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.

```
RangeIndex: 1059 entries, 0 to 1058
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   case                            1059 non-null   int64
 1   cc3                             1059 non-null   object
 2   country                         1059 non-null   object
 3   year                            1059 non-null   int64
 4   systemic_crisis                 1059 non-null   int64
 5   exch_usd                        1059 non-null   float64
 6   domestic_debt_in_default        1059 non-null   int64
 7   sovereign_external_debt_default 1059 non-null   int64
 8   gdp_weighted_default            1059 non-null   float64
 9   inflation_annual_cpi            1059 non-null   float64
 10  independence                    1059 non-null   int64
 11  currency_crisis                 1059 non-null   int64
 12  inflation_crisis                1059 non-null   int64
 13  banking_crisis                  1059 non-null   object
```

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.

	systemic_crisis	exch_usd	domestic_debt_in_default	sovereign_external_debt_default	gdp_weighted_default
count	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000
mean	0.077432	43.140831	0.039660	0.152975	0.006402
std	0.267401	111.475380	0.195251	0.360133	0.043572
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.195350	0.000000	0.000000	0.000000
50%	0.000000	0.868400	0.000000	0.000000	0.000000
75%	0.000000	8.462750	0.000000	0.000000	0.000000
max	1.000000	744.306139	1.000000	1.000000	0.400000

	inflation_annual_cpi	independence	currency_crisis	inflation_crisis
count	1.059000e+03	1059.000000	1059.000000	1059.000000
mean	2.084889e+04	0.776204	0.132200	0.129367
std	6.757274e+05	0.416984	0.349847	0.335765
min	-2.850214e+01	0.000000	0.000000	0.000000
25%	2.086162e+00	1.000000	0.000000	0.000000
50%	5.762330e+00	1.000000	0.000000	0.000000
75%	1.164405e+01	1.000000	0.000000	0.000000
max	2.198970e+07	1.000000	2.000000	1.000000

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).

```
Q1:  0.19535000000000002
Q3:  8.46275
IQR:  8.2674
Q1:  2.0861622595
Q3:  11.644047955
IQR:  9.5578856955
Q1:  0.0
Q3:  0.0
IQR:  0.0
```

```
Number of outlineders in inflation_annual_cpi:  142
Number of outlineders in exch_usd:  200
Number of outlineders in gdp_weighted_default:  30
```

