Customer segmentation report

made by Egor Gladilin

March 4, 2024

Contents

EDA

Data description

1	Feature name	Description	Type
2	ID	Client ID	Categorical
3	INCOME_BASE_TYPE	Income verification	Categorical
4	CREDIT_PURPOSE	Purpose of the loan	Categorical
5	INSURANCE_FLAG	Borrower's insurance when receiving a loan	Categorical
6	DTI	debt-to-income ratio - the ratio of debt to income	Numerical
7	SEX	Floor	Categorical
8	FULL_AGE_CHILD_NUMBER	Number of years of the child	Numerical
9	DEPENDANT_NUMBER	Number of dependents	Numerical
10	EDUCATION	Education	Categorical
11	EMPL_TYPE	Position	Categorical
12	EMPL_SIZE	Salary	Categorical
13	BANKACCOUNT_FLAG	The number of accounts the client has. (0 - no online account, 1 - there is one online account, 2 or more - accessed the online account from another device)	Categorical
14	Period_at_work	Working time (number of days)	Numerical
15	age	Age	Numerical
16	EMPL_PROPERTY	Employer business area	Categorical
17	EMPL_FORM	Organizational and legal form	Categorical
18	FAMILY_STATUS	Family status	Categorical
19	max90days	number of requests to credit bureaus in the last 90 days	Numerical
20	max60days	number of requests to credit bureaus in the last 60 days	Numerical
21	max30days	number of requests to credit bureaus in the last 30 days	Numerical
22	max21days	number of requests to credit bureaus in the last 21 days	Numerical
23	max14days	number of requests to credit bureaus in the last 14 days	Numerical
24	avg_num_delay	Average number of payment delays	Numerical
25	if_zalog	Presence of collateral (apartment, car)	Categorical
1	Feature name	Description	Type
26	num_AccountActive180	number of active accounts accounts for the last 180 days	Numerical
27	num_AccountActive90	number of active accounts accounts in the last 90 days	Numerical
28	num_AccountActive60	number of active accounts accounts in the last 60 days	Numerical
29	Active_to_All_prc	ratio of active accounts to all accounts	Numerical
30	numAccountActiveAll	number of open accounts	Numerical
31	numAccountClosed	number of closed accounts	Numerical
32	sum_of_paym_months	amount of payments for the last month (thousand)	Numerical
33	all_credits	Number of credits	Numerical
34	Active_not_cc	Active credit accounts but no credit card	Numerical
35	own_closed	Number of closed loans	Numerical
36	min_MnthAfterLoan	the minimum number of months that have passed since the last loan was taken, that is, how long ago the last loan was issued to the client	Numerical
37	max_MnthAfterLoan	number of months past since the date of the first loan	Numerical
38	dlq_exist	currently in arrears	Categorical
39	thirty_in_a_year	overdue more than 30 days in the last year	Categorical
40	sixty_in_a_year	overdue more than 60 days in the last year	Categorical
41	ninety_in_a_year	overdue more than 90 days in the last year	Categorical
42	thirty_vintage	overdue more than 30 days, ever	Categorical
43	sixty_vintage	overdue more than 60 days, ever	Categorical
44	ninety_vintage	overdue more than 90 days, ever	Categorical

Figure 1: The description. Categorical features are marked as orange and numerical as blue

Overall one can say that we have a comprehensive dataset, that describes credit story of each customer in details. Thus we have 20 categorical and 23 numerical features (43 features overall). The number of observations is 10243.

Feature name	unique values
INCOME_BASE_TYPE	Поступление зарплаты на счет,2НДФЛ,Форма банка (без печати работодателя),Свободная форма с печатью работодателя
CREDIT_PURPOSE	Покупка недвижимости/ строительство,Покупка автомобиля,Ремонт,Покупка земли,Отпуск,Обучение,Покупка мебели,Покупка бытовой техники,Другое,Лечение
INSURANCE_FLAG	1,0
SEX	мужской,женский
EDUCATION	среднее,высшее,Высшее/Второе высшее/Ученая степень,второе высшее,среднее-специальное,незаконченное высшее,Неполное среднее,ученая степень
EMPL_TYPE	рабочий,менеджер среднего звена,менеджер высшего звена,вспомогательный персонал,специалист,торговый представитель,другое,менеджер по продажам,страховой агент
EMPL_SIZE	>=200,>250,>=100,>100,< 50,>=150,>=50
BANKACCOUNT_FLAG	1.0,0.0,3.0,4.0
EMPL_PROPERTY	Производство,Информационные технологии,Другое,Торговля,Государственная служба,Транспорт,Наука,Финансы,Туризм,Строительство,Сельское и лесное хозяйство,Юридические услуги
EMPL_FORM	ООО,ОАО,Индивидуальный предприниматель,ЗАО,Иная форма,Государственное предприятие
FAMILY_STATUS	холост / не замужем,женат / замужем,гражданский брак,разведен / разведена,повторный брак,вдовец / вдова
if_zalog	0.0,1.0
dlq_exist	0.0,1.0
thirty_in_a_year	0.0,1.0
sixty_in_a_year	0.0,1.0
ninety_in_a_year	0.0,1.0
thirty_vintage	0.0,1.0
sixty_vintage	0.0,1.0
ninety_vintage	0.0,1.0

	count	mean	std	min	25%	50%	75%	max
DTI	10121.0	0.385248	0.135915	0.0	0.280000	0.400000	0.490000	0.6
FULL_AGE_CHILD_NUMBER	10243.0	0.554330	0.785071	0.0	0.000000	0.000000	1.000000	14.0
DEPENDANT_NUMBER	10243.0	0.003710	0.073851	0.0	0.000000	0.000000	0.000000	2.0
Period_at_work	7923.0	66.374479	67.465261	6.0	21.000000	45.000000	87.000000	966.0
age	7923.0	36.324751	8.612645	23.0	29.000000	35.000000	43.000000	62.0
max14days	3921.0	0.523591	1.036584	0.0	0.000000	0.000000	1.000000	15.0
max21days	3921.0	0.635297	1.158869	0.0	0.000000	0.000000	1.000000	15.0
max30days	3921.0	0.850548	1.328305	0.0	0.000000	0.000000	1.000000	15.0
max60days	3921.0	1.140780	1.593697	0.0	0.000000	1.000000	2.000000	18.0
max90days	3921.0	1.585820	1.878678	0.0	0.000000	1.000000	2.000000	18.0
avg_num_delay	3665.0	0.057312	0.103724	0.0	0.000000	0.013889	0.067757	1.0
num_AccountActive60	3676.0	0.094124	0.342632	0.0	0.000000	0.000000	0.000000	4.0
num_AccountActive90	3676.0	0.158324	0.431373	0.0	0.000000	0.000000	0.000000	4.0
num_AccountActive180	3676.0	0.378672	0.679914	0.0	0.000000	0.000000	1.000000	5.0
Active_to_All_prc	3676.0	0.423867	0.288988	0.0	0.222222	0.400000	0.600000	1.0
numAccountActiveAll	3676.0	2.216540	1.682049	0.0	1.000000	2.000000	3.000000	13.0
numAccountClosed	3676.0	3.560664	3.232971	0.0	1.000000	3.000000	5.000000	25.0
sum_of_paym_months	3676.0	83.880033	72.421375	0.0	31.000000	64.000000	117.250000	460.0
all_credits	3676.0	5.777748	4.111024	1.0	3.000000	5.000000	8.000000	27.0
Active_not_cc	3676.0	1.081066	1.052867	0.0	0.000000	1.000000	2.000000	7.0
own_closed	3676.0	0.710555	1.049267	0.0	0.000000	0.000000	1.000000	7.0
min_MnthAfterLoan	3676.0	14.190968	15.302198	-1.0	4.000000	10.000000	18.000000	107.0
max_MnthAfterLoan	3676.0	61.128400	30.323424	0.0	35.000000	66.000000	87.000000	171.0

Figure 2: Unique values of categorical features excluding NaNs and description of numerical features

