Predicting Banking Crises in Africa

Group: GR1

Group members' names and contributions:

- Saniya Abushakimova (saniyaa3, GR) – Abstract, Data Preprocessing, Logistic Regression, Logistic Regression with ElasticNet, Evaluation.
- Jaehyung Kim (jk38, GR) – Introduction, Gradient Boosting, Random Forest, Results.
- Andrew Cao (hanghec2, GR) – Exploratory Data Analysis, SVC, K-NN, Discussion.

Content

1. Abstract

Banking crises, while rare, can cause serious damage to economies, making early detection critically important. In this project, we aimed to build models that could predict the likelihood of banking crises in 13 African countries using macroeconomic indicators from 1970 to 2018. We processed the dataset by cleaning features, adding lagged variables, and splitting it by time to avoid data leakage. Then, we trained and compared six models, including Logistic Regression, SVC, KNN, Random Forest, and Gradient Boosting. Among these, Logistic Regression performed best in terms of recall, detecting 88% of actual crisis cases. Our results suggest that even simple models can be powerful early warning tools when key features like inflation and systemic crises are included. Due to the class imbalance, we prioritized recall as our main evaluation metric. Future work could benefit from more advanced time-series models and a richer set of economic and political features.

2. Introduction

Recently, many countries have faced challenges stemming from unstable trading systems and tariff uncertainties. Given these challenges, governments have started cooperating domestically and internationally to come up with appropriate solutions and protect their economies against various crises. In particular, the financial crisis can cause substantial economic disruption. The financial crises, though infrequent, have severe economic consequences – early detection can inform proactive policy interventions. To address this need, this project aims to **forecast banking-sector financial crises** by leveraging historical macroeconomic indicators.

In terms of the dataset, it represents **annual** financial stability across 13 different African countries from 1860 to 2014. This was collected by Carmen Reinhart, Professor of the International Financial System at the Harvard Kennedy School, and the dataset was obtained from Kaggle for our analysis. For model development, the dataset was split into the following training and testing sets:

- **Training set:** 831 records from 1860 to 2014.
- **Testing set:** 215 records from 1860 to 2014.

The following provides details regarding the feature variables:

- **case:** a unique number that indicates the specific country.
- **cc3:** three-letter country code.
- **country:** name of the country.
- **year:** year of the observation.
- **systemic_crisis:** “0” – systemic crisis did not occur, “1” – systemic crisis occurred.
- **exch_usd:** exchange rate of the country relative to the USD.
- **domestic_debt_in_default:** “0” – sovereign domestic debt default did not occur, “1” – sovereign domestic debt default occurred.
- **sovereign_external_debt_default:** “0” – sovereign external debt default did not occur, “1” – sovereign external debt default occurred.
- **gdp_weighted_default:** total debt in default relative to the GDP.
- **inflation_annual_cpi:** annual CPI inflation rate.
- **independence:** "0" – not an independent country, "1" – an independent country.
- **currency_crises:** "0" – currency crisis did not occur, "1" – currency crisis occurred.
- **Inflation_crises:** "0" – inflation crisis did not occur, "1" – inflation crisis occurred.
- **banking_crisis:** "no_crisis" – banking crisis did not occur, "crisis" – banking crisis occurred.

3. Statistical Methods

3.1. Exploratory Data Analysis

We started our statistical methods by exploring the data. For example, Figure 1 shows that the number of records varies across countries, with Egypt and South Africa having the most data, while countries like Nigeria and the Central African Republic have much less. This imbalance means some countries may influence the model more than others, so we need to be careful when interpreting results or applying the model broadly.