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

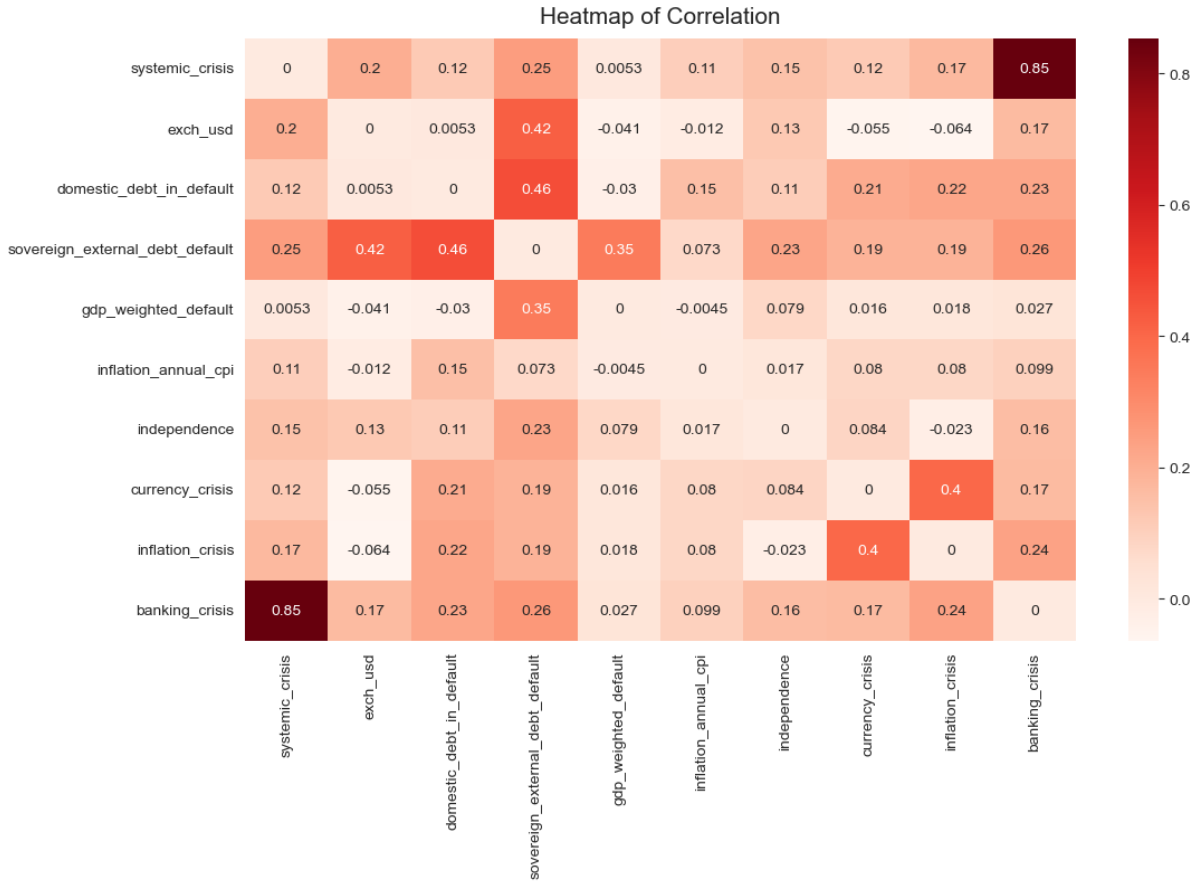

Figure 4: Heatmap of correlation

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.

OLS Regression Results

Dep. Variable:	currency_crisis	R-squared:	0.194
Model:	OLS	Adj. R-squared:	0.187
Method:	Least Squares	F-statistic:	28.10
Date:	Sat, 15 Jul 2023	Prob (F-statistic):	6.07e-44
Time:	13:12:30	Log-Likelihood:	-228.75
No. Observations:	1059	AIC:	477.5
Df Residuals:	1049	BIC:	527.2
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0320	0.020	1.587	0.113	-0.008	0.072
systemic_crisis	-0.0393	0.068	-0.575	0.566	-0.173	0.095
exch_usd	-0.0003	0.000	-3.198	0.001	-0.001	-0.000
domestic_debt_in_default	0.0770	0.060	1.287	0.198	-0.040	0.194
sovereign_external_debt_default	0.1278	0.039	3.278	0.001	0.051	0.204
gdp_weighted_default	-0.3683	0.245	-1.504	0.133	-0.849	0.112
inflation_annual_cpi	1.464e-08	1.4e-08	1.049	0.295	-1.28e-08	4.2e-08
independence	0.0525	0.023	2.275	0.023	0.007	0.098
inflation_crisis	0.3430	0.029	11.631	0.000	0.285	0.401
banking_crisis	0.0861	0.065	1.319	0.187	-0.042	0.214

Omnibus:	408.399	Durbin-Watson:	1.624
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1296.732
Skew:	1.951	Prob(JB):	2.62e-282
Kurtosis:	6.763	Cond. No.	1.80e+07

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.

