Predicting Inflation Crisis in 13 African Countries

Camille Kirongozi

C772: Data Analytics Graduate Capstone

Professor Daniel Smith

January 17, 2022

We often ponder on the question of whether inflation is good or bad. An Economics professor, Mark Kuperberg at Swarthmore College said that decades ago, “there was a view that zero inflation was the right number, and there are some fancy theoretical models that support that. But it’s a minority view among macroeconomists now (Hartman, 2019).” We live in a world where inflation is considered a necessity, but too much inflation can lead to inflation crisis, which can then lead to economic collapse. In this report, I will be analyzing historical economic data for thirteen African countries to predict inflation crisis using a logistic regression model. The countries included in the dataset are Algeria, Angola, Central African Republic, Ivory Coast, Egypt, Kenya, Mauritius, Morocco, Nigeria, South Africa, Tunisia, Zambia, and Zimbabwe. The research question that I address in this report is, which factors are significantly associated with inflation crisis in the 13 African countries observed? In my logistic regression model, I aim to identify the most significant variables in predicting inflation crisis at a 0.05 significance level.

The first assumption of the logistic regression is that the outcome variable is binary. The dependent variable should have values of yes/no or 1/0. The second assumption requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data (Schreiber-Gregory, 2018). The logistic regression does not allow multiple observation of the same rows. The third assumption is the absence of multicollinearity. The fourth assumption is the linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds (Schreiber-Gregory, 2018). These are the assumptions of a logistic regression model.

The outcome variable of the analysis is the inflation\_crises, which is binary with the values 1/0. This aligns with the binary assumption of the logistic regression. Each row of the data represents one year for each country, thus meeting the second assumption. A variance inflation factor of all predictors in the reduced model was produced to evaluate the third assumption of multicollinearity. There were no results higher than 10 in the VIF output, which would indicate multicollinearity. The results can be found below:Text

Description automatically generated with medium confidence

The exch\_usd is the only numeric variable in the final logistic regression model (Model2). To test the fourth assumption, I created a plot for the exch\_usd variable against the log odds of the dependent variable, inflation\_crises. It is evident that the only numeric variable has linear relationship with the log odds of the dependent variable. The plots can be found below:

Chart, line chart

Description automatically generated

The data meets all the logistic regression assumptions. Thus, the logistic regression is an appropriate technique for this analysis.

The african\_crises dataset was downloaded from Kaggle. The dataset is a derivative of Reinhart et. al's Global Financial Stability dataset (Chili, 2019). I used RStudio to analyze the data, which uses the R language. R is optimized for statistical analysis and data visualization (Cloud, 2021). I began the analysis by looking for missing values in the dataset. The image below reveals the code that I used, which evaluates each variable for missing values. Looking at the results below, there were no missing values found in the raw data.

A picture containing graphical user interface

Description automatically generated

The data preparation steps taken before the analysis consisted of removing unnecessary variables, altering the banking\_crisis variable to have values 1/0 instead of crisis/no crisis, and removing some rows bases on the outliers found in the exch\_usd and currency\_crises variables. I removed the first three columns which consisted of, case number, three letter country code, and country name. These variables were listed 1-3. I also removed the independence variable which indicated the year of a country’s independence. This variable is a binary categorical variable with values 1/0. I removed the variable because I believe independence for these countries was imperative and I did not want to account for it as a predictor for inflation crises.

I also removed the inflation\_annual\_cpi variable due to a perfect predictor error. When I included the inflation\_annual\_cpi variable in the logistic regression model, I receive a warning message which I included in the screenshots below. This warning message is due to a perfect prediction, indicating that one or more predictors can predict the outcome variable perfectly (Rulphus, 2018). To address the error, I removed the variable from the dataset. These steps reduced the number of variables from fourteen to nine and the number of observations from one thousand fifty-nine to one thousand fifty-three.

