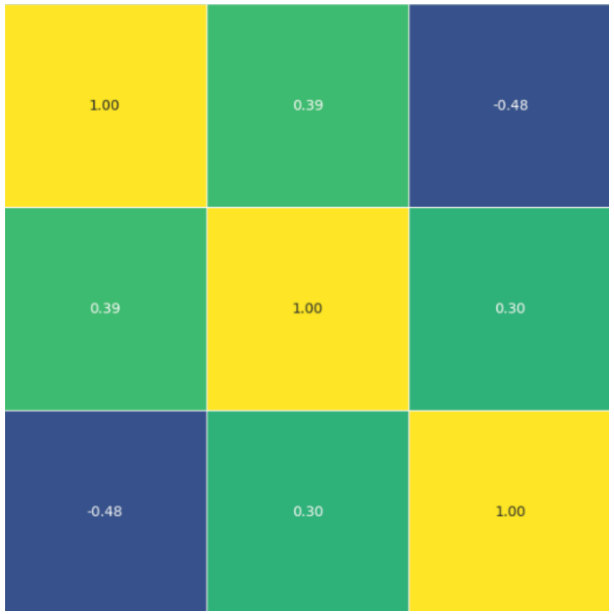
num_AccountActiveN [number of active accounts in the past N days] (N $\in 60, 90, 180$) - highly positive correlation. The reason for this is the same as for the last class - every value for a higher N cannot be lower than the one for a lower N



Correlation matrix

Active_to_All_pre (num of active accounts divided by total number of accounts), numAccountActiveAll (number of active accounts) and numAccountClosed (number of closed accounts).

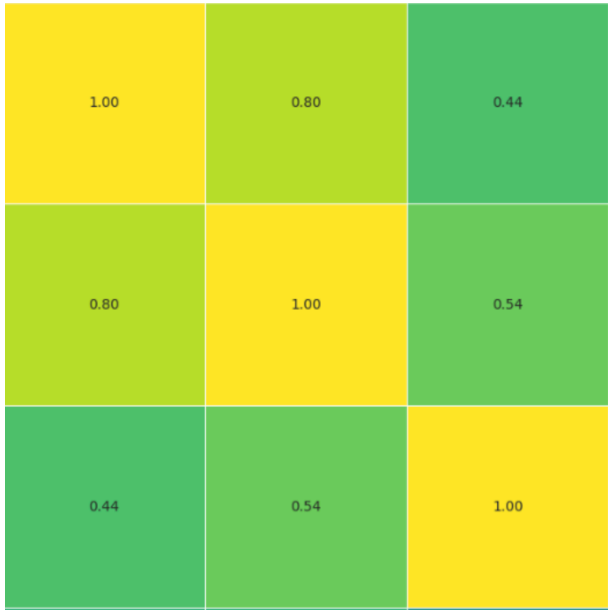
There's a positive correlation between Active_to_All_pre and numAccountActiveAll, and a highly negative correlation between Active_to_All_pre and numAccountClosed, which makes sense, as both of these make up Active_to_All_pre in the form of a fraction ((numAccountActiveAll)/(numAccountActiveAll+numAccountClosed)).



Correlation matrix

Finally, there's a strong postive correlation between the total payments (sum_of_paym_months) and the number of credits (all_credits) - due to basic summation. The more credits you have - the more you pay.

And, additionally, there's a strong positive correlation between the number of credits and the number of active non-credit card credits (Active_not_cc). This also makes sense - the more credits one has, the higher their total number of credits.



Data distribution of categorical variables

Based on the categorical variables, a very approximate profile of our "average" customer has been constructed:

- Male
- Has attained at least higher education
- Is a specialist
- Has a salary of 250 or higher, which gets sent to his bank account
- Has insurance
- Is seeking a loan for renovations
- Does NOT have an online account
- Works in a Sales-related company, which is an "OOO" legally
- Is married
- Doesn't have collateral for the loan
- Is currently in arrears, but has never been overdue for more than 30 days

Average Profile

Additionally, some other assumptions about the “average” client were made using numerical values:

- Roughly 36 years old
- Has no children
- Has been at their current workplace for 55 days
- Has 6 credits, one of which is active

Segmentation

Preparing for segmentation

There is still something to consider before proceeding with segmentation - categorical variables stored in non-numeric dtypes.

Boolean variables are already in numeric form, ordinal variables will have to be transformed with relation to their order (e.g., Pioneer -> 1, Specialist -> 2, Professional -> 3), while other categoricals will just have to be encoded using one-hot encoding.

Booleans:

```
['thirty_vintage',
 'sixty_vintage',
 'ninety_vintage',
 'thirty_in_a_year',
 'sixty_in_a_year',
 'ninety_in_a_year',
 'dlq_exist',
 'if_zalog',
 'SEX',
 'INSURANCE_FLAG']
```

Non-ordinal:

```
['EMPL_TYPE',
 'FAMILY_STATUS',
 'EMPL_PROPERTY',
 'INCOME_BASE_TYPE',
 'EMPL_FORM',
 'CREDIT_PURPOSE']
```

Ordinals not in need of conversion:

```
['BANKACCOUNT_FLAG']
```

Ordinals in need of conversion:

```
['EDUCATION', 'EMPL_SIZE']
```

Preparing for segmentation

Finally, let's remove outliers like in WORKSHOP 2 of the course.

After everything has been done, 3430 records (roughly 30% of the initial data) and 75 variables (thanks to one-hot encoding) remain.

```
df.shape
```

```
(3430, 75)
```

Final data mart

INSURANCE_FLAG
DTI
SEX
FULL_AGE_CHILD_NUMBER
DEPENDANT_NUMBER
EDUCATION
EMPL_SIZE
BANKACCOUNT_FLAG
Period_at_work
age
max90days
max60days
max30days
max21days
max14days
avg_num_delay
if_zalog
num_AccountActive180
num_AccountActive90
num_AccountActive60

Active_to_All_prc
numAccountActiveAll
numAccountClosed
sum_of_paym_months
all_credits
Active_not_cc
own_closed
min_MnthAfterLoan
max_MnthAfterLoan
dlq_exist
thirty_in_a_year
sixty_in_a_year
ninety_in_a_year
thirty_vintage
sixty_vintage
ninety_vintage
EMPL_TYPE_вспомогательный персонал
EMPL_TYPE_другое
EMPL_TYPE_менеджер высшего звена
EMPL_TYPE_менеджер по продажам
EMPL_TYPE_менеджер среднего звена
EMPL_TYPE_рабочий