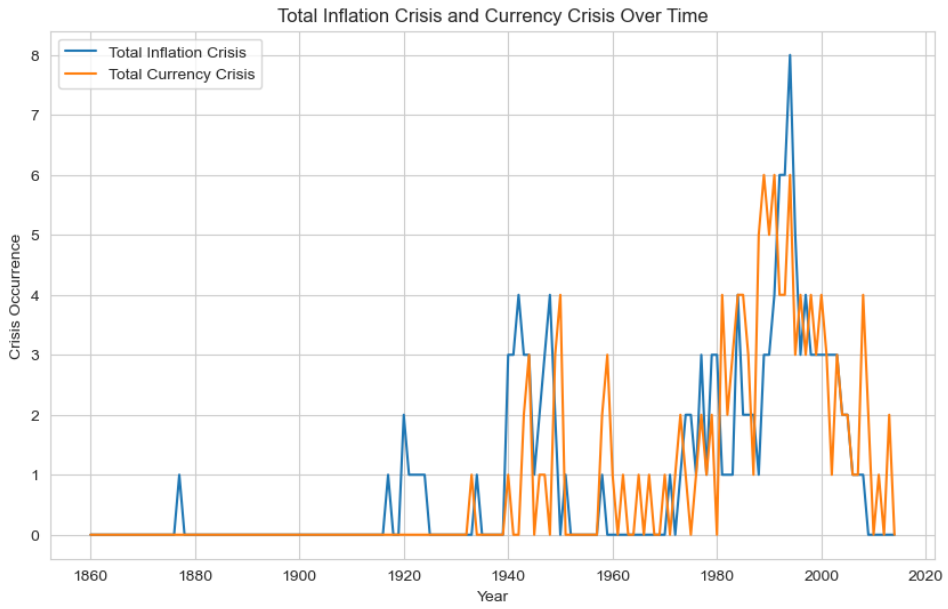

Figure 6: Total inflation crisis and currency crisis over time.

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

Logit Regression Results

Dep. Variable:	inflation_crisis	No. Observations:	1059
Model:	Logit	Df Residuals:	1057
Method:	MLE	Df Model:	1
Date:	Sat, 15 Jul 2023	Pseudo R-squ.:	0.8142
Time:	02:35:21	Log-Likelihood:	-75.789
converged:	True	LL-Null:	-407.91
Covariance Type:	nonrobust	LLR p-value:	1.789e-146

	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.9710	0.866	-10.359	0.000	-10.668	-7.274
inflation_annual_cpi	0.4336	0.045	9.539	0.000	0.344	0.523

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).

	Variable	Coefficient
0	exch_usd	0.003023
1	domestic_debt_in_default	0.000000
2	sovereign_external_debt_default	0.000000
3	gdp_weighted_default	0.000000
4	inflation_annual_cpi	0.000064
5	independencecurrency_crisis	0.000000
6	inflation_crisis	0.000000
7	banking_crisis	5.437468

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.

```
Out sample accuracy_score| 0.9773480083857442
In sample accuracy_score| 0.9775978407557354
```