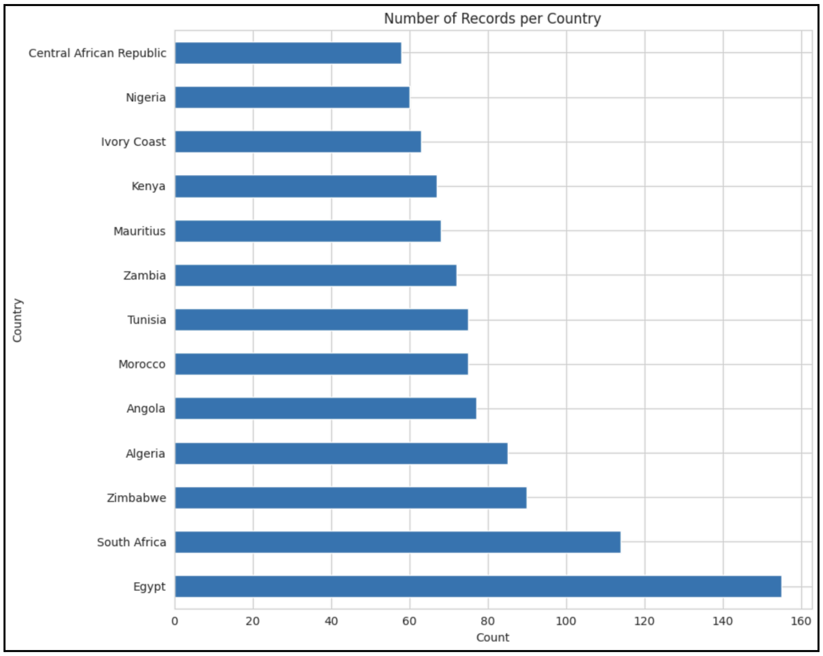

Figure 1: Number of Records per Country

Figure 2 shows inflation rates alongside marked inflation crises for each country. In most cases, sharp spikes in inflation clearly align with crisis periods. However, in countries like Angola and Zimbabwe, the extreme inflation values distort the scale, making crisis years harder to interpret visually.