Missing values

	num_of_unqiue	unique_percenage	num_of_null	null_percentage	num_of_zero	zero_percentage
ID	10243.0	1.000000	0.0	0.000000	0.0	0.000000
INCOME_BASE_TYPE	5.0	0.000491	57.0	0.005565	0.0	0.000000
CREDIT_PURPOSE	10.0	0.000976	0.0	0.000000	0.0	0.000000
INSURANCE_FLAG	2.0	0.000195	0.0	0.000000	4104.0	0.400664
DTI	62.0	0.006126	122.0	0.011911	1.0	0.000099
SEX	2.0	0.000195	0.0	0.000000	0.0	0.000000
FULL_AGE_CHILD_NUMBER	7.0	0.000683	0.0	0.000000	6154.0	0.600801
DEPENDANT_NUMBER	3.0	0.000293	0.0	0.000000	10214.0	0.997169
EDUCATION	9.0	0.000879	0.0	0.000000	0.0	0.000000
EMPL_TYPE	10.0	0.000977	9.0	0.000879	0.0	0.000000
EMPL_SIZE	9.0	0.000889	123.0	0.012008	0.0	0.000000
BANKACCOUNT_FLAG	5.0	0.000631	2320.0	0.226496	6169.0	0.778619
Period_at_work	360.0	0.045437	2320.0	0.226496	0.0	0.000000
age	41.0	0.005175	2320.0	0.226496	0.0	0.000000
EMPL_PROPERTY	13.0	0.001641	2320.0	0.226496	0.0	0.000000
EMPL_FORM	7.0	0.001754	6253.0	0.610466	0.0	0.000000
FAMILY_STATUS	7.0	0.001754	6253.0	0.610466	0.0	0.000000
max90days	20.0	0.005101	6322.0	0.617202	1082.0	0.275950
max60days	19.0	0.004846	6322.0	0.617202	1558.0	0.397348
max30days	17.0	0.004336	6322.0	0.617202	1988.0	0.507014
max21days	17.0	0.004336	6322.0	0.617202	2377.0	0.606223
max14days	16.0	0.004081	6322.0	0.617202	2589.0	0.660291
avg_num_delay	1126.0	0.307231	6578.0	0.642195	1595.0	0.435198
if_zalog	3.0	0.000816	6567.0	0.641121	2462.0	0.669750
num_AccountActive180	7.0	0.001904	6567.0	0.641121	2619.0	0.712459
num_AccountActive90	6.0	0.001632	6567.0	0.641121	3179.0	0.864799
num_AccountActive60	6.0	0.001632	6567.0	0.641121	3377.0	0.918662
Active_to_All_prc	97.0	0.026387	6567.0	0.641121	479.0	0.130305
numAccountActiveAll	13.0	0.003536	6567.0	0.641121	461.0	0.125408
numAccountClosed	23.0	0.006257	6567.0	0.641121	417.0	0.113439
sum_of_paym_months	337.0	0.091676	6567.0	0.641121	12.0	0.003264
all_credits	28.0	0.007617	6567.0	0.641121	0.0	0.000000
Active_not_cc	9.0	0.002448	6567.0	0.641121	1236.0	0.336235
own_closed	9.0	0.002448	6567.0	0.641121	2115.0	0.575354
min_MnthAfterLoan	102.0	0.027748	6567.0	0.641121	113.0	0.030740
max_MnthAfterLoan	132.0	0.035909	6567.0	0.641121	8.0	0.002176
dlq_exist	3.0	0.000816	6567.0	0.641121	1606.0	0.436888
thirty_in_a_year	3.0	0.000816	6567.0	0.641121	3132.0	0.852013
sixty_in_a_year	3.0	0.000816	6567.0	0.641121	3354.0	0.912405
ninety_in_a_year	3.0	0.000816	6567.0	0.641121	3431.0	0.933351
thirty_vintage	3.0	0.000816	6567.0	0.641121	3566.0	0.970076
sixty_vintage	3.0	0.000816	6567.0	0.641121	3623.0	0.985582
ninety_vintage	3.0	0.000816	6567.0	0.641121	3628.0	0.986942

Figure 3: The number and percentage of missing values for each feature

Looking at the tables below, one can sees that many features has the same number of missing values and the percentage is over 60%. My hypothesis is that, those people who has missing values in feature *all credits*, simply have never had a credit. And the graph below proves it, there is no ”0” value and I assume that in this case NaN value is sames as ”0”.

all_credits	
2.0	475
3.0	460
4.0	450
5.0	390
1.0	342
6.0	335
7.0	236
8.0	212
9.0	174
10.0	140
11.0	112
12.0	81
13.0	66
14.0	49
15.0	40
16.0	31
18.0	16
17.0	16
19.0	13
20.0	11
22.0	6
23.0	6
24.0	6
21.0	5
25.0	2
27.0	1
26.0	1

Figure 4: Values of *all credits*

So, I can conclude that in our initial data there are 2 major groups. People how took a loan at least once and had a credit account and those who has never did it. So I think it would be reasonable to prepare and explore datasets for both groups separately. Further I will remove all NaNs in categorical features and substitute NaNs for numerical with statical values.

Remove and fill missing and incorrect values

After removing missing data for *all credits*, there is also no missing data in all categorical, at the same the number of observations was reduced down to 3676, that is around 40% from initial data. Although we still have several missing values in numerical features, we can conclude that we exactly removed all cutomers who haven't used credit accounts, as *all credits* ranges from 1 to 28 (in our data). I have also found a incorrect value **n.a.** in several features, it is a text value which implies NaN, rows with this value were also removed.

	num_of_nulls
Номер варианта	0
ID	0
INCOME_BASE_TYPE	0
CREDIT_PURPOSE	0
INSURANCE_FLAG	0
DTI	0
SEX	0
FULL_AGE_CHILD_NUMBER	0
DEPENDANT_NUMBER	0
EDUCATION	0
EMPL_TYPE	0
EMPL_SIZE	0
BANKACCOUNT_FLAG	0
Period_at_work	0
age	0
EMPL_PROPERTY	0

EMPL_FORM	0
FAMILY_STATUS	0
max90days	2
max60days	2
max30days	2
max21days	2
max14days	2
avg_num_delay	11
if_zalog	0
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
dlq_exist	0
thirty_in_a_year	0
sixty_in_a_year	0
ninety_in_a_year	0
thirty_vintage	0
sixty_vintage	0
ninety_vintage	0

	num_of_*n.a.*
ID	0
INCOME_BASE_TYPE	0
CREDIT_PURPOSE	0
INSURANCE_FLAG	0
DTI	0
SEX	0
FULL_AGE_CHILD_NUMBER	0
DEPENDANT_NUMBER	0
EDUCATION	16
EMPL_TYPE	0
EMPL_SIZE	1

Figure 5: Missing values after update. Incorrect values.

Next, we're going to apply data imputation methods to deal with remaining missing values in numerical feature. Common ways is to substitute NaNs with mean or median value of the feature. As for categorical features there are no remaining ones with missing values, thus we will keep working with them a bit later. As we can see the distributions below they are skewed to the left and mean and median do not coincide, but as we have insignificant number of NaNs the choice of imputation approach won't influence the distributions much, so lets fill NaNs with mean values.

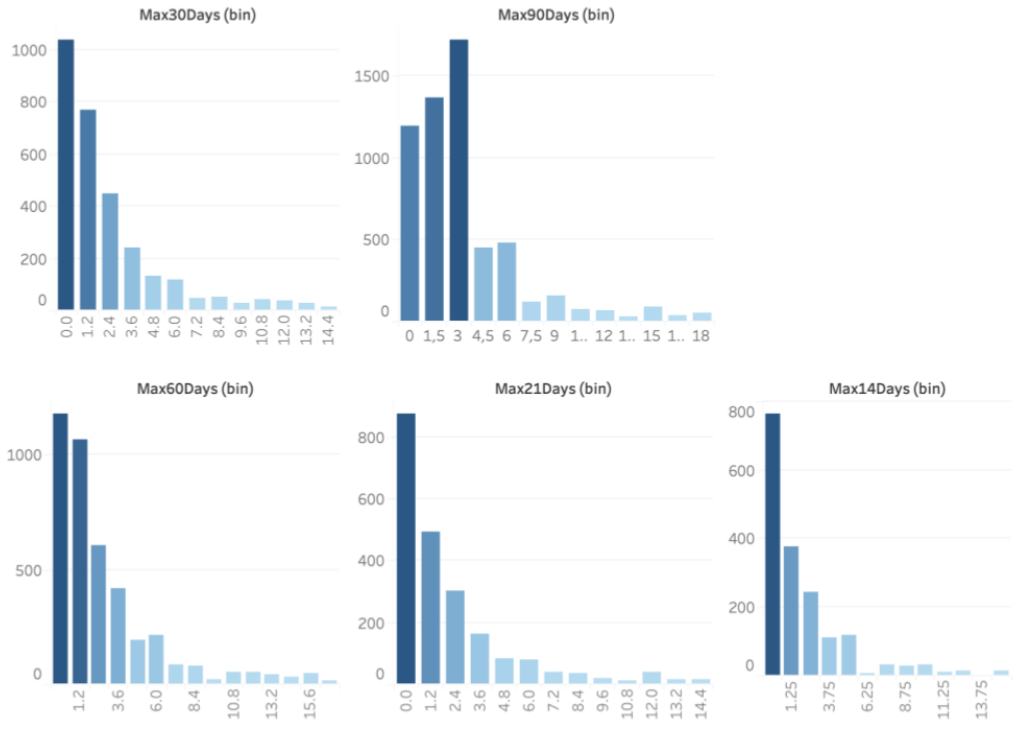


Figure 6: Feature distributions with missing values.

Customer portrait

Figure 7: Updated categorical features.

Based on the histogram above we can roughly describe a portrait of the customer who is likely to take a loan or use a credit account. It is person who works in retail sphere as Specialist in LLC (Limited Liability Company) and earns over 250k rubles. The person is likely to have a bachelor degree and unlikely to have children and dependants, he or she is likely to be married, but also can be single and take a loan to make a renovation. The customer do not leave a deposit and prefer to take insurance on a loan. As calculations show the average age of the customer is 36 years and do not overdue loan payments.