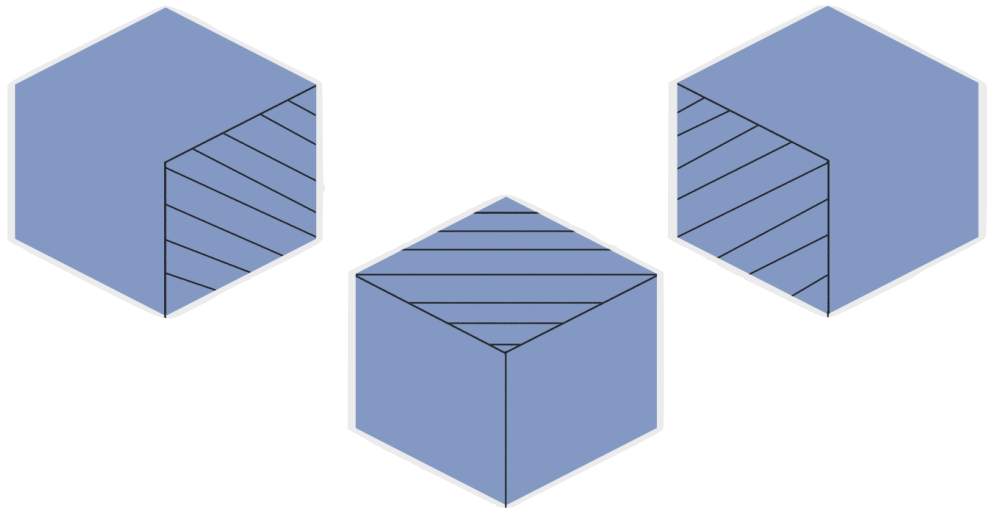
EMPL_TYPE_специалист
EMPL_TYPE_торговый представитель
FAMILY_STATUS_вдовец / вдова
FAMILY_STATUS_гражданский брак
FAMILY_STATUS_женат / замужем
FAMILY_STATUS_повторный брак
FAMILY_STATUS_разведен / разведена
FAMILY_STATUS_холост / не замужем
EMPL_PROPERTY_Другое
EMPL_PROPERTY_Информационные технологии
EMPL_PROPERTY_Сельское и лесное хозяйство
EMPL_PROPERTY_Торговля
EMPL_PROPERTY_Юридические услуги
INCOME_BASE_TYPE_2НДФЛ
INCOME_BASE_TYPE_Поступление зарплаты на счет
INCOME_BASE_TYPE_Свободная форма с печатью работодателя
INCOME_BASE_TYPE_Форма банка (без печати работодателя)
EMPL_FORM_Государственное предприятие
EMPL_FORM_ЗАО
EMPL_FORM_Иная форма
EMPL_FORM_Индивидуальный предприниматель

Final data mart

```
EMPL_FORM_OAO
EMPL_FORM_OOO
CREDIT_PURPOSE_Другое
CREDIT_PURPOSE_Лечение
CREDIT_PURPOSE_Обучение
CREDIT_PURPOSE_Отпуск
CREDIT_PURPOSE_Покупка автомобиля
CREDIT_PURPOSE_Покупка бытовой техники
CREDIT_PURPOSE_Покупка земли
CREDIT_PURPOSE_Покупка мебели
CREDIT_PURPOSE_Покупка недвижимости/ строительство
CREDIT_PURPOSE_Ремонт
```

Segmentation: RFM

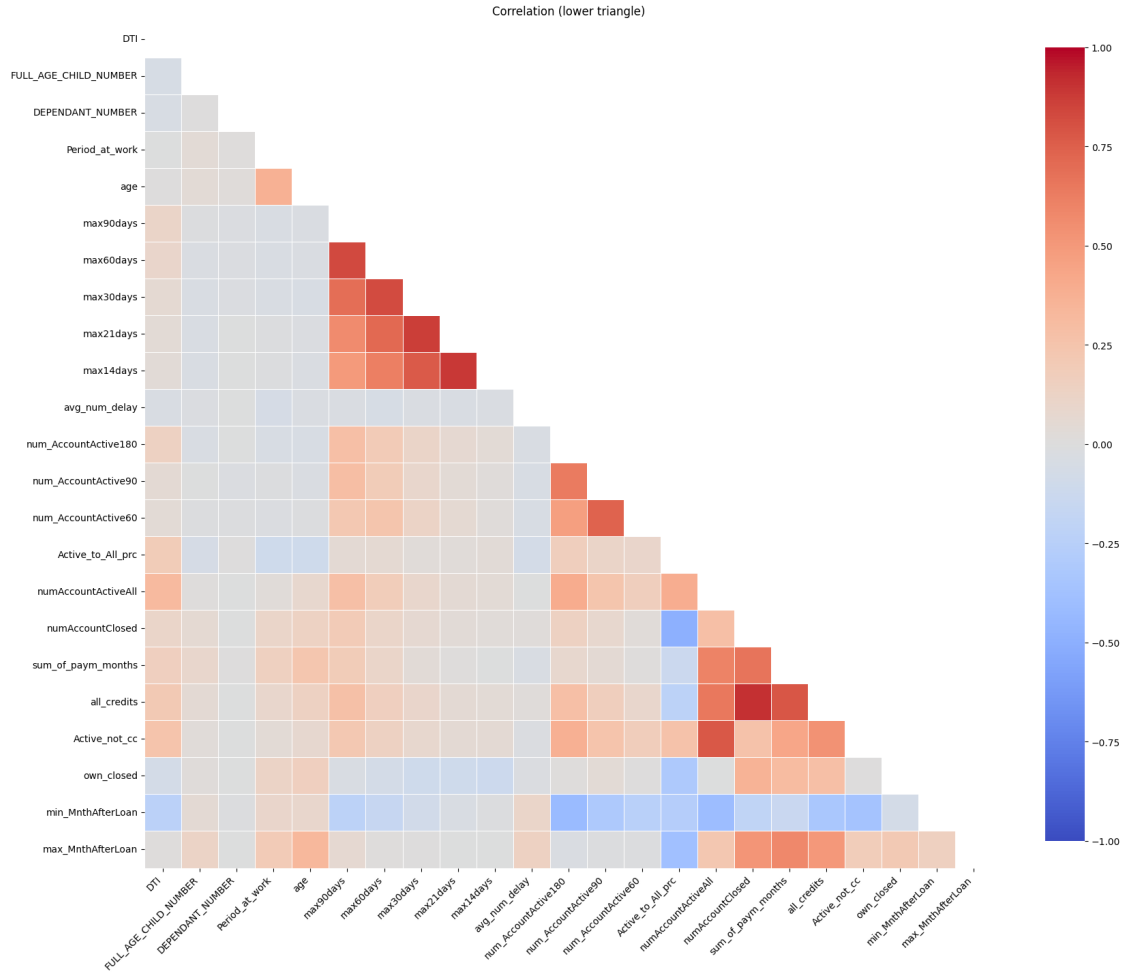
RFM is quite a useful method for assessing the segments the bank currently has and modifying the strategy based on the findings, as by its nature it already provides the analyst (me, in this case) with information about the importance of each of the clusters to the company.



