Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/92/20230491

Predicting Models of Financial Crises

Yao Lu[1,a,*]

[1]*School of Mathematics, Nanjing Audit University, Nanjing, China*
a. 213080520@stu.nau.edu.cn
**corresponding author*

Abstract: This study used data from 1870 to 2008 to test the role of logistic regression models in predicting financial crises. This article first dealt with missing values in the data and removed time trends. Then, logistic regression models were used to predict and explore the occurrence of financial crises. During the exploration process, this article adopted principal component analysis and the method of eliminating multicollinearity interference for model optimization, and ultimately found that the area under the curve (AUC) in logistic regression was the highest, at 0.86. This indicates that logistic regression models provide more reliable predictions for financial crises. Based on the results, we further speculate that the total asset value and GDP are very important factors determining whether a financial crisis will occur. While there is room for improvement, the conclusion for this study still provides valuable insights into the connection between financial crises and specific banking factors.

Keywords: financial crisis, predicting model, differencing method, Logistic Regression

1. Introduction

The financial crisis has been paid a lot of attention in the latest literature. During the past few decades, the financial crisis has caused many obstacles for humans to have development. A typical example is the global financial crisis in 2007-08 which significantly caused a drop in economic development within the whole world. Due to the sudden decrease in financial assets, policymakers and enterprises may not have a quick response to the crisis toward the crisis, the financial situation would be definitely worse. To prevent this happens, prediction of the financial crisis is necessary for pre-noticing decision-makers in the economy.

Different features of the financial crisis have been studied by economists in order to identify an accurate methodology for predicting financial crisis using the data from historical years [1,2]. The prediction aims to forewarn the decision-makers within the economy to prevent or reduce the damage caused by the crisis as much as possible by taking advanced actions. Due to the impact of decision-making, the accuracy of prediction is fatal in the model. Although the existing literature has already suggested multiple methodologies for predicting the financial crisis, the increasing complexity of the market continuously increased the difficulty of prediction. Trying to overcome the difficulties, the study aims to use the existing data on monetary policy, Leverage policy, cycles, and financial crisis from 1870 to 2008 to construct a regression model to get an initiatory model for the prediction of the financial crisis. By comparison among the random forest model, the xtlogit model and the logistic regression model, the study suggested a possible prediction model for the financial crisis. Besides, in the wake of AI boom, machine learning and deep learning have been in use for financial crisis prediction.

Although the existing literature has already suggested multiple methodologies for predicting the financial crisis, the increasing complexity of the market continuously increased the difficulty of

DOI: 10.54254/2754-1169/92/20230491

prediction. Trying to overcome the difficulties, the study aims to use the existing data on monetary policy, Leverage policy, cycles, and financial crisis from 1870 to 2008 to construct a regression model to get an initiatory model for the prediction of the financial crisis. By comparison among the random forest model, the xtlogit model and the logistic regression model, the study suggested a possible prediction model for the financial crisis.

2. Data Processing

2.1. Data Collection and Variable Setting

Thanks to the work of Schularick and Tylor's (2009) team, our data cover 14 countries from the year 1870 to 2008. The dataset they built is one of the most detailed and comprehensive ones recording macroeconomic indicators for such a long span of time. Table 1 displays the definitions of the core concepts in the original research. The data source can be viewed in https://www.openicpsr.org/openicpsr/project/112505/version/V1/view?path=/openicpsr/112505/fcr:versions/V1/Credit-BoomsAER_data_replication&type=folder.

Table 1: Definitions of core concepts

concept	definition
bank loan	the end-of -year total of unpaid domestic currency lending by domestic banks to domestic households and non-financial firms(lending within the financial system excluded)
bank asset	the end-of -year total balance sheet assets of all banks with national residency(foreign currency assets excluded)
money	official statistical publications like All Bank Statistics by the U.S. Federal Reserve, etc. with referencing the research of specific economist historians.