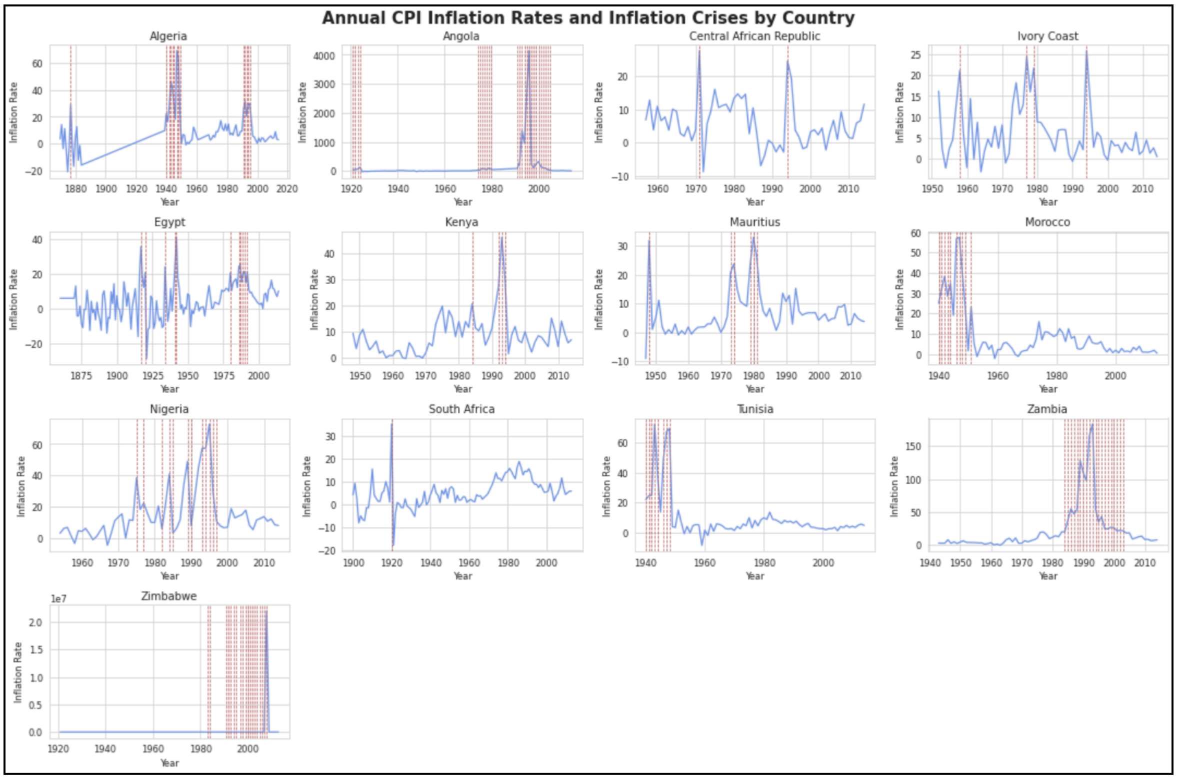

Figure 2: Inflation Rates and Crisis Periods Across Countries

Figure 3 shows the exchange rate trends alongside currency crises across countries. In most cases, a spike or persistent rise in exchange rates is followed by or coincides with periods marked as currency crises (red lines). However, this relationship is not visually consistent across all countries. For instance, countries like Kenya or Egypt show significant exchange rate jumps aligning with crisis periods, while others like Zimbabwe or Tunisia show flatter trends with little visual alignment. This highlights that while exchange rate volatility can be a contributing factor, it alone may not explain the onset of currency crises and must be considered alongside other indicators.

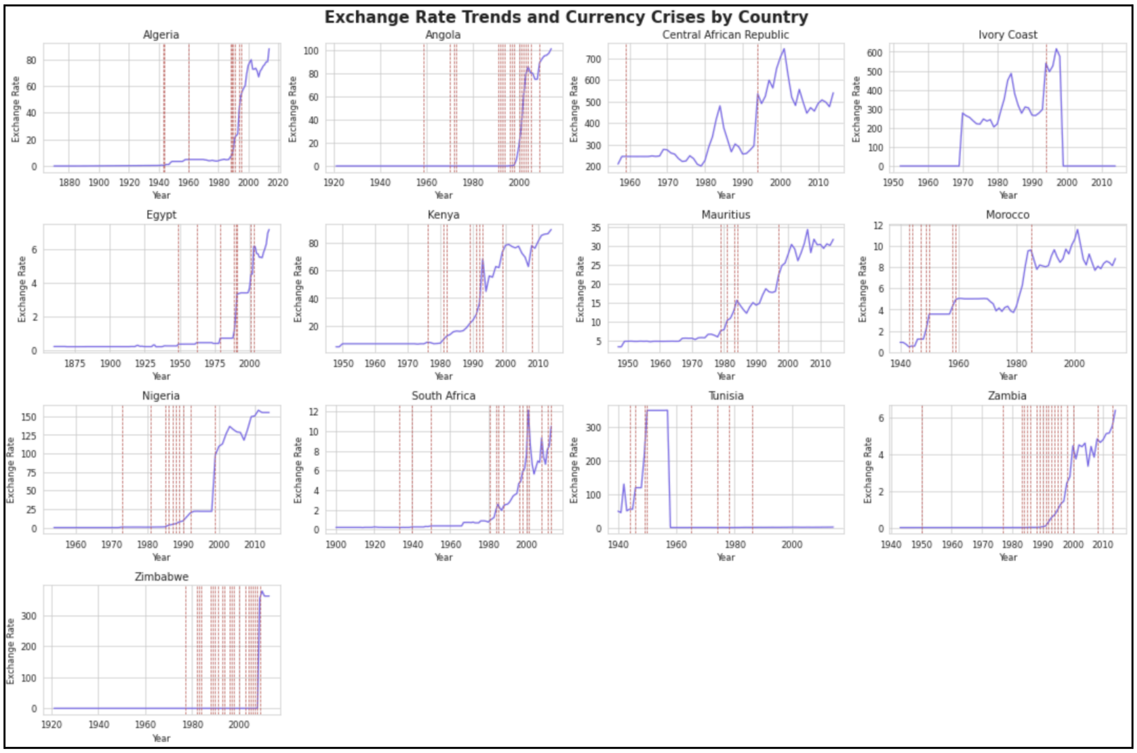

Figure 3: Exchange Rate Movements and Currency Crisis Markers Across Countries

Figure 4 shows the percentage of years each country experienced a banking crisis during the analysis period (1860–2014). The Central African Republic stands out with over 30% of its years marked by crisis, followed by Nigeria and Zimbabwe. On the other hand, countries like South Africa, Morocco, and especially Mauritius have faced relatively few banking crises. This variation highlights how financial stability differs significantly across African nations, with some being more prone to recurring banking system stress.