Segmentation: RFM

The first step before starting segmentation would be to choose the R, the F and the M.

In this particular case, it was decided that (due to their high correlation + business logic) that:
- Recency -> min_MnthAfterLoan (how recent was the last time the client was given a loan)
- Frequency -> numAccountActiveAll (the number of currently active accounts, for the lack of a better frequency variable)
- Monetary -> sum_of_paym_months (the current number of monthly payments)



Segmentation: RFM

In order to choose the segments, the chosen metrics were converted into a 5-point scale (since 5^3 is divisible by 5 [the minimum amount of segments we are allowed to have], it made sense to use this to look for potential obvious segments in the data).

After failing to find segments that would make sense from a business point of view, the following approach to segmentation was thought up:

- VIP - those who have frequent, recent and pricey payments (all >= 4)
- Core - those who have at least average (=3) values for all the metrics, yet are not in VIP
- Potential - those who have made recent payments, yet who were lacking in either frequency or monetary value
- At risk - those who have not made any payments in a while (<= 2), yet whose payments are either semi-frequent or at least somewhat valuable in monetary terms.
- Churn - everyone else

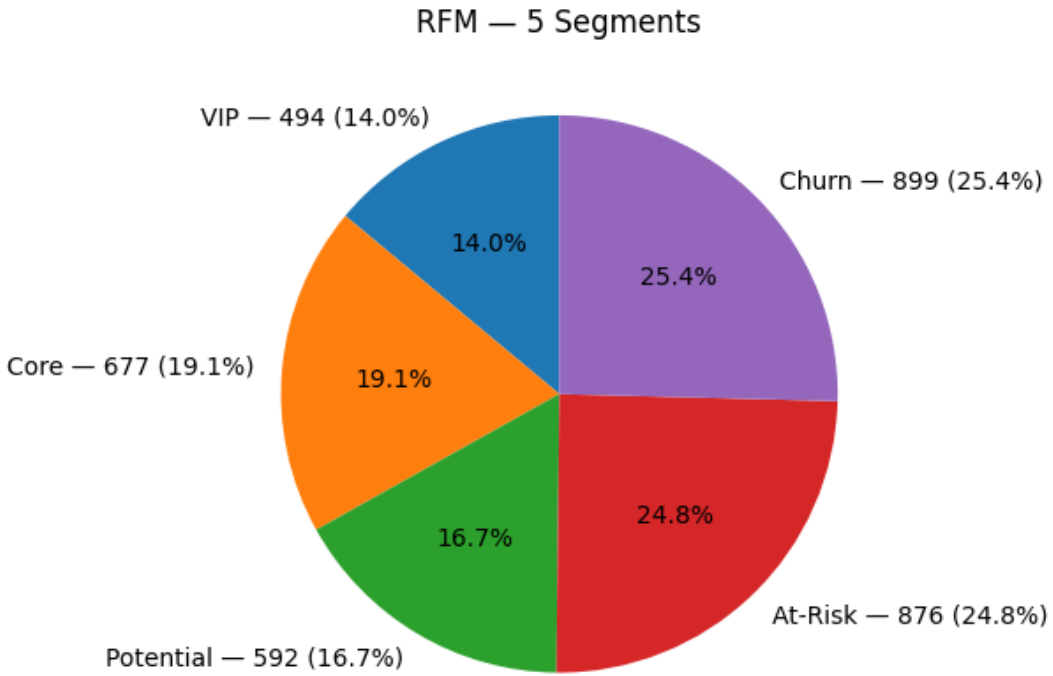
Segmentation: RFM

And here's a breakdown of the bank's potential goals with each segment:

- VIP are the customers we want to keep. We don't need to ask them to spend more or take out loans more often, we just want to stop them from potentially moving somewhere else
- Core are the customers we still want to keep, maybe sometimes incentivising them to be more active
- Potential are those who have potential, being somewhat recent clients, yet do not spend as much or as frequently as we want
- At risk are those who used to be our core, but for some reason haven't taken out a loan in a while
- Finally, churn (traditionally) contains clients lacking in all aspects simultaneously. We do not need to target them, as there is largely no profit in doing so

Segmentation: RFM

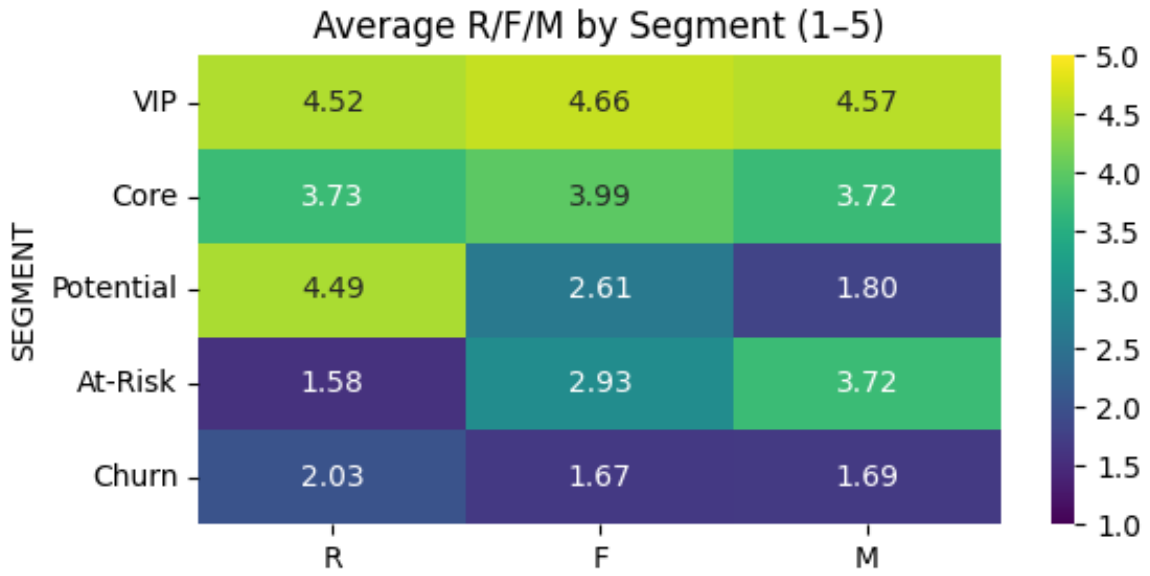
As a result, 5 segments of an almost equal size were obtained, the largest of them (1/4 of the data), unfortunately, being Churn.