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

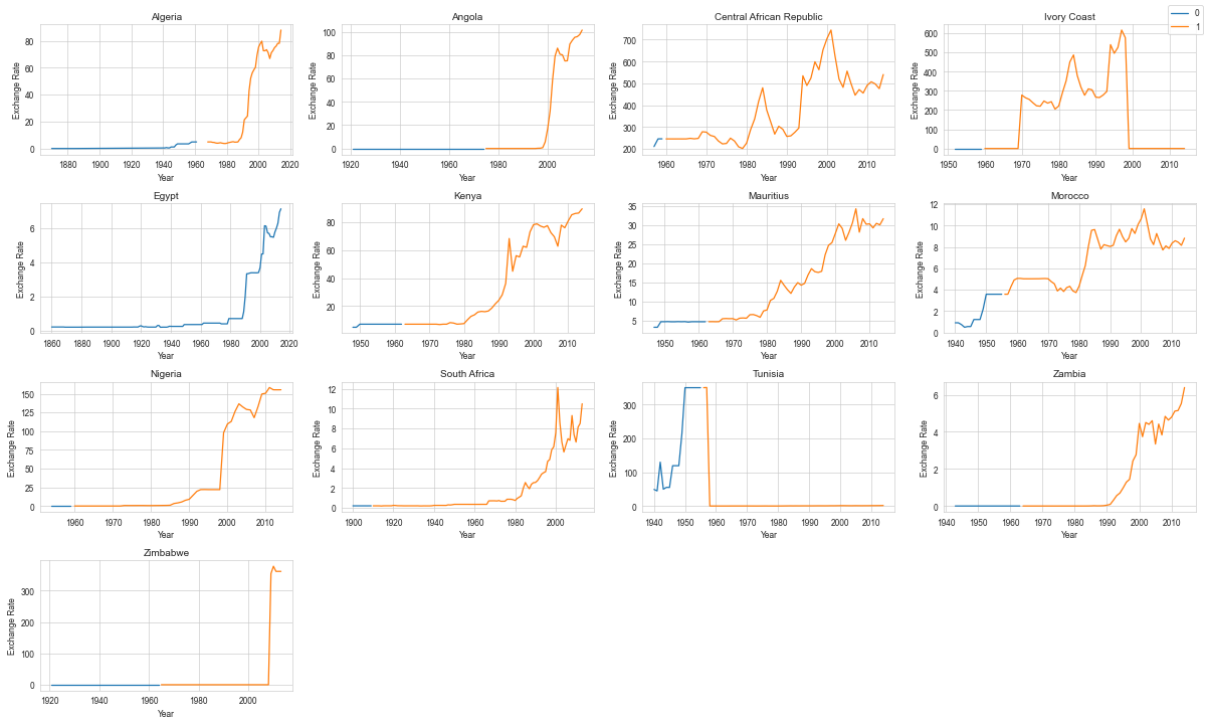

Figure 10: Exchange rate after and before Independency.

(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.

