Content

[Week 1 Assignment Report 1](#_Toc82360966)

[Data Overlook 1](#_Toc82360967)

[Data Cleaning 2](#_Toc82360968)

[Logit regression on single variables 3](#_Toc82360969)

[Logit regression on multiple variables 6](#_Toc82360970)

# 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):

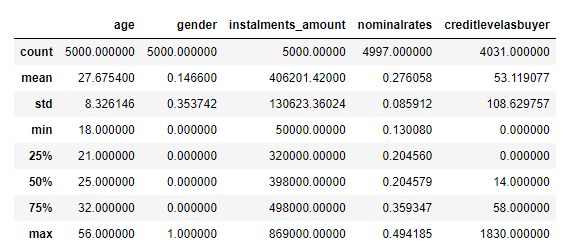
We found that the average age of the sample is about 27.5 which means it is a rather “young sample”. At the same time, after we transform gender into numeric value (0: Male, 1 : Female), the statistics indicates that the proportion of males in the sample is much higher. To be more specific, gentlemen 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.

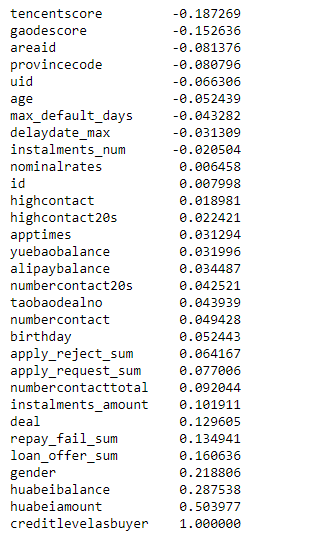
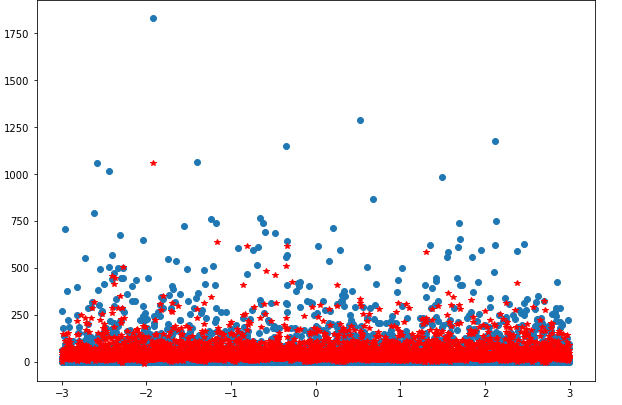


Figure (2)

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

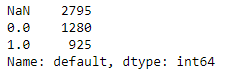
Default variable will work as dependent variable later, since its data type is Boolean, we transform it into numeric value 0/1.

## Logit regression on single variables

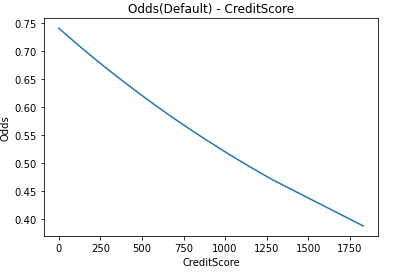
1. Default vs Credit score.

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,



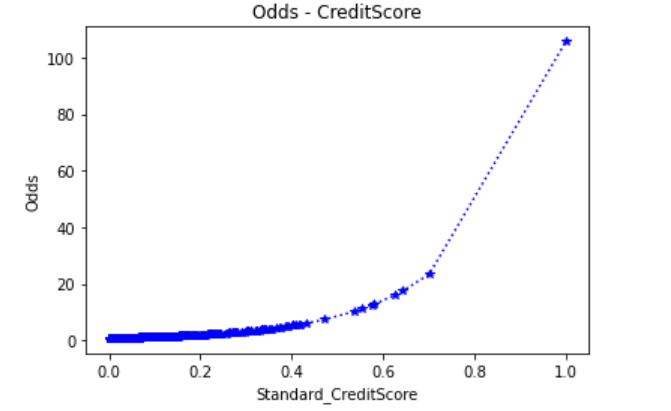
Then we conduct the logit regression with Default variables and credit score. We get the coefficient = -0.00035419. To illustrate the result better, we use odds to show the result.



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.

I run logit regression on Deal and Credit score, then we get the result of coefficient is 0.00304548. And the model scores 0.5794. To illustrate the result better, we illustrate odds and with standardized credit score. From the line, we learn that when the individual’s credit score is higher, his application for loan is more likely to be approved. Further, compared its influence on the likelihood of default, it obviously influenced the likelihood of approval much stronger, ranging from odds at 0 to 100.