Segmentation: RFM

As a result, 5 segments of an almost equal size were obtained, the largest of them (1/4 of the data), unfortunately, being Churn.

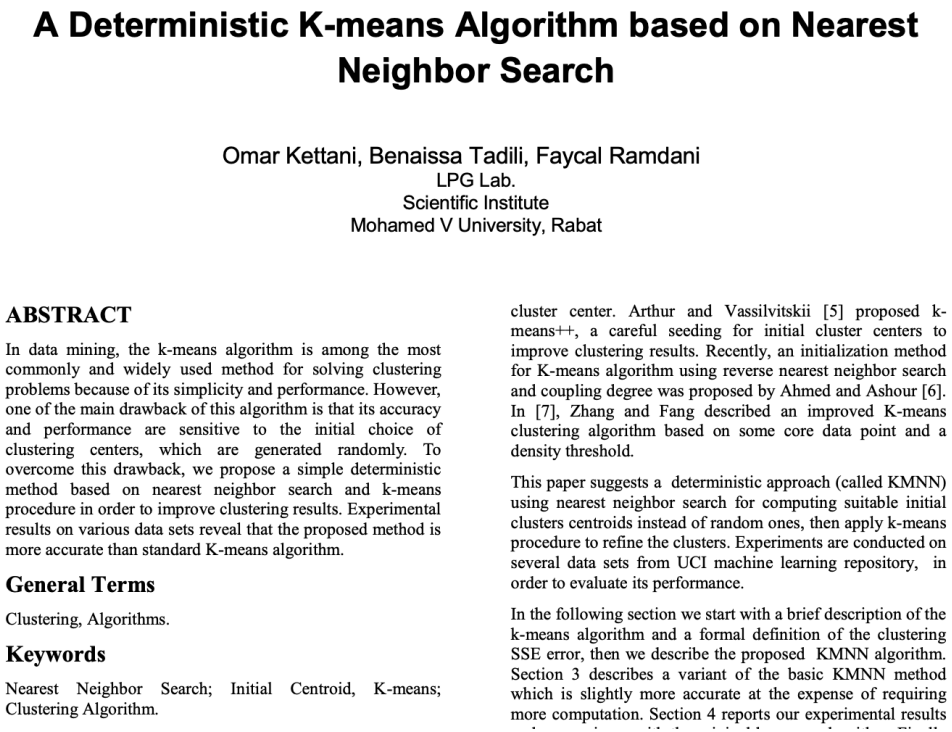
Additionally, a table of the means of R, F and M for each segment was constructed for additional observations. It does not present much new information, as it largely shows the rules for the construction of each segment, and yet some conclusions could be drawn from it - e.g., the bank's VIPs and Potentials have almost the same mean Recency measure.



Segmentation: Unsupervised (K-means)

K-means based segmentation ended up quite a useful tool, albeit the existing algorithms for it use a lot of randomness, so unless the dataset is perfectly segmentable, the results may vary from run to run.

Therefore, an algorithm loosely based on this paper had to be implemented. Unfortunately, it led to a significantly longer runtime for the notebook, but at least the results are now mostly reproducible



A Deterministic K-means Algorithm based on Nearest Neighbor Search

Omar Kettani, Benaissa Tadili, Faycal Ramdani
LPG Lab.
Scientific Institute
Mohamed V University, Rabat

ABSTRACT

In data mining, the k-means algorithm is among the most commonly and widely used method for solving clustering problems because of its simplicity and performance. However, one of the main drawback of this algorithm is that its accuracy and performance are sensitive to the initial choice of clustering centers, which are generated randomly. To overcome this drawback, we propose a simple deterministic method based on nearest neighbor search and k-means procedure in order to improve clustering results. Experimental results on various data sets reveal that the proposed method is more accurate than standard K-means algorithm.

General Terms

Clustering, Algorithms.

Keywords

Nearest Neighbor Search; Initial Centroid, K-means; Clustering Algorithm.

cluster center. Arthur and Vassilvitskii [5] proposed k-means++, a careful seeding for initial cluster centers to improve clustering results. Recently, an initialization method for K-means algorithm using reverse nearest neighbor search and coupling degree was proposed by Ahmed and Ashour [6]. In [7], Zhang and Fang described an improved K-means clustering algorithm based on some core data point and a density threshold.

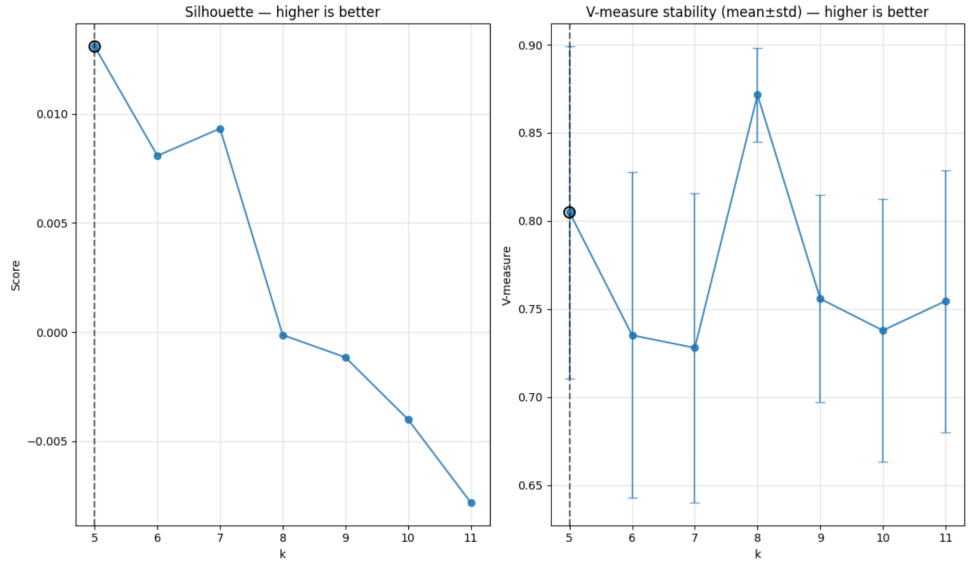
This paper suggests a deterministic approach (called KMNN) using nearest neighbor search for computing suitable initial clusters centroids instead of random ones, then apply k-means procedure to refine the clusters. Experiments are conducted on several data sets from UCI machine learning repository, in order to evaluate its performance.

