Andrii Osipa

Dataset Analysis Report

I selected "*Default of credit card clients Data Set*" (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#).

Dataset information: It contains 30000 records of clients with their education, age, marital status, credit limit, payments for the 6 months and bills for 6 months. Every record has a label "default" which corresponds to the failure of a client to do his monthly payment in the next month after 6, that are recorded in the dataset.

Variables Description:

default (Yes = 1, No = 0)

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

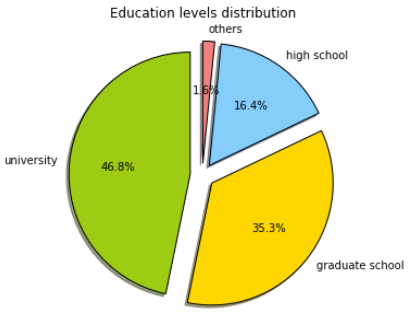
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

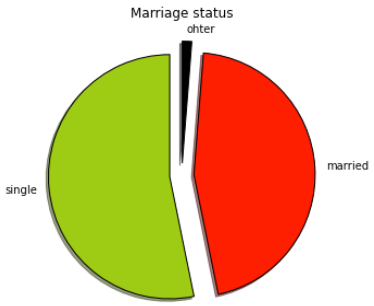
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

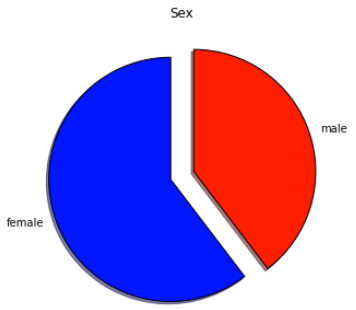
Goals:

- collect some statistical indicators about credit card owners;
- try to connect some personal information to the default of payment;
- try to find groups which are most and least risky to have a default.

Firstly we will look at the statistics about credit card owners in such perspectives: by education level, by sex, by marital status, by age and also look at average and median of those parameters and of credit limit of all clients.

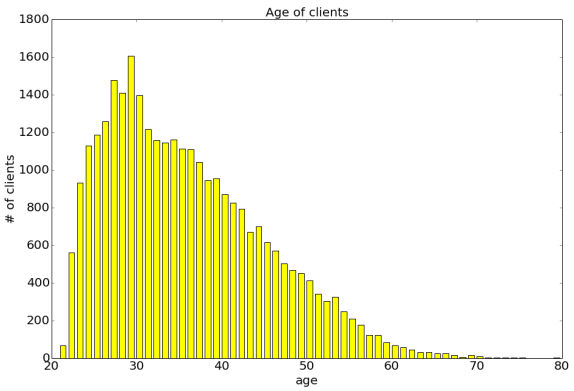






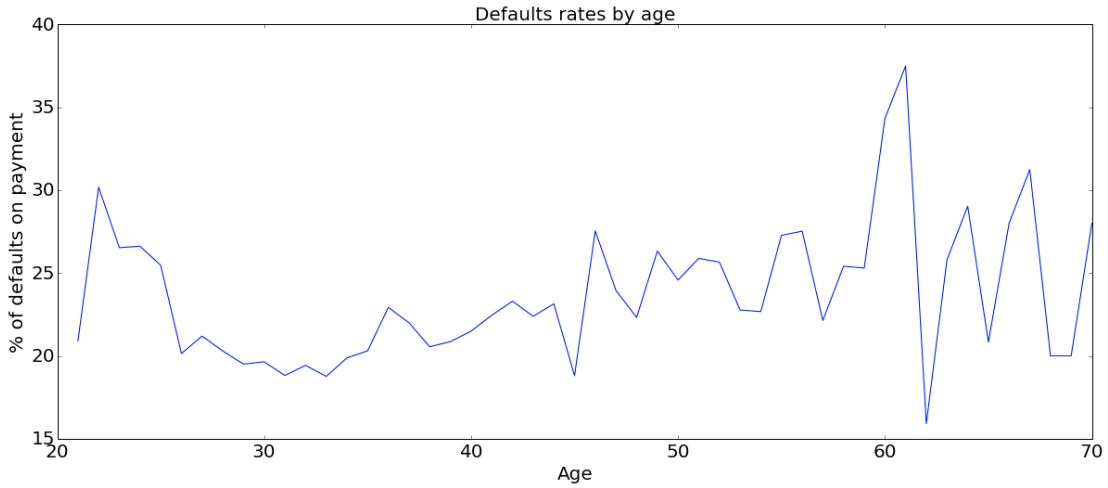
So here we see that mostly card owners are women by sex, single by marital status and graduated from university. In the age graphic we see that mostly clients are young people and the

median age is 34(median for males is 35 and median for females is 33), while the most popular one is 29. Total range of age is 21-79.