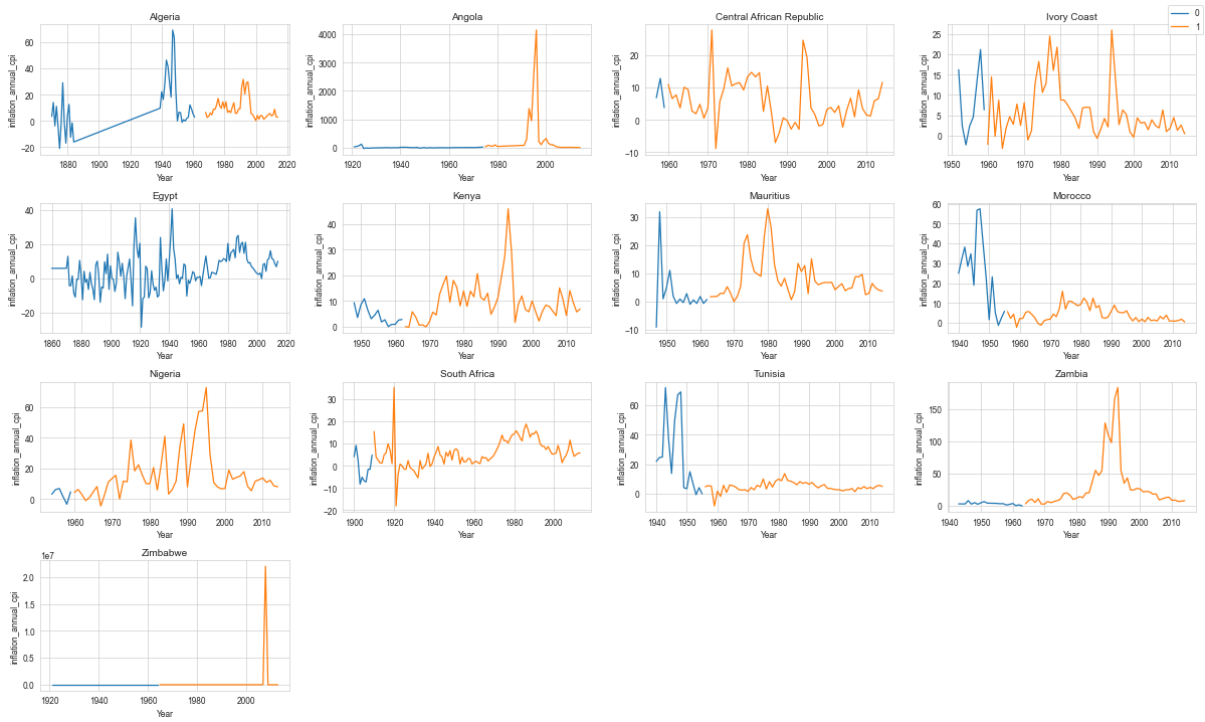

Figure 11: CPI after and before Independency

(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.

