Predicting Financial Instability in Africa

Peter Morian

Professional Certificate in Data Science - 2020

Contents

1 Introduction **2**

1.1 Motivation . . . 2

1.2 Executive Summary . . . 2

2 Method & Analysis **2**

2.1 Background . . . 2

2.2 Initial Data Inspection . . . 2

2.3 Data Cleaning . . . 3

2.3.1 Reformat Banking Crisis . . . 4

2.3.2 Exchange Rate & Inflationary Annual Changes . . . 4

2.3.3 Reformat Currency Crisis . . . 4

2.3.4 Other Procedures . . . 4

2.4 Data Analysis . . . 5

2.5 Model Building . . . 9

2.5.1 GLM . . . 10

2.5.2 Decision Tree . . . 10

2.5.3 Random Forest . . . 11

3 Results **12**

4 Conclusion **13**

4.1 Summary of Findings . . . 13

4.2 Suggested Improvements . . . 13

5 Acknowledgements **13**

1 Introduction

1.1 Motivation

The purpose of this report is to develop a machine learning model that is able to predict the financial stability of an African country, based off its economic factors. The overall motivation of this report is to better undertsand the complexity and the sensitivity of the African financial system. The database used is a subset of the Reinhart et. al's Global Financial Stability dataset which can be found here. Column names have been renamed before any cleaning or analysis was be conducted, to make references to columns shorter. This report is a part of the Professional Certificate in Data Science program by HarvardX & edX.

Before developing the model that will be used for predicting financial instability, this report will firstly provide a high-level overview of the sample data, as well the discuss key characteristsics of the data and justify any trasnformations. We will then build multiple machine learning models to predict systemic crises within the African economy, and determine which of these models is the best at predicting financial instability.

1.2 Executive Summary

After analysing the relationships between specific variables, the final model for predicting a systemic crisis was built using a Random Forest model, which takes into account the Exchange rate, Inflation rate & GDP weighted debt of an economy, as well as indicators of Banking and Soverign debt crises. When compared to the test dataset, the accuracy of the Random Forest model is 0.981.

2 Method & Analysis

2.1 Background

The collected data will be refered to as the African Crises dataset throughtout this report. The following sub-sections will provide insights into the structure of the African Crises dataset, as well as explain key transformations and insights. After this, we will then proceed into the model building stage.

The African Crises data was paritioned into a 80-20 split. The 80% partion will be referred to as the training dataset, which will be the main focus of the model building stage in this report. The remaining 20% partion will be refered to as the test dataset, which will be used the the Results section of this report.

2.2 Initial Data Inspection

The African Crises dataset contains 1059 rows and 14 columns, with the economic information of 0 different African countries over 155 years, from 1860 to 2014. A summary of what each column reprsents is as follows:

- *caseA*: Number denoting a specific country.
- *cc3A*: Three letter abbreviation of the country name.
- *Country*: Name of the country.
- *Year*: The year of observation.
- *Syst*: "0" means that no systemic crisis occurred in a given year, whilst "1" means that a systemic crisis occurred. A systemic crisis is defined as an economy-wide stress which typically leads to the breakdown of financial institutions. This variable can be essentially viewed as a "recession" indicator, and will be the focus of this report.
- *EX*: The exchange rate of the country against the USD, at the year of observation.
- *Dom*: "0" means that no sovereign domestic debt default occurred in a given year, whilst "1" means that a sovereign domestic debt default occurred.

- *Sov*: "0" means that no sovereign external debt default occurred in a given year, whilst "1" means that a sovereign external debt default occurred.
- *GDP_w*: The total debt in default relative to the economy's GDP.
- *CPI*: The annual CPI Inflation rate.
- *Ind*: "0" means the economy was not independent in a given year, whilst "1" means that it was independent.
- *Curr*: "0" means that no currency crisis occurred in a given year, whilst "1" means that a currency crisis occurred.
- *Infl*: "0" means that no inflation crisis occurred in a given year, whilst "1" means that an inflation crisis occurred.
- *Bank*: "no_crisis" means that no banking crisis occurred in a given year, whilst "crisis" means that a banking crisis occurred.

To give the reader a clear understanding of the dataset, the first few rows are shown below.

Table 1: First few rows of the African Crises dataset

case	cc3	Country	Year	Syst	EX	Dom	Sov	GDP_w	CPI	Ind	Curr	Infl	Bank
1	DZA	Algeria	1870	1	0.052264	0	0	0	3.441456	0	0	0	crisis
1	DZA	Algeria	1871	0	0.052798	0	0	0	14.149140	0	0	0	no_crisis
1	DZA	Algeria	1872	0	0.052274	0	0	0	-3.718593	0	0	0	no_crisis
1	DZA	Algeria	1873	0	0.051680	0	0	0	11.203897	0	0	0	no_crisis
1	DZA	Algeria	1874	0	0.051308	0	0	0	-3.848561	0	0	0	no_crisis
1	DZA	Algeria	1875	0	0.051546	0	0	0	-20.924178	0	0	0	no_crisis

Since the purpose of this project is to predict financial instability, below is a brief summary of the number of systemic crisies recorded per country.