Correlations

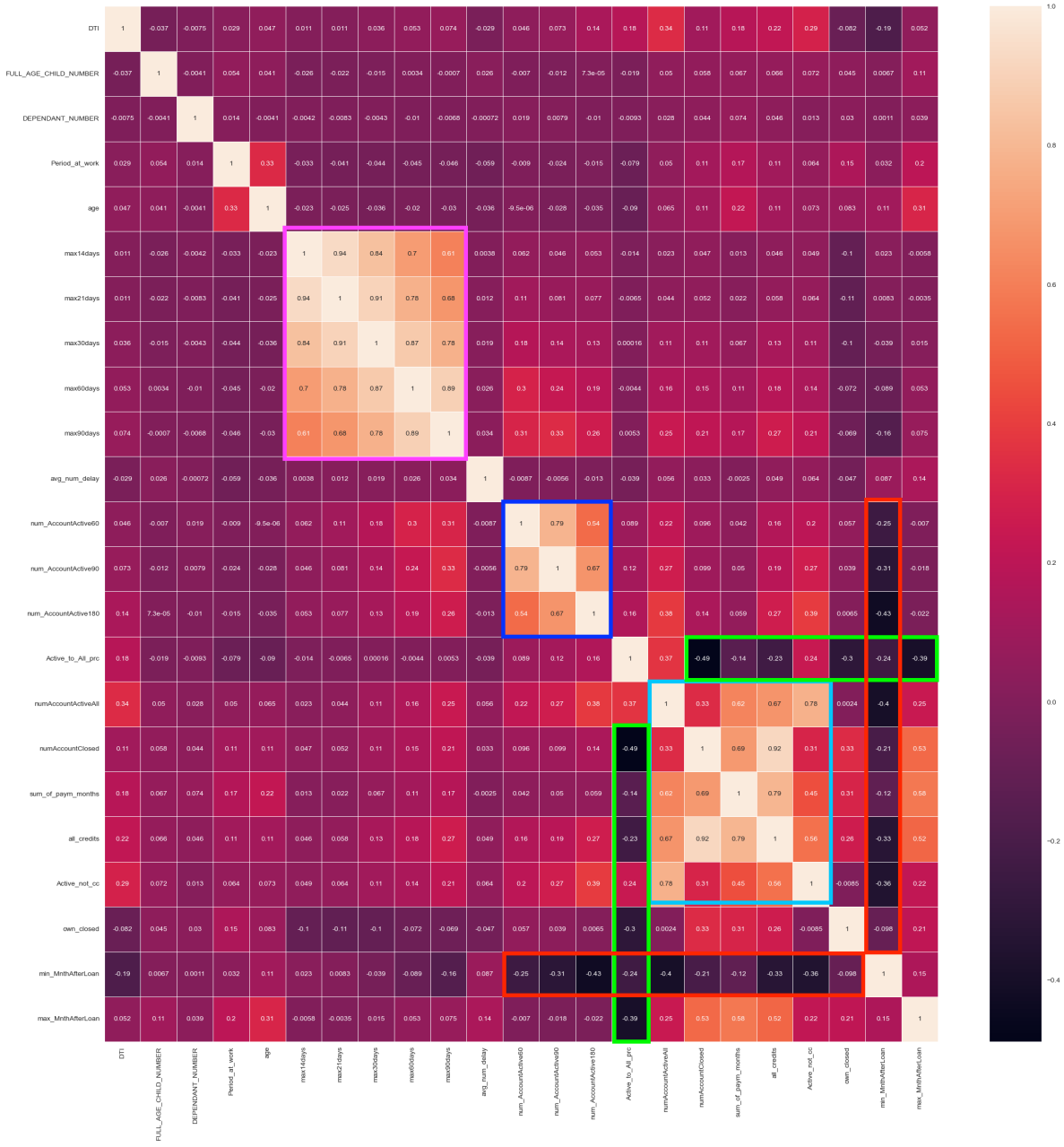


Figure 8: Updated categorical features.

Now lets have a look at the correlation matrix and try to interpret positive and negative correlations of bunch of features. We can easily see 4 major groups of features that have correlations over 0.65 and up to 0.95 and 2 features that are anti-correlated with a set of other ones. All other ones ranges from -0.4 to 0.4, but probably have non-linear dependencies.

Positive correlations

1. max14days, max21days, max30days, max60days, max90days:
These feature refers to the number of days since the last request to the credit bureaus. Common sense helps us to explain the dependency. As the feature depends on time, features with smaller time-delta will have closer values, thus higher correlation.

2. num AccountActive60, num AccountActive90, num AccountActive180:
This set of features indicates the number of active accounts over the last n-days.As features in pink box, these ones have a similar nature and depends on time, thus through the time they have close values what provide high correlations.

4. numAccountActiveAll, sum of paym months, all credits, active not cc, numAccountClosed:
-*numAccountActiveAll* and *active not cc* are really similar, the only difference is that *active not cc* only considers active accounts without credit cards, thus as this feature grows, *numAccountActiveAll* grows as well. Similarly, to this *all credits* is highly correlated with *numAccountActiveAll*, because any credit is linked to an account, so these features grows simultaneously. By the same logic, *sum of paym months* increases, with growth of your active account your monthly payments rises.
-As was mentioned before, *numAccountActiveAll* and *active not cc* have similar nature, that is why *active not cc* and *sum of paym months* are highly correlated.
-High correlation between *all credits* and *numAccountClosed* is also obvious. All open credits will be inevitably closed, so with growth of 1st feature we have growth of the 2nd. The correlation between *numAccountClosed* and *sum of paym months* is clear. Since open new account, our monthly payments rises and right after closing the account our payments goes down and number of closed accounts increases. So, they increases together, but with a short delay.

Negative correlations

1. numAccountClosed, all credits, own closed, max MnthAfterLoan, Active to All prc:
-Obviously the more closed accounts a customer has the less percentage of active ones.
- Feature *own closed* refers to closed loans, not closed credit accounts. But we can assume that a customer is less likely to open credit account having loans and that is why the ratio of active accounts to all does not increase and we have negative correlation.
- As the number of months since first loans grows (*max MnthAfterLoan*) the ratio of active accounts is likely to decrease, as a customer still takes new loans, thus we have negative correlation.

2. min MnthAfterLoan, num AccountActive60, num AccountActive90, num AccountActive180, numAccountActiveAll, all credits, own closed:
- Feature *own closed* refers to closed loans, not closed credit accounts. But we can assume that a customer is less likely to open credit account having loans and that is why the ratio of active accounts to all does not increase and we have negative correlation.
- As the number of months since first loans grows (*max MnthAfterLoan*) the ratio of active accounts is likely to decrease, as a customer still takes new loans, thus we have negative correlation.

Encode categorical features

Ultimately, we have 23 numerical features and 19 numerical ones, excluding *ID*, as this feature has 100% unique values it is not informative for us.
8 categorical features already have binary value, and it remains to encode other 11 ones. Lets a closer look at all of them, one can notice that we have 2 ordinal feature: *BANKACCOUNT FLAG* and *EDUCATION*.
-According to description of *BANKACCOUNT FLAG*, values 2 or greater means the same, thus we can substitute 3 and 4 with 2.
-Values of *EDUCATION* are of text type and we can logically assume that the higher degree a customer has, the more likely it will get a loan or credit account. So lets encode the feature in the following way:

- 1 - ”Неполное среднее”
- 2 - ”среднее”
- 3 - ”среднее-специальное”
- 4 - ”незаконченное высшее”
- 5 - ”высшее”
- 6 - ”второе высшее”
- 7 - ”ученая степень”

. Let’s encode ”Высшее/Второе высшее/Ученая степень” as 6 - mean value of 5, 6, 7.
Thus, we are left with 9 features: *INCOME BASE TYPE*, *CREDIT PURPOSE*, *EMPL TYPE*, *EMPL SIZE*, *EMPL PROPERTY*, *EMPL FORM*, *FAMILY STATUS*. Let’s apply standard *One-hot-encoding* approach to them. Thus, we get dataset of the shape (3676, 77).

Standartization

Finally, we apply a standardisation method to all numerical feature to make the same of the same range.Many machine learning algorithms are sensitive to the scale of the input features. Standardizing features ensures that they are on a similar scale, preventing certain features from dominating others simply due to their larger magnitude. For our purposes standardisation is important because algorithms that rely on distances between data points, such as k-nearest neighbors or support vector machines, can be influenced by the scale of features. Standardization helps to ensure that distances are computed accurately.We will use a basic Standard-scaling method, as many features a skewed to the right MinMax-scaling will take into account outliers (max values) and won’t consider feature distributions, which is not effective.

Segmentation

Clustering

Dimension reduction

In the context of cluster visualization, tSNE (t-distributed stochastic neighbor embedding) and PCA (Principal Component Analysis) can both be used, but they have different features that should be considered depending on the specific task.