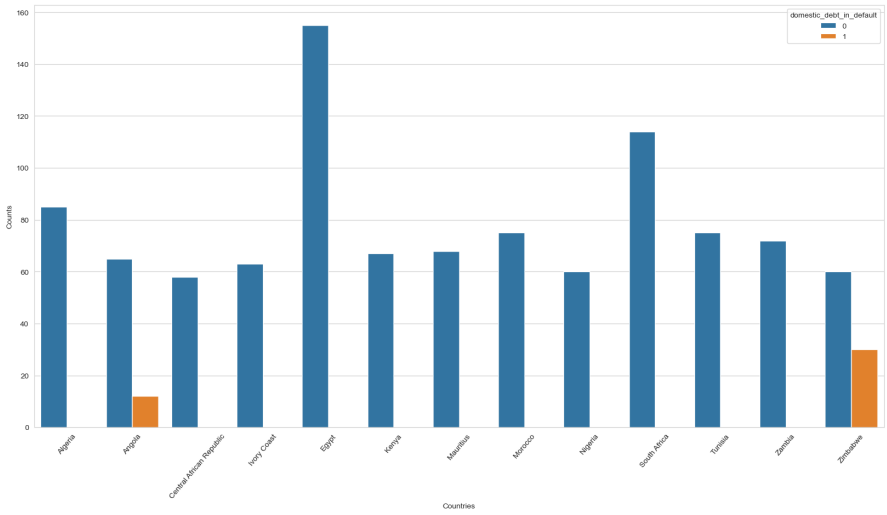

Figure 12: The default the domestic debt

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

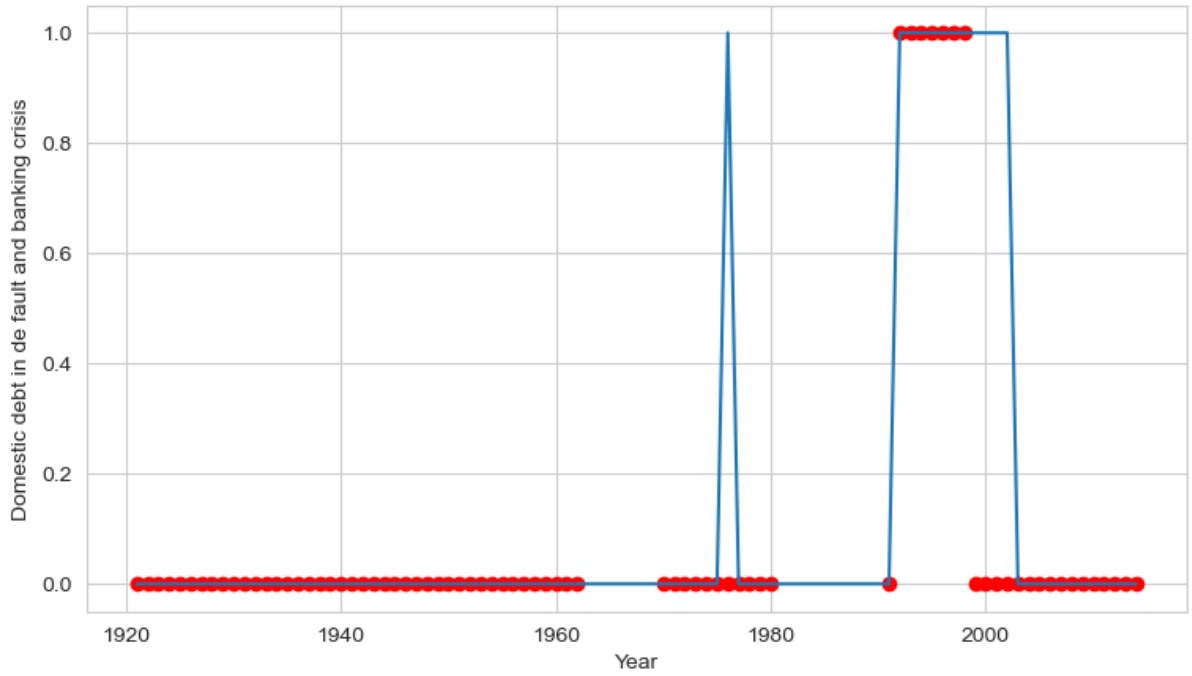

Figure 13: Domestic debt in de fault and banking crisis in Angola

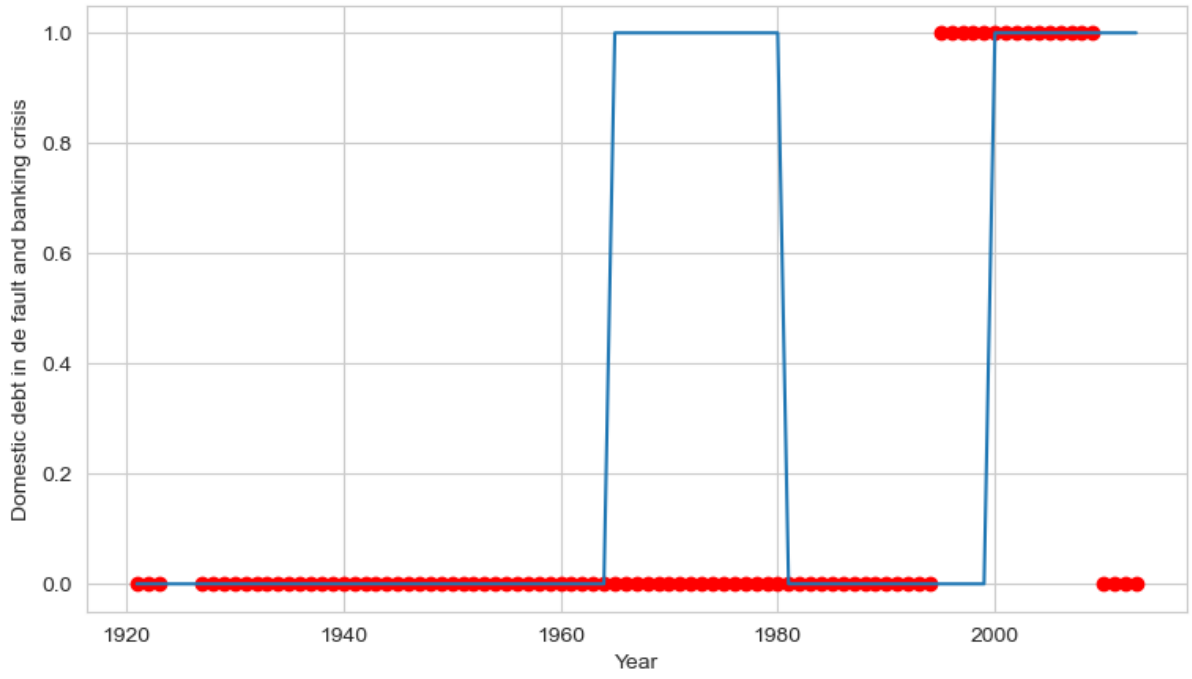

Figure 14: Domestic debt in de fault and banking crisis in Zimbabwe

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.

```
Bootstrap iteration: 100

Final Model Ranking:
                     Recall  F1 Score  Accuracy  Precision
Bagging            0.960610  0.960564  0.988270   0.962725
DecisionTree       0.958983  0.959141  0.987862   0.961344
RandomForest       0.949412  0.958210  0.987925   0.969513
LogisticRegression 0.905297  0.874844  0.936667   0.873251
Bayes              0.722447  0.707077  0.944245   0.766917
KNeighbors         0.622418  0.654007  0.921761   0.754444
SVC                0.506872  0.492011  0.919088   0.587446
```

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.

```
In-sample Accuracy: 0.9851551956815114
Out-of-sample Accuracy: 0.8867924528301887
```

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.

```
def calculate_iqr(data):
    Q1 = data.quantile(0.25)
    Q3 = data.quantile(0.75)
    IQR = Q3 - Q1
    print('Q1: ', Q1)
    print('Q3: ', Q3)
    print('IQR: ', IQR)
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return lower_bound, upper_bound
```

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

```
y_systemic_crisis = df.iloc[:,4:5].values.ravel()
x_factors = df[["exch_usd", "inflation_annual_cpi","banking_crisis"]].values
x_train_bag, x_test_bag, y_train_bag, y_test_bag = train_test_split(x_factors,y_systemic_crisis, test_size= 0.3)

in_sample_pred_bag = model_BAG.predict(x_train_bag)
in_sample_accurancy_bag = accuracy_score(y_train_bag, in_sample_pred_bag)
print(f'In-sample Accuracy: {in_sample_accurancy_bag}')

out_sample_pred_bag = model_BAG.predict(x_test_bag)
out_of_sample_accuracy_bag = accuracy_score(y_test_bag, y_pred)
