Predicting Models of Financial Crises

Abstract

In this study, three models are examined in the prediction of the financial crisis by using the data from 1870 to 2008. The first model is the Random Forest model (RF model), the second is the panel logistic model (xtlogit model) and the third is the logistic regression model. Various consequences are found in different models. We found the highest Area Under the Curve (AUC) of 0.86 in the logistic regression model. This indicates that the logistic regression model provides a more reliable prediction for the financial crisis.

Keywords financial crisis | predicting model | differencing method | Random Forest | SMOTE | Xtlogit | Logistic Regression

Chapter 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. According to Piruna Polsiri (2009), improving a well-prepared early-warning system for a financial crisis could notice the probability of an immediate recession in an economy. A successful prediction system could be a strong signal for the government and enterprises to prevent bankruptcy or sudden collapse.

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. 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.

This paper is organized as follows: Section 2 briefly introduces previous literature related to the prediction of the financial crisis, Section 3 is the data descriptions, Section 4 is the 3 methodologies we used to predict the financial crisis, Sections 5 to 7 detail the build of the random forest model, Xtlogit model and logistic regression model respectively, Section 8 discussed the consequence and limitations on the models that we examined, and Section 9 is the final conclusion.

Chapter 2

Literature Review

Financial crises have been outlined in several research. Kaminsky and Reinhart (1999) propose the well-known “twin crisis”, depicting the interaction between banking and currency crises. Considering currency crises do not necessarily impact the financial sector, Schularick and Taylor (2009) specify that financial crises occur when a nation’s banking system experiences bank runs, a significant increase in default rates, *etc*. Similarly, Jordà *et al.* (2011) regard banking crises as equivalent to financial crises to recognize 79 major events in 14 countries from 1870 to 2008. These authors mainly use event-analysis to define the dependent variable. More recently, the Financial Stress Index (FSI), initially introduced by Mark Illing and Ying Liu (2003), has been used by Maryam *et al.* (2022). With the FSI combining the US dollar interest rate, exchange rate, and foreign exchange reserves, the authors select an appropriate threshold value for the criterion to identify crisis conditions from January 1990 to December 2021 in Indonesia. Whether a crisis is found or not largely depends on the parameter and the variables in the formula they predetermine.

There are numerous methods employed to figure out the primary factors of financial crises. Probit and Logit Regressions are widely used among traditional methodologies. Schularick and Taylor (2009) give up the Linear Probability model for its identical variation in a probability outcome with changes to independent variables in different unit intervals, and turn to the Logit model. The authors conclude that a slowdown of credit growth booming is a reliable predictor for financial crises in coming years. Gourinchas and Obstfeld (2012) utilize Logit model to confirm their event study results. Their results corroborate the ideas of Schularick and Taylor (2009), and discover two additional robust determinants: real currency appreciation and the level of foreign exchange reserves, especially for crises in Emerging Market Economies.

In the wake of AI boom, machine learning and deep learning have been in use for financial crisis prediction. Giovanis (2010) has challenged the setting of dummy binary crisis variable in traditional models. He distinguishes between economic recessions and depressions, which have different effects on GDP, and believes they should be given separate values rather than both being given 1. Similar treatment should be taken between post-economic recovery periods and economic expansion period. He then finds that an ANFIS model provides a superior signal compared with Logit and Probit models in the prediction of 2007/08 crisis in USA. The last decade has seen a growing trend towards the approach of Artificial Neural Network (ANN). Through this technique, while studies such as the one conducted by Aydin and Cavdar (2015) report that foreign exchange regime is playing an increasingly significant role in occurrence of crises in the 21st century, Nik *et al.* (2016) find that domestic credit to private sector is of the greatest importance among all the variables. The research could have been more persuasive if the foreign exchange market had been included in their model. A hybrid model with ANN and Particle Swarm Optimization (PSO) is built by Maryam *et al.* (2022). Under the fittest weights and thresholds, their model attains a 96.8% accuracy that forecasts a crisis correctly in the test set.

Chapter 3

Methodology

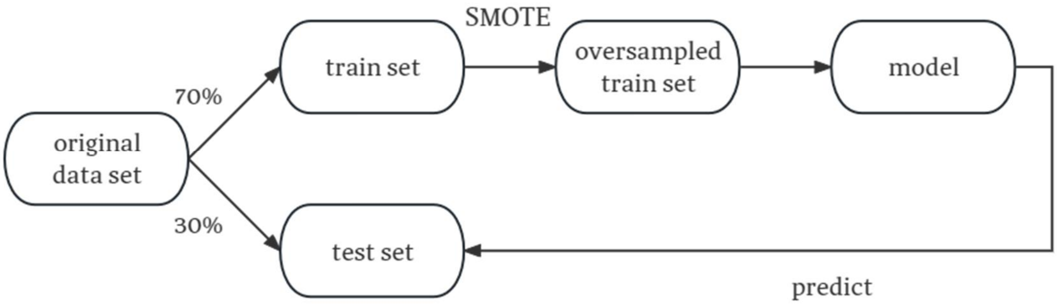
This research paper aims at the prediction of the occurrence of a financial crisis based on several factors in terms of banking and finance. Our original database is from the paper "Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008" (Schularick and Taylor, 2012), which contains data on economic and banking factors in 14 countries over 100 years. The details of the data will be described in the section of the data description.

Since the dependent variable of whether a financial crisis will happen is only consisted of 2 possibilities: a financial crisis can either happen or not happen, which is labeled as 0 and 1 in the data set, so, it is obvious that we are facing a classification problem, and to address this problem, different classification methods should be applied.

Here, to examine the occurrence of a financial crisis, we employed 3 machine learning methods: random forest, logistic regression, and panel logistic regression with fixed effect.

3.1 Random Forest and SMOTE

As a popular machine learning algorithm, Random Forest (RF) is widely used in various fields due to its versatility and robustness. The foundation of RF is built upon decision trees. With the RF, decision trees are produced by randomly selecting data from a dataset and replacing it. The best features from a subset of several features randomly selected from the retrieved dataset are then utilized for tree splitting. To create the final training model, the RF mixes all of the decision trees. When the individual trees in the RF are good at classification, its prediction accuracy increases.



In machine learning, the Synthetic Minority Over-Sampling Technique (SMOTE) is a well-liked method of data augmentation, notably in the area of imbalanced classification, describing a situation in which a dataset's classes are not evenly represented, with one class having considerably less samples than another. In the case we meet, the variable "crisisST" takes the value of 1 (indicating the occurrence of a financial crisis in a specific country in a certain year) only around 5% of the time, which constitutes a serious case of classification imbalance. Therefore, we suggest using the SMOTE approach to overcome this issue.

3.2 The Panel Logistic Regression

We also consider a panel logistic regression since the data we currently have can be grouped based on the year and the country, therefore it is obvious that a regression, which is used specifically for the analyzation of panel regression, can be employed here. Different from the traditional logistic regression, here, in the panel logistic regression we must consider an additional parameter of time t due to the existence of the time series.

Generally, as a model with time series, the panel logistic regression can be divided into 2 types: the one with the fixed effect and the one with the random effect.
Here, we employed the panel logistic regression with the fixed effect, which is based on the following formula:

$$y = \frac{e^{\alpha_i + \beta x_{it}}}{1 + e^{\alpha_i + \beta x_{it}}}$$

It is clear that the parameter α does not rely on the time parameter t, suggesting that each country will have an individual effect of α, that does not vary through time. This is called the fixed effect, the β, however, will be determined by both time t and the country code. Similar to the traditional logistic regression, the panel logistic regression also gives out the probability of the dependent variable being 1.