The variables are set as Table 2 shows:

Table 2: Variable settings

Variables	Description
crisisST	A dummy of 0-1 for a financial crisis in country i in year t
iso	country identifier
ccode	country code
loansgdp	bank loans/gdp
credgdp	bank assets/gdp
moneygdp	broad money/gdp
loansmoney	bank loans/broad money
credmoney	bank assets/broad money
lloansmoney	log of loans/money ratio
lcredmoney	log of credit/money ratio
lrgdp	log real gdp

DOI: 10.54254/2754-1169/92/20230491

Table 2: (continued).

lpc	log of CPI price level
lnm	log of narrow money
lm	log of broad money
lloans	log bank loans
lcred	log bank assets

2.2. Missing value analysis

After our initial observations, 12 of these variables were missing. We first performed an analysis of missing values, obtaining the following column for the t-value of the independent variance t-test table, which is shown in Table 3 below:

Table 3: Independent variance t-test table

crisisST	
variables	t
loansgdp	-.7
credgdp	-.8
moneygdp	-.7
loansmoney	-1.0
credmoney	-1.3
lloansmoney	-1.0
lcredmoney	-1.3
lm	-.7
lnm	.3
lloans	-.8
lcred	-1.0

It can be noticed that the columns of t-values for all variables are between [-2, -2], so we can exclude that the missing values are completely non-random.

When no manual missing values were filled in for the data, we simply removed the missing rows and did nothing else. When the data were filled manually with missing values, we used linear trend interpolation of neighbouring points and compared this with the effect of not filling the missing values [3]. Based on this data, we obtained descriptive statistics for the 13 columns of data, including: means, standard deviations, and indicators of significance in relation to "whether a financial crisis occurred". After filling in the missing values using the 'linear trend at neighbouring points', the resulting line follows the linear trend and is a good fit. We still believe that it is a better decision not to use missing value filling. Therefore, we will simply delete the rows containing the null values before examining the follow-up questions.

2.3. Elimination of time series trends

The simplest way to de-trend a time series is by differencing it. We have done some first and second order differencing of the characteristic columns. The formula and code are as follows (equation 1):

DOI: 10.54254/2754-1169/92/20230491

$$value(t) = observation(t) - observation(t-1) \quad (1)$$

Some of the data after de-trending are shown below in Table 4:

Table 4: Data after de-trending

d_loansgdp	*d_credgdp*	.....	*d_lcred*	*d_lrgdp*	*d_lpc*
0.039898053	0.018583864	.....	0.251627445	0.00769043	-0.015503883
0.103350654	0.089733094	.....	0.345909119	0.07328701	0
.....	.....	.....	.....	.....	.....
0.024058104	0.026351511	.....	0.08613205	0.0187006	0.021673203
0.011889726	0.021480918	.....	0.0532341	0.033379555	0.007641315

3. Logistic regression model

3.1. Improvement of the panel logistic regression model

3.1.1. Model improvement 1: Principal component analysis to reduce the dimensionality of variables

We conducted principal component analysis for variables other than moneygdp and lrgdp. First, solve for the correlation coefficient matrix of the dependent variable x matrix $R = \frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2 \sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}}$:

The results of the KMO test show that the value of KMO is 0.61 and KMO>0.6, which indicates that there is a correlation between the variables of the question items, which meets the requirements of the principal component analysis [4]. Also, the results of the Bartlett's spherical test showed a significance p-value of 0.000*** [5].

Contribution of each principal component $= \frac{\lambda_i}{\sum_{k=1}^{P}\lambda_k}(i-1,2,\ldots,p)$. We calculate that setting 4 principal components is appropriate (the first 4 principal components account for 76.839%, close to 80%). And we can also see by the gravel plot below that the slope becomes smaller after the fourth principal component.