Next we are going to look at statistics of default. Let introduce default ration as number of defaults on payment divided by total number of people in the group. We will look at default ratio for each age available. It is possible to see that rate is quite higher for young people but the difference is not so big. On the right part of the plot we see an anomaly and that part of plot is not really informative. This cause by very small number of clients in age groups after 60. And when we look at age > 70 then we have really a few people in a group and this is not informative at all because selection is very small. To make it look more realistic last age group is 71+.

And also lets take a look at defaults ratio by sex and by education level. Here we notice that fact that default ratio for males is 24% and for females is 20%. General default ratio is 22%, which is average of the two ratios above. From here we can say that lower default ratio for women probably causes that fact that mostly credit cards owners are women.



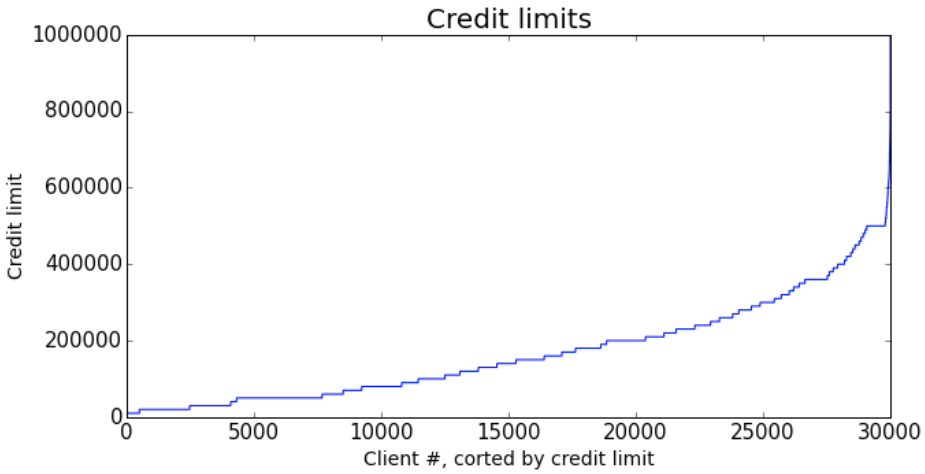
If we look at default ratios by education level we will see that lowest ratio have a group with “others” mentioned in educational level - only 7%. For group with only high school education default ratio is the highest - 25%; for people, who graduated from the university - 22%; for graduates - 19%.

Now I am going to identify group of people, that has the lowest default ration but still have enough representatives(>50), so the information from that group is quite reliable. Group will be defined by all the parameters mentioned above: fixed age, sex, education level and marital status. With all possible variations of the parameters we have 175 significant groups. Group with the smallest default ratio is single females, who graduated from university and are 38 years old.This group has ratio 7.8% only and this group has 76 representatives among 30000 provided records. This default ratio is significantly smaller than all the next ones. Next group is married females, who graduated from grad school and are 31 years old. For this group ratio is 8.7% and this group has 80 representatives. And 9 of top 10 such groups are female groups, and 6 of them are single females gropus.

Groups with the highest default ratios are: married female, who only finished high school and 46 years old(66 representatives and ratio is 37.8%) or 50 years old(62 representatives, ratio is 33.8%); single males, who graduated from the university and 22 years old(82 representatives, 34.9%).

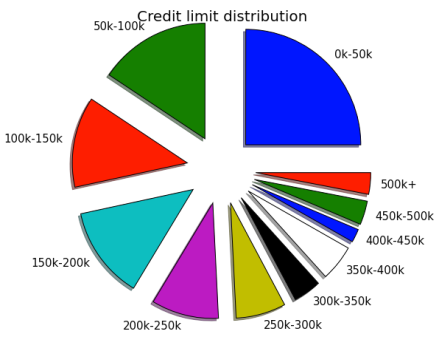
Also it makes sense to look at groups whose age is not limited to only one value but belongs some interval. I will look at intervals of length 5. Logically here we set that significant enough group must have al least 250 members. With this kind of analysis we get that group with the lowest payment default ration is single females, aged between 26 and 30, who finished grad school. This group has size 2210, which is almost 7.4% of all credit cards owners. Payment default ration in this group is 16.16%. Next group is also single female, who graduated from university and are aged 31-35. In the group there are 1005 members and default ratio is 16.31%. It seems quite logical to unite these two groups together, since the only different parameter is age. Therefore we can conclude that people with that set of parameters are best one to give them credit cards. This conclusion is much more informative than one given in the previous paragraph because group here is nearly 10 times bigger and really represents 10% of all credit cards owners(according to the current records).