Table 2: Systemic Crises by African Country

Country	Number of Systemic Crises
Algeria	4
Angola	0
Central African Republic	19
Ivory Coast	4
Egypt	6
Kenya	13
Mauritius	0
Morocco	2
Nigeria	10
South Africa	0
Tunisia	5
Zambia	4
Zimbabwe	15

2.3 Data Cleaning

As shown in Table 1, the current form of some columns are not ideal for analysis. Additional columns are reqired to look at country-specific impacts. This section describes the procedures made to the African Crises data to "clean" these columns.

Besides *Country*, non-numeric & non-integer columns were then removed from the dataset in order to make the size of the data smaller and to make the model building process more efficient. Duplicate rows were also removed. After cleaning, the dataset that will be used for model building now conatins 1059 rows and 14 columns.

2.3.1 Reformat Banking Crisis

The indicator column for a banking crisis, *Bank*, was originally a charater column. To make the modelling process for efficient with a smaller file side, numeric values were used to replace the existing character values. Now, "0" means that no banking crisis occurred in a given year, whilst "1" means that a banking crisis occurred.

2.3.2 Exchange Rate & Inflationary Annual Changes

Since the exchange rate and the inflation rate can vary significantly by country, it may be hard to include these variables in a model are their values are not standardised to the entire dataset. Thus, two additonal columns were created that measure the year-by-year percentage change in rates per country, which are called *EX_change* & *CPI_change* respectively.

Table 3: Exchange rate and Inflationary Annual Changes

Country	Year	EX	EX_change	CPI	CPI_change
Algeria	1870	0.052264	0.0000000	3.441456	0.00000
Algeria	1871	0.052798	1.0114019	14.149140	75.67728
Algeria	1872	0.052274	-1.0024104	-3.718593	480.49713
Algeria	1873	0.051680	-1.1493808	11.203897	133.19017
Algeria	1874	0.051308	-0.7250331	-3.848561	391.11915
Algeria	1875	0.051546	0.4617235	-20.924178	81.60711

2.3.3 Reformat Currency Crisis

An issue that was spotted in the African Crises dataset was that the *Curr* column had some values of "2", which is not a valid option for this indicator variable.

Table 4: Rows with Currency Crisis 2

Curr	case	cc3	Country	Year	Syst	EX	Dom	Sov	GDP_w	CPI	Ind	Infl	Bank	EX_change	CPI_change
2	2	AGO	Angola	1995	0	0.005692	1	1	0.00	2672.230000	1	1	1	91.05302	64.457737
2	2	AGO	Angola	1999	0	5.579920	1	1	0.00	248.248000	1	1	0	87.51774	56.725130
2	56	ZAF	South Africa	1967	0	0.709300	0	0	0.00	2.151508	1	0	0	49.65459	-69.107501
2	63	TUN	Tunisia	1958	0	0.419700	0	1	0.06	5.216941	1	0	0	-83292.89969	-5.518621

Whilst this was only a small number of obvservations with this error, it is still worthwhile to clean these observations. Each row with this error was checked against the original Reinhart et. al source data, and it was confirmed that each of these rows were valid currency crises. This means that "1" should be used instead of "2". The adjustments were made, and now all values of *Curr* are either "0" or "1".

2.3.4 Other Procedures

It should be noted that the *case* & *cc3* columns were removed from the African Crises dataset, as the represent the same information as other columns, and are thus redundant.

Futhermore, whilst there was a check conducted to remove any duplicate observation rows from the cleaned dataset, it should also be noted that there were no duplicate rows found.

After all the previously discussed procedues were run, the cleaned African Crises dataset has 1059 rows and 14 columns. Below is a brief view of the cleaned data, which is now ready for analysis and modelling.

Table 5: First few rows of the cleaned African Crises dataset

Country	Year	Syst	EX	Dom	Sov	GDP_w	CPI	Ind	Curr	Infl	Bank	EX_change	CPI_change
Algeria	1870	1	0.052264	0	0	0	3.441456	0	0	0	1	0.0000000	0.00000
Algeria	1871	0	0.052798	0	0	0	14.149140	0	0	0	0	1.0114019	75.67728
Algeria	1872	0	0.052274	0	0	0	-3.718593	0	0	0	0	-1.0024104	480.49713
Algeria	1873	0	0.051680	0	0	0	11.203897	0	0	0	0	-1.1493808	133.19017
Algeria	1874	0	0.051308	0	0	0	-3.848561	0	0	0	0	-0.7250331	391.11915
Algeria	1875	0	0.051546	0	0	0	-20.924178	0	0	0	0	0.4617235	81.60711

2.4 Data Analysis

In this section, we will derive some preliminary insights from the cleaned African Crises data.