The definitions of the four categories were collated as follows:

$F1 = 0.1749 \times credgdp + 0.1131 \times lcred + 0.1420 \times credmoney$. We have named F1 "asset". $F2 = 0.1170 \times loansgdp + 0.0623 \times lcredmoney + 0.1230 \times lloansmoney + 0.1191 \times lloans$ .We have named F2 the "liability ". $F3 = -0.1534 \times lm - 0.1120 \times lnm + 0.1382 \times lcredmoney + 0.1292 \times lloansmoney + 0.1336 \times loansmoney + 0.1275 \times credmoney$. We have named F3 the "Currency ". $F4 = -0.3170 \times loansgdp - 0.2454 \times credgdp + 0.3244 \times lpc$. We have named F4 the "gross product ".

We added the original two variables moneygdp, and we can see that the principal component variables of **"asset class"** also show a degree of significant correlation.

3.1.2. Removing the interference of "multicollinearity"

Multicollinearity refers to the high degree of correlation between the independent variables in a multiple linear regression model, which can lead to inaccurate estimates of the regression coefficients and deviations from the true values, thus making the model results unstable.

DOI: 10.54254/2754-1169/92/20230491

We solved this problem by manually moving out the covariates. First, we did a correlation analysis of the four variables for which we had already reached a "significance" conclusion, resulting in the following matrix of correlation coefficients (equation 2):

$$\begin{pmatrix} 1 & 0.003 & 0.022 & 0.173 \\ 0.003 & 1 & 0.001 & 0.006 \\ 0.022 & 0.001 & 1 & 0.009 \\ 0.173 & 0.006 & 0.009 & 1 \end{pmatrix} \ldots\ldots\ldots \quad (2)$$

From the above matrix, R^2 are < 0.7, which excludes the correlation within the independent variables. In addition, the analysis of the results of the $F-test$ can be obtained that the significance $P-value$ is 0.000***, which presents significance at the level and rejects the original hypothesis that the regression coefficient is 0. Therefore, the model basically meets the requirements. And VIF all $=1$, for the performance of variable co-linearity, VIF all less than 10, so the model does not have the problem of multiple co-linearity, the model is well constructed [6].

3.2. Determining the final fit curve and associated factors

In addition to the above considerations, we also added the square term to better fit the nonlinear relationship. After we have gone through the above series of adjustments, we get the final results as shown in Table 5 below:

Table 5: Significance level and results of various indicators

Item	*B*	*S.E.*	*Wald*	*Sig.*	*Exp(B)*
REGR factor score 1 for analysis 1	-0.170	0.113	2.266	0.100*	0.844
lpcsquare	-0.219	3.613	0.004	0.052**	0.803
d_lrgdp	-11.583	2.995	14.954	0.000***	0.000
d2_moneygdp	-4.569	2.567	3.169	0.075*	0.010
Constant	-3.001	0.134	498.734	0.000***	0.050

where each $\hat{\beta}_i$ value is the "regression coefficient" column. Therefore, by substituting the data $\hat{\beta}_0 = -3.001, \hat{\beta}_1 = -0.170, \hat{\beta}_2 = -0.219, \hat{\beta}_3 = -11.583, \hat{\beta}_4 = -4.569$, the fitted curves between the dependent variable "the occurrence of a financial crisis" and the four significant independent variables (i.e. the principal component variables F1, Zlpc_square - the square of the log of the CPI price level, lrgdp - the log of real GDP and moneygdp - broad money/gdp) are: $P(y_i = 1|x) = \frac{e^{-3.185-0.4x_1+0.032{x_2}^2+0.022{x_3}^2+0.195x_4}}{1+e^{-3.185-0.4x_1+0.032{x_2}^2+0.022{x_3}^2+0.195x_4}}$ [7]. Where, the expression F1 = 0.1749credgdp+0.1131lcred+0.1420credmoney, where F1 represents the principal component of the "asset class".x_1 indicates CPI price level, x_2 denotes logarithm of real GDP, x_3 denotes broad money/gdp.

3.3. Analysis of the fitting effect

Finally, we have analyzed the effectiveness of the model implementation. The effectiveness of the clustering implementation of logistic regression is further measured by quantitative metrics. We

DOI: 10.54254/2754-1169/92/20230491

derived accuracy, recall, precision, F1 values and AUC values respectively (as shown in Table 6), all very close to 100% and effective.