In the following section we start with a brief description of the k-means algorithm and a formal definition of the clustering SSE error, then we describe the proposed KMNN algorithm. Section 3 describes a variant of the basic KMNN method which is slightly more accurate at the expense of requiring more computation. Section 4 reports our experimental results

Segmentation: Unsupervised (K-means)

In order to assess how many segments to use, multiple approaches were employed and applied iteratively one after another:

1. maximise silhouette method score
2. maximise v-measures*

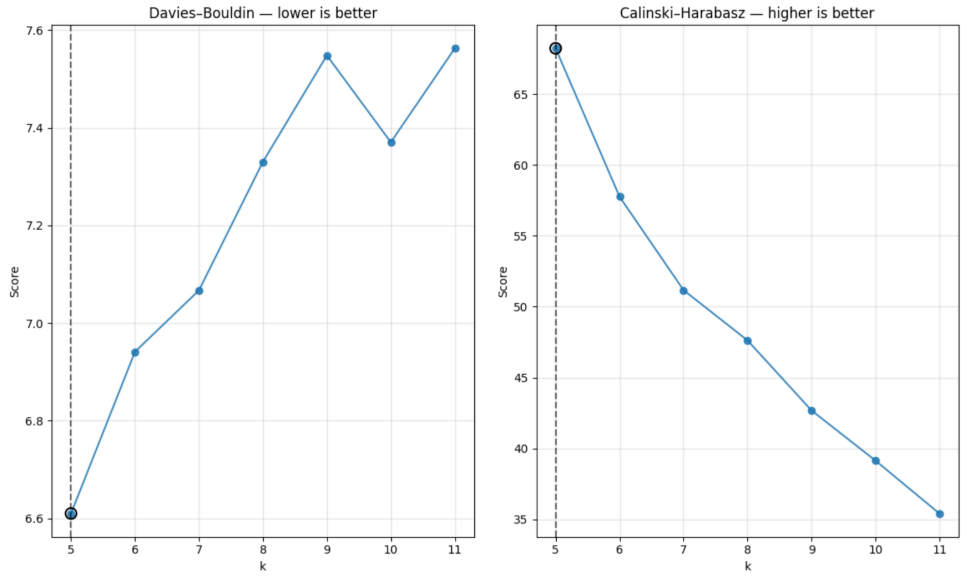


* This method was added in order to make the choice of k more stable due to its symmetry and rigidity. Previously, the optimal value of k "jumped" around

Segmentation: Unsupervised (K-means)

In order to assess how many segments to use, multiple approaches were employed and applied iteratively one after another:

1. maximise silhouette method score
2. maximise v-measures*
3. minimise David-Boldin score
4. maximise Calinski-Harabasz score



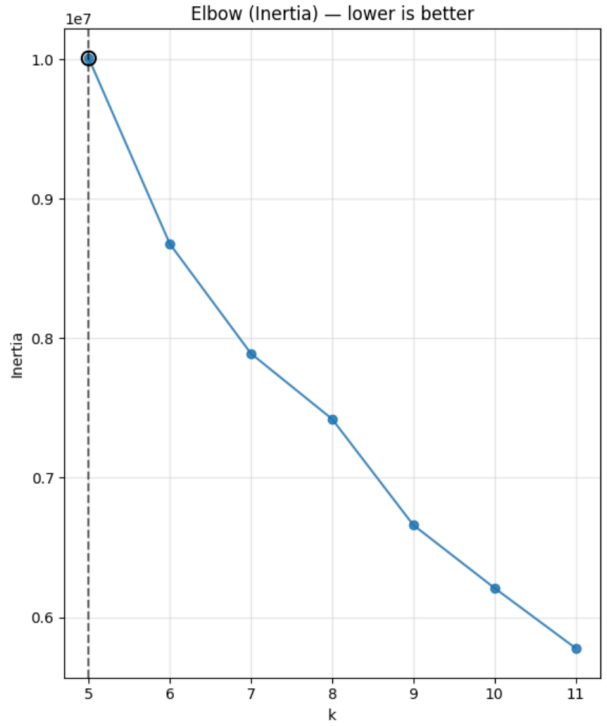
* This method was added in order to make the choice of k more stable due to its symmetry and rigidity. Previously, the optimal value of k "jumped" around

Segmentation: Unsupervised (K-means)

In order to assess how many segments to use, multiple approaches were employed and applied iteratively one after another:

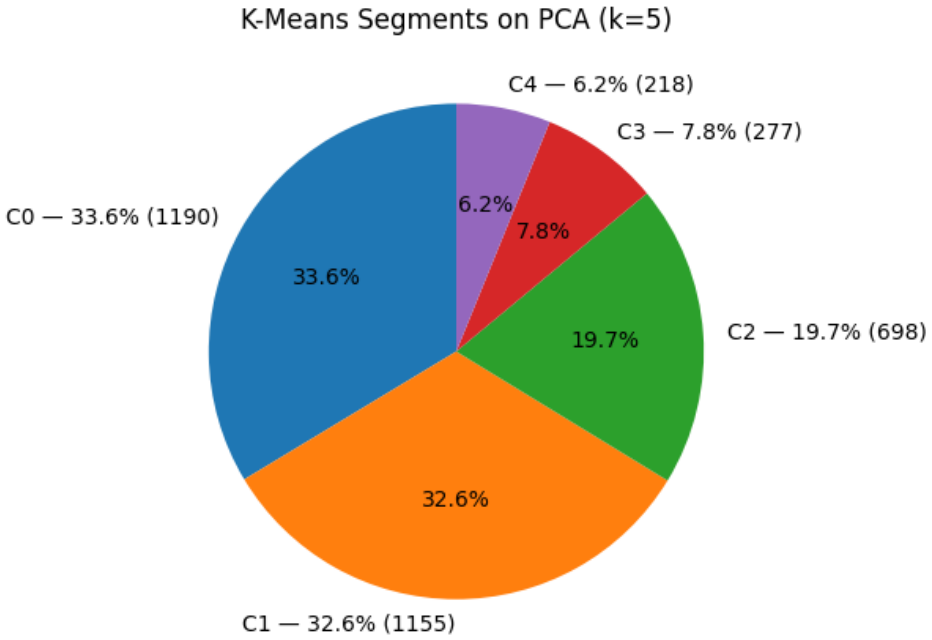
1. maximise silhouette method score
2. maximise v-measures*
3. minimise David-Boldin score
4. maximise Calinski-Harabasz score
5. minimise elbow method score