Figure 5 shows a strong class imbalance in banking crisis status across countries — the number of non-crisis years (yellow) significantly outweighs crisis years (blue). Because our data is grouped by country and ordered chronologically, we can't apply SMOTE, which relies on randomly sampling and generating synthetic points. Instead, we use **class weighting**, which adjusts the model to give higher importance to the minority class (crisis) during training. This helps the model learn patterns of rare events without artificially altering the time structure of the data.

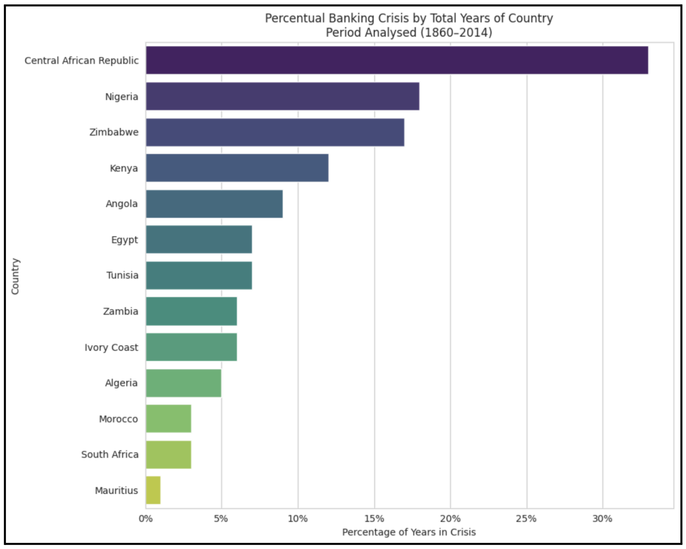

Figure 4: Share of Banking Crisis Years by Country

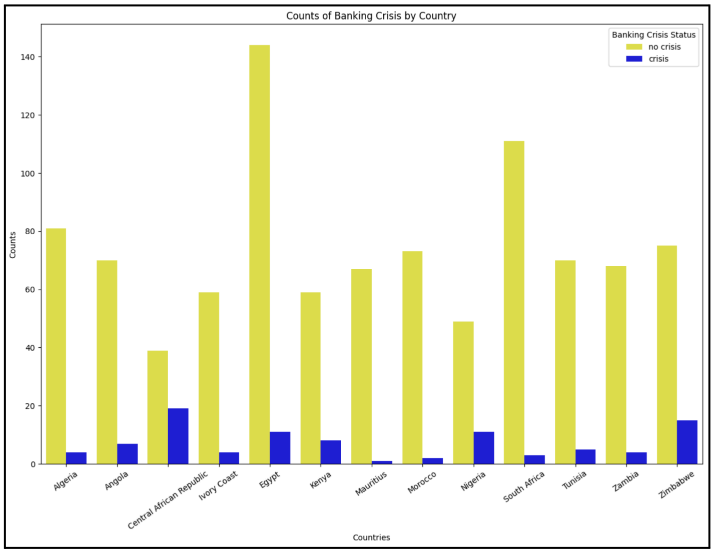

Figure 5: Class Imbalance in Banking Crisis

3.2. Data Processing

Our preprocessing steps focused on cleaning the dataset, creating more informative features, and preparing it for time-based modeling.

1. <u>Removing Unbalanced Features:</u>

We began by dropping two features: **domestic_debt_in_defaul**t and **gdp_weighted_default**. From Figures 6 and 7, we can see that these were extremely imbalanced, with very few positive cases compared to negatives, which could have skewed the models. Removing them helped reduce noise and made the training more stable.

