FIN3210 Week 1 Assignment Report

Ma Kexuan 120090651

**Abstract**

This report provides a descriptive summary statistic for the dataset provided, and construct several regressions to discover how the digital footprints affect the outcome of debt collection, including loan approval likelihood and delinquency likelihood.

**Data Preprocessing**

First, select the relevant columns in the dataset to be included to the latter analysis, in my code, I select 'age', 'gender', 'instalments\_amount', 'nominalrates', 'tencentscore', 'highcontact20s', 'deal', 'default' as the columns. Then I do some type transfer of the original columns, also take natural log of instalment\_amount and tencentscore in the regression part, in order to eliminate the scale difference between the variables to ensure the reliability of the later results.

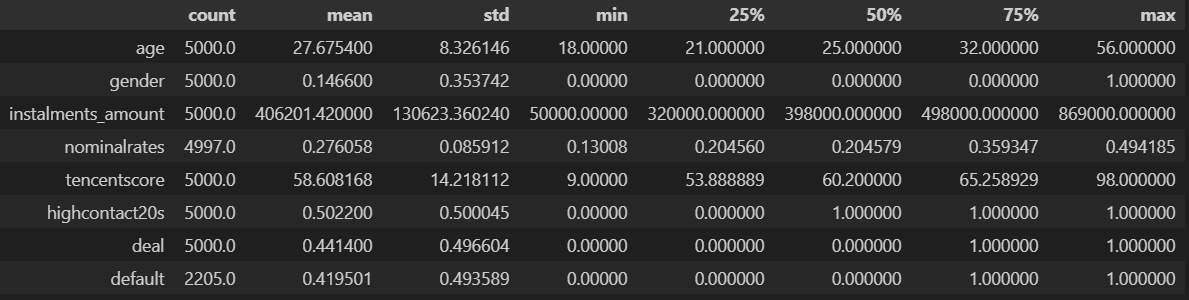
In the following tasks, it’s worth noting that I use tencentscore as the credit score, this is because the creditlevelsbuyer has too many missing values, and gaodescore has too little scale to distinguish the difference between different borrower.

Also, I choose highcontact20s instead of highcontact as variable because there’s a probability that the highcontact person is some market salesman, which a person will hang up the phone quickly, so selecting the phone call duration more than 20 seconds is more convincing.

**Questions**

1) Present a table of summary statistics for the key variables including the borrower’ age, gender, loan amount, interest rate, credit scores, a dummy whether the borrower has a frequent contact, approval dummy, and delinquency dummy.

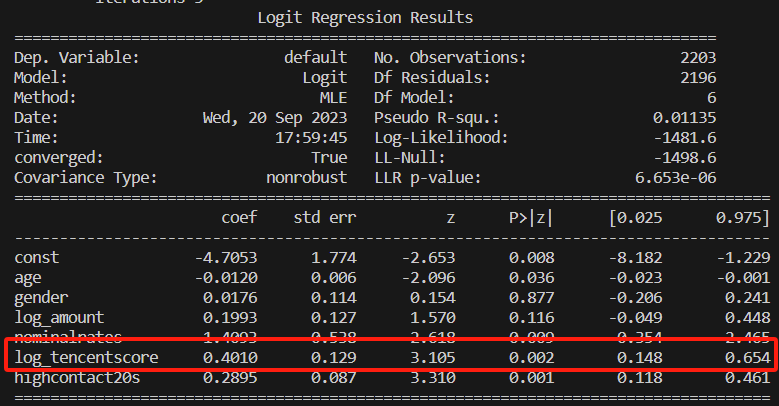
The summary statistics table is shown below as the chart. The descriptions of variables are shown in the footnote. From the statistics, I find that there exists some imbalance in the distribution of genders, and the default column has plenty of missing values, which may cause some issues in the following regression tasks.



From this question on, in the following 3 questions, I choose 'age', 'gender', 'log\_amount', 'nominalrates', 'log\_tencentscore', and 'highcontact20s' as the independent variables, intentionally to control other characteristics of loan and borrower.

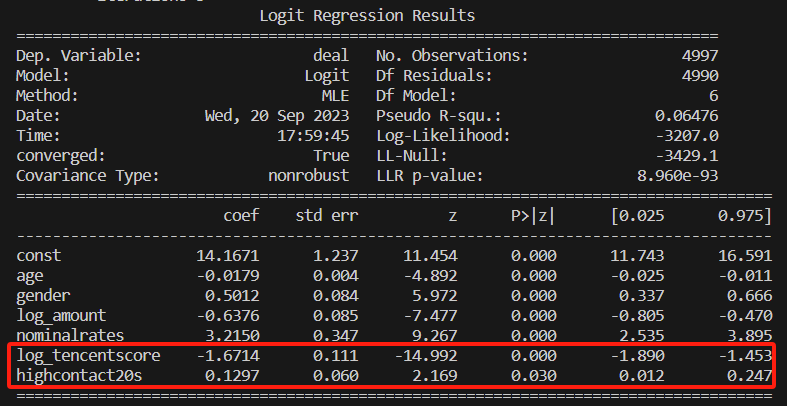
2) Perform a logit regression and examine the relation between the delinquency likelihood and credit scores

The result is provided below for this question. The default column has already been modified by dropna function. From the result, the coefficient is 0.4010 and the p-value is 0.002, which implies that the variable is significant. we know that there’s a positive correlation between the delinquency variable and the tencentscore, since if tencentscore is larger, it indicates there’s a higher risk for the borrower.



3) Perform a logit regression and examine the relation between the loan approval likelihood and credit scores.

The result is provided below for this question. From the result, the coefficient is -1.6714 and the p-value is 0, which implies that the variable is in 99.9% significance level. A borrower with a higher risk profile (log\_tencentscore) is less likely to be approved for his/her loan application.



4) Perform a logit regression and examine the relation between the loan approval likelihood and the dummy whether the borrower has a frequent contact.

In this question, the regression model is the same as in question 3, since we need some more variables to control the other effect in the whole regression. The result shows that the coefficient is 0.1297 and the p-value is 0.03, which implies that the more frequent contact a borrower has, he is more likely to be given an approval. A suitable implementation is that if the borrower has more friends or relatives, their likelihood to default is lower since they can usually find someone to fill the vacancy even if they don’t have enough money themselves.