3.3 Logistic Regression

Through our research, we found out that it is difficult to predict a financial crisis since a financial crisis happened rarely in the last century. Therefore, on the one hand, there are only limited data available, especially the one with the value of the dependent variable of 1 (occurrence of a financial crisis). On the other hand, the damage of a financial crisis is massive, and the financial factor therefore cannot rely on a model that cannot give out an exact prediction. However, after the research, we found out that, especially with the 2 logistic regressions, by applying some data preprocessing strategies in advance, we have gained high precision and recall scores, and the AUC (the area under the ROC curve) is significantly higher than 0.5, suggesting that with enough data and a various kind of machine learning techniques, the financial crisis is not entirely unpredictable. During the regression, we can also see whether some of the factors have a strong positive impact on the probability of the occurrence of a financial crisis, which means a lot for the banking factors and the government, because these factors can be a sign of the financial crisis, and the growth of those factors can be detrimental to the whole economy.

Chapter 4

Data Description

4.1 Preliminary Data Processing

4.1.1 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. This table displays the definitions of the core concepts in the original research

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 data source can be viewed in detail in https://www.openicpsr.org/openicpsr/project/112505/version/V1/view?path=/openicpsr/112505/fcr:versions/V1/CreditBoomsAER_data_replication&type=folder.

We grouped the data into the |credit crisis_data_for_project file. The variables are set as follows:

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

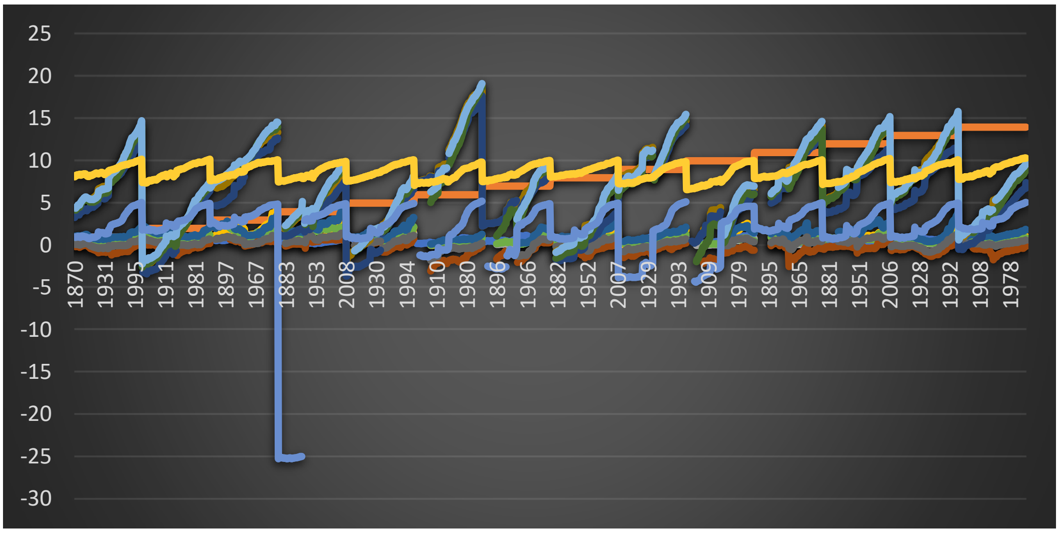
4.1.2 Missing Value Analysis

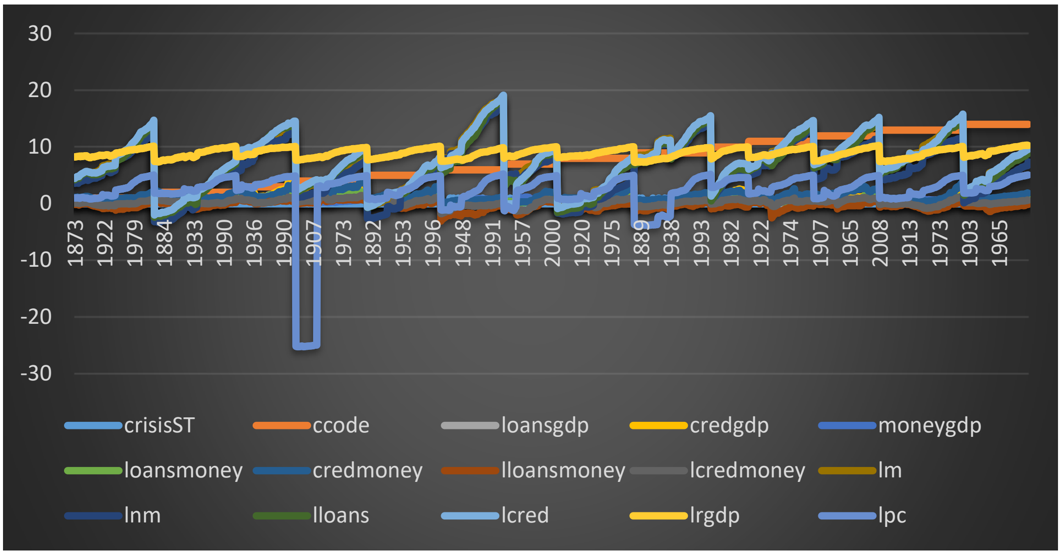
After our initial observations, 12 of these variables were missing. We first performed an analysis of missing values, obtaining the following column for the t-value of the independent variance t-test table:

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

In general, a larger t-value indicates a higher degree of confidence in a statistical difference. In general, a significant difference is considered to exist when the t-value is greater than 2. 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.

As least squares is not concerned with any data gaps, it simply fills in the relevant independent and dependent variable values for matrix calculations. Therefore, for random missing values, we have explored comparisons as to whether or not they need to be filled in. 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. 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", as shown below.





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.

not fill in missing values					fill in missing values				
	N	Mean	Std. Deviation	Sig.		N	Mean	Std. Deviation	Sig.
loansgdp	1377	0.5188	0.4030	0.3986	TREND(loansgdp)	1736	0.4933	0.3735	0.9954
credgdp	1359	0.8902	0.6041	0.0961	TREND(credgdp)	1736	0.8843	0.5496	0.3197
moneygdp	1377	0.6133	0.2336	0.2649	TREND(moneygdp)	1736	0.5933	0.2211	0.2226
loansmoney	1377	0.8301	0.4714	0.5691	TREND(loansmoney)	1736	0.8053	0.4375	0.7860
credmoney	1359	1.4433	0.6925	0.6439	TREND(credmoney)	1736	1.4519	0.6292	0.7206
lloansmoney	1377	-0.3646	0.6381	0.5948	TREND(lloansmoney)	1736	-0.4100	0.6149	0.5696
lcredmoney	1359	0.2472	0.5107	0.2801	TREND(lcredmoney)	1736	0.2529	0.4622	0.9388
lm	1377	7.2778	4.3663	0.012	TREND(lm)	1736	6.8872	4.1280	0.4968
lnm	1377	5.9669	4.4032	0.0129	TREND(lnm)	1736	5.5189	4.1942	0.0610
lloans	1377	6.9132	4.3618	0.2684	TREND(lloans)	1736	6.5585	4.0947	0.2684
lcred	1359	7.5569	4.4304	0.9857	TREND(lcred)	1736	7.4317	4.0262	0.9857
lpc	1377	2.0471	4.5418	0.0133	TREND(lpc)	1736	1.4936	4.8176	0.0106

Before and after filling in the missing values, descriptive statistics such as means and standard deviations were generally similar, but differed significantly from the significance indicator of "whether or not a financial crisis occurred", and even made variables that were previously significant insignificant. This suggests that when a large amount of missing data is filled in, it may replace the original relationship.

In summary, 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.

4.1.3 Elimination of Time Series Trends

The simplest way to de-trend a time series is by differencing it. Specifically, a new series is constructed by calculating the difference between the previous observation and the observation on the basis of equal time steps. We have done some first and second order differencing of the characteristic columns. The formula and code are as follows:

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

CODING

```
from pandas import read_csv
from pandas import datetime
from matplotlib import pyplot

def parser(x):
    return datetime.strptime(x, '%Y-%m')

series = read_csv('1.csv', header=0, parse_dates=[0], index_col=0, squeeze=True,
date_parser=parser)
X = series.values
diff = list()
for i in range(1, len(X)):
    value = X[i] - X[i - 1]
    diff.append(value)
pyplot.plot(diff)
pyplot.show()
```

* Note: In the data, the beginning of d_ is the first order difference, and the beginning of d2_ is the second order difference, where such as loansgdp and taking the natural logarithm of lloansmoney, as the second order difference and taking the logarithm after the first order difference can be regarded as the rate of change, so we only keep the first and second order difference of loansmoney, and the logarithm of it is not made difference again.

