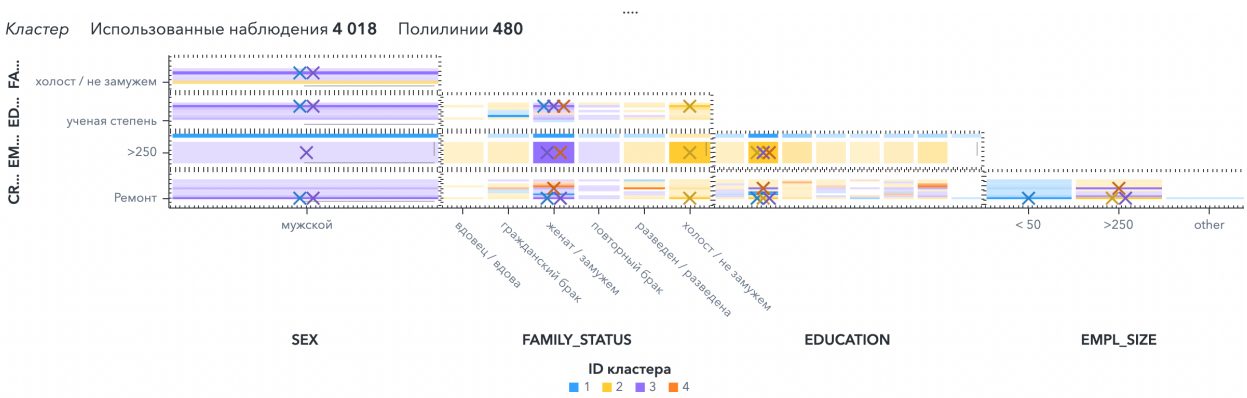
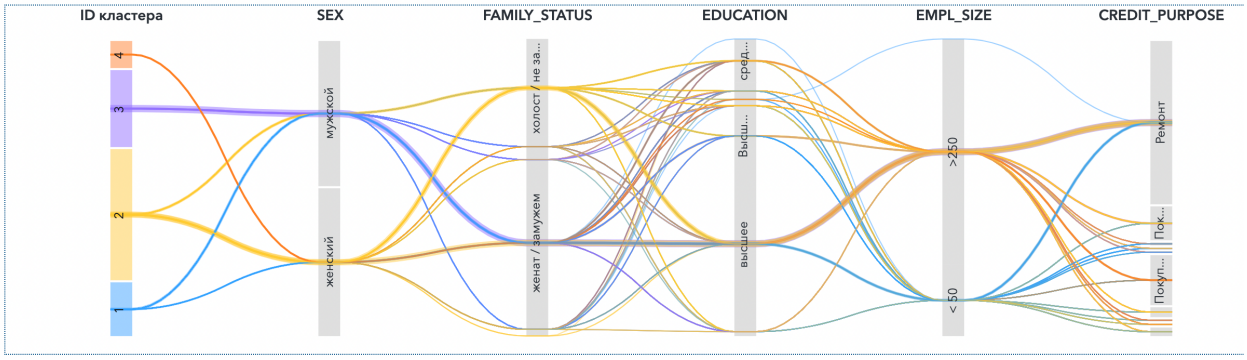
2 Task Fishman Maxim

A good visualization came out in teacherless learning - klasterisation:

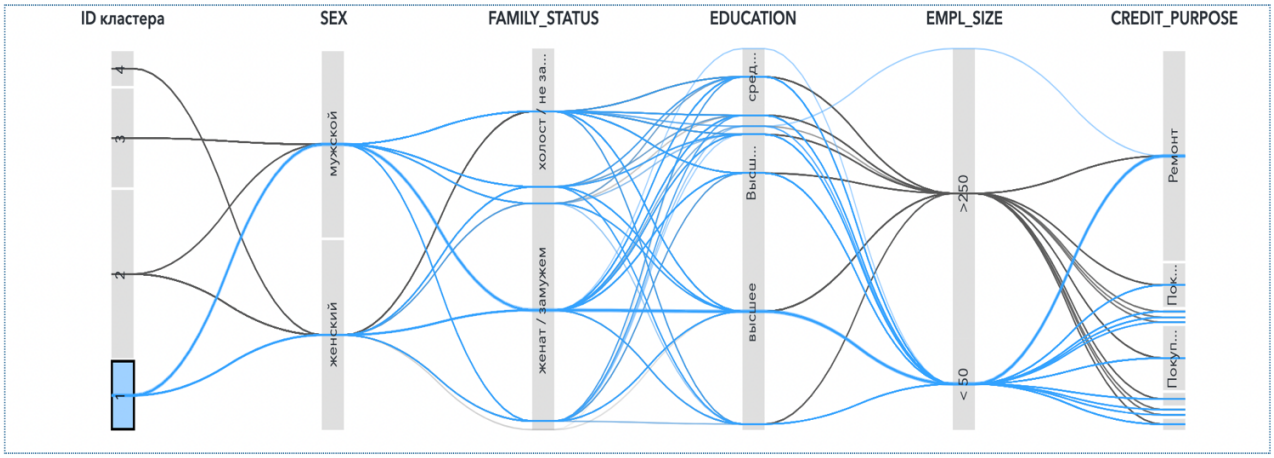
I tried different sets of variables and к number of кlasters. This combination turned out to be the most successful