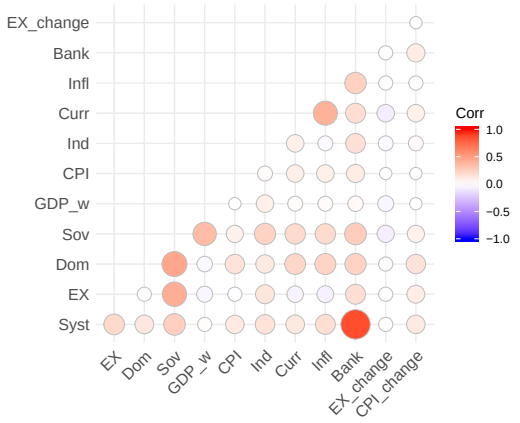
Firstly, we will look at the proportions of all types of financial crisies recorded. From the table below, we can see that the highest proportion of crises comes from sovereign debt in default, currency crisis, and inflation crises.

Table 6: Proportion of Crises

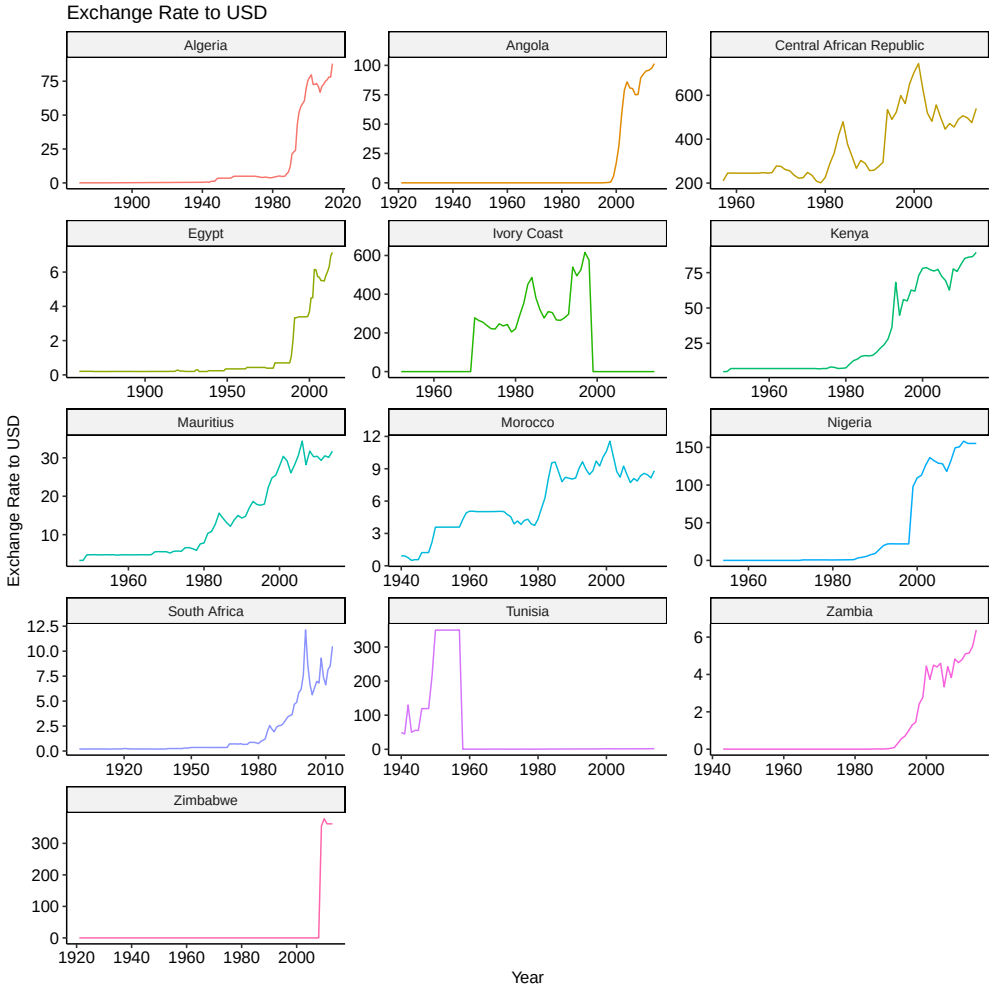
Column	Value	Count	Proportion
Systemic Crisis	0.00	977	92%
	1.00	82	8%
Banking Crisis	0.00	965	91%
	1.00	94	9%
Inflation Crisis	0.00	922	87%
	1.00	137	13%
Currency Crisis	0.00	923	87%
	1.00	136	13%
Domestic Debt in Default	0.00	1017	96%
	1.00	42	4%
Sovereign Debt in Default	0.00	897	85%
	1.00	162	15%
GDP Weighted Default	0.00	1029	97%
	0.06	7	1%
	0.23	6	1%
	0.40	6	1%
	0.13	6	1%
	0.36	5	0%
Independence	1.00	822	78%
	0.00	237	22%