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

d_loansgdp	d2_loansgdp	d_credgdp	......	d_lcred	d_lrgdp	d_lpc
0.039898053		0.018583864	......	0.251627445	0.00769043	-0.015503883
0.014537349	-0.025360703	0.021353453	......	0.267232895	0.07444191	-0.048009336

0.103350654	0.088813305	0.089733094	......	-0.345909119	0.07328701	0
......	......	......	......	......	......	......
0.024058104	0.000296801	0.026351511	......	0.08613205	0.0187006	0.021673203
0.028136998	0.004078895	0.040108621	......	0.086397171	-0.004910469	0.040363789
0.011889726	-0.016247272	0.021480918	......	0.0532341	-0.033379555	0.007641315

Chapter 5

Random Forest Model

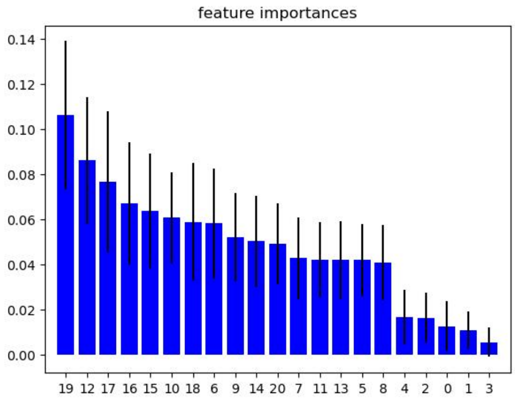
Our first model is Random Forest. Alessi and Detken (2011) and Joy *et al.* (2017) employed the RF method to predict financial crises in earlier studies, with promising findings. In our first RF model, we use all differential variables, the year of 1945 as the threshold, and the continents as features. To start off, we split the original dataset into a training set and a test set with a ratio of 7:3. Due to the issue of class-imbalance in the dataset we've mentioned earlier, we used the SMOTE method to balance the number of 0's and 1's in the "crisisST" variable. Based on python 3.11 environment, we use the automated tuning function "GridSearchCV" from the sklearn and then select the optimal hyperparameter values:

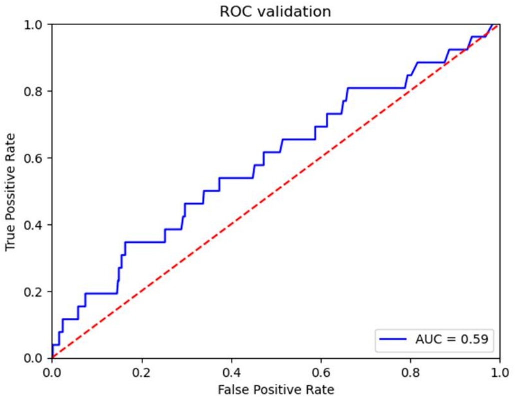
$$'criterion' = 'gini'$$
$$'min\_samples\_leaf' = 1$$
$$'min\_samples\_split' = 2$$
$$'n\_estimators' = 700$$

We sort the feature importance values after determining the best hyperparameters, then compute the accuracy, precision, recall, F1 score, and AUC, plotting the AUC curve as well. The figures below display the outcomes.

No.	Variables	Importance
19	d_lrgdp	0.106201
12	d2_loansmoney	0.086017
17	d_lloans	0.076739
16	d_lnm	0.067028
15	d_lm	0.063803
10	d2_moneygdp	0.060640
18	d_lcred	0.058864
6	d2_loansgdp	0.058189
9	d_moneygdp	0.052030
14	d2_credmoney	0.050242
20	d_lpc	0.049237
7	d_credgdp	0.042700
11	d_loansmoney	0.042006
13	d_credmoney	0.041904
5	d_loansgdp	0.041827
8	d2_credgdp	0.040768
4	pre45	0.016663
2	continent_eu	0.016311
0	continent_oa	0.012621
1	continent_na	0.010648

3	continent_as	0.010648





Accuracy	Recall	Precision	F1	AUC
91.4%	91.4%	91.2%	91.3%	59%

From the results, we find that the two variables, d_lrgdp and d2_loansmoney, have the greatest impact on the outcome. They respectively represent the growth rates of real GDP and the ratio of loans to money. This result is reasonable. When there is a financial crisis, a nation's real GDP growth rate frequently experiences a significant setback. When the growth rate of GDP shows a confusing decline in a certain year, it may indicate that a financial crisis occurred that year.

Besides, a notable drop in the growth rate of the ratio of bank loans to broad money is observed, indicating a substantial decrease in credit activity during that year, which offers a strong explanation for financial crises. Conversely, the classifications of years before 1945 or not and continents exhibit the weakest ability to anticipate financial crises, which supports that the operational structures of economies and societies in different ages and across different nations are fundamentally alike, leading to minimal diversification in their reactions to financial crises.

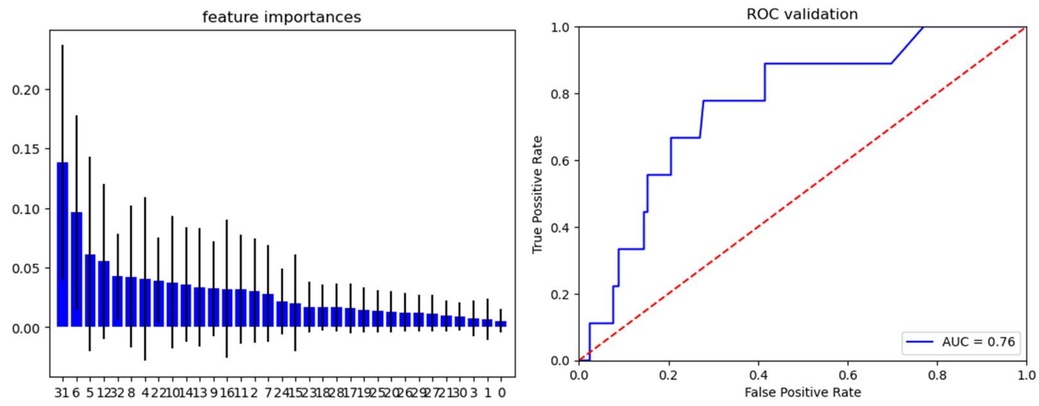
However, the model does not perform well, because the AUC value is only 0.59. We believe that this is not only due to the limited number of variables with strong explanatory power in the selected features, but also because we only adopt differential variables and remove all the original values. We do this because RF algorithm does not allow for null values, unlike in STATA where a sample including null values is automatically ignored.

In the data description section, it has been noted that certain countries lack available information regarding the pre-1945 values of various indicators. There is no compelling reason to fill these absent values; thus, we solely substitute the null values of differential variables with 0. This operation will have no bearing on the direction (positive or negative) when investigating the influence of a particular variable on the financial crisis. On the contrary, the elimination of these fundamental values would result in a loss of valuable information necessary for forecasting crises. This is incongruous with what an RF model wants: as many feature variables as possible.

We finally choose a compromise solution. Considering that the influence of the era is not strong in prediction, and that the original data from all countries in each year after 1945 is complete, we select variables as the feature values for our second RF model, using merely post-1945 data. Using the same operation steps as the first RF model, we obtain the results of the second one.