- tSNE aims to preserve the relative distances between points in the original space. This means that close objects in the source data must remain close after being projected into a smaller space.
- tSNE uses a probabilistic neighborhood representation, which determines the probability that two points will be neighbors in a space of a smaller dimension. These probabilities are calculated based on the Gaussian distribution around each point.
- PCA calculates the main components, which represent the directions in the feature space along which the data has the greatest variance.
- PCA performs a linear transformation of the data, transforming it from the original feature space into a new space where the axes represent the main components.

Since our data is non-linear and we aim to preserve the shape of clusters we will use tSNE to reduce data dimension and then will apply clusteting models.

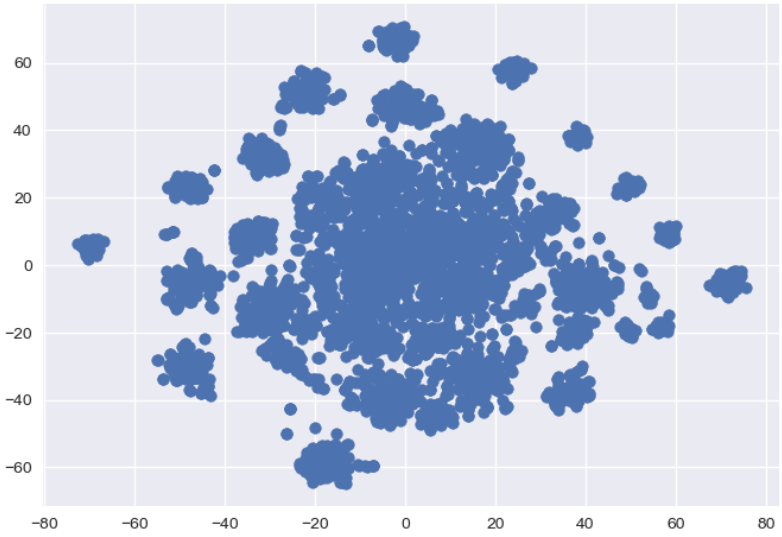


Figure 9: tSNE reduction (perplexity = 30)

On the plot we can clearly see many visualised clusters, tSNE helped us to plot both large and small ones on 2-dimension plane. Looking at plot we can generally identify one big cluster in the middle, several medium ones close to it and many small clusters along the edges.

Clustering algorithms

So, we aim to find the clustering algorithm that would comprehensively describe the data plotted on the previous page. Actually, the first idea that is obviously comming is to use clustering algorithms that are based on density estimation like DBSCAN or OPTICS. Indeed these methods works well with our data and identify all large and small clusters. So, from the plot it is seen

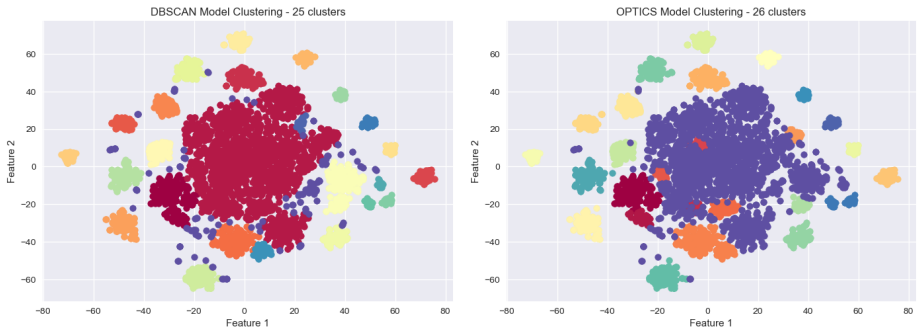


Figure 10: DBSCAN and OPTICS clustering

that both algorithm were capable to identify the big cluster in the middle and all other medium and small ones, nevertheless there are many points that were not identified as members of any class, OPTICS left significantly more such data points than DBSCAN. In fact even of we remove non-labeled points in DBSCAN, the problem is that we comes up with over than 30 ones. We would like to use the algorithms that would be able to generalise the clusters, thus reduce their number.
So, I applied 4 other well known clustering methods, one can observe the results below.

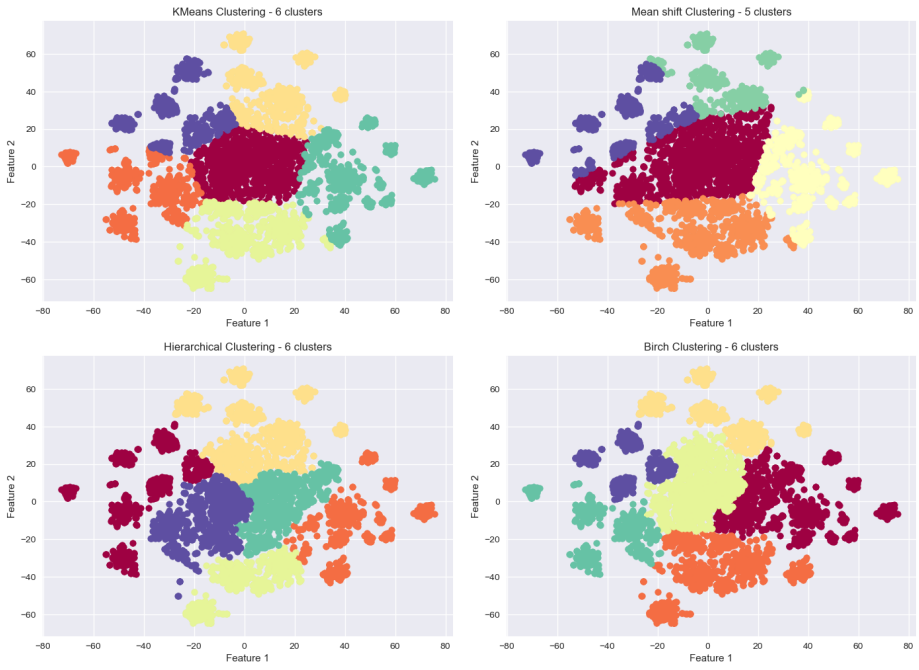


Figure 11: KMeans, MeanShift, Hierarchical and Birch clusterings

Finally, it was chosen to use KMeans clustering (0-orange cluster, 1-purple cluster, 2-red cluster, 3-yellow cluster, 4-green cluster, 5-light green cluster). So, after clustering it is time to choose the best algorithm. In the experiments, the common parameters for every method were tuned and the best were found. There were faced the problems in all cases: either the generalisation ability was poor and split was pretty weird or it was found too small or too big number of clusters. Ultimately, it was decided to choose kmeans clustering as it deals most effectively. It identifies big cluster in the middle and unite small clusters with medium ones making 4 other cluster around the central one. So, now we are left with choosing the number of clusters. As one can see it was chosen 5 clusters with the help of elbow and silhouette methods.

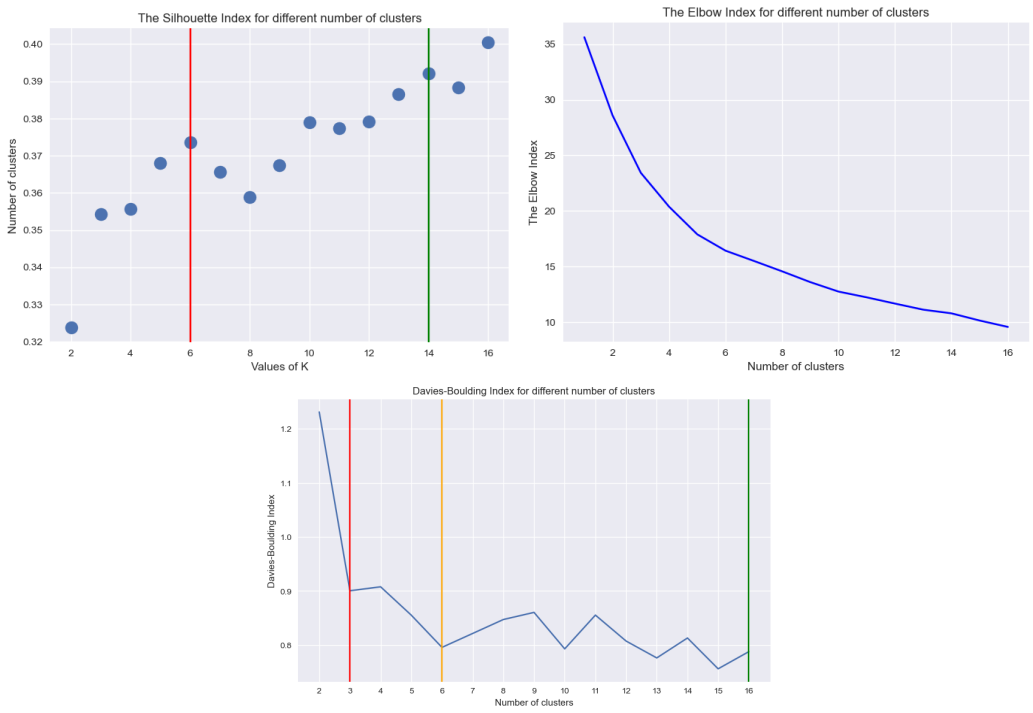


Figure 12: Silhouette and elbow methods for KMeans

Initially it was chosen to use the elbow index to determine the optimal number of clusters, however one can see that slope of the curve is smooth and it is hard to identify significant decline in the index value. So, as alternative way to determine number of clusters there were applied Davies-Boulding and Silhouette indexes. Looking at the silhouette plot above one can see several local maximums at the points 6, 14 and the values raises significantly after point 6. The Davies-Boulding index indicates several dramatic drop downs at points 4, 6, 16. So, it was chosen to take 6 clusters as larger number of clusters have weak generalization ability (it detects too much clusters), whereas 4 clusters do not cover enough information about the clusters.