* This method was added in order to make the choice of k more stable due to its symmetry and rigidity. Previously, the optimal value of k "jumped" around



Segmentation: Unsupervised (K-means)

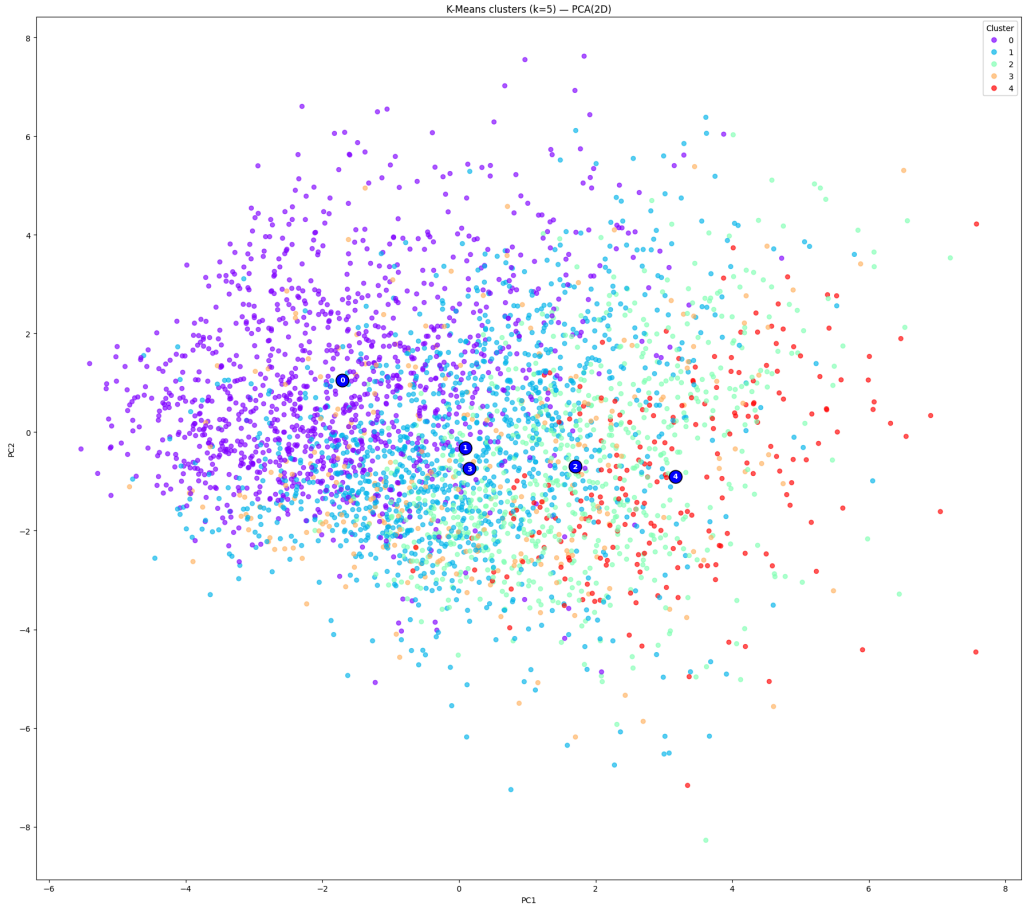
Once both the k-choice and the clustering algorithm were run, there were 5 clusters, 3 of which were larger and took 86% of the data, whereas the remaining 2 only took up 14%.



Segmentation: Unsupervised (K-means)

Once both the k-choice and the clustering algorithm were run, there were 5 clusters, 3 of which were larger and took 86% of the data, whereas the remaining 2 only took up 14%.

In a 2-dimensional PCA projection, 4 of the 5 were well-defined and almost didn't overlap, although one of the clusters was spread through the entire graph. Alas, that is mostly due to the limitations of dimension reduction.

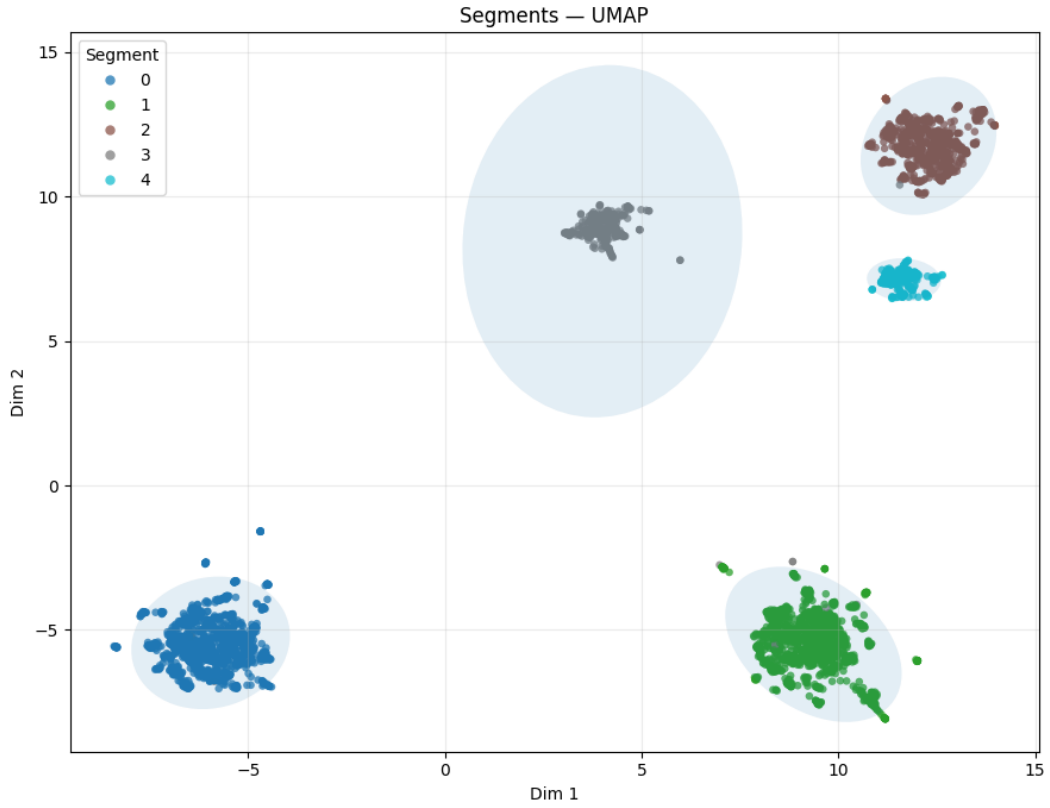


Segmentation: Unsupervised (K-means)

Once both the k-choice and the clustering algorithm were run, there were 5 clusters, 3 of which were larger and took 86% of the data, whereas the remaining 2 only took up 14%.