No.	Variables	Importance
31	d_lrgdp	0.138236
6	moneygdp	0.096321
5	credgdp	0.061093
12	lnm	0.055143
32	d_lpc	0.042593
8	credmoney	0.042456
4	loansgdp	0.040503
22	d2_moneygdp	0.039016
10	lcredmoney	0.037432
14	lcred	0.035786
13	lloans	0.033233
9	lloansmoney	0.032272
16	lpc	0.032203
11	lm	0.031914
2	continent_eu	0.030508
7	loansmoney	0.028287
24	d2_loansmoney	0.021915
15	lrgdp	0.020064
23	d_loansmoney	0.017028
18	d2_loansgdp	0.016706
28	d_lnm	0.016481
17	d_loansgdp	0.015823
19	d_credgdp	0.014765
25	d_credmoney	0.013479
20	d2_credgdp	0.012873
26	d2_credmoney	0.012524
29	d_lloans	0.011884
27	d_lm	0.011724
21	d_lcred	0.009646
30	d_lcred	0.008629
3	continent_as	0.007467

1	continent_na	0.006635
0	continent_oa	0.005360



Accuracy	Recall	Precision	F1	AUC
94.9%	94.9%	93.1%	94.0%	76%

We are pleasantly surprised to learn that the model's predictive power has significantly improved from the findings of the second RF model. The addition of original variables has enhanced the quantity of information that can be used for prediction, which is largely responsible for this improvement. It is clear from the feature importance ranking table that, in addition to the variable d_lrgdp, which is still the most significant feature, variables like moneygdp (the ratio of money to GDP) are also quite essential. This result is consistent with study conducted by Schularick and Taylor (2011).

By comparing the results of the two models, we can conclude that including more variables in the RF model contributes to improving its predictive ability. Although our second model only utilizes data from after 1945, if the original dataset had more abundant data, we might have obtained even more compelling conclusions.

Indeed, both of these models have their limitations. Firstly, there are few features with strong explanatory power. In the RF model, a single variable with strong explanatory power often surpasses multiple variables with weak explanatory power. However, in both models, we can only assert the outstanding performance of the d_lrgdp variable, while the importance of other variables is not stable enough. Secondly, our models do not consider lagged factors related to the years. We only used macro variables and their differentials for the same year to indicate whether a financial crisis occurred. However, in practical model applications, we deal with data from previous years and aim to predict the occurrence of financial crises in future years. Our models are not sufficiently adapted to real-world applications. Finally, we did not handle the time series characteristics effectively. Many data exhibit trends over time, and when we include the initial variables in the model, the trend factor can significantly impact the accuracy of predictions.

Chapter 6

Xtlogit Model

As already mentioned in the data description section, the data we have is across 14 countries with an observed period of over 100 years, therefore, the data set is a so-called panel data set, and there are several techniques that are used for the analysis of time series and panel data and can therefore be applied here.

In this section, we will try to forecast the possibility of the happening of a financial crisis with a logistic model with a fixed effect.

6.1 Model Constructing

Here, we applied a model called logistic model with fixed effect, which has the formula as already mentioned in the methodology section:

$$P(y_{it} = 1) = \frac{e^{\alpha_i + x_{it}\beta}}{1 + e^{\alpha_i + x_{it}\beta}}$$

Here, the fixed effect α is from the fixed effect of each country, and the β are various among all the observations. In STATA, this regression can be simply applied using the command xtlogit with the option fe. To evaluate the model performance, we used the criterion of the AUC, namely the area under the ROC curve as other classification methods.

6.2 Selection of Independent Variable

6.2.1 Backward Stepwise Selection

To determine which variable has the most significant effect, a method called backward stepwise regression can be applied. With backward stepwise regression, the variables, that do not have a massive impact on the independent variable will be gradually ignored in the further process.

Here is the result of panel logistic regression with a fixed effect with independent variables that are considered significant by the backward stepwise selection technique:

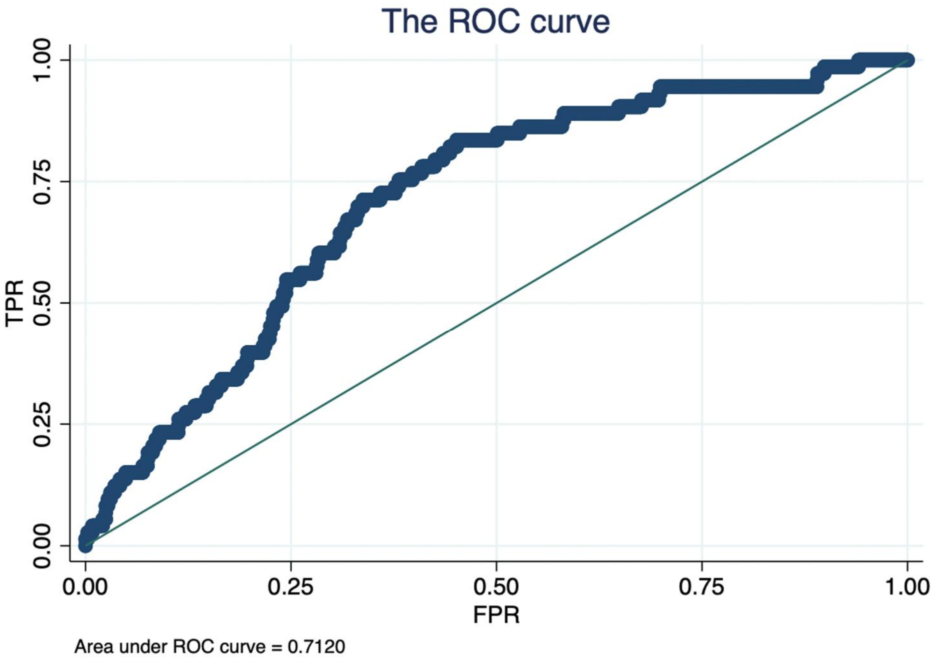
	B	S.E.	z	P-value	0.95-confidence interval	
pre45	.806	.276	2.92	0.004***	.264	1.348
d_lrgd	-8.490	2.898	-2.93	0.003***	-14.178	-2.815
d_moneygdp	8.809	3.936	2.24	0.025**	1.094	16.524
d2_moneygdp	-11.288	3.244	-3.48	0.001***	-17.646	-4.929
d2_loansmoney	-4.296	1.497	-2.87	0.004***	-7.230	-1.362

Note: ***, **,* represent 1%, 5%, 10% level of significance, respectively

We only considered the variables pre45, d_lrgdp, d_moneygdp, d2_moneygdp, and d2_loansmoney, since they have shown significance in the regression.

Furthermore, it is obvious that some of the significant terms here is the term with the second difference, this is might because after 2 years the change in the variable has become larger.

The ROC curve of this model is presented in figure 1, and this model's area under the ROC curve is about 0.71.



6.2.2 Adding Square Terms

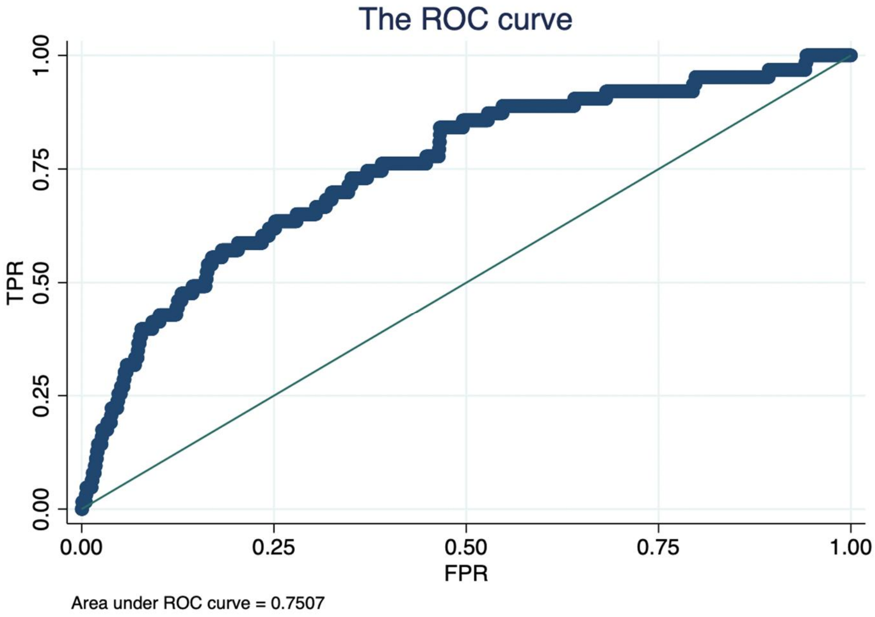
Since it is possible that the square term of some variables can also be significant, during the next backward stepwise selection, we try to add the square terms of each variable. As it turns out, only the square terms of d_credgdp_2 and d_lm are significant, implying the relationship between those two parameters and the probability of the occurrence of a financial crisis might not be linear. The coefficients of our new model are shown as follows:

	B	S.E.	z	p-value	0.95 confidence interval	
pre45	.900	.360	2.50	0.012**	.195	1.61
d2_loansmoney	-5.157	1.75	-2.95	0.003***	-8.59	-0.730
d_credgdp^2	18.120	5.276	3.43	0.001***	7.779	28.462
d_lpc	12.826	4.235	3.03	0.002***	4.525	21.127
d_lm^2	-18.480	5.693	-3.25	0.001***	-29.638	-7.322

d_moneygdp	23.969	5.727	4.19	0.000***	12.745	35.193
d2_moneygdp	-13.07	3.713	-3.52	0.000***	-20.347	-5.794
d_lpc^2	1.374	.689	2.00	0.046**	.024	2.724
d_lm	-10.852	3.153	-3.44	0.001***	-17.033	-4.673

Note: ***, **,* represent 1%, 5%, 10% level of significance, respectively

Moreover, the area under the ROC curve has slightly increased by 0.04 to approximately 0.75.



6.3 Results Checking and Interpretation

As the table illustrates, of all the independent variables (to the first power) we kept, only the coefficients of pre45, d_lpc, and d_moneygdp are positive, meaning that a financial crisis was more likely before 1945, which is possible due to the construction of the Bretton woods and the improvement of the banking system. A higher level of price growth rate may imply that the central bank is losing control of the pricing system because for some countries, controlling inflation is also part of the objective of the central bank. Broad money is less liquid compared to narrow money and therefore is less likely to be transferred to cash. In this case, if there is a huge amount of broad money in this economy, and the major of depositors want to retrieve their money from the banks, then the banks might not be able to pay back, which ultimately leads up to bankruptcy and financial crisis in this economy.

On the other hand, a larger amount of broad money could mean that in this economy there is enough money in this system, and therefore a large amount of narrow money that can deal with the bank liquidity problem, so the d_lm has a negative coefficient.

The loans-to-money ratio also has a negative coefficient here, which could be attributed to the fact that with more loans, the bank can gain more payback to repay the depositors to avoid bankruptcy.

Chapter 7

Logistic Regression Model

In order to study which key factors and patterns are associated with the occurrence of financial crises, for cases where the dependent variable is a categorical variable, we can use logistic regression for this purpose. The logistic regression model is a commonly used categorical model that can be used to predict binary or multivariate dependent variables. After initial processing of the data, we performed principal component analysis to downscale the data in order to prevent redundancy in the independent variables and to simultaneously ensure that the downscaled variables are still relevant for the study. Logistic regression models usually include three steps: feature selection, model training, and parameter tuning. Considering y as the probability of event occurrence, y ⩾ 0.5 means occurrence; y < 0.5 means no occurrence. After that, we add the squared term and interaction term to the model to explore the impact of more factors on “whether or not the financial crisis occurs”. After the final results are obtained, the model needs to be evaluated to determine its accuracy, usually using ROC curves, AUC and other metrics to assess the model performance.

7.1 First Build of a Logistic Regression Model

7.1.1 Significance Level Results

	B	S.E.	Wald	df	Sig.	Exp(B)
d_loansgdp	-3.759	8.546	0.193	1	0.660	0.023
d2_loansgdp	-2.088	6.572	0.101	1	0.751	0.124
d_credgdp	2.248	5.385	0.174	1	0.676	9.472
d2_credgdp	1.588	4.154	0.146	1	0.702	4.893
d_moneygdp	**15.494**	**7.533**	**4.230**	**1**	**0.040****	**5355292.909**
d2_moneygdp	**-13.404**	**6.067**	**4.882**	**1**	**0.027****	**0.000**
d_loansmoney	6.589	5.357	1.513	1	0.219	726.965
d2_loansmoney	-3.262	4.172	0.611	1	0.434	0.038
d_credmoney	-0.872	3.774	0.053	1	0.817	0.418
d2_credmoney	-2.011	3.007	0.447	1	0.504	0.134
d_lm	-1.022	1.513	0.456	1	0.499	0.360
d_lnm	-0.252	1.179	0.046	1	0.831	0.777
d_lloans	-1.547	0.868	3.175	1	0.075	0.213
d_lcred	0.472	0.944	0.251	1	0.617	1.604
d_lrgdp	**-10.033**	**3.327**	**9.096**	**1**	**0.003*****	**0.000**

d_lpc	1.020	2.628	0.151	1	0.698	2.773
Constant	**-3.135**	**0.168**	**349.745**	**1**	**0.000*****	**0.043**

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

Conclusion: moneygdp, lrgdp and the constant term have a significant relationship with “the occurrence of a financial crisis”. The other variables do not show significance at the level of the original hypothesis, and therefore d_lpc does not have a significant effect on the dependent variable.

Fitting results: The final determination of the fitted curve. When the logistic regression model is operational and well realised, the following formula is generally used.

$$P(y_i = 1|x) = S(x_i^{'}\hat{\beta}) = \frac{\exp(x_i^{'}\hat{\beta})}{1+\exp(x_i^{'}\hat{\beta})} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+...+\hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+...+\hat{\beta}_k x_{ki}}}$$

where each $\hat{\beta}_i$ value is the “regression coefficient” column. Therefore, by substituting the data $\hat{\beta}_0 = -3.135, \hat{\beta}_1 = -13.404, \hat{\beta}_2 = -10.033$, the fitted curve between the dependent variable “the occurrence of a financial crisis” and the two significant independent variables (i.e., lrgdp - the logarithm of real GDP and lpc - the logarithm of the CPI price level) is: $P(y_i = 1|x) = \frac{e^{\text{-3.135-13.404}x_1 - 10.033x_2}}{1+e^{\text{-3.135-13.404}x_1 - 10.033x_2}}$.

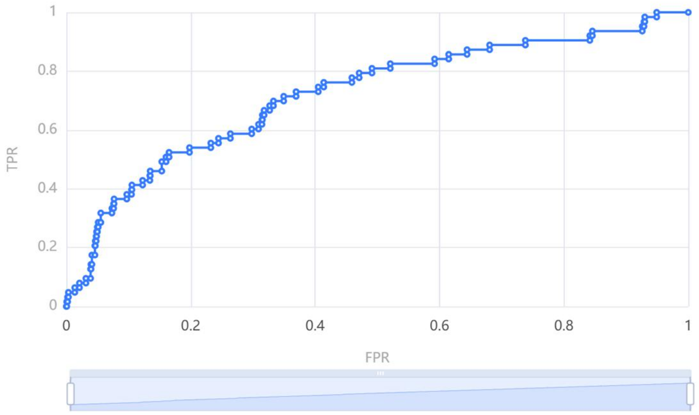
7.1.2 Analysis of the Effect of the Model Implementation

Finally we have analysed the effectiveness of the model implementation. The effectiveness of the clustering implementation of logistic regression is further measured by quantitative metrics. We derived accuracy, recall, precision, F1 values, and AUC values respectively.