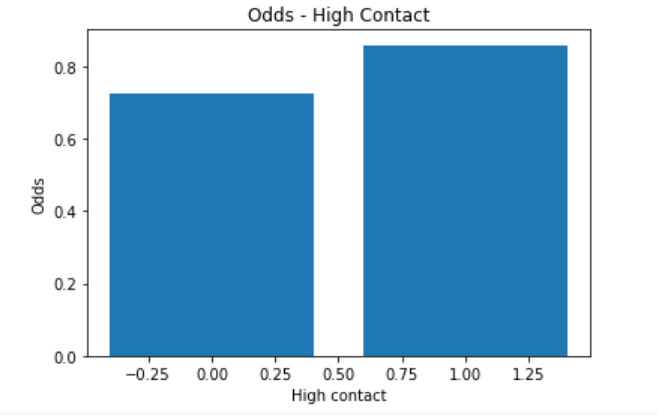


(3) Deal vs frequent contact.

There are two types of values (True/False) in the contact variable and true value stands for the object has a frequent contact. Based on that, we only need to create one dummy variable for contact variable (0: False, 1: True).

After we run the logit regression on these two variables, we got a model score at 0.5586 which is a bit lower than the last logit regression. Besides we get the model coefficient at around 0.16685. Then we also draw a line of odds based on the regression.

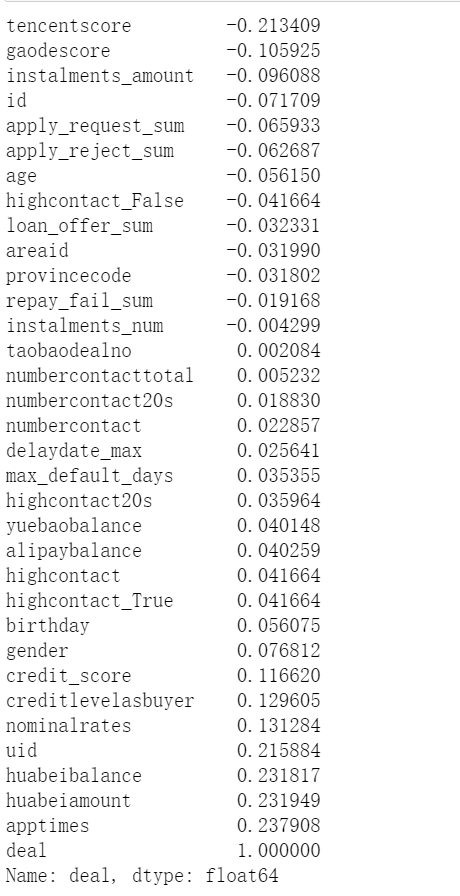
According to the line, we can learn that when some has a frequent contact his loan is slightly more likely to be approved. Nevertheless, two value of odds is less than 1 which means their loans is more likely to be rejected. In other words, contact frequency only has small connection with the likelihood of loan approval.



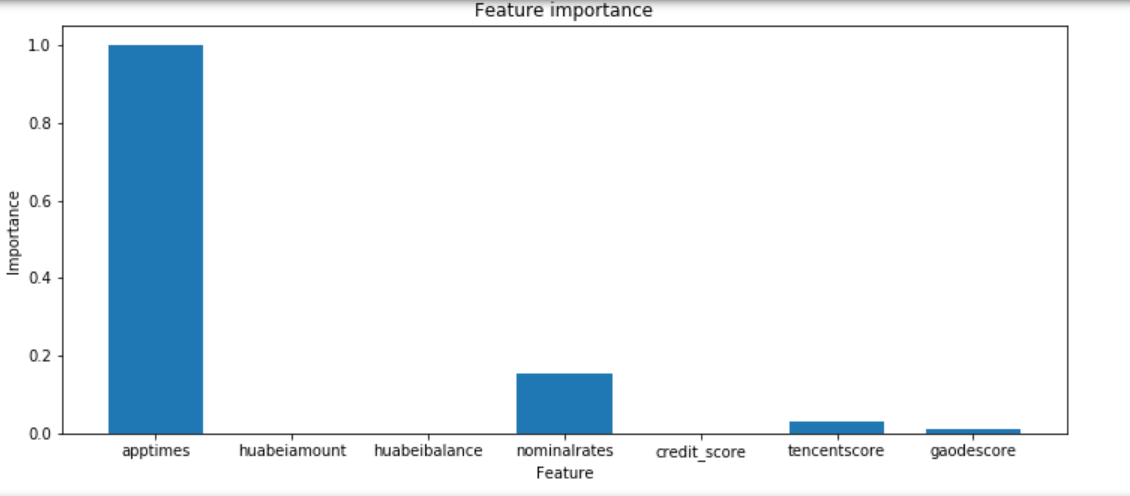
## Logit regression on multiple variables

The two logit regression models on single variable both end up at a pretty low score. Therefore, I then involve more variables to enhance the model.