print(f'Out-of-sample Accuracy: {out_of_sample_accuracy_bag}')
```

Appendix 3: Bootstrap to find the best classification method.

```
model_LR = LogisticRegression(max_iter=1000)
model_DT = DecisionTreeClassifier()
model_RF = RandomForestClassifier(n_estimators = 10)
model_KN = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
model_SVM = SVC(probability=True)
model_BAG = BaggingClassifier(base_estimator=None, n_estimators=100, max_samples=1.0, max_features=1.0,bootst
model_Bayes = GaussianNB()

models_all = [model_LR, model_DT, model_RF,model_KN,model_SVM,model_BAG,model_Bayes]
model_names = ['LogisticRegression', 'DecisionTree', 'RandomForest','KNeighbors','SVC','Bagging','Bayes']

criteria = ['Recall', 'F1 Score', 'Accuracy', 'Precision']

bootstrap_iteration = 100
model_scores = {name: {criterion: [] for criterion in criteria} for name in model_names}

for i in range(bootstrap_iteration):
    print(f"Bootstrap iteration: {i+1}")

    X_resample, y_resample = resample(x_factors, y_systemic_crisis, n_samples = len(x_factors))
    x_train, x_test, y_train, y_test = train_test_split(X_resample,y_resample, test_size= 0.3)

    for idx, model in enumerate(models_all):
        model.fit(x_train, y_train)
        y_pred = model.predict(x_test)

        model_scores[model_names[idx]]['Recall'].append(recall_score(y_test, y_pred, average='macro'))
        model_scores[model_names[idx]]['F1 Score'].append(f1_score(y_test, y_pred, average='macro'))
        model_scores[model_names[idx]]['Accuracy'].append(accuracy_score(y_test, y_pred))
        model_scores[model_names[idx]]['Precision'].append(precision_score(y_test, y_pred, average='macro'))


result_df = pd.DataFrame()

for model_name, scores in model_scores.items():
    for criterion, score_list in scores.items():
        result_df.loc[model_name, criterion] = np.mean(score_list)

result_df.sort_values(by=['Recall','F1 Score','Accuracy','Precision'],ascending=False,inplace=True)
```