<table>
<tr><th>Accuracy</th><th>Recall rate</th><th>Accuracy</th><th>F1</th><th>AUC</th></tr>
<tr><td>0.959</td><td>0.959</td><td>0.919</td><td>0.938</td><td>0.721</td></tr>
<tr><td colspan="5">Accuracy: the proportion of correct predicted samples to the total sample, the greater the accuracy the better.</td></tr>
<tr><td colspan="5">Recall: the proportion of results that are actually positive samples that are predicted to be positive samples, the larger the recall the better.</td></tr>
<tr><td colspan="5">Accuracy: the proportion of results predicted to be positive that are actually positive, the greater the accuracy the better.</td></tr>
</table>

<table>
<tr><th>Accuracy</th><th>Recall rate</th><th>Accuracy</th><th>F1</th><th>AUC</th></tr>
<tr><td colspan="5">F1: The summed average of precision and recall. Precision and recall affect each other, and although a high level of both is the desired ideal, in practice it is often the case that a high precision rate results in a low recall rate, or a low recall rate results in a high precision rate. If a balance is needed between the two, then the F1 metric can be used.</td></tr>
<tr><td colspan="5">AUC: The closer the AUC value is to 1 the better the classification is. Apart from the AUC value, we found that the model was very accurate in its implementation and the AUC was not too far from 1.</td></tr>
</table>

The table above shows the classification evaluation metrics and further measures the classification effectiveness of logistic regression through quantitative metrics. We can see from the table that, apart from the AUC value, we find that the model is generally accurate in its implementation. But the AUC is a little far from 1.



We used SPSS to draw the ROC plot, shown in the figure. the ROC plot combines sensitivity (TPR) and specificity (FPR), allowing both relationships to be measured. Ideally, TPR should be close to 1 and FPR should be close to 0. Where sensitivity is the proportion of results with actual positive samples that are predicted to be positive, and specificity is the proportion of results with actual negative samples that are predicted to be positive. However, we can find that the TPR and FPR of most of the scatters are concentrated between 0 and 0.4, so we can only say that the logistic regression model is implemented in an average way.

7.2 Second Build of the Logistic Regression Model

The first model fit was mediocre and could have been due to the following reasons:

- Too many independent variables, with variable redundancy and overfitting problems.
- No study of equation-fitting relationships or interactions between variables other

than multivariate primary.

7.2.1 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:

$$R=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2\sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}}:$$

$$\left(\begin{array}{ccccccccccccc}
1.000 & 0.599 & 0.653 & 0.306 & 0.501 & 0.284 & 0.271 & 0.091 & 0.144 & 0.037 & 0.367 & 0.134 & 0.001 \\
0.599 & 1.000 & 0.382 & 0.577 & 0.300 & 0.470 & 0.144 & 0.179 & 0.109 & 0.035 & 0.256 & 0.086 & 0.006 \\
0.653 & 0.382 & 1.000 & 0.609 & 0.183 & 0.094 & 0.488 & 0.286 & 0.144 & 0.063 & 0.204 & 0.264 & -0.010 \\
0.306 & 0.577 & 0.609 & 1.000 & 0.091 & 0.146 & 0.329 & 0.479 & 0.090 & 0.038 & 0.128 & 0.170 & 0.001 \\
0.501 & 0.300 & 0.183 & 0.091 & 1.000 & 0.598 & 0.666 & 0.378 & -0.116 & -0.104 & 0.450 & 0.101 & 0.006 \\
0.284 & 0.470 & 0.094 & 0.146 & 0.598 & 1.000 & 0.393 & 0.620 & -0.084 & -0.079 & 0.274 & 0.013 & -0.014 \\
0.271 & 0.144 & 0.488 & 0.329 & 0.666 & 0.393 & 1.000 & 0.648 & -0.142 & -0.096 & 0.228 & 0.222 & -0.004 \\
0.091 & 0.179 & 0.286 & 0.479 & 0.378 & 0.620 & 0.648 & 1.000 & -0.107 & -0.075 & 0.116 & 0.097 & -0.021 \\
0.144 & 0.109 & 0.144 & 0.090 & -0.116 & -0.084 & -0.142 & -0.107 & 1.000 & 0.739 & 0.126 & 0.173 & 0.041 \\
0.037 & 0.035 & 0.063 & 0.038 & -0.104 & -0.079 & -0.096 & -0.075 & 0.739 & 1.000 & 0.097 & 0.193 & 0.156 \\
0.367 & 0.256 & 0.204 & 0.128 & 0.450 & 0.274 & 0.228 & 0.116 & 0.126 & 0.097 & 1.000 & 0.469 & 0.068 \\
0.134 & 0.086 & 0.264 & 0.170 & 0.101 & 0.013 & 0.222 & 0.097 & 0.173 & 0.193 & 0.469 & 1.000 & 0.349 \\
0.001 & 0.006 & -0.010 & 0.001 & 0.006 & -0.014 & -0.004 & -0.021 & 0.041 & 0.156 & 0.068 & 0.349 & 1.000
\end{array}\right)$$

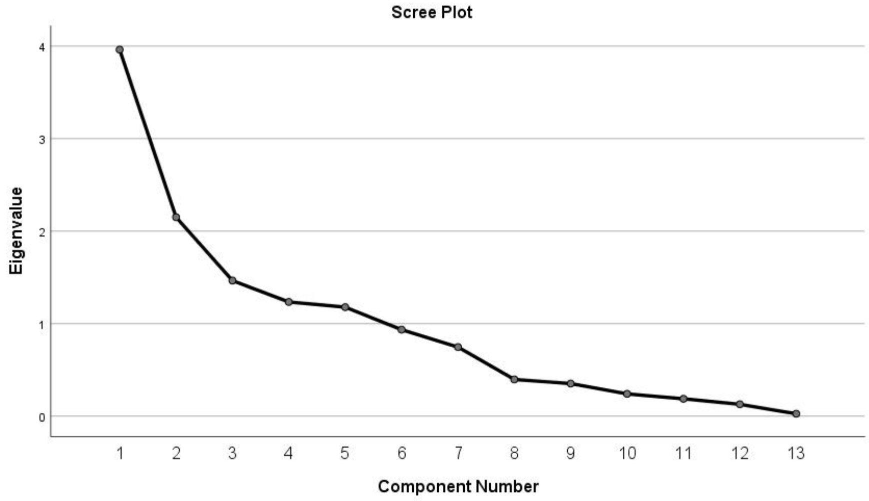
The KMO test and Bartlett's test are obtained from the matrix above.

KMO		0.61
Bartlett's test of sphericity	Approximate cardinality	12502.272
	df	78
	P	0.000***

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

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. Also, the results of the Bartlett's spherical test showed a significance p-value of 0.000***, presenting significance at the level, rejecting the original hypothesis that there is a correlation between the variables and that the principal component analysis is valid to the extent that it is appropriate.

Contribution of each principal component $=\frac{\lambda_i}{\sum_{k=1}^{P}\lambda_k}(i-1,2,...,p)$. This leads to the cumulative contribution rate. The 1st, 2nd,..., mth principal component corresponding to the eigenvalue with a cumulative contribution of more than 80% is generally selected. 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 i-th principal component is: (terms with too small coefficients are omitted). Thus, principal component analysis is also somewhat of an interactive variable setting. The definitions of the four categories were collated as follows：

- F1=0.1749×credgdp+0.1131×lcred+0.1420×credmoney

We have named F1 the “asset class”.

- F2=0.1170×loansgdp+0.0623×lcredmoney+0.1230×lloansmoney+0.1191×lloans

We have named F2 the “liability class”.

- F3=-0.1534×lm-0.1120×lnm+0.1382×lcredmoney+0.1292×lloansmoney+0.1336×loansmoney+0.1275×credmoney

We have named F3 the “Currency Category”.

- F4=-0.3170×loansgdp-0.2454×credgdp+0.3244×lpc

We have named F4 the “gross product category”.