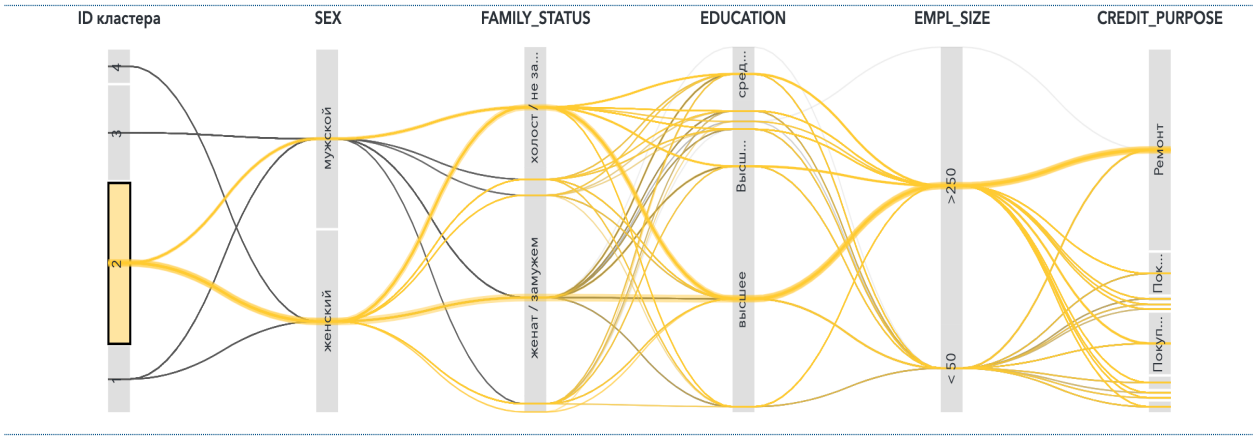
using these variables it is possible to make customer profiles for targeted advertising of favourable loan terms based on the most preferred area of credit for people in a particular cluster. In this way Loan companies and banks can send targeted, advantageous offers to clients or potential clients.

Let's look at each cluster individually:



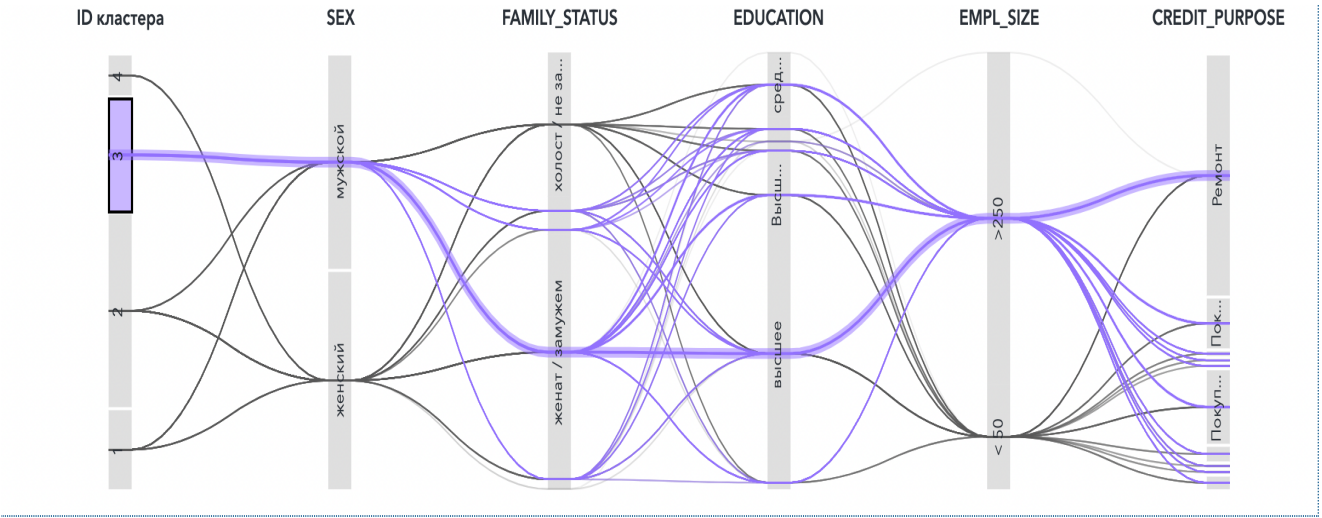
1) The first cluster is general, just based on the data from the date we uploaded it gives us an idea that мввen are more likely to take a loan than women, also a person can have different marital status and any education (often higher education) and a salary under 50 - such a person will take money for repairs more often than for other areas.