Table 6: Effectiveness of the clustering implementation of logistic regression

Accuracy	Recall rate	Accuracy	F1	AUC
0.955	0.955	0.912	0.933	0.891

At this point, the logistic regression model is completed.

3.4. Analysis of results

Given that we propose this model: $P(y_i = 1|x) = \frac{e^{-3.001-0.170F_1-0.219{x_1}^2-11.583x_2-4.569x_3}}{1+e^{-3.001-0.170F_1-0.219{x_1}^2-11.583x_2-4.569x_3}}$,We explore whether there are economic implications for these four variables.

When F1 tends to infinity, the probability will go to zero. This means that an increase in F1 will help to avoid financial crises. We define F1 as the principal component variable of the "asset class". The graph shows that the two folds overlap to a high degree (shown in Figure 1 below). And, from common sense economics, we know that an increase in total assets indicates that: the faster a company expands the scale of its asset operations over a certain period of time, the better the business is doing. It confirms our suspicion.

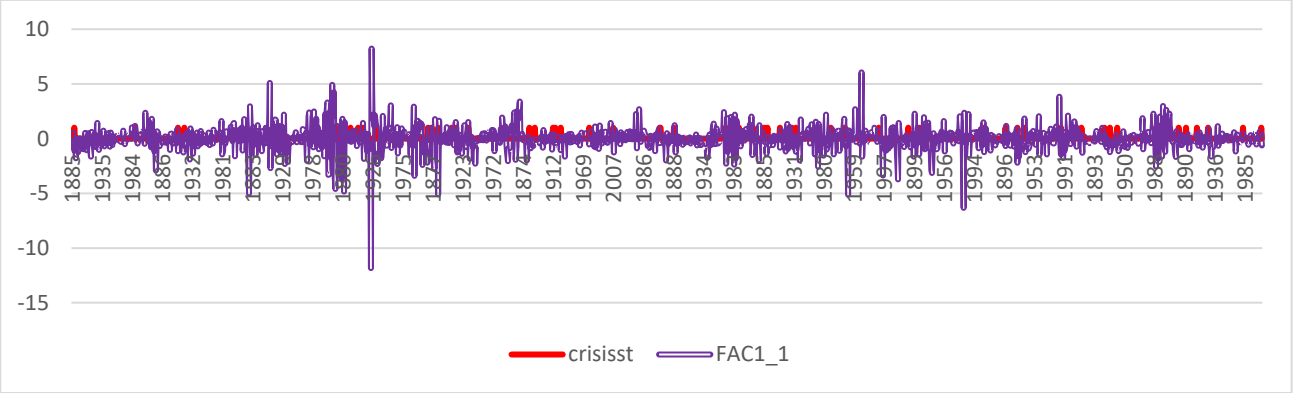

Figure 1: Trend chart of F1 and crisis occurrence

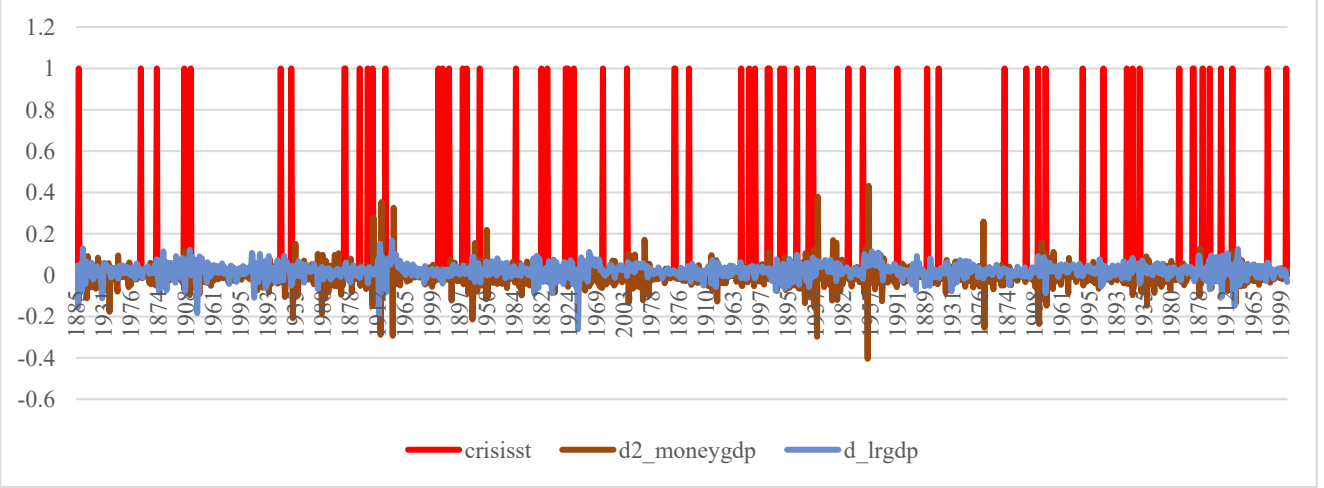

Figure 2: Trend chart of other 2 variences and crisis occurrence

DOI: 10.54254/2754-1169/92/20230491

Similarly, a line plot of the other three variables against the occurrence of a financial crisis was made and a least squares linear trend was explored (shown in Figure 2 above). We also find that lrgdp fits very well for the occurrence of a financial crisis, while the other two variables do not have a linear trend, the magnitudes are not consistent with the dependent variable. Lrgdp represents the natural logarithm of true GDP. That is, GDP growth suppresses financial crises.

4. Conclusion

In this paper, we aim to accurately forecast a financial crisis's occurrence by using adaptive mathematical model. Our approach involves the utilization of a logistic regression mode. By employing the most advanced data preprocessing technique, our study successfully attained the highest Area Under the Curve (AUC) score of over 0.85 using the logistic regression model. This indicates a significant level of accuracy in our predictions.