Since the focus of model building is to predict a systemic crisis, we will now look into the correlations of different crises against *Syst*. From the correlation matrix below, we can see that Banking crisies are highly

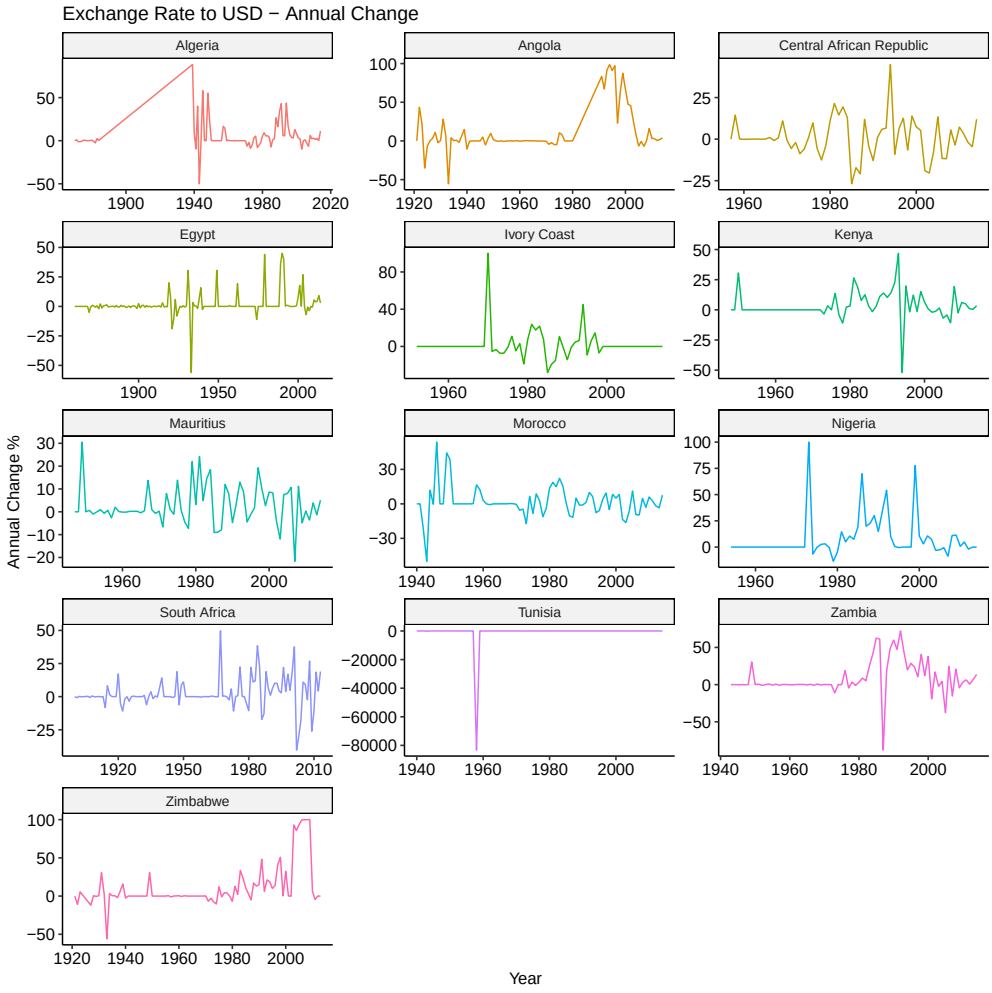
correlated to Systemic crises. Similarly, Currency & Inflationary crisies are highly correlated with each other, just like Domestic & Sovereign Debt crises.



We will now look into the exchange rates against the US Dollar for each Country. The charts below show the movement in each country's exchange rate over time. Besides Tunisia and the Ivory Coast, we can say that African Countries have had large currency depreciations against the US Dollar, with most of these depreciations coming in towards the end of the 20th century.