In a 2-dimensional PCA projection, 4 of the 5 were well-defined and almost didn't overlap, although one of the clusters was spread through the entire graph. Alas, that is mostly due to the limitations of dimension reduction.

In a UMAP-based projection, one can more clearly see that the clusters, indeed, exist and have reasons to be separated as such.



Customer Profiles

Data Science and
Business Analytics

Client Segmentation

Customer Profiles

RFM

Segment 1: VIP / Engaged premium clients

Clients in this group, on average:

- Tend to be middle-aged (37 y.o.) men with a child, who have been at their new workplace for ~64 days by now
- They receive salary via a bank form
- They extensively use their account, showcased by them having ~10 credits, 4 of which are active.
- They rank in payments of 152 each month, successfully close the overwhelming majority of their loans, and they have at least one active loan opened over the last 6 months

In terms of risk:

- They have an average Debt-to-Income ratio of 0.437, which means that their income is just twice their debt (the worst DTI of the segments)
- Delinquencies in terms of payments are quite common, boasting a score of 79.4%. Yet there's only a 17% chance that they will remain unpaid for over a month, and only 6% of them delay payments
- The majority lack collateral, but do have insurance

Segment 1: VIP / Engaged premium clients

Thus, these clients are important and have good financial activity, yet are likely to carry higher risks and delay payments (which is critical due to their payments also being the largest on average)

As a result, the bank ought to keep their high value while trying to control their high risk. Additional offers could be provided, such as bundles (credit card + insurance + etc) while also setting stricter limits on delays, as well as issuing early warnings about incoming delays to the customers who are the most prone to them.

Variables to pay attention to in this segment: monthly payments, 30/60/90 day delays over the last year, DTI

Segment 2: Core / Reliable mainstay

Clients in this group, on average:

- Tend to be middle-aged (35.6 y.o.) married men, who have been at their new workplace for ~54 days by now
- They get salary sent straight to their bank account
- Their usage is less extensive: ~7 credits, 3 of which are currently active.
- They pay 2/3 of what VIP pays each month - ~100, successfully close only 0.80 of their loans, and they have, on average 0.45 active loans opened over the last 6 months

In terms of risk:

- They have an average Debt-to-Income ratio of 0.432, which means that their income is also appx. twice their debt
- Delinquencies in terms of payments are less common, with a score of 67.1%. There's a 14% chance that they will remain unpaid for over a month, and 6% of them delay payments
- They're even less likely than VIP to have collateral, but more likely to have insurance

Segment 2: Core / Reliable mainstay

While also having relatively high financial activity, this group carries less risk than VIP.

Therefore, while there's room for growth in this segment, the bank still has to pursue it safely. E.g., they could be sold adjacent products to get them more interested in the ecosystem of the bank. They could also be provided various loyalty rewards in order to transfer some of the members to the VIP segment. In that regard, their loan limit could also be periodically increased after streaks of successful on-time payments.

Variables to pay attention to in this segment: active credits divided by all credits, 30/60/90 day delays over the last year, on-time payments

Segment 3: Potential / Younger professionals

Clients in this group, on average:

- Are early 30s (~33 y.o.) married women, who have been at their new workplace for ~44 days
- They also get salary sent straight to their bank account
- Their usage is relatively tame: ~4 credits, 2 of which are active.
- They pay ~32.5 each month, successfully close just 0.529 of their loans, yet, on average, they have 0.81 active loans opened over the last 6 months, which is larger than that of Core

In terms of risk:

- They have the lowest DTI in the top-3: 0.382
- Only delay their payments in 35.8% of cases. There's just a 7% chance that they will delay one for over a month
- Only 1/5 of them have collateral, but the majority has insurance

Segment 3: Potential / Younger professionals

Despite significantly lower payments and closures, this group is engaged and not prone to significant delays.

The bank, thus, has to keep an eye on this specific group, as it is relatively easy to see them being converted to either of the previous two segments in the long term. Cashback and various onboarding offers could be introduced in order to nurture the relationship with the bank. Several improved micro-loan incentives could also be added, as that seems to be the primary use case of the bank by this group.

Variables to pay attention to in this segment:
all_credits, payment growth, 30/60/90-day delays

Segment 4: At-risk / Recently inactive, yet acquainted

Clients in this group, on average:

- 37.4 y.o. men with children (the most likely group to have a child), who have been at their new workplace for just over 2 months
- They obtain their salary via a bank form
- Their usage is generally average: ~6 credits, just 2 of which are active.
- However, they pay ~102.1 each month and successfully close 0.872 of their loans. This signifies that they used to be in the Core sector, but have dwindled in terms of activity, having opened no loans over the past 6 months, with their last loan being opened **23.5** months ago

In terms of risk:

- They have an even lower DTI - 0.379
- Often delay their payments (66.9% cases). There's 16.2% chance of a delay of over 30 days, and 7% of them delay their payments, making this particular part of risk assessment almost as significant as that of VIP
- 40% of them have collateral and 58% have insurance, which is also similar to VIP.

Segment 4: At-risk / Recently inactive, yet acquainted

In spite of the similarities to Core and VIP, this group is an active risk to the bank and has to be tackled thoroughly.

One potential approach is to win back these clients by offering them due-day shifts and reducing interest. However, autopay should be required and no new credit should be approved until a pattern of on-time payments emerges. After “reactivation”, they should be incentivised to go after small, low-risk loans first.

Variables to pay attention to in this segment: recent activity over the last 60/90/180 days, average payment delay

Segment 5: Churn / Small wallet AND low engagement

Clients in this group, on average:

- 34.3 y.o. women with children, who have been at their new workplace for 51 days
- They get their salary sent to their bank account
- Their usage is nearly non-existent: ~2-3 credits, less than 1 of which is active (=> some of them do not have active accounts).
- They only pay 30.2 each month and successfully close less than half of their loans. Additionally, there's no recent activity whatsoever

In terms of risk:

- They have the lowest DTI - 0.342
- Rarely delay their payments (34.2% cases). There's only an 8.6% chance of a delay of over 30 days, and 6% of them delay their payments. While this is good, it could be linked to them having less payments overall
- Only 20.7% of them have collateral, but 61.5% have insurance

Segment 5: Churn / Small wallet AND low engagement

This is a segment which would prove to be barely profitable to pursue, but a few “nudges” could be attempted nonetheless.

