# Week 1 Assignment Report

## Data Overlook

There are total 5000 records with 47 columns. Based on past research on likelihood of loan delinquency and loan approval, we do some statistic summary on the key variables on the sample space, including age, gender, loan amount, interest rate, credit scores.

Based on the result, we can define these key variables into three types:

(1) Personal properties (such as age and gender):

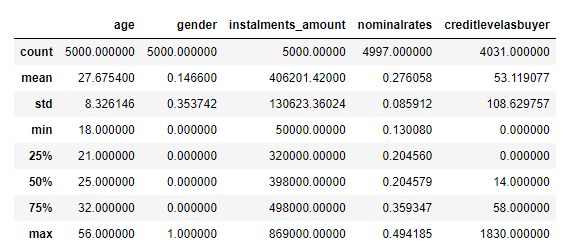
We found that the average age of the sample is about 27.5 which means it is a rather “young sample”. At the same time, after we transform gender into numeric value (0: Male, 1 : Female), the statistics indicates that the proportion of males in the sample is much higher. To be more specific, gentlemen are six times more than lady in the sample space. Based on the result it occurs to me that gender is likely to be a useless result in the sample.

(2) Properties of the loan (such as: instalments and rates):

As for these two variables, data in instalments is fairly clean, and we can easily get the statistics of instalments. However, there are three records (0.06%) missing in rates. Since the number of missing records is limited, we will simply fill the value with mean value of the sample.

(3) Personal digital footprint (such as credit as buyers, contact information)

As for the credit score, there are 969 missing records in the sample, we then try several ways to fill the value.



Figure(1)

## Data Cleaning

1. Creditlevelasbuyer

I firstly conduct the correlation test on it and get the result in Figure (2), then I select highly related variables involvingTENCENT score, GAODE score, HUABEI balance and HUABEI amount and conduct a linear regression on the data.

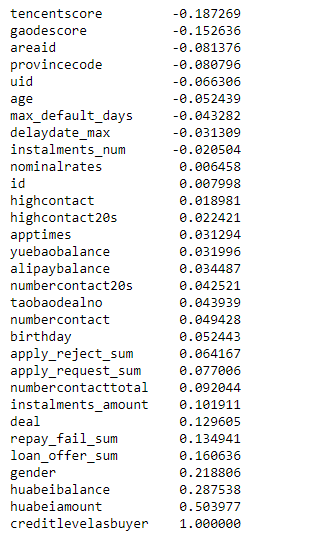
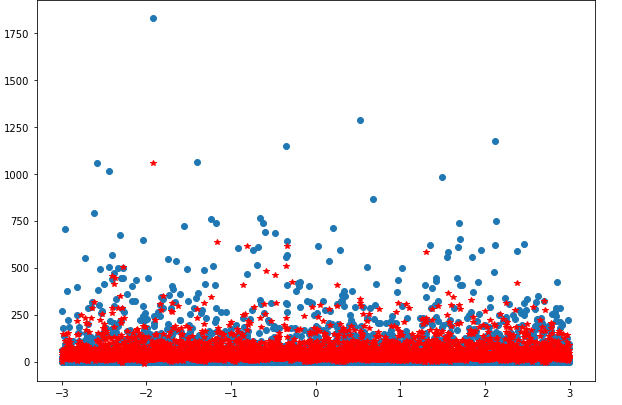


Figure (2)

Nevertheless, the regression result is fairly pleasant compared to using average value. Although due to its high variance the regression result can only cover part of special values, it istill has better performance than average number.



1. Default

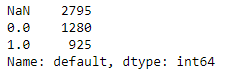
Default variable will work as dependent variable later, since its data type is Boolean, we transform it into numeric value 0/1.

## Logit regression on single variables

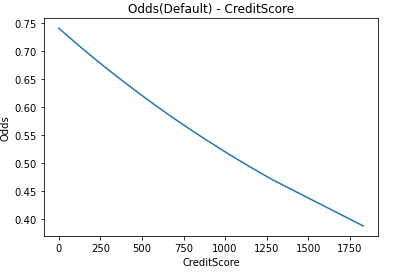
1. Default vs Credit score.

There are above half of the Default records are missing. Because it is the dependent variable, we just simply drop the missing records.

As for Default records,



Then we conduct the logit regression with Default variables and credit score. We get the coefficient = -0.00035419. To illustrate the result better, we use odds to show the result.



According to the figure, we can conclude that with higher credit score, there is lower likelihood of default. But its effect is not significant because with the range from 0 to 1830, it only impacts the possibility of default slightly (0.75-0.40).

1. Deal vs Credit score.

## Logit regression on multiple variables