1st cluster. Promising employees.

- Age: 34 - 35 years
- Sex: Male
- Job: Specialist / commercial representative
- Business sphere: retail
- Salary: over 250k rubles / under 50k rubes
- Avg number of credits:
- Avg number months since 1st loan: 58 months
- Avg payments for last month: 77.6k rubles

The first cluster mainly include male representatives of retail business, who can either have low or high salary. As we can see these people can occupy different position at companies. Also, we see that they have taken their first loan recently regarding other customers and spent significantly lower amount of money on their last payments. So we can conclude that these people relate either to wealthy or promising people and the goal of our bank is stimulate their purchases.

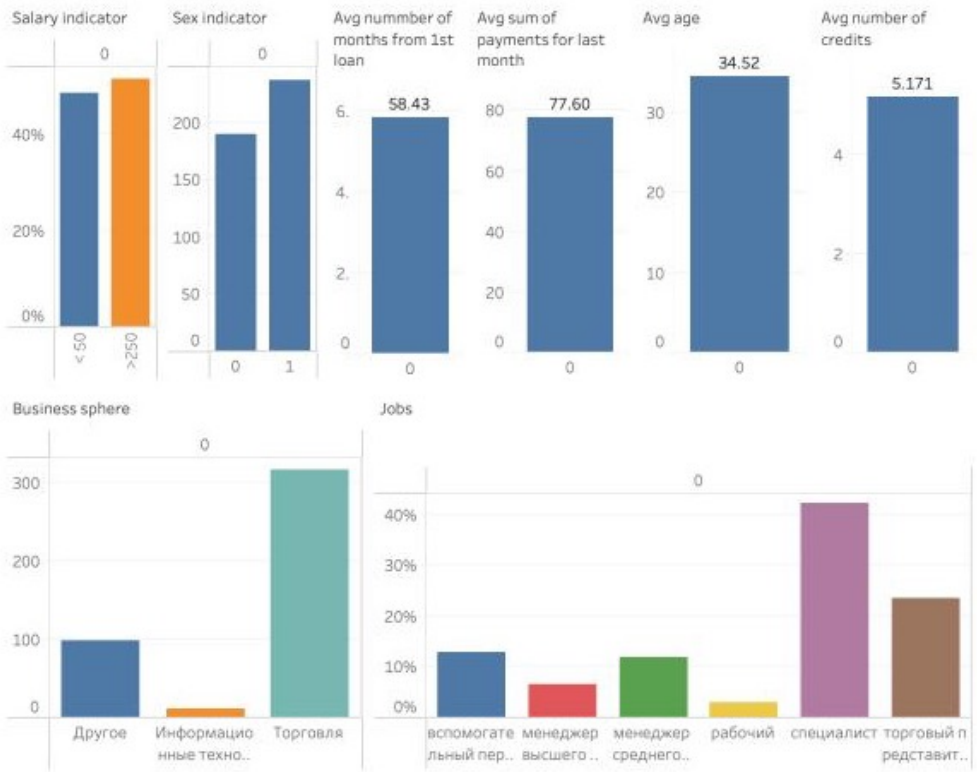


Figure 13: 1st cluster statistics

2nd cluster. Wealthy managers.

- Age: 34 - 35 years
- Sex: Male
- Job: Specialist
- Business sphere: IT / retail
- Salary: over 250k rubles
- Avg number of credits: 5
- Avg number months since 1st loan: 57.4 months
- Avg payments for last month: 76.25k rubles

The second cluster looks similar to the first one, as pattern of taking loans looks almost the same - first loan is taken recently and low money for last payment. However, this group of people are different in the portrait. First of all, most of them have high salary and they work other in IT or retail. We can also notice that the share of specialist and middle managers is higher and we also face high share of PR managers in this group. Thus, in this group most people can relate to wealthy managers, but not really active in taking loans. So, we should also encourage them to take more loans.

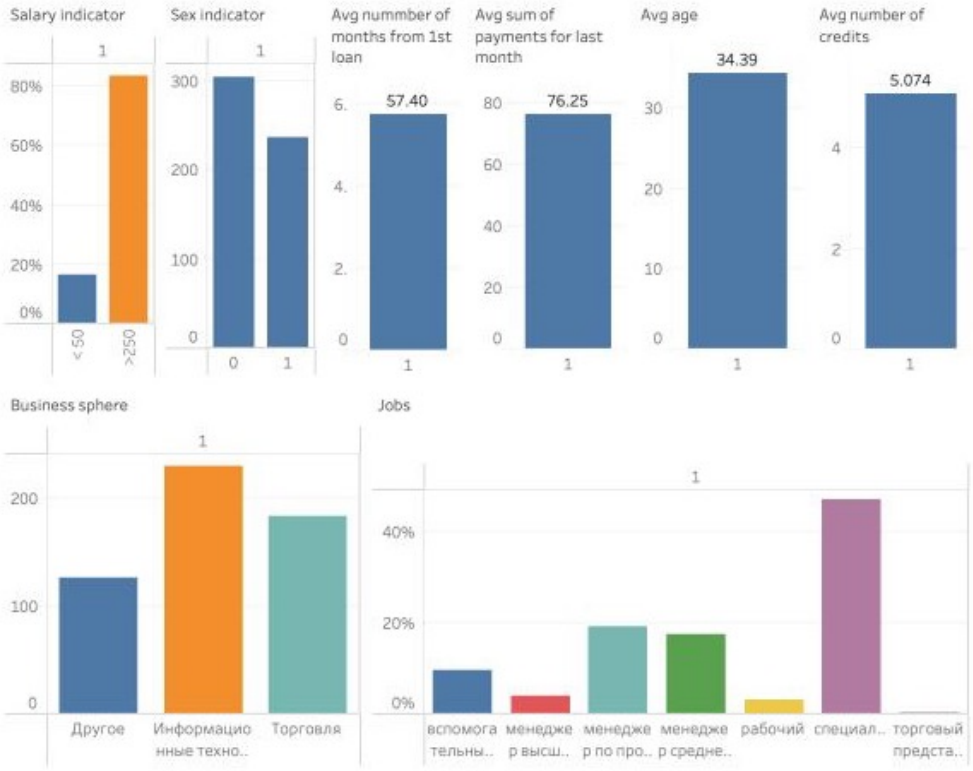


Figure 14: 2nd cluster statistics

3rd cluster. Wealthy females.

- Age: 34 - 35 years
- Sex: Female
- Job: specialist / staff member / middle manager
- Business sphere: over 250k rubles / under 50k rubles
- Salary: over 250k rubles
- Avg number of credits: 5
- Avg number months since 1st loan: 55 months
- Avg payments for last month: 72.4k rubles

This group is pretty similar to the previous one, it also include wealthy retail managers, who is not active into taking loans. The main difference is that majority of our customers are female. We can also notice that females who earn much work not only retail, but in other business spheres as well. Unexpectedly, we can see that significant share of women work as support staff, that differs much from first 2 clusters. To sum up this cluster contains wealthy female managers.