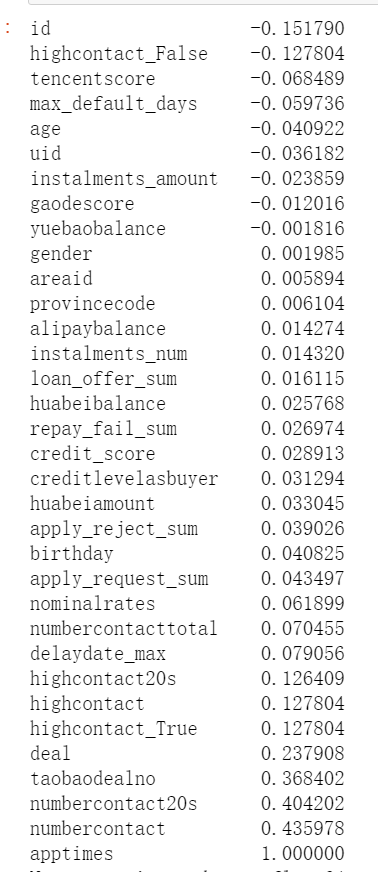
Since there are many variables are available in the data sample, I simply run a correlation test on Deal in the sample. Then I pick variables which has high absolute value of correlation. To be more specific I pick variables rank high in absolute value of correlation and reaches at lease 0.1, including: "apptimes","huabeiamount","huabeibalance","nominalrates","credit\_score","tencentscore","gaodescore".



Next, I run the logit regression on the sample, the model has much better performance, rating 0.7080. Then I conduct a feature importance test on the features. We notice that “apptimes” is most important. And “Huabei Amount” and “Huabei Balance” are least significant. So I improve the model with other variables.



To pick new variable in variable left, I run a correlation test on the most important variable “apptimes”. I then use the variable which is least relevant to “apptimes” to involve new data angle into the regression. Therefore, I replace “Huabei Amount” the least important variable with “gender” the least relevant variable to “apptimes”



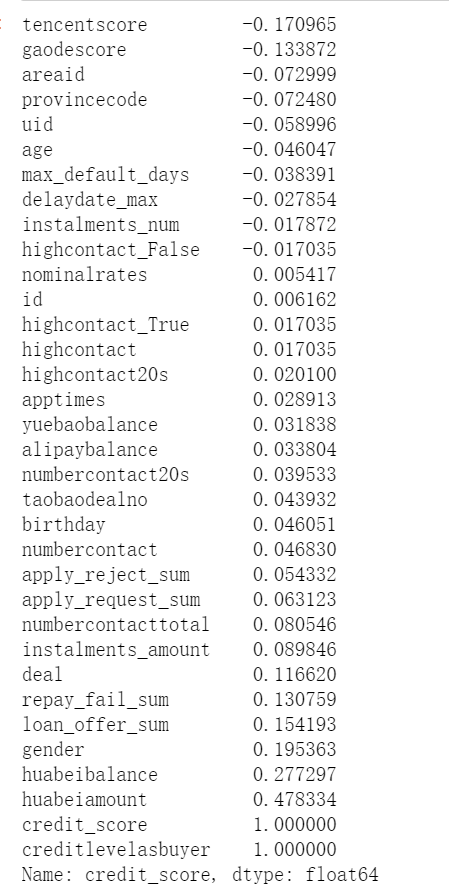
After I rerun the logit regression with new features. To control the number of features, I replace the least important feature (“Huabei Amount”) with new feature (“Gender”). And the model improves with score reaching 0.7096.

|  |  |
| --- | --- |
| Model | Score |
| model ~ (huabei amount,```) | 0.7080 |
| model ~(gender,```) | 0.7096 |

The new coefficient list:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| apptimes | huabeibalance | nominalrate | Creditscore | tencentscore | gaodescore | gender |
| 1.18021588e+00 | 6.59119053e-05 | 4.39638322e-04 | 1.37616783e-01 | -1.235210e-04 | -3.46274704e-02 | 9.37269120e-03 |

Then we repeat the method, finding the least relevant variables to credit score. However we found that the “nominalrates” is already involved in the model.



The third important feature “gaodescore” is least relevant with “apptimes” which is also involved in the model. At the same time other variables has small coefficient so I stop selecting new feature and I believe the we have an elegant solution now.