The bank could attempt to reinitiate contact via push/ e-mail (possibly also SMS, but that may be too expensive). They could also offer things like fee-free periods and cashback. If none of that would reactivate the client, it would be best to limit communications to a minimum from then onwards.

Variables to pay attention to in this segment: activity over the last 60 days, recency of the latest transaction

Data Science and
Business Analytics

Client Segmentation

Customer Profiles

K-Means

Segment 0: Young & New

Clients in this group, on average:

- Tend to be young (31.8 y.o.) married women who have been at their new workplace for just over a month so far
- They receive salary to their bank account
- They use their account lightly, having 3 total credits, 1.5 of which are active
- They have low monthly payments of ~27, successfully close less than half of their loans

In terms of risk:

- They have a low-ish DTI of 0.376
- Delinquency is uncommon in this group: 35.5%. Moreover, there's only a 4.9% chance that they will not pay their interest for over a month
- Less than 1/5 of them have collateral, but 64.1% has insurance

Segment 0: Young & New

As can be seen, these clients are early-stage, low-revenue but already reasonably active and mostly lack risk. They bring average short-term value while having high future potential.

Therefore, the bank should focus on growth and habit formation: bundling (debit+credit), autopay activation, behavior-based limit ramps, simple offers. However, access to higher credit tiers should probably be restricted before an on-time payment streak emerges.

Variables to pay attention to in this segment: Active_to_All_prc, num_AccountActive180/90/60, sum_of_paym_months, first 30-day delay, DTI.

Segment 1: Growing middle segment with arrears

Clients in this group, on average:

- Tend to be middle-aged (35.6 y.o.) men with children, who have been at their new workplace for ~49 days by now
- They get salary sent straight to their bank account
- Their usage is medium: ~5 credits, 2 of which are currently active.
- They have moderate payments of ~63, successfully close 0.675 of their loans

In terms of risk:

- They have an average DTI of 0.381
- Delinquencies in terms of payments are semi-common - 58.9%. There's a 15.9% chance that their loans will remain unpaid for over a month
- A third of them has collateral, but two thirds have insurance

Segment 1: Growing middle segment with arrears

This group has a higher-than-average payment rate, yet already shows signs of delays here and there. Therefore, the bank should take measures to somehow minimise said delays.

E.g., align due dates with a client's pay day, require autopay for any change of limit, and convert existing balances to structured installments. Return access to some of the more sensitive (to delays) features after a streak of on-time payments.

Variables to pay attention to in this segment: avg_num_delay, 30/60/90-day delays (lifetime), num_AccountActive180, sum_of_paym_months, DTI.

Segment 2: Average risk, High reward

Clients in this group, on average:

- Older middle aged (~37.2 y.o.) men with children, who have been at their new workplace for ~49 days
- They get salary via a bank form
- Their usage is extensive: ~8.48 credits, 3 of which are active.
- Incredibly, they pay ~138 each month, successfully close 1.11 of their loans, making them highly rewarding for the bank

In terms of risk:

- They have one of the highest DTI out of the segments - 0.410
- But they delay their payments in 77.2% of cases, with a worrisome 18.8% chance that they will delay a payment for over a month
- 44.3% of them have collateral, 60% has insurance

Segment 2: Average risk, High reward

This group drives seriously strong revenue for the bank, yet carries over the problem of heightened risk.

The bank should, thus, balance increasing value and control: offer bundles while simultaneously introducing dynamic limits and early warnings for emerging arrears, as well as leveraging collateral, which this group has more of than others.

Variables to pay attention to in this segment: sum_of_paym_months, all_credits / numAccountActiveAll, 30/60/90-day delays, dlq_exist, DTI.

Segment 3: Disengaged legacies

Clients in this group, on average:

- Are older (the oldest of all segments), 43.7 y.o. women with children, who have been at their new workplace for approximately 190 days (the longest)
- They obtain their salary via a bank form
- They, unlike every other segment, are employed in an “Other” industry instead of “Sales”
- Their usage is generally average: ~5 credits, just 2 of which are active
- They pay a moderate ~77 each month and successfully close 0.809 of their loans

In terms of risk:

- They have a low DTI - 0.38
- Sometimes delay their payments, sometimes do not (50.5% cases). There’s only a 9% chance of a delay of over 30 days
- A third of them has collateral, two thirds have insurance

Segment 3: Disengaged legacies

While this group provides higher than average revenue to the bank and is the most experienced, its engagement is not as high as the bank would probably like it to be.

It could, for example, try to reactivate these accounts with targeted offers and cashback. Then, once these account have been reactivated to a more appropriate degree, other methods could be employed to generate more revenue.

Variables to pay attention to in this segment: num_AccountActive180/90/60, Active_to_All_prc, sum_of_paym_months, first 30-day delay, DTI.

Segment 4: High-risk power users

Clients in this group, on average:

- Mature (40.4 y.o.) men with children, who have been at their new workplace for 70 days
- They get their salary sent to their bank account
- Their usage is the largest between segments: ~12-13 credits, 5 of which are active.
- They pay a whopping 257 each month and successfully close 1.275 of their loans.

In terms of risk:

- They have the actual highest DTI - 0.432
- Incredibly often delay their payments (91.3% cases). There's a serious **1 in 5** chance of a delay of over 30 days
- Over half of them have both collateral and/or insurance

Segment 4: High-risk power users

This is both an extremely good and an extremely bad segment for the bank. On the one hand, it brings in the most revenue. On the other hand, these clients almost never bring in the money on time.

This is not good for the bank, as it needs loan money to manage other items in its portfolio, and these payments are by far the most impactful. Therefore, the bank should most likely tighten the restrictions, aggressively warning the client of late payments and decreasing the level of tolerance to said late payments.

Variables to pay attention to in this segment: sum_of_paym_months, all_credits / numAccountActiveAll, 30/60/90-day delays (lifetime), dlq_exist, DTI

Conclusion

Based on personal observations, it seems as if out of the two segmentation methods, RFM is better for quickly assessing business needs and categorising large groups of customers into broad segments, whereas K-Means is better at finding peculiar, interesting (and possibly uneven) segments of customers. Ideally, if both had to be used in some real task, it would make the most sense to separate customers by RFM and then look for subsegments in each group using K-Means.


HSE UNIVERSITY
EST 1992