Africa Economic, Banking and Systemic Crisis Data

Hoang Long Luu

Student ID:220932091

# **Africa Economic, Banking and Systemic Crisis Data**

This paper was conducted by analysing the data about the financial crisis of 13 African countries from 1860 to 2014. Through the data, the author aims to investigate crisis context of countries in Africa. Python was the technical used for data analysis in this research.

# Question 1

The data set has data from 13 African countries including Algeria, Angola, Central African Republic, Ivory Coast, Egypt, Kenya, Mauritius, Morocco, Nigeria, South Africa, Tunisia, Zambia, and Zimbabwe. These nations were once under the control of major superpowers, which is why their subsequent independence is considered a factor. Other factors are systemic crisis, exchange rate, sovereign domestic debt default, sovereign external debt default, total debt in default as a proportion of total GDP, annual consumer price inflation (CPI) rate, currency crisis, inflation crisis, and banking crisis. The data is totally sufficient.

A screenshot of a computer program

Description automatically generated

Figure 1: Description of the data

Each country in 13 countries has difference in the timeline including the data before and after independent. Systemic crisis, sovereign domestic debt default, sovereign external debt default, sovereign domestic debt default, sovereign external debt default, currency crisis, inflation crisis, and banking crisis are binary variables. In detail, “number 1” means year or crisis while “number 0” represents for no or no no-crisis.

A screenshot of a computer

Description automatically generated

A table of numbers and symbols

Description automatically generated

Figure 2: The description of the data

This table illustrate the general details of each variable in the data set. Based on the max value of currency\_ crisis, we can find out the problem that the value 2 is out of range of binary data type. The binary data only contain two outcomes which are normally 0 and 1, yes and no, or Buy and Sell (Kolanovic & Krishnamachari, 2017). Based on the information of the dataset owner, Harvard Business School, the value “2” is considered “crisis” (Harvard Business School, 2016).To fix this problem, we replaced value “2” to become value “1” which represent for “yes”. The character binary of banking crisis also converted to numeric binary for computational convenience.

The IQR method of identifying outliers was used to identify the outliers in the data set. The method was applied for non-binary variables including “inflation\_annual\_cpi”, “exch\_usd” and “gdp\_weighted\_default”. To find outliers using the IQR, the IQR need to calculate firstly by subtracting the value of the third quartile (Q3) to value of the first quartile (Q1) and then any data point that falls below Q1 - 1.5IQR or above Q3 + 1.5IQR is considered an outlier (Saleem, et al., 2021).

A white background with black numbers

Description automatically generatedA close up of words

Description automatically generated

Figure 3: IQR method result

The result shows that “inflation\_annual\_cpi”, “exch\_usd” and “gdp\_weighted\_default” have 142, 200, and 30 outliers respectively. When analyzing crisis events that are inherently rare or extreme, outliers can often contain very valuable information (Tongyu Wang, 2021). Hence, the outliers should not be removed.

# Question 2

A screenshot of a graph

Description automatically generated

*Figure 4: Heatmap of correlation*

This map shows the Pearson correlation among the variables. The data show that currency crisis has positive correlation with all variables excluding exchange rate. It is noticeable that inflation crisis is the variables highest correlation with currency crisis which is equal 0.4. Nevertheless, the Pearson P-value is 1.1341794159781767e-41 with inflation crisis variable which is significantly high. Hence, we cannot confidently confirm the strong relationship between currency crisis and inflation crisis.

A screenshot of a computer

Description automatically generated

Figure 5: OLS regression result of currency crisis with others

According to the table, currency crisis has positive significant relationship with exchange US rate, sovereign external debt default, and inflation crisis. Meanwhile, the inflation crisis has highest coefficient index with 0.3430.

A graph showing the growth of the currency

Description automatically generated

Figure 6: Total inflation crisis and currency crisis over time.

The line charts have illustrated the movement of total number of Total inflation crisis and currency crisis overtime in 13 counties. The similar pattern of two variable was seen from 1930s to 2000.

All in all, we can conclude that inflation crisis has the most associated with currency crises.

Logistic regression is applied to test the relationship between inflation crisis and CPI since both variables are binary variable. Logistic regression is a classification method that utilized to handle with binary output. Although the main purpose of Logistic regression is classification task, it still can investigate the relationship between variables (Kolanovic & Krishnamachari, 2017).

# Question 3

A screenshot of a computer

Description automatically generated

*Figure 7: Logistc regression results*

The table illustrates the logistic regression test between the inflation crisis and CPI in 13 countries in general. The result shows that the two variable has positive significant relationships. Specifically, the p-value is equal to 0 while the coefficient index is 0.4336.

# Question4

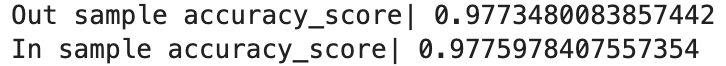
It is true that not all variables contain the useful information for prediction. Many irrelevant independent variables can lead to noisy, overfitting, unnecessary complexity and difficulty in computation, thereby reduce the performance of the model (Andersen & Bro, 2010).

A screenshot of a computer

Description automatically generated

*Figure 8: Lasso regression for logistic regression*

To select the best (possible) model to predict the systemic crisis, Lasso regularization (L1) is applied to perform variable selection. The regularization method is used with the logistic regression to find the best variables. The result shows that exchange rate, CPI, and Banking crisis are the important variable for the prediction model of systemic crisis variables. Other unimportant variables are strung down to zero by the Lasso regularization.