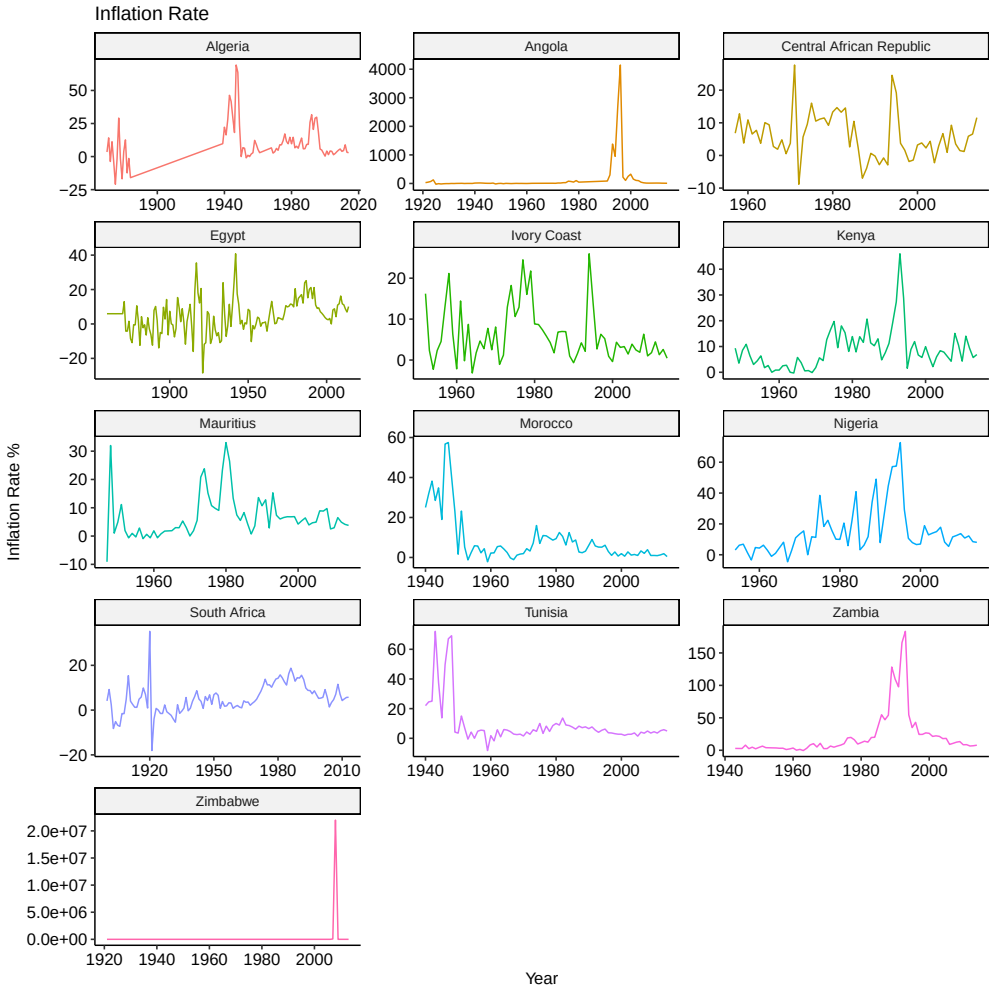


To expand on this, we will now look into the annual changes in these exchange rates per Country. We can clearly see the above comments on currency depreciation across all Countries is validated. We can also see the severity of each of these currency depreciations. In the case of Zimbabwe, the below graphs provide a clearer understanding in their exchange rate history.