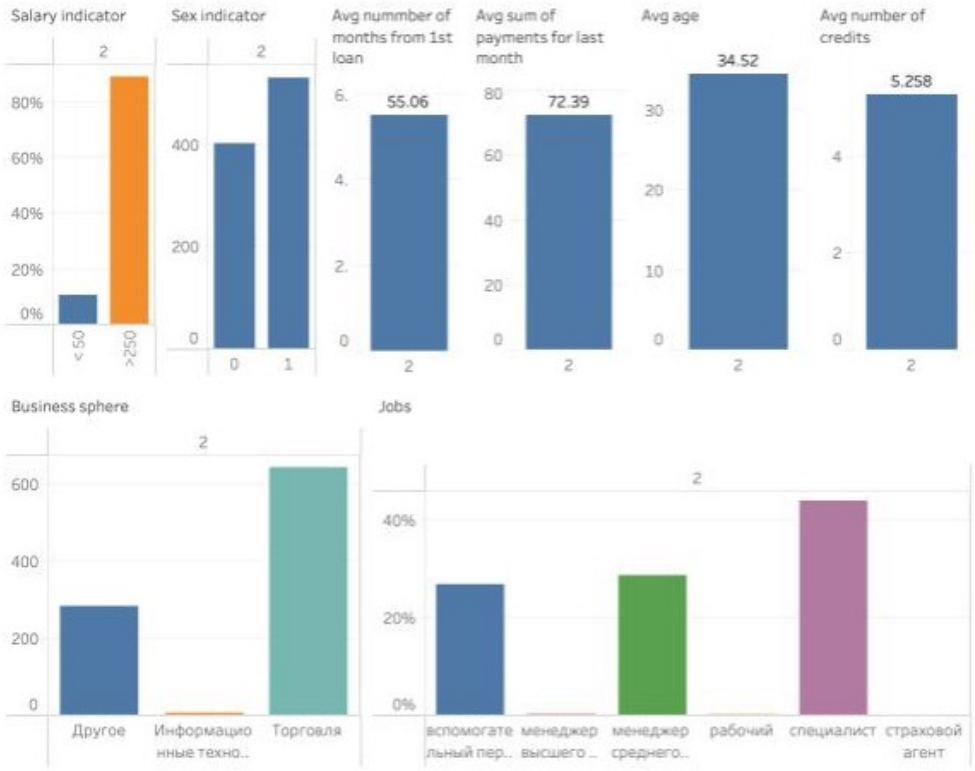


Figure 15: 3rd cluster statistics

4th cluster. Wealthy long-time customers.

- Age: 37
- Sex: Female / male
- Job: specialist / staff member / middle manager
- Business sphere: retail / other sphere
- Salary: over 250k rubles
- Avg number of credits: 7
- Avg number months since 1st loan: 72
- Avg payments for last month: 102.2k rubles

The 4th cluster looks similar to the 3rd one, it also includes wealthy manager, who works mainly in retail.However, we can observe that the average age of customers is 2 years older and the patterns of taking loans are much different. The average number of credits is higher by 2 credits, the number of months since first loans is 72 months and the amount of payments for the last months has grown up to 102.2k rubles. So, the group can described as cluster of wealthy retail employees who take loans for many years and spend much money on it. The strategy of our bank regarding these customers should be keeping their engagement.

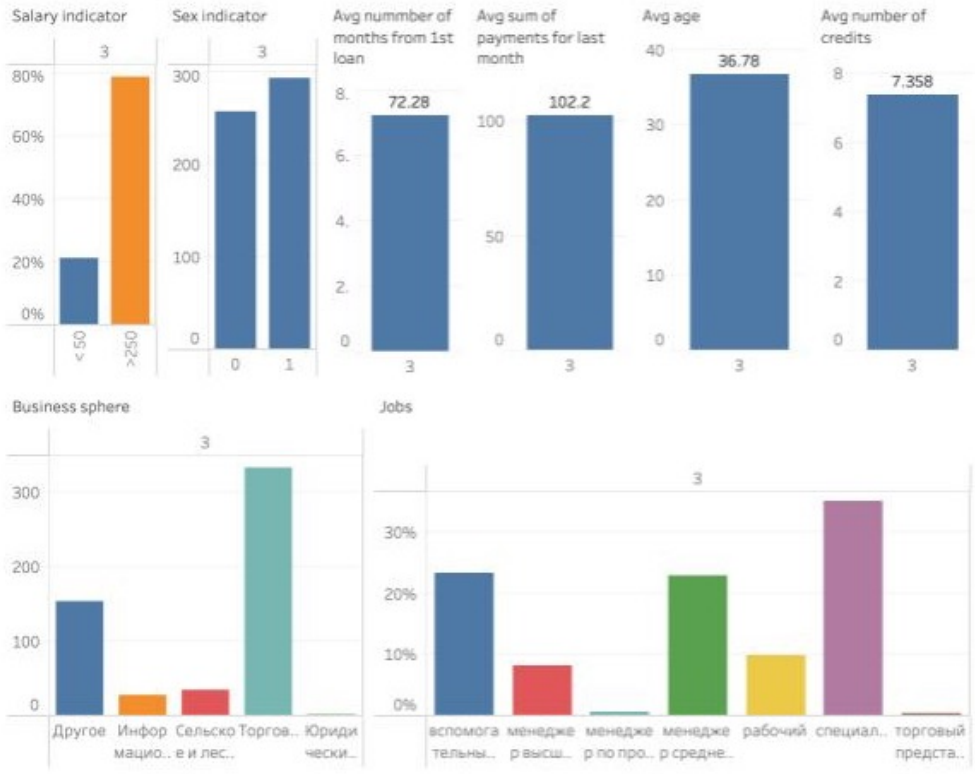


Figure 16: 4th cluster statistics

5th cluster. Much savings owners.

- Age: 39 years
- Sex: Male
- Job: retail / other sphere
- Business sphere: top or middle manager
- Salary: under 50k rubles / over 250k rubles
- Avg number of credits: 6 - 7 credits
- Avg number months since 1st loan: 68 months
- Avg payments for last month: 108k rubles

The 5th cluster also include customers who use our services for a long time, pay much for them and take many credits, but the portrait is different. This cluster includes mostly male top and middle managers in retail, however we see that many customer earn relatively little money. This can be explained by the fact that average age increased much thus we can assume that many customers already have significant savings which they spend on our services. So, we should keep these customers engaged as well.

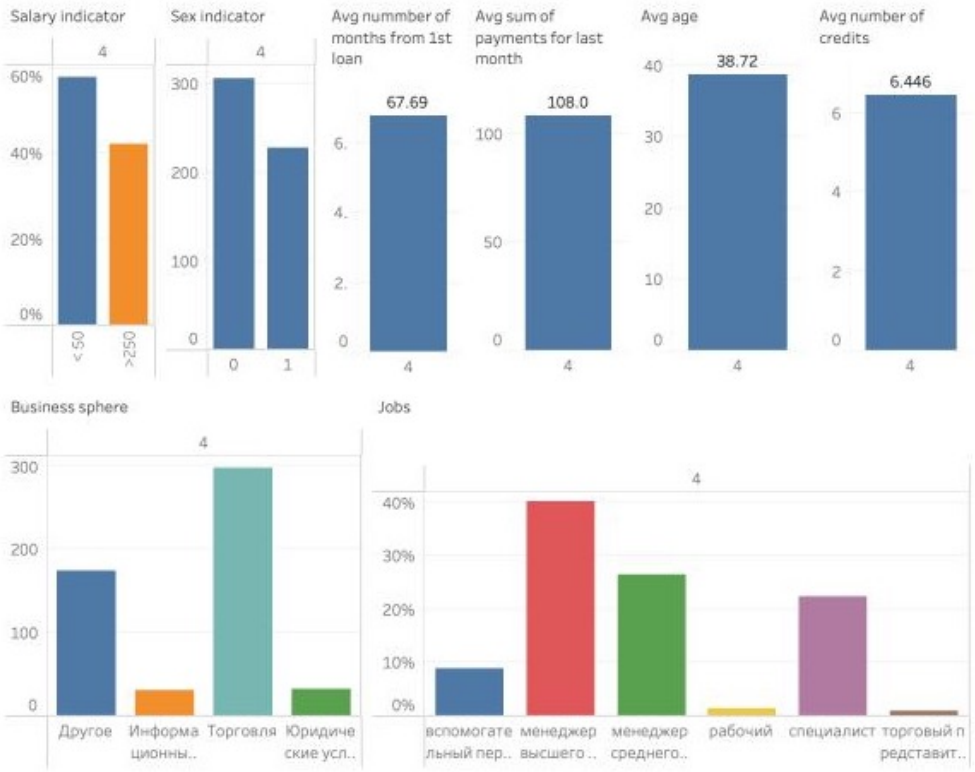


Figure 17: 5th cluster statistics

6th cluster. Workers and specialists.

- Age: 37 years
- Sex: Male
- Job: Specialist / worker
- Business sphere: Other sphere / retail
- Salary: over 250k rubles
- Avg number of credits: 5 - 6 credits
- Avg number months since 1st loan: 60 months
- Avg payments for last month: 76.32k rubles

The last cluster looks similar to the first 3 ones as these customers take loans recently and do not spend much, but the portrait differs much. These customers are older and works in different sphere as workers or specialists. The bank also aims to make these customers take more loans and spend more money.