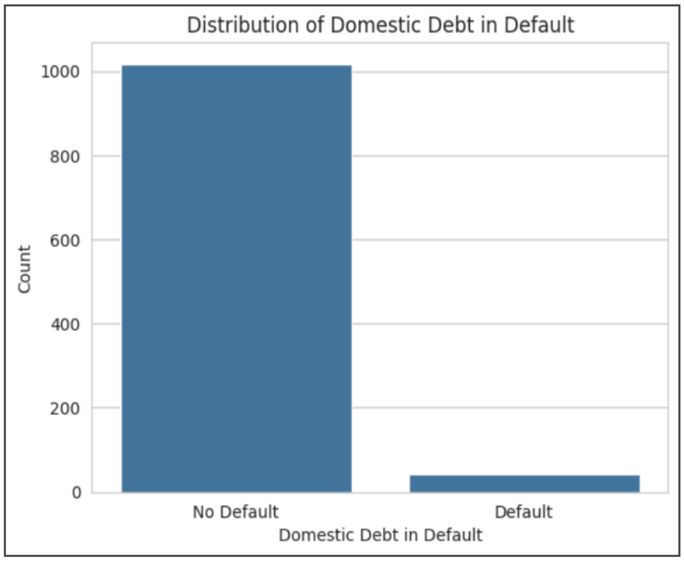

Figure 6: Domestic Debt in Default distribution

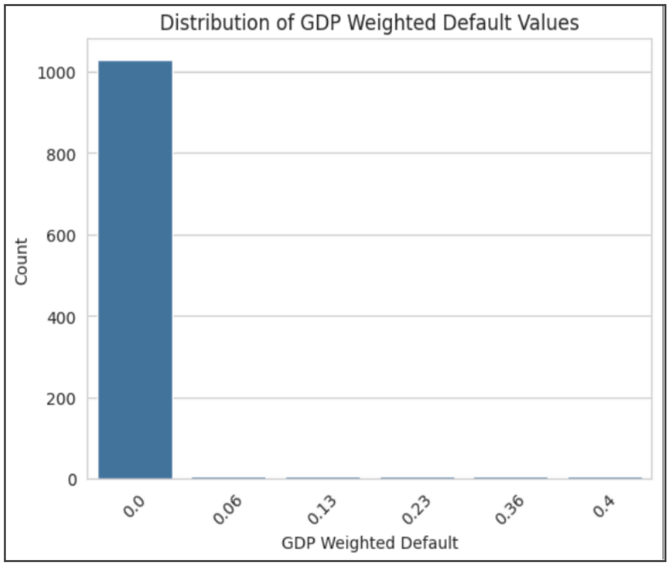

Figure 7: GDP Weighted Default distribution

2. Adding Lagged Variables:

To help the model understand how past events influence current banking crises, we created lagged versions of selected variables. The data was first grouped by country and sorted by year to ensure the lagged features aligned correctly. The lagged variables we added were:

- **banking_crisis_lag1:** If there was a crisis last year, there's a higher chance of continued instability this year.
- **currency_crises_lag1:** Currency crashes often trigger banking stress due to foreign reserve losses and devalued assets.
- **inflation_crises_lag1:** A past inflation crisis can shake financial systems and weaken bank balance sheets, increasing this year's crisis risk.
- **inflation_annual_cpi_lag1:** Persistently high inflation from the previous year can signal ongoing economic stress, which may continue to weaken the banking sector in the current year.

These lagged features allow the model to consider how recent macroeconomic stress contributes to current financial vulnerabilities.

3. Splitting Train and Test Sets:

Since this is time series data, we divided the dataset chronologically. We made sure that each country's data was split such that the earliest 80% went into training and the most recent 20% into testing. This ensures that each country is present in both sets and prevents the model from seeing future information when training.

