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.

	count	mean	std	min	25%	50%	75%	max
age	5000.0	27.675400	8.326146	18.00000	21.000000	25.000000	32.000000	56.000000
gender	5000.0	0.146600	0.353742	0.00000	0.000000	0.000000	0.000000	1.000000
instalments_amount	5000.0	406201.420000	130623.360240	50000.00000	320000.000000	398000.000000	498000.000000	869000.000000
nominalrates	4997.0	0.276058	0.085912	0.13008	0.204560	0.204579	0.359347	0.494185
tencentscore	5000.0	58.608168	14.218112	9.00000	53.888889	60.200000	65.258929	98.000000
highcontact20s	5000.0	0.502200	0.500045	0.00000	0.000000	1.000000	1.000000	1.000000
deal	5000.0	0.441400	0.496604	0.00000	0.000000	0.000000	1.000000	1.000000
default	2205.0	0.419501	0.493589	0.00000	0.000000	0.000000	1.000000	1.000000

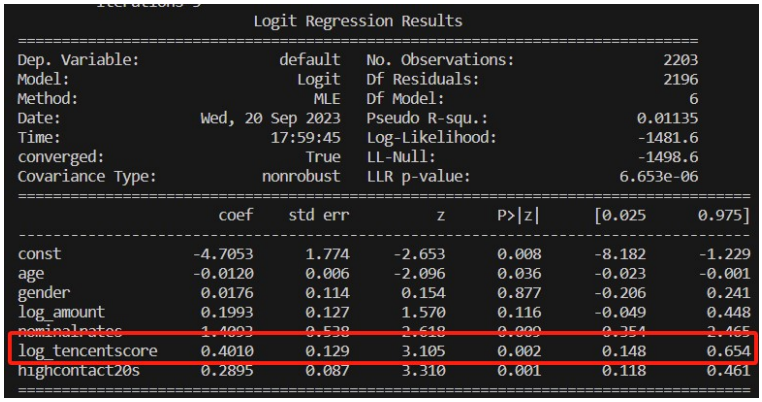
From this question on, in the following 3 questions, I choose 'age', 'gender', 'log_amount', 'nominalrates', 'log_tencentscore', and 'highcontact20s' as the independent variables, intentionally to control other characteristics of loan and borrower.

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

(1)Gender: 1 for Female, 0 for male; (2)instalments_amount: of loan principal in thousands of Chinese Yuan; (3)Nominalrates: interest rate of loan on annual basis; (4)Tencentscore: the credit score, and logarithm value is taken in regression analysis; (5)Highcontact20s: An indicator variable equal to one if a borrower has at least 3 frequent contact (last for more than 20 seconds), and zero otherwise.

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.



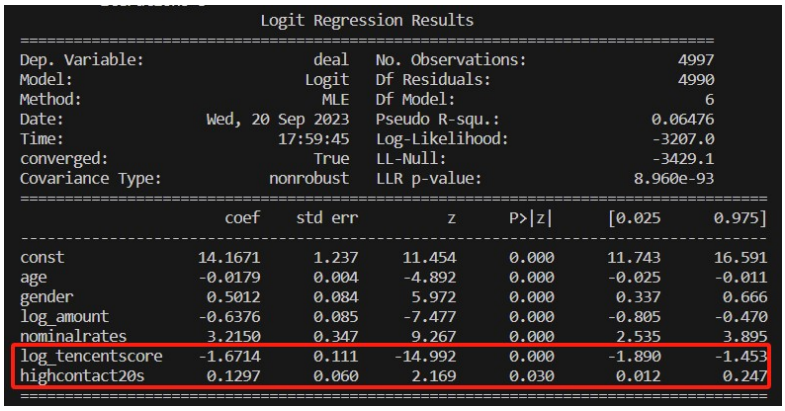
Logit Regression Results

Dep. Variable:	default	No. Observations:	2203
Model:	Logit	Df Residuals:	2196
Method:	MLE	Df Model:	6
Date:	Wed, 20 Sep 2023	Pseudo R-squ.:	0.01135
Time:	17:59:45	Log-Likelihood:	-1481.6
converged:	True	LL-Null:	-1498.6
Covariance Type:	nonrobust	LLR p-value:	6.653e-06

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.7053	1.774	-2.653	0.008	-8.182	-1.229
age	-0.0120	0.006	-2.096	0.036	-0.023	-0.001
gender	0.0176	0.114	0.154	0.877	-0.206	0.241
log_amount	0.1993	0.127	1.570	0.116	-0.049	0.448
nominalrates	1.4093	0.538	2.618	0.009	0.354	2.465
log_tencentscore	0.4010	0.129	3.105	0.002	0.148	0.654
highcontact20s	0.2895	0.087	3.310	0.001	0.118	0.461

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.



Logit Regression Results

Dep. Variable:	deal	No. Observations:	4997
Model:	Logit	Df Residuals:	4990
Method:	MLE	Df Model:	6
Date:	Wed, 20 Sep 2023	Pseudo R-squ.:	0.06476
Time:	17:59:45	Log-Likelihood:	-3207.0
converged:	True	LL-Null:	-3429.1
Covariance Type:	nonrobust	LLR p-value:	8.960e-93

	coef	std err	z	P>|z|	[0.025	0.975]
const	14.1671	1.237	11.454	0.000	11.743	16.591
age	-0.0179	0.004	-4.892	0.000	-0.025	-0.011
gender	0.5012	0.084	5.972	0.000	0.337	0.666
log_amount	-0.6376	0.085	-7.477	0.000	-0.805	-0.470
nominalrates	3.2150	0.347	9.267	0.000	2.535	3.895
log_tencentscore	-1.6714	0.111	-14.992	0.000	-1.890	-1.453
highcontact20s	0.1297	0.060	2.169	0.030	0.012	0.247

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.