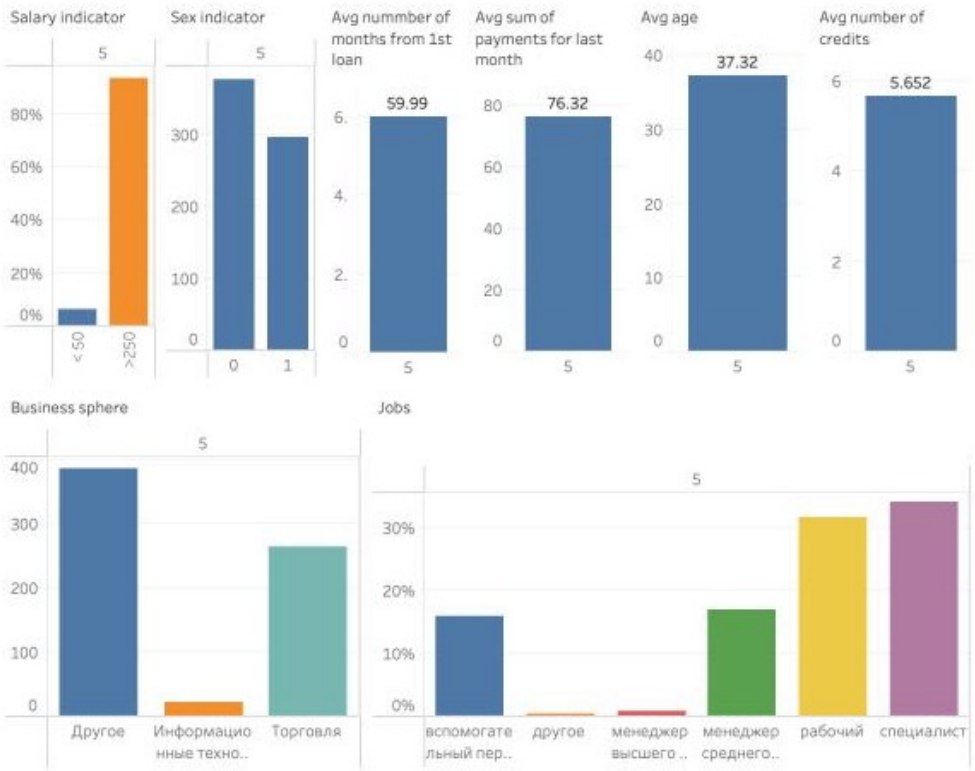


Figure 18: 6th cluster statistics

RFM-analysis

RFM analysis (Recency, Frequency, Monetary) is a method of analyzing customer behavior that is used to segment the customer base based on three key characteristics:

1. **Recency**: - Evaluates how recently the customer made his last purchase. The closer to the current moment, the higher the value of this parameter.
2. **Frequency**: - Determines how often the customer makes purchases over a certain period of time. This parameter evaluates how often a customer becomes your customer.
3. **Monetary (Monetary volume)**: - Measures the total amount of money that the client has spent over a certain period of time. This parameter reflects the monetary value of the customer for your business.

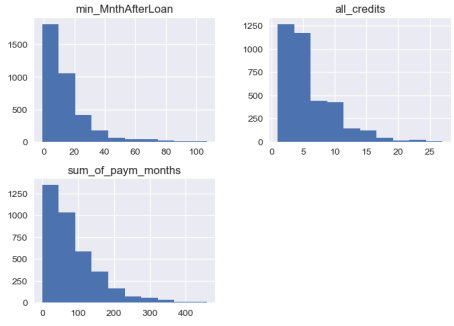
Advantages of RFM analysis:

1. **Customer segmentation**: - RFM analysis allows you to divide the customer base into segments depending on their behavior and business value. This can help in creating more targeted marketing strategies.
2. **Customer prioritization**: - Based on RFM analysis, it is possible to determine the priority of customers. For example, more valuable customers (the top segment) may be exposed to more personalized marketing efforts.
3. **Instant overview of the customer base**: - RFM analysis provides an instant overview of the customer base, allowing you to quickly identify the most active and valuable customers.

In general, RFM analysis is an effective tool for understanding and managing customer behavior, which can lead to increased effectiveness of marketing strategies and increased business profits. Each client needs to be assigned a score from 1 to n for each of the signs — recency (R) , frequency (F) and monetary (M). In our case, as we attemtpts to segment customers with regard to their credit activity:

- recency - min MnthAfterLoan,
- frequency - all credits
- monetary - sum of paym months

Next, accorfing to RFM, we have to split somehow values of R,F and M in several classes. To do this, there were calculated the quantiles for each feature.



	min_MnthAfterLoan	all_credits	sum_of_paym_months
0.10	2.0	2.0	13.0
0.25	4.0	3.0	31.0
0.50	10.0	5.0	64.0
0.75	18.0	8.0	118.0

Figure 19: Quantiles and distributions for RFM

So, we mark min MnthAfterLoan values in ascending order (if value lower than quantile, we mark with smaller value) and mark all credits and sum of paym months values in descending order (if value lower than quantile, we mark with greater value). So, it was chosen to have 5 classes for R, F and M, because our data is skewed and we should also take into account long tails. So lets explore the obtained classes.

1st cluster. Most active customers.

- Age: 35
- Sex: Female
- Salary: over 250k rubles
- Purpose: renovation
- Job: specialist / middle manager
- Number of credits: 8
- Number of months after previous loan: 1
- Last month payments: 95k rubles

The 1st cluster contains most reliable customers, these ones are most likely to take loans, they pay most of all and take many loans. In this cluster most customers have RFM score = 111 - this is the best RFM score, what means they have highest priority and we should make most advantageous offers. Our goal is keeping these customers.

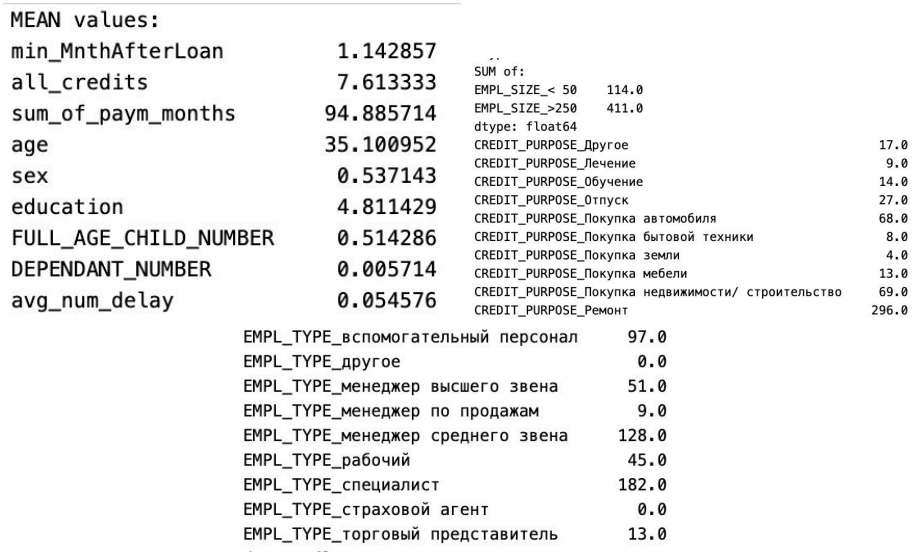


Figure 20: 1st cluster description (mean,sum,sum)

2nd cluster. Active customers.

- Age: 36
- Sex: Female
- Salary: over 250k rubles
- Job: specialist / middle manager
- Number of credits: 7
- Number of months after previous loan: 3
- Last month payments: 86k rubles
- Purpose: renovation

The 2nd cluster still contains reliable customers, who is ready to pay pretty much and take many loans as well, however these customers take loans more rarely. In this cluster most customers have RFM score = 211 and 222, this means that customers from 2nd cluster also have high priority in the bank.

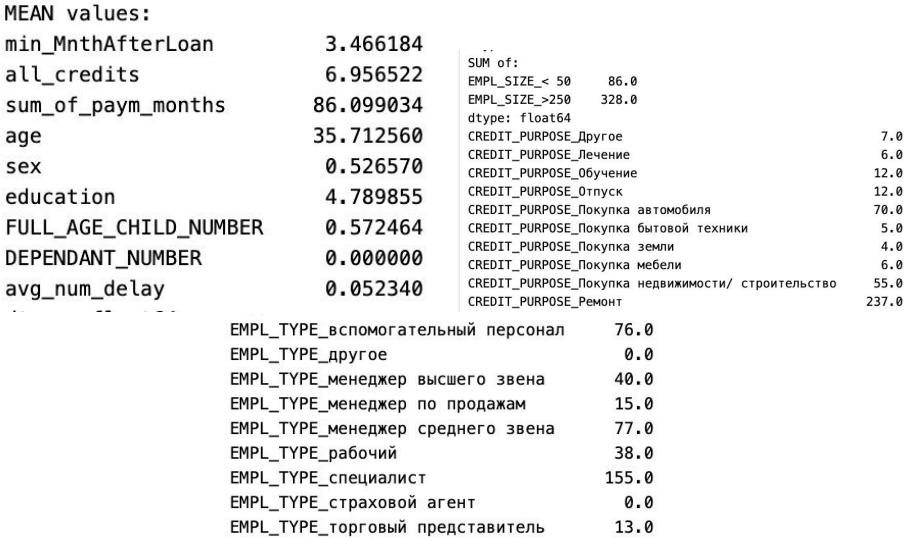


Figure 21: 2ns cluster description (mean,sum,sum)

3rd cluster. Neutrally-active customers.