We added the original two variables moneygdp, lrgdp and these four principal component variables to the logistic regression model to obtain the following table:

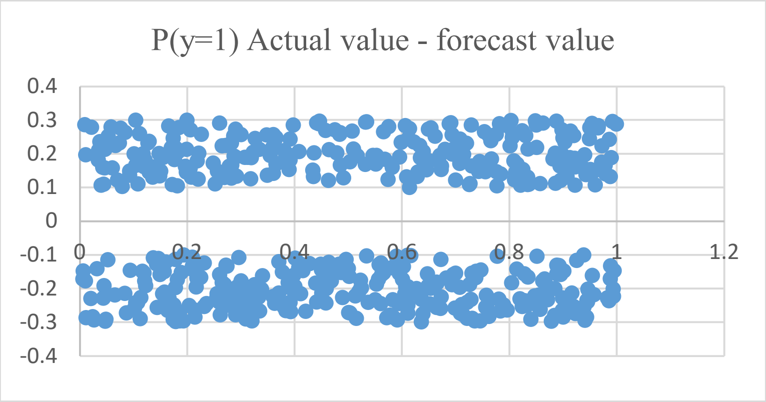
Item	B	S.E.	Wald	Sig.	Exp(B)
REGR factor score 1 for analysis 1	-0.149	0.111	1.817	0.100*	0.861
REGR factor score 2 for analysis 1	-0.025	0.129	0.037	0.847	0.976
REGR factor score 3 for analysis 1	-0.142	0.137	1.074	0.300	0.867
REGR factor score 4 for analysis 1	0.070	0.119	0.345	0.557	1.072
d2_moneygdp	-5.365	3.266	2.699	0.100*	0.005
d_lrgdp	-11.309	2.996	14.249	0.000***	0.000
Constant	-3.019	0.135	497.767	0.000***	0.049

As can be seen, the principal component variables of "asset class" also show a degree of significant correlation.

7.2.2 Model Improvement 2: Adding a Squared Term

We try to study equation-fitting relations other than multivariate primary. The inclusion of the squared term can, on the one hand, help us to fit the non-linear relationship better. If the fit of the model is significantly improved by adding the squared term, it suggests that there may be some degree of non-linearity between the independent and dependent variables, increasing the strength of the fit to some extent. On the other hand, there are currently too few significant variables, and doing so expands the number of variables we have that are significant, fitting results similar to the This is also a common way of setting up the model for the fit.

Firstly, there is indeed a non-linear relationship between x and y. If there is a non-linear relationship, there is often a tendency for the two ends of the prediction to be under (or over) and the middle to be over (or under). The following residual plot results using a partial 30% of the test set predictions fit just such a scenario.



The variable columns were squared and saved as new columns. The significance table is then derived in the same way as shown below (insignificant and non-squared terms are omitted here).

Item	Regression coefficient	Standard error	Wald	P	OR	OR value 95% confidence interval	
						Upper limit	Lower limit
Constant	-3.56	0.246	210.131	**0.000*****	0.028	0.018	0.046
lloansmoney_ square	0.206	0.102	4.056	**0.044****	1.229	1.006	1.501
lpc_square	0.03	0.018	2.863	**0.091***	1.031	0.995	1.067
Dependent variable: whether (i.e. the class variable of whether a financial crisis has occurred)							

Note: ***, **, * represent 1%, 5%, 10% level of significance respectively

The square of lloansmoney was found to be significantly correlated with the occurrence of a financial crisis, but not as well as the sig value before the square. The squared term of lpc, on the other hand, was more significantly correlated than lpc itself (0.698). Therefore, the squared term of lpc was included together in the final variable consideration.

7.2.3 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.

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:

$$\begin{pmatrix} 1 & 0.003 & 0.022 & 0.173 \\ 0.003 & 1 & 0.001 & 0.006 \\ 0.022 & 0.001 & 1 & 0.009 \\ 0.173 & 0.006 & 0.009 & 1 \end{pmatrix}$$

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.

7.2.4 Determining the Final Fit Curve and Associated Factors

After we have gone through the above series of adjustments, we get the final results as shown in the table below:

Variables in the Equation

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

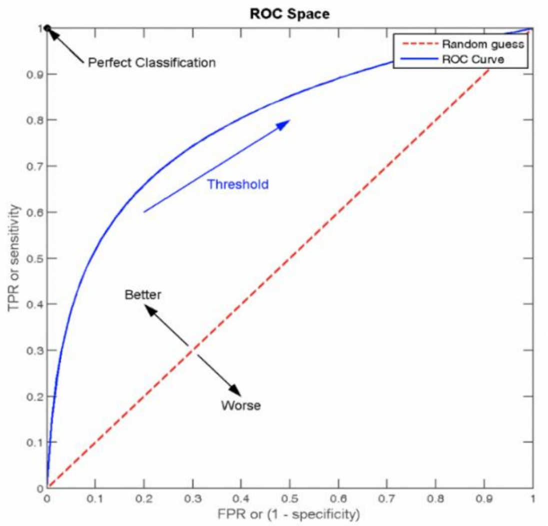
where each $\hat{\beta}_i$ value is the “regression coefficient” column. Therefore, by substituting the data $\hat{\beta}_0 = -3.001, \hat{\beta}_1 = -0.170, \hat{\beta}_2 = -0.219$, $\hat{\beta}_3 = -11.583$, $\hat{\beta}_4 = -4.569$, the fitted curves between the dependent variable “the occurrence 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: $P(y_i = 1|x) = \frac{e^{-3.185-0.4x_1+0.032x_2^2+0.022x_3^2+0.195x_4}}{1+e^{-3.185-0.4x_1+0.032x_2^2+0.022x_3^2+0.195x_4}}$. 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.

7.2.5 Analysis of the Fitting Effects

Finally we have analysed the effectiveness of the model implementation. The effectiveness of the clustering implementation of logistic regression is further measured by quantitative metrics. We derived accuracy, recall, precision, F1 values and AUC values respectively, all very close to 100% and effective.

Accuracy	Recall rate	Accuracy	F1	AUC
0.955	0.955	0.912	0.933	0.891



We used R studio to draw the ROC plot, curve the scatter, and mark and embellish the image as shown above. Ideally, the TPR should be close to 1 and the FPR should be close to 0. We can see that the curve is very close to the ideal implementation boundary compared to the last implementation of the model. We can roughly assume that the model is currently fitting well.

At this point, the logistic regression model is completed.

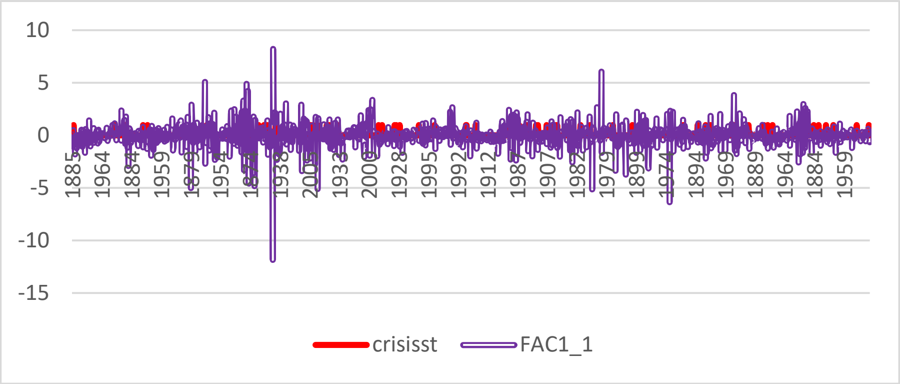
7.2.6 Analysis of Results

Given that we propose this model:

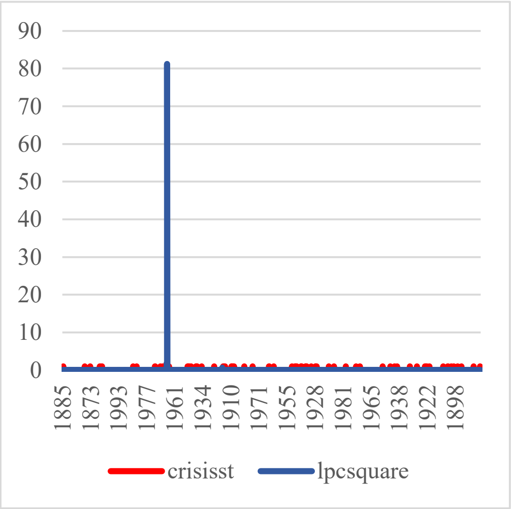
$$P(y_i = 1|x) = \frac{e^{-3.001-0.170F_1-0.219x_1^2-11.583x_2-4.569x_3}}{1+e^{-3.001-0.170F_1-0.219x_1^2-11.583x_2-4.569x_3}}$$

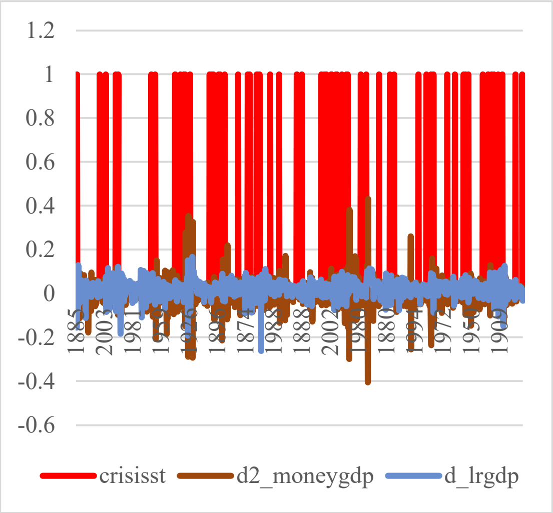
We explore whether there are economic implications for these four variables.