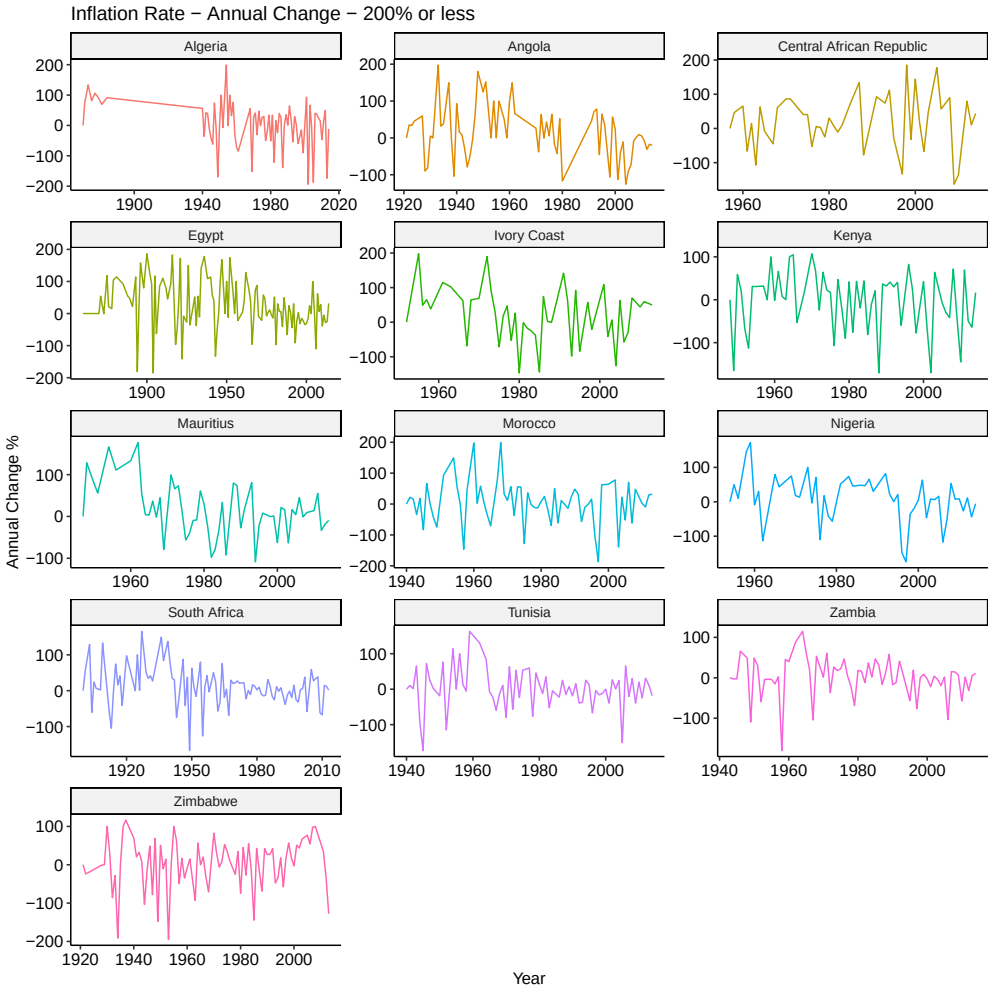


With regards to Inflation, the charts below show the movement in each country's inflation CPI rate over time. Overall, it is clear to see moments of Hyper-inflation and Hyper-deflation across all economies. Also, it appears that the 1990's were a common period of high inflation across many countries, whilst other countries had inflation spikes post-WWII (late 1940's and 50's).

Unlike the previous exchange rates plots, Tunisia and the Ivory Coast have clear information when looking at their respective charts. However, countries with extreme Hyper-inflation like Zimbabwe and Angola are harder to decipher.



Similarly, we will now look into annual changes in the inflation rates of each African country. In the first run of these charts, extreme values tended to skew the visuals. Thus, the below chart exclude absolute annual inflationary movements of 200% or larger, in order to provide a clearer understanding of the magnitude of CPI movements.



2.5 Model Building

Noting the analysis that we done in previous sections of this report, in this section we will now look into bulding a few machine learning models to predict Systemic Crises in African Economies. To produce the training & testing data, the cleaned Afrian Crisies dataset was split 80/20. There are 846 rows in the training data and 208 rows in the testing data. Models were build using the training set, and then compared to the test set via the Accuracy measure from the Confusion Matrix, in order to determine if additional parameters are need before a final model can be selected.

In deteriming the success of a machine learning model, the Confusion Matrix is a summary tool that tabulates the each combination of predictions and actuals. It shows the number of true & false positive outcomes, and the number of true & false negative outcomes. The ability of a model to correctly predict true outcomes is called the Sensitivity of a model (number of true positives over the sum of true positives and false negatives), whilst the ability of model to correctly not predict false ouctomes is called the Specificity (number of true negatives over the sum of true negatives and false positives).

The Accuracy measure combines Sensitivity and Specificity into a single metric for assessing a model's predictive ability. It is calcuated as the sum of true positives & true negatives, over the sum of all positives and negatives. The higher the Accuary score, the better the model is at predicting outcomes.

2.5.1 GLM

The first Machine learning method that was trialed was the Generalised Linear Model (GLM). The initial model that was run is using *Country* as the only predictor which is shown in the formula, where c, y refers to each Country & Year pairing. Whislt is it not expected to use the Country's name as a predictor for a systemic crisis, this model was run as a baseline model.

$$Model_1 : \hat{Syst}_{c,y} = \alpha + \beta_c Country_c$$

The second GLM model was run using Banking Crisis (*Bank*) and Annual Echange rate changes (*EX_change*) as predictors. These variables were selected after inspecting the correlation matrix and selecting parameters that are highly correlated to *Syst*.

$$Model_2 : \hat{Syst}_{c,y} = \alpha + \beta_1 Bank_{c,y} + \beta_2 \Delta EX_{c,y}$$

The third and final GLM model was run using Banking Crisis (*Bank*) and Currency Crisis (*Curr*) as predictors. Similary, these variables are selected as they were highly correlated to *Syst*. However, *Curr* has slighly higher correlation than *EX_change*.

$$Model_3 : \hat{Syst}_{c,y} = \alpha + \beta_1 Bank_{c,y} + \beta_2 Curr_{c,y}$$

The accuracies of these three GLM models were compared against the test dataset, with Model 3 having the highest accuracy overall.