Text

Description automatically generated

Warning message caused by the inflation\_annual\_cpi variable in the logistic regression model:



Text

Description automatically generated

Removing the observations with the 2 values brought the total number of observations from 1059 to 1055. The following boxplot revealed two extreme values in the exch\_usd variable:

A picture containing graphical user interface

Description automatically generated

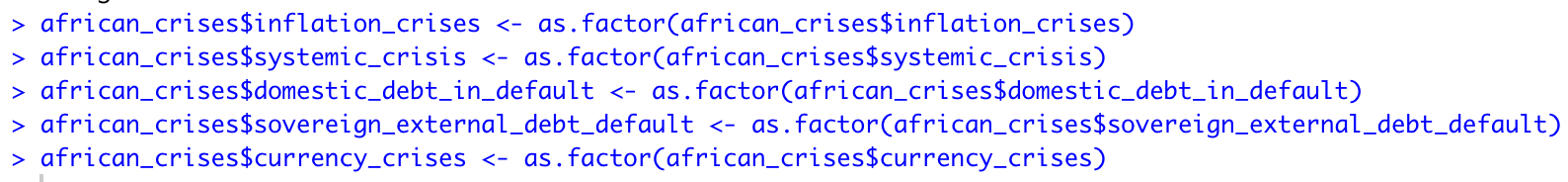
I removed the observations with the extreme exch\_usd values, reducing the total number of observations to 1053:



I altered the categorical variable types from numeric to factors prior to beginning the analysis:

Text, letter

Description automatically generated



The changes to the data were made before the summary statistics was produced. The summary statistics can be found in the image below. The summary statistic was generated using the “describe” command in R. The data consists of nine variables and a sample size of one thousand fifty-three. The target variable for this analysis is the “inflation\_crises” variable. The variable is a binary categorical variable with values 1/0. Most of the predictor variables are factors with binary outcomes. These variables are distinguished with a star next to the variable name on the printout of the describe command. All the categorical variables have standard deviations ranging from 0.27 to 0.36, with means ranging from 1.04 to 1.15. The year, exch\_usd, and gdp\_weighted\_default variables are the numeric variables. The first regression model I produced consisted of all independent predictor variables, I later reduced the variables for the final model.

Summary statistic of the cleaned data:

Table

Description automatically generated

Here are my univariate plots for the variables that I used in the final regression model and the dependent variable. The independent variables are exch\_usd, sovereign\_external\_debt\_default, currency\_crises, and banking\_crisis. The plots consist of frequency plots and a box plot for the wxch\_usd variable which is numeric. The variable captures the exchange rate of the countries to the US dollar. The boxplot revealed numerous outliers in the exch\_usd variable. I decided to ignore the findings because of the significance of the US dollar in the world’s economy.

A picture containing text, screenshot, display

Description automatically generatedChart, histogram, scatter chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart

Description automatically generated

Here are my bivariate plots for the variables that I used in the reduced regression model. In this section, the variables are plotted against the dependent variable, inflation\_crises, with 0 meaning no inflation\_crises occurred and 1 meaning an inflation crises occurred. I produced a box plot for the exch\_usd variable and included the observations in the plot to show the difference between observations where inflation crises occurred and where it did not. The categorical variables are shown using frequency plots.

Chart, scatter chart

Description automatically generatedChart, bar chart

Description automatically generated

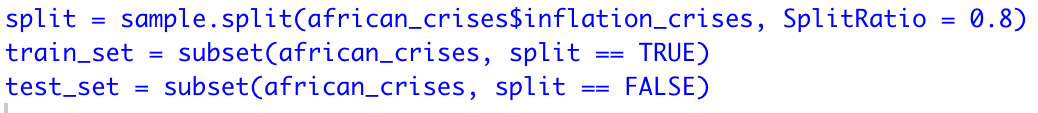
Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

The prepared data used in the analysis will be included in the submission. Following the changes made to the data, I split the data into a training and testing set using a 0.8 ratio, 80 percent as the training set and 20 percent as a testing set. I then constructed the initial logistic regression model, which I named “Model,” using R’s glm command, “Model <- glm(inflation\_crises ~.,data = train\_set, family = binomial(link = "logit")),” I then printed a summary of the model using the print command, “print(summary(Model)).” The codes and summary of the initial model can be found below.

Splitting the data into a train/test set:



Summary of the initial Logistic regression model:

Table

Description automatically generated

Looking at the initial model’s summary, it is evident that not all the variables are statistically significant. The model resulted in four variables with p-values less than .05, which was the significance level that I determined for the analysis.

I created a reduced logistic regression model using the variable ranking method. This method eliminates unnecessary variables, only keeping statistically significant variables, i.e., variables with the lowest p-values. This method is very useful because models with many variables can be difficult to interpret. The excess of variables unrelated to the outcome can influence the model, it can also lead to overfitting of the training set (Vettoretti, 2021). I established an alpha level of 0.05 and only selected variables with P-values equal to or less than 0.05. The reduced model is named, “Model2.” The model includes both categorical and numeric variables. The image below displays the summary of Model2, it also includes the code used to generate the model. When comparing the Akaike Information Criterion (AIC) of both models, the second model has a lower AIC, indicating that it is the better model.