Let's call this type - **unpredictable**, because the person does not have little money and does not depend on other variables and may need a different type of credit, but the most popular will still be a repair based on the date



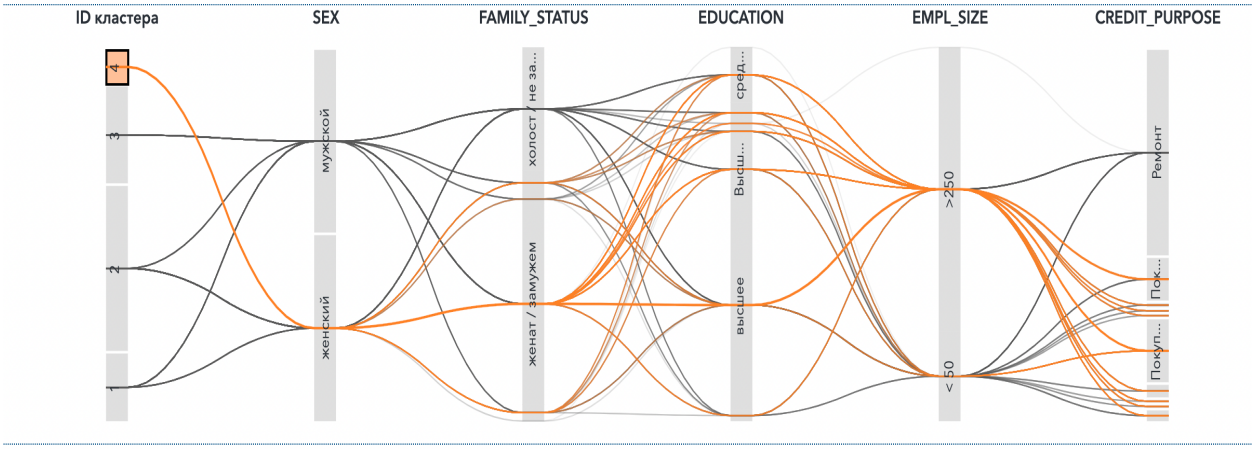
2) The second cluster will be women, who, regardless of their marital status with higher education, will take out loans for repairs (much more often than for other needs). There is a correlation that if the salary is higher, loans are taken more often because of the possibility to pay them back, and if the salary is small, loans for other needs are taken less often than if a woman has a large salary

Let's call this type - **a woman with a high salary**
this type is inclined to take out loans for repairs and in most cases has high education, such people take out loans more often than women with low salary.



3) in this cluster there is a much stronger correlation between marital status, so we see that exclusively men who are not single and have high salaries also take out a renovation loan.

Let's call this cluster - **the rich familyman**, most likely men who are over middle-aged, who work long hours and therefore have high salaries, take a home repair loan because they have nothing else to worry about in their lives.



4)

The last cluster is exclusively women, mostly married and with a high paycheck (Can be compared with cluster 2) do not take money for repairs, which is very strange as this is the preferred type of credit for the whole database.

Let's call this type of cluster - women who do not like repairs

For this type of precisely do not offer the terms of a repair loan, because none of them take credit and advertising will be wasted, because repair loans are likely to have taken their men from the previous cluster)

Pros and cons of such clustering

Pros

1)It is possible to see clearly the dependencies of the final variable on all at once gradually passing through each variable

2) clearly arranged clusters divided by special characteristics - it is possible to analyze a date at once even without going to calculation and without looking closely at the date

3) This way is very suitable for small databases when there are not many values and they directly correlate with each other.

