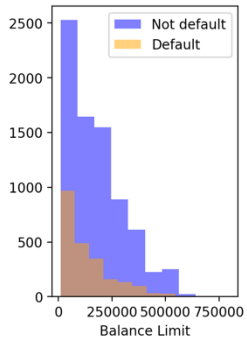
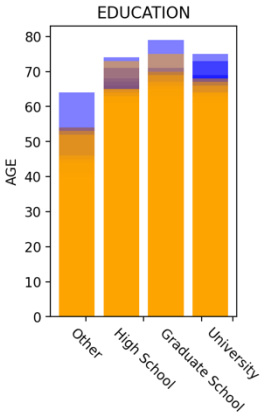
Martinez P. M May 17th, 2021

Exploring the Credit One database





Exploring and cleaning the database, I identified two main groups of people: those who default on loans (~22% of clients in the database, 6,682/30,0201 people) and those that do not default (~78% of clients in the database). Hereafter, I. will call them the *default* group and *not-default* group. Each of these groups can be classified according to their *sex, education, marital status, age*, and balance limit (*limit_balance*). Comparing these features, I found that, for example, the oldest people never default (see Figure 1). The balance limit of the *not-default* group is, on average, ~1.3 times the balance limit of the *default* group. As an example, in Figure 1, I show the comparison among the balance limit histograms of both groups.

Then, I calculated the correlation and covariance matrices and noted some interrelationships, for example, between the *limit_balance* and *age, education, marital status, sex*, bill (*BILL_AMT* tags), and bill payments (*PAY_AMT* tags), see the jupyter notebook file. I also made some box plots for the *limit_balance* variable, grouping the data by *sex, age, marriage, education*, and *default payment next month (DPNM)*. The *default* group has women and men, both with a university, high school, graduate school education, and a single or married *marital status*. Particularly, I noted that 57% are women, that the majority have a university education (3,303 people), and that 3,351 people are married. I made a scatter plot among the *limit_balance* vs *PAY_* variables and noted that both groups (*default* and *not-default*) tend to delay payments when the *limit_balance* is fewer than ~300,000$. However, when the *limit_balance* is higher than ~300,000$, people who default tend to delay the payments since the first month (April), while people who do not default start to delay the payments two months later.

After that, I also grouped the data by the *limit_balance*. Then, I compared the *limit_balance* with the *age, marital status, sex*, and *education* variables. I noted that the more significant differences among *default* and *not default* groups occur when their *limit_balance* is higher than 500,000$. For example, on average, when the *limit_balance* of people is between 700,000$ and 800,000$, the *default* group get a fewer bill than people who do *not default*. However, the latter pay a more significant amount in the corresponding month (April, for example). Similar behavior is observed in people with a *limit_balance* between 600,000$ and 700,000$. Curiously this group is composed only of women.
On the other hand, the *default* people with a *limit_balance* between 500,000$ and 600,000$ get the highest bill. Curiously, this group make the highest payment. Therefore, based on this result, I would suggest increasing the bill for people who tend to default. Although, a more exact and depth analysis have to be done assuming a machine learning algorithm, that allow us to do better predictions.