The next what we are going to look at statistics about credit balance. Credit balance in these records is between 10000 and 1000000. Next plot shows how credit limit is distributed among the clients. We can see that it grows almost linearly at the quite big part of plot. It seems to be very unlined in the part which corresponds to quite big credit limits(250.000 TN dollars and more). Also it can be seen that there are some credit limits that are the most common. They are



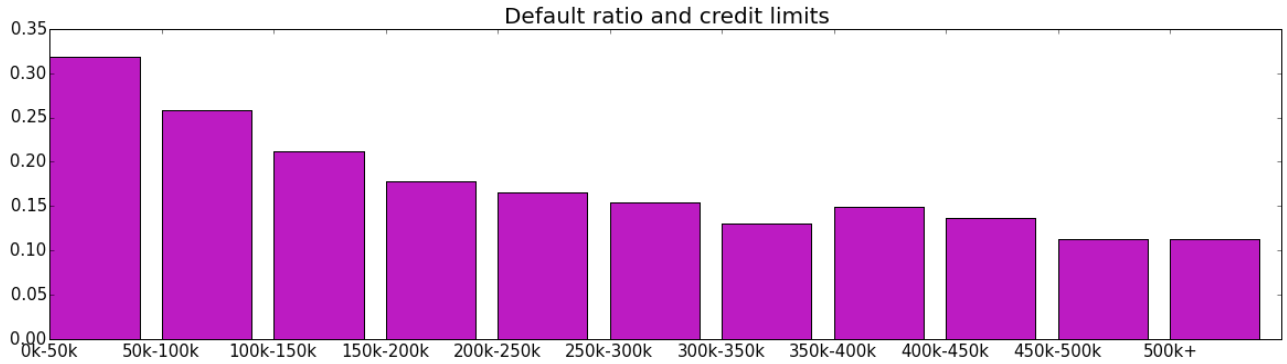
20.000, 30.000, 50.000, 80.000, 100.000, 150.000, 200.000 NT dollar. All these limits appear from 1.000 to 3.300 times in the list of 30000 records. This is quite logical that mostly credit limits are standard.

Lets now look at groups with lowest and highest median credit limit. Groups are defined as it was stated above: 5 years interval of age; education level, marital status and sex are fixed to one value for each parameter. The top three groups are married males, who finished graduate school and are 46-50, 36-40, 41-45 years old. The median credit balance for those groups is 240.000-250.000 TN dollars. All these three groups have in total 1300 representatives. This information is not very surprising.



People from these groups probably have to cover huge expenses of the family, that include expenses of their children.

Next point that I want to discover is how default ratio depends on the credit limit. For this task all credit limits are divided into groups, each of them represents interval of credit limits, which is equal to 50000 TN dollars. Considering that fact, that there a only a few clients with very high credit limit, I made last group much wider: there is everybody with credit limit greater than 500.000 TN dollars. From the bar plot below it is obvious that larger credit limits implies lower default ratio. But also from this plot it can be interfered that dependency between these parameters is not linear. It seem that default ratio is stabilizes after some credit limit and starts to fluctuate near this average default ratio.



Conclusion. After providing such statistical research on a given dataset I may conclude that even from this not very deep analysis we can get some valuable knowledge in a quite hard question of credit rating of people. I got propositions for client classification by given parameters, such as education level, age and marital status. This information is enough to approximate default ratio that may be expected from a new client. Therefore, even simple statistics is enough sometimes to do some predictions in business. Other interesting fact, that is not really obvious is that mostly credit card users are women. Also simple statistical analysis showed that default ratio is not really depend a lot on age, but is has more clear dependency on educational level. Moreover it appeared that default rate for clients with rather high credit card limits is pretty similar and it not seems to go down to zero as credit limit grows up. Doing this stuff without appropriate tools will be quite hard because amount of data is quite big and it requires some automation. By applying some technics from data mining we can achieve such valuable results.