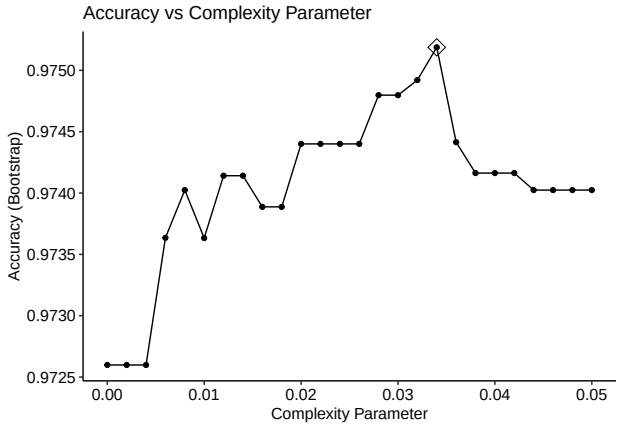
Table 7: GLM Accuracy

Model	Accuracy
1	0.918
2	0.966
3	0.971

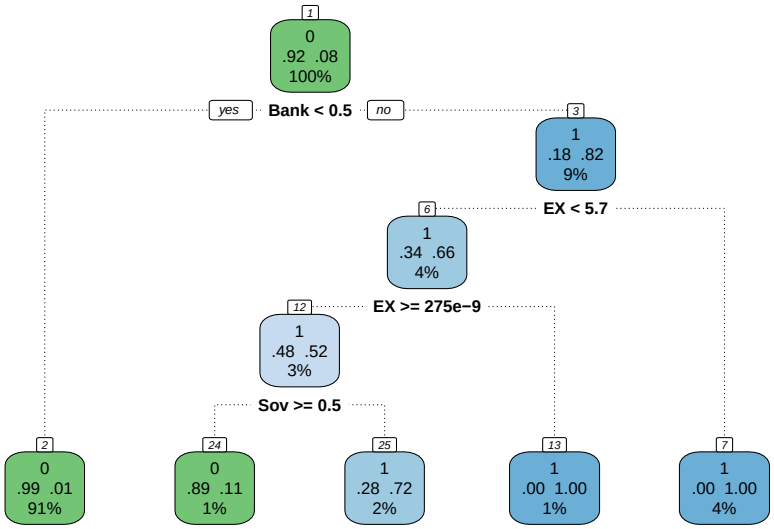
2.5.2 Decision Tree

The Decision Tree is a machine learning technique that is classifies predictions based on the most common outcomes. It is ideal for when the outcome is categorical, and since *Syst* can only be classified as "0" or "1", this technique is suitable. It can be thought of as a series of "gates" that observations must go through, before being clustered into final prediction buckets, where each "gate" is a filter based on a parameter of the tree.

Once component of Decision Tree modelling that must be taking into account is the Complexity Parameter (cp) - the threshold for the number of nodes that can exist in a tree. The more nodes that are in a tree, the better the model is at fitting the data, but too many nodes can lead to over-training. Thus, a suitable Complexity Parameter must be selected that can make the Decision Tree accurately predict outcomes, but not overfit the data. In the case of this report, the below chart shows different complexity parameters against their produced accuracy against the test dataset, with cp = 0.034 resulting in the highest accuracy. This Complexity Parameter value was used to build the Decision Tree.



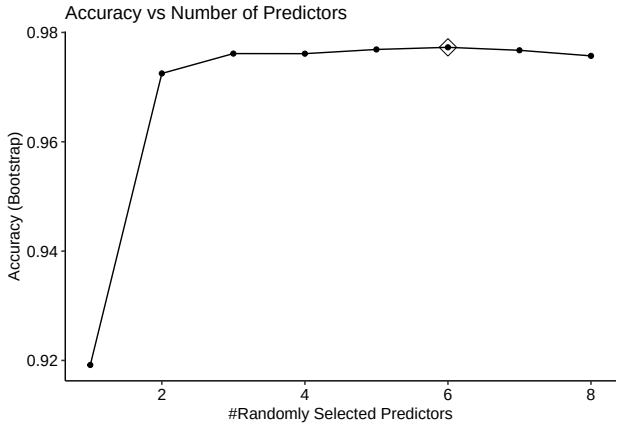
The final decision tree is shown below, with *Bank*, *EX* and *Sov* selected as the parameters of our tree. The majority of observations can be determined using just the *Bank* variable, with the remaining being filtered on exchange rates higher than 5.7 or if a Country is in a sovereign debt crisis. The accuracy of this Decision Tree is discussed in the Results section of this report.