4. Normalization:

We normalized features for models that are sensitive to scale, like Logistic Regression, SVC, and KNN. This helped ensure that variables with larger numerical ranges didn't dominate the learning process. For tree-based models (Random Forest, Gradient Boosting), normalization wasn't applied since those models aren't scale-sensitive.

3.3. Models

3.3.1. Logistic Regression

To start, we implemented a standard Logistic Regression model due to its simplicity, interpretability, and effectiveness in binary classification tasks. Since Logistic Regression is sensitive to the scale of input features, we normalized the data to ensure that all variables contributed equally to the model. Given the heavy class imbalance in our target variable – banking crises are relatively rare – we applied *class_weight* = *'balanced'* to give more weight to the minority class during training.

We used **Recursive Feature Elimination (RFE)** to reduce the number of features and improve model focus and generalization. RFE works by iteratively fitting the model, ranking features by their importance (in this case, based on the magnitude of coefficients), and removing the least significant ones until the desired number of features remains. We selected the top 8 features using this method.

Among these, the top three features with the highest absolute coefficients were:

- **inflation_annual_cpi:** High inflation often signals macroeconomic instability. It can reduce consumer purchasing power and investor confidence, leading to lower bank revenues, defaults on loans, and liquidity problems, all of which raise the risk of a banking crisis.
- **systemic_crisis**: System-wide financial crises (e.g., stock market crashes or currency collapses) can trigger bank runs, capital outflows, and broader panic in the financial sector. This feature directly reflects episodes where the financial system is under severe stress.
- **independence**: Countries undergoing major political changes, like gaining independence, may face temporary institutional voids, policy shifts, or public unrest. These transitions often bring economic volatility, which increases the risk of instability in the banking sector.

Together, these variables capture both short-term economic shocks and longer-term structural disruptions. Their prominence in the model supports their importance as early warning signals for banking crises.

3.3.2. Logistic Regression with LASSO

We also implemented a Logistic Regression model with ElasticNet regularization. This approach combines both L1 (Lasso) and L2 (Ridge) penalties, allowing for variable selection while still maintaining model stability. To balance both penalties, we set the l1_ratio to 0.5. We also increased the number of iterations to 5000 to ensure proper convergence of the optimizer.

Like in standard Logistic Regression, we normalized the data since ElasticNet is sensitive to feature scale. To address the issue of class imbalance in our target variable, we again used *class_weight = 'balanced'*, which adjusts weights inversely to class frequencies, giving more importance to the rare banking crisis cases.

For feature selection, we relied on the built-in effect of Lasso regularization, which shrinks less important feature coefficients toward zero. As a result, ElasticNet automatically selected a subset of informative variables without the need for additional RFE. Interestingly, the top three features selected – **inflation_annual_cpi**, **systemic_crisis**, and **independence** – matched those identified in the standard Logistic Regression model. This consistency reinforces their importance as strong early warning indicators of banking crises across different modeling approaches.

3.3.3. Support Vector Classifier (SVC)

The third model we used was a Support Vector Classifier (SVC) with a linear kernel. SVC aims to find the optimal separating line (or hyperplane) that distinguishes between crisis and non-crisis cases using the most relevant features. We applied Recursive Feature Elimination (RFE) beforehand to reduce the number of input variables and focus on the most informative ones. To address the class imbalance in our dataset, we used the *class_weight = 'balanced'* parameter, which helps the model give more attention to the underrepresented crisis cases. The top three features that influenced the SVC's predictions were: ***inflation_annual_cpi***, which shows how high inflation can hurt financial stability; ***inflation_annual_cpi_lag1***, which captures the lingering effects of inflation from the previous year; and

systemic_crisis, which highlights how broader financial system disruptions can raise the risk of a banking crisis.

3.3.4. K-NN

For our fourth model, we used a K-Nearest Neighbors (KNN) classifier. Since KNN relies on distance between data points, we first standardized all the features so they would contribute equally to the distance calculations. We then used five-fold cross-validation on the training set to find the best number of neighbors, k, by testing a range of possible values. Once we found the optimal k, we retrained the model on the full-scale training data and evaluated it on the test set. Unlike previous models, we didn’t use class weighting or RFE here – KNN doesn’t support class weighting directly, and feature selection with RFE isn’t ideal for instance-based models like KNN because they don’t have coefficients to rank feature importance. This model is simple and interpretable, but it can struggle when the class distribution is imbalanced, as is the case with our data.