*Figure 9: In-sample accuracy vs out-sample accuracy*

To test how well a new logistic regression model works in-sample versus out-of-sample, you typically split your data into a training set and a test set. The model is trained on the training set and tested on both the test set and train set. In other words, they are in sample accuracy and out sample accuracy rate. These accuracy scores show the idea of how well the model works in-sample versus out-of-sample. In this model, two indexes are almost equal which means that the model is highly not overfitting or underfitting the data.

Question 5  
The result from the data analyze show several information about some countries

A graph of different colored lines

Description automatically generated

*Figure 10: Exchange rate after and before Independency*.

(P/S: Egypt actually was independent since 1860, but the line still blue due to the limit of “ hue” attribution in sns.lineplot)

The plots illustrate the fact almost countries in Africa had good exchange rates after independence. While Nigeria, Mauritius, Algeria, Angola, Kenya, and Morocco experienced stable growth, the Central Africa Republic and Ivory Coast increased with some fluctuations. Besides, Egypt, Zimbabwe, Zambia, and South Africa took a long time after independence to significantly raise. The unprecedented phenomenon is seen in the case of Tunisia when the exchange rate dropped dramatically after its independence. This phenomenon can be the consequence of issuing new currency (Tunisian dinar) and miss management the economic after independence from France in 1958 (Calamitsis, 1970). Nevertheless, Tunisian can keep a stable inflation rate after that even.

A group of graphs showing different types of data

Description automatically generated

*Figure 11: CPI after and before Independency*

(P/S: Egypt actually was independent since 1860, but the line still blue due to the limit of “ hue” attribution in sns.lineplot)

Tunisian has relatively low inflation when compared with other North African neighbors like Egypt, and Algeria. In general, almost countries have fluctuated movements of inflation which should that their currency is unstable.

A graph with blue bars

Description automatically generated

*Figure 12: The default the domestic debt*

Angola and Zimbabwe are the only countries that default the domestic debt.

A graph with red dots and blue lines

Description automatically generated

Figure 13: Domestic debt in de fault and banking crisis in Angola

A graph with a line graph

Description automatically generated

Figure 14: Domestic debt in de fault and banking crisis in Zimbabwe

The figures indicate that when these two countries encountered a banking crisis, they also failed to fulfill their sovereign debt obligations.

# Question 6

To find the most suitable classification model for the systemic crisis, the data is tested with several classification methods including Logistic Regression, Random Forest, K-Neighbours, SVC, Bagging, and Gaussian. The model only contains 3 selected variables including exchange rate, CPI, and Banking crisis. The metrics used to evaluate the model performance are accuracy, F1 Score, recall, and precision. Recall criteria are important when the consequences of missing false negatives are high the false positive. A high recall ensures that the model can detect most of the actual crises, minimizing the chances of overlooking events (Powers, 2008). Hence the recall metric is considered first when ranking the model. Bootstrapping method is used to enhance the robustness of the result and reduce the impact of randomness. It is particularly useful when the underlying data distribution is unknown or when the sample size is limited. The number of tests only reach 100 times due to the limitation in the user's computer capability.

A screenshot of a computer

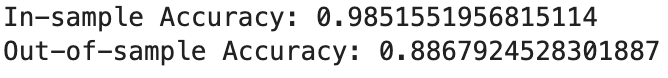
Description automatically generated

*Figure 15: Classification index table*

The scores in the figures are the average score after a 100-time test. In general, Logistic regression, random Forest, decision tree and bagging classifier had very high performance.

Nonetheless, the winner is Bagging Classifier with the highest point in three criteria including recall, F1 Score, and accuracy.

The next step is testing again the performance of the model between out-sample and in-sample based on the accuracy.



*Figure 16: In-sample accuracy vs out-sample accuracy*

Considering the in-sample accuracy is higher than the out-of-sample accuracy, it suggests a degree of overfitting in the model. Nonetheless, the difference is not too large.

# Bibliography

Andersen, C. M. & Bro, R., 2010. Variable selection in regression—a tutorial. *Special Issue: Herman Wold Medal Winners 2007–2009,* 24(11-12), pp. 728-737.

Calamitsis, E. A., 1970. Stability problem and policies in Tunisia. *Finance and Development ,* 7(3), p. 43.

Harvard Business School, 2016. *Data: Global Crises Data by Country.* [Online]   
Available at: https://www.hbs.edu/behavioral-finance-and-financial-stability/data/Pages/global.aspx  
[Accessed 15 July 2016].

Kolanovic, M. & Krishnamachari, R. T., 2017. *Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing.* s.l.:J.P.Morgan.

Powers, D. M. W., 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies,* 2(1), pp. 37-63.

Saleem, S., Aslam, M. & Shaukat, M. R., 2021. A REVIEW AND EMPIRICAL COMPARISON OF UNIVARIATE OUTLIER DETECTION METHODS. *Pak. J. Statist.,* 37(4), pp. 447-462.

Tongyu Wang, S. Z. G. Z. H. Z., 2021. A machine learning-based early warning system for systemic banking crises. *Applied Economics,* 53(26), pp. 2974-2992.

# Appendixes

Appendix 1: Function to identify outliers.

A screenshot of a computer code

Description automatically generated

Appendix 2: Testing in-sample vs out-sample accuracy

A screenshot of a computer code

Description automatically generated

Appendix 3: Bootstrap to find the best classification method.

A screenshot of a computer code

Description automatically generated