2.5.3 Random Forest

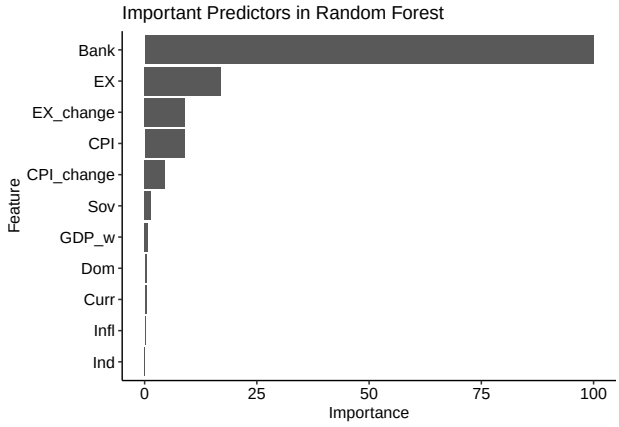
In simple terms, a Random Forest is a machine learning technique where many Decision Trees are created, with each tree being different to one another due to random bootstrapping of the training set, and the average outcome of all trees is selected as the final model. It can be viewed as taking the average of a "forest" of trees, where each tree has a unique combination of parameters and nodes. This machine learning technique is more stable than a single Decision Tree since the average of a large sample of outcomes is chosen, instead of relying on one parameter set.

In our model, 100 trees were created in the Random Forest. From the chart below, we can see that model with 6 randomly selected predictors resulted in the highest accuracy in our Random Forest. This is selected as the final model and the outcomes are discussed in the Results section of this report.



Whilst a Decision Tree is easier to interpret/visualise than a Random Forest, one metric that we can use to better understand this Random Forest is to look at Importance - a measure of how many times a variable has been applied in the trees of a Random Forest.

Below is a chart that shows how important each variable is in the training set. *Bank* is by far the most important parameter as it was used across all trees in the Random Forest - this is expected as it was highly correlated to *Syst* as shown in the Data Analysis section of this report. Exchange rates (values & annual changes), Inflation (values & annual changes), Sovereign debt and GPD weighted debt, are also important parameters in our forest. It is interesting to note that the Independence of an African Economy was not used in any of the 100 trees in the Random Forest.



3 Results

The final model that will be used is the Random Forest model, which accounts for the stability of economy's banking system, the volume of soverign debt, the exchange rate and the inflation rate. As shown in the table below, this Random Forest Model was selected as it had the highest Accuaracy measure against the test dataset out of the three machine learning models used.

Table 8: Accuracy of all models

Model	Accuracy
GLM	0.971
Decision Tree	0.971
Random Forest	0.981

4 Conclusion

4.1 Summary of Findings

Accross all models, a Banking Crisis is by far the strongest indicator of a Systemic Crisis in an African Country. It is also important to note that other economic indicators such as the exchange rate to the US Dollar, inflation rate and amount of soverign debt, are also useful predictors of Systemic Crises.

Based of the findings of the final Random Forest Model, it should be noted that African economies which are independent, are not necessarily more protected against financial instability than those which are not independent. Also, the amount of domestic debt is not as strong of an indicator of financial instability than soverign (foreign) debt.

4.2 Suggested Improvements

As the focus of this report is only on the African contient, the most notiable suggestion for future improvements would be to gather a larger dataset with information from other counties outside of Africa, as this will allow us to idenfity if the discussed economic indicators of crisies are only unique to African economies. However, since the focus of this report was to only focus on the African economy, expanding the datset was beyond the initial objective. With that being said, it is believed that the general findings of this reports can still be applied to all economies. An additional future suggestion would be to apply other types of machine learning technicques like logistic regression, LDA, K-nearest neighbour, etc.

5 Acknowledgements

Chiri. (2019) Africa Economic, Banking and Systemic Crisis Data: Data on Economic and Financial crises in 13 African Countries (1860 to 2014). [online] Available at: https://www.kaggle.com/chirin/africa-economic-banking-and-systemic-crisis-data [Accessed: 20 April 2020]

Reinhart, C., Rogoff, K., Trebesch, C. and Reinhart, V. (2019) Global Crises Data by Country. [online] Available at: https://www.hbs.edu/behavioral-finance-and-financial-stability/data/Pages/global.aspx [Accessed: 20 April 2020].