3.3.5. Gradient Boosting

One of the boosting methods we implemented is the Gradient Boosting Machine (GBM) model. Using the gradient of the loss function, it creates the model sequentially by forecasting the residuals from the previous model. These steps reiterate and predict the accurate status of the banking crisis.

Selecting influential features is significant for improving the model’s performance. Accordingly, as described above, we utilized the **Recursive Feature Elimination (RFE)** method to select meaningful features. By ranking features, we opted for the top 8 features before fitting the model. Furthermore, we used a hyperparameter tuning method, *GridSearchCV*, to find the best combination of the parameters. According to the tuning and cross-validation process, we could determine the parameters of the best-performing GBM model: **learning_rate: 0.01, max_depth: 1, n_estimators: 100.** Lastly, we applied the *compute_sample_weight* method to our model to address the class distribution bias.

For feature selection to enhance the performance of the model, we conducted a **feature importance analysis** (Figure 8a)**.** Consequently, we attempted to identify the top three significant features, but only two features, **systemic_crisis and banking crisis_lag1,** were obtained among the top 8 features**.** The systemic crisis denotes a failure of major institutions across the country that can significantly affect a banking crisis. Also, the banking crisis status of the previous year indicates lingering economic effects on the overall economy and a decline in investor confidence.

3.3.6. Random Forest

For the bagging method, we used the Random Forest (RF) tree-based model. Before fitting the model, we addressed the class imbalance issue. Accordingly, we opted for the ‘Class_weight’ parameter: This parameter places greater weights on the minority class (‘crisis’ class) in the loss function during the training procedure.

As mentioned above, we utilized the **Recursive Feature Elimination (RFE)** method to select meaningful features. By ranking features, we also opted for the top 8 features before fitting the model. Furthermore, we used a hyperparameter tuning method, *GridSearchCV*, to find the best combination of the parameters and conduct a cross-validation process. According to the tuning process, we could determine the parameters of the best-performing RF model: ‘**class_weight': 'balanced', 'max_depth': 5, 'n_estimators': 100.**

For feature selection to enhance the performance of the model, we conducted a **feature importance analysis** (Figure 8b)**.** Consequently, the top three significant features, **systemic_crisis, banking_crisis_lag1,** and **exch_usd,** were obtained among the top 8 features**.** The exchange rate against the USD specifies that the volatility in the exchange rate can significantly affect a banking crisis. (The effect of *systemic_crisis* and *banking_crisis_lag1* was mentioned above)