Due to the low probability of occurrence and insufficient data, the financial crisis is always considered unpredictable. In this study, we have discovered that by leveraging data preprocessing techniques and machine learning algorithms on specific factors related to banks, we can uncover a certain level of predictability. While there is room for improvement, they still provide valuable insights into the connection between financial crises and specific banking factors. This knowledge holds significant importance for both banks and the broader public economics.

References

[1] *Kaminsky, G.l. and Reinhart, C.M.(1999) The Twin Crises The Causes of Banking and Balance of Payments Problems. American Economic Review, 89, 473-500.*

[2] *Maryam Maryam, Dimas Ayro Anggoro, Muhibah Fta Tika and Fitri Cahya Kusumawati.(2002) An Intelligent Hybrid Model Using Artificial Netural Networks and Particle Swarm Optimization Technique For Financial Crisis Prediction. Pakistan Journal of Statistics and Operation Research, Vol.18 No.4 2022 pp 1015-1025.*

[3] *Yuan Zhongru. Acomparison of the effects of missing data filling methods in multiple linear regression models [D]. Zhongnan University, 2008.mis*

[4] *Chenyi Yang; - 《Proceedings of 2022 International Conference on Agriculture, Forestry and Economic Management (AFEM 2022)》- 2022-05-28*

[5] *Yanyan Han;Ke Xu;Jiayin Qin; - 《Proceedings of 2023 International Conference on Mathematical Modeling, Algorithm and Computer Simulation (MMACS 2023)》- 2023-02-25*

[6] *Haonan Yu;Hao Cheng;Wanting Zhan; - 《Proceedings of 5th International Conference on Chemical Engineering and Advanced Materials (CEAM 2022)》- 2022-09-24*

[7] *Ruan Hongfang. Research on financial early warning of manufacturing industry based on principal component analysis-logistic model [D]. Anqing Normal University, 2022. Doi:10.27761/d.cnki.gaqsf.2022.000018.*