- Age: 35
- Sex: Female
- Salary: over 250k rubles
- Purpose: renovation
- Job: specialist / middle manager / support staff
- Number of credits: 6
- Number of months after previous loan: 7
- Last month payments: 86.4k rubles
- Purpose: renovation / car purchase

The 3rd cluster includes neutrally-active customers. Despite that their payments are still high, they rather less often take loans, the evidence of this assumption is doubled number of months since last loan and decreased number of loans at all. One can notice that the customer portrait has changed a bit, which allows to assume that behavior of customer in this cluster differs from customers in the first 2 clusters. It would be reasonable for our bank to investigate deeper the difference between customers and stimulate customers in 3rd clusters to take more loans.

```
MEAN values:
min_MnthAfterLoan          7.184925
all_credits                6.437186
sum_of_paym_months        86.467337
age                       35.250251
sex                        0.525628
education                  4.735678
FULL_AGE_CHILD_NUMBER      0.475377
DEPENDANT_NUMBER           0.002010
avg_num_delay              0.054493
```

```
SUM of:
EMPL_SIZE_< 50    220.0
EMPL_SIZE_>250    775.0
dtype: float64
CREDIT_PURPOSE_Другое                                  47.0
CREDIT_PURPOSE_Лечение                                 11.0
CREDIT_PURPOSE_Обучение                                14.0
CREDIT_PURPOSE_Отпуск                                  48.0
CREDIT_PURPOSE_Покупка автомобиля                     176.0
CREDIT_PURPOSE_Покупка бытовой техники                 17.0
CREDIT_PURPOSE_Покупка земли                           18.0
CREDIT_PURPOSE_Покупка мебели                          17.0
CREDIT_PURPOSE_Покупка недвижимости/ строительство     91.0
CREDIT_PURPOSE_Ремонт                                 556.0
```

```
EMPL_TYPE_вспомогательный персонал    177.0
EMPL_TYPE_другое                       0.0
EMPL_TYPE_менеджер высшего звена      67.0
EMPL_TYPE_менеджер по продажам        33.0
EMPL_TYPE_менеджер среднего звена    221.0
EMPL_TYPE_рабочий                     76.0
EMPL_TYPE_специалист                 389.0
EMPL_TYPE_страховой агент              0.0
EMPL_TYPE_торговый представитель      32.0
```

Figure 22: 3rd cluster description (mean,sum,sum)

4th cluster. Inactive customers.

- Age: 36
- Sex: Male
- Salary: over 250k rubles
- Job: specialist / middle manager / support staff
- Number of credits: 5
- Number of months after previous loan: 13
- Last month payments: 84.4k rubles
- Purpose: renovation / car purchase / property purchase

The 4th cluster contains inactive customers. The of payments decrease and the number of months after last loan doubled, this could be the consequence of changed purpose for taking loans. As we can see customer take loans to purchase car or property and we can assume that these purchases a quite rare, that is why they take loans much rare. The bank can think of providing special offers to satisfy needs of these customers and push them into being more active.

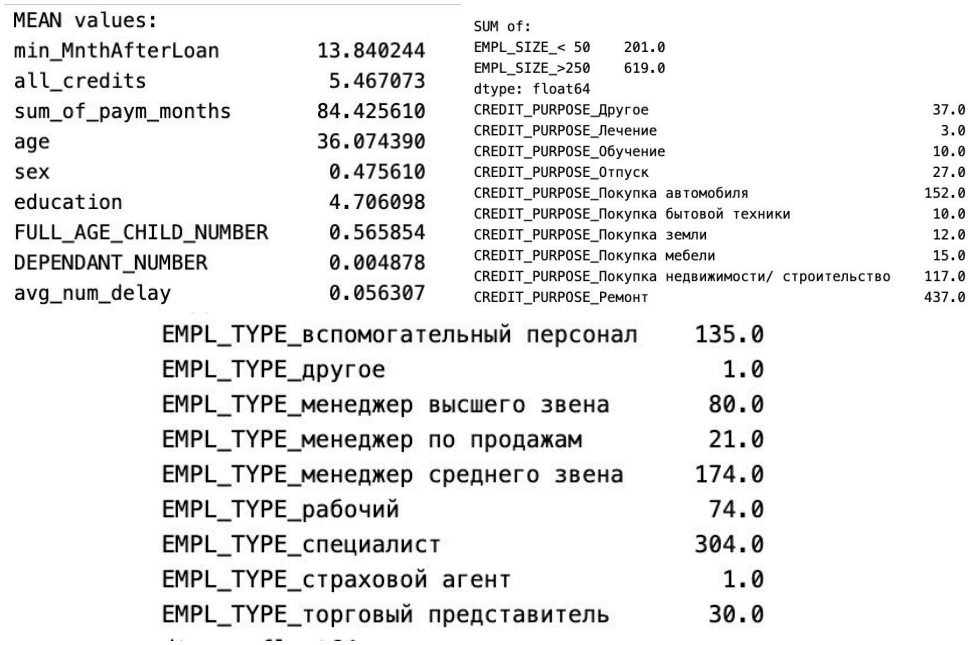


Figure 23: 4th cluster description (mean,sum,sum)

5th cluster. Totally inactive customers.

- Age: 35
- Sex: Male
- Salary: over 250k rubles
- Job: specialist / middle manager / support staff
- Number of credits: 4
- Number of months after previous loan: 34
- Last month payments: 73.5k rubles
- Purpose: renovation / car purchase / property purchase

The 5th cluster contains totally inactive customers. The of payments significantly low and the number of months after last loan much greater in comparison with all other customer clusters. The total number of credits has also decreased. All these facts allow to conclude that these group of customers take loans extremely rare and prefer not to do it in case it is possible. Thus, our bank may not put much efforts in atracting these customers back and focus on providing services for the previous groups.

```
MEAN values:
min_MnthAfterLoan          34.686188
all_credits                 3.750276
sum_of_paym_months         73.533702
age                        37.293923
sex                         0.451934
education                   4.731492
FULL_AGE_CHILD_NUMBER       0.548066
DEPENDANT_NUMBER            0.004420
avg_num_delay               0.064959
```

```
SUM of:
EMPL_SIZE_< 50    242.0
EMPL_SIZE_>250    663.0
dtype: float64
```

```
CREDIT_PURPOSE_Другое                                 34.0
CREDIT_PURPOSE_Лечение                                12.0
CREDIT_PURPOSE_Обучение                               10.0
CREDIT_PURPOSE_Отпуск                                 35.0
CREDIT_PURPOSE_Покупка автомобиля                    156.0
CREDIT_PURPOSE_Покупка бытовой техники                10.0
CREDIT_PURPOSE_Покупка земли                           7.0
CREDIT_PURPOSE_Покупка мебели                         17.0
CREDIT_PURPOSE_Покупка недвижимости/ строительство    132.0
CREDIT_PURPOSE_Ремонт                                 492.0
```

```
EMPL_TYPE_вспомогательный персонал    154.0
EMPL_TYPE_другое                       2.0
EMPL_TYPE_менеджер высшего звена      79.0
EMPL_TYPE_менеджер по продажам        28.0
EMPL_TYPE_менеджер среднего звена    195.0
EMPL_TYPE_рабочий                     74.0
EMPL_TYPE_специалист                 353.0
EMPL_TYPE_страховой агент              0.0
EMPL_TYPE_торговый представитель      20.0
```

Figure 24: 5th cluster description (mean,sum,sum)

Conclusion

In conclusion, by leveraging advanced analytics techniques like tSNE, KMeans clustering, and RFM analysis, the bank gains the power to make data-driven decisions that are both nuanced and customer-centric. This strategic approach fosters increased customer satisfaction, loyalty, and ultimately, business growth.**
While both RFM and ML-clustering offer valuable insights into customer segmentation, they cater to distinct objectives.** RFM, with its focus on recency, frequency, and monetary value, provides a customer-centric perspective, enabling the bank to:

- **Identify and prioritize high-value customers**: By segmenting customers based on their purchasing behavior, the bank can focus its resources on retaining and nurturing its most valuable customers.
- **Predict customer churn**: RFM scores can be used to identify customers at risk of churning, allowing the bank to implement targeted retention strategies.
- **Measure the effectiveness of marketing campaigns**: RFM analysis can be used to track changes in customer behavior resulting from specific marketing campaigns.

On the other hand, ML-clustering algorithms offer a more comprehensive understanding of the customer base. By considering a wider range of data points beyond just purchasing habits, ML-clustering helps the bank:

- **Discover hidden patterns and segments**: ML algorithms can uncover previously unknown customer segments with distinct characteristics, enabling the bank to tailor its offerings accordingly.
- **Identify potential opportunities**: By understanding the characteristics of different customer groups, the bank can identify potential cross-selling or upselling opportunities.
- **Develop targeted marketing campaigns**: ML-derived insights can be used to create highly targeted marketing campaigns that resonate with specific customer segments, leading to increased campaign effectiveness

Therefore, the optimal choice between RFM and ML-clustering depends on the specific goals of the bank. When the focus is on customer-centricity and performance evaluation, RFM's targeted approach proves advantageous. However, for a deeper understanding of the customer base and the identification of strategic opportunities, ML-clustering's comprehensive analysis is more suitable.**