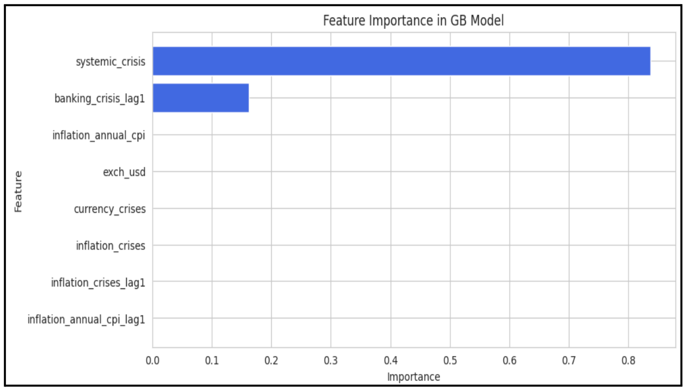

Figure 8a: Feature Importance in GB

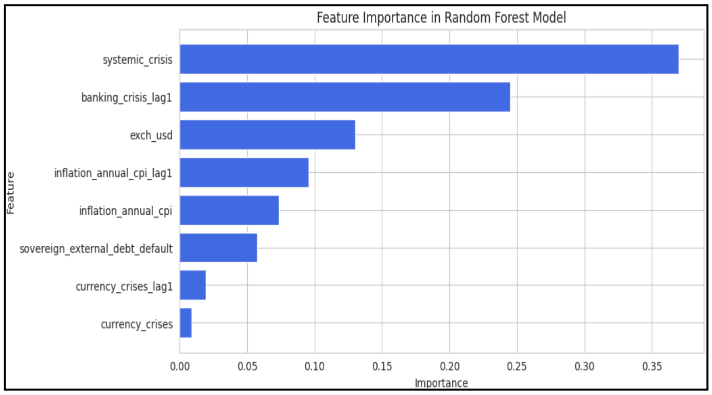

Figure 8b: Feature Importance in RF

3.4. Evaluation

To evaluate the performance of our models, we used several metrics: Accuracy, F1 Score, ROC-AUC, and Recall. Accuracy measures the overall correctness of predictions, but it can be misleading in imbalanced datasets where the majority class dominates. The F1 Score balances precision and recall, making it more informative when false positives and false negatives both matter. ROC-AUC evaluates how well the model separates the positive and negative classes across all thresholds and is especially useful for comparing models.

However, our main focus was on **Recall**, which measures the proportion of actual banking crises that the model correctly identified. Since banking crises are rare but highly impactful, missing a true crisis (false negative) could lead to serious consequences. Prioritizing recall helps ensure that most real crisis events are flagged, even if it comes at the cost of some false alarms. This makes recall the most relevant metric in the context of early warning systems for financial instability.

4. Result

Model	Accuracy	Recall	F1	ROC - AUC
Logistic Regression	0.9674	**0.8846**	0.8679	0.9904
Logistic Regression with ElasticNet	0.9581	0.8462	0.8302	0.9878
SVC	0.9581	0.7692	0.8163	0.9556
K-NN	**0.9721**	**0.7692**	0.8696	0.9449
Gradient Boosting	0.9721	0.7692	0.8696	0.9783
Random Forest	0.9628	0.7692	0.8333	**0.9908**

The table above denotes the performance of multiple classification models. The Logistic Regression model showed the highest recall, making it the most effective at identifying crises. Tree-based models like Random Forest, the highest ROC-AUC, and Gradient Boosting also performed well; however, in terms of recall, they slightly underperformed compared to Logistic Regression.

Even though KNN has the highest accuracy score, its recall is comparatively low. Moreover, KNN does not offer class imbalance adjustment through parameters, making its results less interpretable and less reliable.

Lastly, as mentioned in the Evaluation section, prioritizing the Recall is significant for providing early warnings of the financial instability in advance. Accordingly, we selected **Logistic Regression** as the most appropriate model to achieve this goal.

5. Discussion

We analyzed data from 13 African countries between 1970 and 2018 to understand what signals might help predict banking crises. Surprisingly, a simple Logistic Regression model performed better than more complex models, achieving the highest recall (0.8846), meaning it was good at identifying actual crisis years. The most important features across countries were:

1. **Annual inflation (CPI)** – higher inflation was often linked to banking instability,
2. **Systemic crisis indicator** – if there was already a broader financial crisis, a banking crisis was more likely to follow, and
3. **Lagged banking crisis** – if a crisis happened last year, there's a higher chance it could continue this year.

This shows that even simple models like Logistic Regression can be powerful early warning tools when built with carefully selected economic indicators.

However, our approach has some limitations. First, since banking crises are rare, there's still a risk that our model misses some of them despite good recall. Second, combining all countries into one model might hide country-specific patterns because each country has its own policies and financial systems. Third, our dataset lacked some important features like GDP growth, foreign exchange reserves, and governance quality, which could help the model make better predictions.

In the future, we suggest creating new features that measure volatility in inflation and exchange rates using models like ARCH/GARCH. We also recommend including more economic and political indicators, and trying time-based models like LSTM or Hidden Markov Models (HMMs) that can capture trends over time. Finally, tools like SHAP could help us better understand how the model is making decisions, which is especially useful for policy-making.