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”. So, does having enough assets mean that the probability of a financial crisis will decrease?



The graph above shows that the two folds overlap to a high degree. 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.





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. 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 too different and the fluctuations are not consistent with the dependent variable.

Lrgdp represents the natural logarithm of true GDP. That is, GDP growth suppresses financial crises. This confirms the importance, rationality and popularity of the GDP indicator, which is a reflection of the fact that the higher the GDP, the more developed the industry, the higher the level of the economy, the more tax revenues are guaranteed and the more tax revenues, the more money the treasury will have on hand to invest in infrastructure, health care, education, etc., and the more welfare is guaranteed. In such a situation, common sense dictates that a financial crisis is unlikely.

Chapter 8

Discussion

Financial crises are often hard to predict. Fortunately, all three of our models—especially the upgraded versions—have performed admirably.

The yearly growth rate of the real GDP (d_lrgdp) in a particular year has been found by all three models to be a substantial predictor of the occurrence of a financial crisis. During a financial crisis, there are societal issues such as bank runs, numerous business failures, and an increase in unemployment. This causes the capacity of society to produce to decrease, which causes a slowdown in GDP growth or even negative growth. Hence, this finding holds statistical and economic significance.

In addition, traditional regression models are not inferior to machine learning models in terms of predictive power. In fact, the way that financial crises are described is not entirely dependent on changes in macroeconomic variables. The reality is that a variety of difficult-to-quantify factors, including a nation's political environment, social dynamics, and external forces, frequently play a crucial part in fostering financial crises. Because economic variables have a limited capacity for explanation, econometric models and Random Forest models show largely similar predictive abilities.

However, our study also has room for improvement.

Firstly, as mentioned earlier, we use economic variables in all three models from one year to determine whether a financial crisis occurred in that year. This leads to a lack of actual predictive power in our models. However, the results we have achieved can still be used to determine whether a country experienced a financial crisis in a particular year in the past, as long as we have access to the relevant data. In this regard, our three models still hold value.

Second, to eliminate trends in our data, we employed the most basic differencing method in the data processing stage. This technique shows how a variable has changed from the prior year. However, we discovered that the Seasonal-Trend decomposition using LOESS approach is also a useful tool for preprocessing time series data. Unfortunately, we were forced to give up this approach since we lacked the necessary theoretical understanding and programming skills. Therefore, we process our data using the differencing approach alone, without comparing it to any other methods.

Finally, because the three models are built independently, we do not fully examine the connections between them. We optimize and enhance each model's ability to predict. However, the next model is not built using the insightful lessons learned from the earlier model. Our models' predictive abilities are as a result only moderately strong, and we haven't developed a powerful predictive model that combines the benefits of each one.

Chapter 9

Conclusion

In this paper, we aim to accurately forecast a financial crisis's occurrence by using machine learning techniques. Our approach involves the utilization of three distinct models: a random forest model, a logistic regression model, and a time series-enhanced logistic regression model. 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.

Bibliography

[1] Polsiri, P. (2009)Corporate Distress Prediction Models Using Governance and Financial Variables: Evidence from Thai Listed Firms during the East Asian Economic Crisis, Journal of Economics and Management, 2009, Vol.5, No.2, 273-304.

[2] 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.

[3] Schularick, Moritz, and Alan M. Taylor. 2012. "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008." American Economic Review, 102 (2): 1029-61.

[4] Berge, Travis J., and Òscar Jordà , 2011, "Evaluating the Classification of Economic Activity into Recessions and Expansions," American Economic Journal: Macro- economics, Vol. 3, No. 2, pp. 246–77.

[5] Mark Illing, Ying Liu (2003) Measuring financial stress in a developed country: An application to Canada. Staff Working Papers from Bank of Canada.

[6] Maryam Maryam, Dimas Aryo Anggoro, Muhibah Fata Tika and Fitri Cahya Kusumawati. (2022) An Intelligent Hybrid Model Using Artificial Neural 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.

[7] Gourinchas, Pierre-Olivier and Maurice Obstfeld (2012) "Stories of the Twentieth Century for the Twenty-First," American Economic Journal: Macroeconomics, 4(1), 226-65.

[8] Eleftherios Giovanis. (2010) Study of Discrete Choice Models and Adaptive Neuro-Fuzzy Inference System in the Prediction of Economic Crisis Periods in USA. Economic Analysis and Policy, Volume 42, Issue 1, Pages 79-95.

[9] Alev Dilek Aydin and Seyma Caliskan Cavdar. (2015) Prediction of Financial Crisis with Artificial Neural Network: An Empirical Analysis on Turkey. International Journal of Financial Research, Vol. 6, NO. 4: 36-45.

[10] Nik, P.A. & Jusoh, M. & Shaari, A.H. & Sarmdi, T.. (2016). Predicting the probability of financial crisis in emerging countries using an early warning system: Artificial neural network. 37. 25-40.

[11] Wu YI, Yi DONGYUN. Data preprocessing research for measurement error smoothness test[J]. Journal of the National University of Defense Technology,1996(02):130-134.

[12] Yuan Pingping, Yu Jianling, Shang Pengmian. Multiple fractal elimination trend analysis of stock market time series[J]. Journal of Beijing Jiaotong University,2007(06):69-72.

[13] Yuan Zhongru. A comparison of the effects of missing data filling methods in multiple linear regression models[D]. Zhongnan University, 2008.

[14] Alessi, Lucia and Carsten Detken (2011) "Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity", European Journal of Political Economy, Vol. 27, No. 3, pp. 520–533.

[15] Mark Joy, Marek Rusnák, Kateřina Šmídková and Bořek Vašíček (2017) "Banking

and currency crises: Differential diagnostics for developed countries”, International Journal of Finance & Economics, Vol. 22, No. 1, pp. 44–67.

[16] Wang Xia. Application of technical indicator analysis in securities investment - based on factor analysis and logistic regression analysis [J]. Logistics engineering and management,2016,000(004):P.155-157

[17] He Xiaoqun. Multivariate statistical analysis. Beijing: People's University of China Press, 2012.

[18] 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.

[19] Robert I-Kabacoff. R language in action (2nd ed.). Beijing:People's Post and Telecommunications Publishing House,2020.

[20] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO.(Version 1.0.11) [Online Application Software]. Retrieved from https://www.spsspro.com.

[21] Online Platform. Retrieved from https://www.zhihu.com/question/266537608

Contribution

Aibo:
Abstract, Introduction, reading Literature.

Simon:
Literature Review, Methodology (Random Forest and SMOTE part), Data Description (write 10% and do the data-preprocessing work), Random Forest Model, Discussion, consolidating the first draft.

Steffi:
Methodology (beginning and the Panel Logistic Regression part), arrange the Methodology section, Xtlogit Model, Conclusion.

Laura:
Methodology (Logistic Regression part), Data Description (write 90%), Logistic Regression Model.