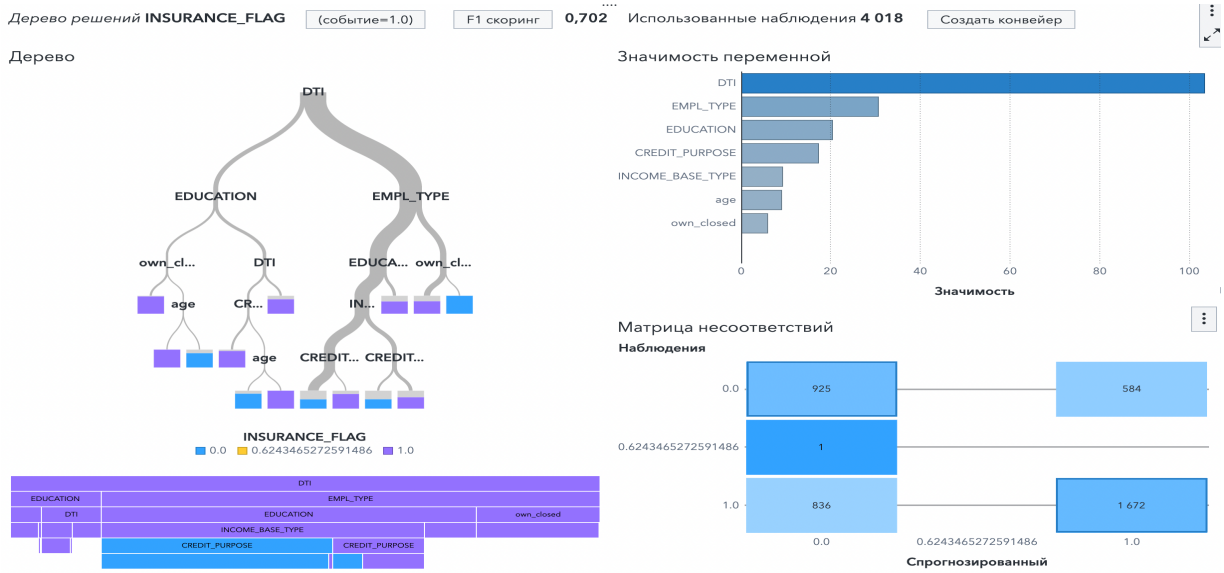
Disadvantages

1) Sometimes such a clustering as in our case may not always be perfect, because our table is likely to be glued together from several and not all data depend on each other and eventually we can get a completely strange clusters that are very similar to each other and do not differ

2) Does not work well with large data, you get too many lines and have to cut the graph

3) It is necessary to think over which variables follow each other to make the clustering more visible and understandable.

Moving on to the second method, the decision tree, I decided to look at what factors affect the borrower's credit, and this method is ideal for that



As you can see - the main dependency is the DTI, well this makes sense because if a person has a high enough value and no big income or prospect of paying back the debts, they will not get insurance.

Further the interesting fact is that while choosing and choosing variables the salary does not correlate in any way with insurance from what we can conclude that these two variables are from different tables and it means that the model will not give out many F1 soon and there will be false positives as well as negative ones.

As there are quite a few nodes, let's look at some of the highlights

Идентификатор узла:	14
Число:	5
EMPL_TYPE:	вспомогательный персонал or рабочий
DTI:	#ПОЛЕ!, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22, 0.23, 0.24, 0.45, 0.46, 0.47, 0.48, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.61, 0.66, фев.33, or Missing
own_closed:	5.0
0.0:	5 (100,00%)

Let's call it an **auxiliary worker cluster**, such a cluster of people are not given credit insurance no matter how many loans they have closed, even though there are 5 in the date

Идентификатор узла:	24
Число:	33
EMPL_TYPE:	менеджер высшего звена, менеджер по продажам, менеджер среднего звена, специалист, стр
CREDIT_PURPOSE:	Покупка бытовой техники or Покупка земли
EDUCATION:	*n.a.*, второе высшее, высшее, Высшее/Второе высшее/Ученая степень, or ученая степень
DTI:	#ПОЛЕ!, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.45, 0.46, 0.47, 0.48, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.61, 0.66, фев.33, or Missing
INCOME_BASE_TYPE:	2НДФЛ, Поступление зарплаты на счет, or Свободная форма с печатью работодателя
0.0:	6 (18,18%)
1.0:	27 (81,82%)

The second cluster of people is called **educated people**, in this cluster there are people with higher education or two higher educations who have a sufficiently high position and such people are approved for insurance in 81 percent of cases all because they have a sufficiently high potential for this insurance to not apply in lending

Идентификатор узла:	21
Число:	6
age:	25.0, 27.0, 35.0, or 39.0
CREDIT_PURPOSE:	Отпуск, Покупка бытовой техники, or Покупка мебели
EDUCATION:	высшее or Высшее/Второе высшее/Ученая степень
DTI:	0.59
0.0:	5 (83,33%)
1.0:	1 (16,67%)

as seen in the third cluster, **unemployed people with a high level of education** are not given credit insurance regardless of other factors very decisive is the position in the company

Pros and cons of this method

Advantages

1) the decision tree can input quite a lot of data and safely process them, in contrast to the usual clustering by attributes, because nodes are used here

2) Variety of nodes - the nodes themselves are more accurate and show the dependence on different variables with the outcome of these dependences

3)You can visually look at the outcome of binary values

disadvantages

1) this variety of nodes may not always be a plus, because sometimes the full picture is important, or the tree throws out quite important variables in some nodes, which makes such nodes irrelevant for consideration

2) you have to read every node with a large amount of data in some variables, it is not always clear what node it is and what variables are in it, so it is long and tedious to examine this graph