Table

Description automatically generated with medium confidence

To evaluate the models, I performed an accuracy test for both models, comparing the predictions with the actual values in the test set. The test resulted in the same accuracy score for both models. I also produced a confusion matrix for each model and the results were identical, with the same true positives and true negatives.

Accuracy score of the initial model:

Text

Description automatically generated

Accuracy score of the reduced model:

Text

Description automatically generated

Confusion matrix for both models:

Graphical user interface, text, application, email

Description automatically generated

Although both models resulted in the same accuracy score, the reduced model provides the thirteen African countries with a narrower approach in preventing inflation crises. Here is the logistic regression equation for Model2: Log(p/1-p) = inflation\_crises + exch\_usd \* (X1) + sovereign\_external\_debt\_default \* (X2) + currency\_crises \* (X3) + banking\_crisis \* (X4). The log(p/1-p) is the link function between the probability of success and failure (1/0). Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way (Vidhya, 2015). The coefficients for the variables are included in the image below.

A picture containing text

Description automatically generated

The purpose of this analysis was to identify which variables are the most significant in predicting inflation crises in the thirteen African countries. Based on the logistic regression analysis performed with the data provided, the variables used in Model2 produce the best predictions. Although we are limited in making predictions at a one hundred percent accuracy rate, the final logistic regression model provided predictions at an accuracy rate of ninety percent. My recommendation is to understand the model’s equation to determine each coefficient’s effect on inflation crises. Understanding the equation will allow the countries to target each category/variable respectively to prevent inflation crises.

Works cited

“A Box and Whiskers Plot (in the Style of Tukey) - geom\_boxplot.” *- geom\_boxplot • ggplot2*, 2021, https://ggplot2.tidyverse.org/reference/geom\_boxplot.html.

Chiri. “Africa Economic, Banking and Systemic Crisis Data.” *Kaggle*, 21 July 2019, https://www.kaggle.com/chirin/africa-economic-banking-and-systemic-crisis-data.

Cloud Team, IBM. “Python vs. r: What's the Difference?” IBM, 23 Mar. 2021, [www.ibm.com/cloud/blog/python-vs-r#:~:text=R%20is%20an%20open%20source,elegant%20tools%20for%20data%20reporting.&text=Creating%20visualizations,learning%20and%20deep%20learning%20algorithms](http://www.ibm.com/cloud/blog/python-vs-r#:~:text=R%20is%20an%20open%20source,elegant%20tools%20for%20data%20reporting.&text=Creating%20visualizations,learning%20and%20deep%20learning%20algorithms).

Gelman, Andrew. “Statistical Modeling, Causal Inference, and Social Science.” *Statistical Modeling Causal Inference and Social Science*, 10 Jan. 2020, https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/.

Hartman, Mitchell. “Why Is Inflation Necessary?” *Marketplace*, 12 Sept. 2019, <https://www.marketplace.org/2019/09/12/why-is-inflation-necessary/>.

Rulphus, Adam, and Roman Ahmed. “Logistic Regression Model: " Glm.fit: Fitted Probabilities Numerically 0 or 1 Occurred ".” *RStudio Community*, 19 June 2018, <https://community.rstudio.com/t/logistic-regression-model-glm-fit-fitted-probabilities-numerically-0-or-1-occurred/9828/15>.

Schreiber-Gregory, Deanna. *Logistic and Linear Regression Assumptions: Violation Recognition and Control*. Henry M Jackson Foundation for the Advancement of Military Medicine, 2018, https://www.lexjansen.com/wuss/2018/130\_Final\_Paper\_PDF.pdf.

Vettoretti, Martina, and Barbara Di Camillo. “A Variable Ranking Method for Machine Learning Models with Correlated Features: In-Silico Validation and Application for Diabetes Prediction.” *Applied Sciences*, vol. 11, no. 16, 2021, p. 7740., https://doi.org/10.3390/app11167740.

Vidhya, Analytics. “Logistic Regression R: Introduction to Logistic Regression.” *Analytics Vidhya*, 1 Nov. 2015, https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/.