Stanford | Pre-Collegiate Summer Institutes

MACHINE LEARNING
SYSTEMIC CRISIS PREDICTION

Andrey Gromyko

introduction

This project leverages a dataset spanning 1870 to 2014, focusing on 13 African countries' financial variables. By employing machine learning techniques, I aim to develop a robust model that **predicts the occurrence of financial crises**. By identifying historical patterns and key indicators, this study contributes to effective policies and interventions, fostering economic stability and sustainable development across the African continent.

INDEPENDENT VARIABLES

- The year of the observation
- The exchange rate of the country vis-a-vis the USD
- Whether domestic debt default occurred in the year
- Whether sovereign external debt default occurred in the year
- The total country's debt
- The annual CPI Inflation rate
- Whether the country is considered independent
- Whether currency crisis occurred in the year
- Whether inflation crisis occurred in the year

DEPENDENT VARIABLES

- Whether a systemic crisis occurred in the year

Since **systemic crisis is the outcome** that I am predicting

methods

First of all, I have randomly split the entire dataset into two - one for training and the other one for testing.

I used both regular **classification methods** and **boosting methods** to predict whether a crisis would occur in an African country.

CLASSIFICATION METHODS

- Classification methods are used to predict the class membership of a data point, in this case whether a country will experience a financial crisis or not.
- Algorithms used: (tuned) decision tree, (tuned) bagging classifier, (tuned) random forest

BOOSTING METHODS

- Boosting methods are a type of ensemble learning algorithm. They work by iteratively training a sequence of models on the same data, with each model being trained to correct the mistakes of the previous models. This process results in a model that is more accurate than any of the individual models.
- Algorithms used: (tuned) ada boost, (tuned) gradient boost, (tuned) xgb boost

data analysis

The data analysis also showed that there was a significant association between some of the factors. For example, year, annual inflation cpi and exchange rate were found to be associated the most with both banking crises and systemic crises.

The findings of this data analysis can be used to **develop machine learning model** that can predict whether a financial crisis will occur in an African country.

PEARSON'S CORRELATION

Pearson's correlation is a measure of the linear relationship between two variables. It can be used to determine whether there is a significant association between two variables, and the direction of the association (positive or negative).

	year	systemic_crisis	exch_usd	domestic_debt	external_debt	gdp_weighted	inflation_annual_cpi	independence	currency_crises	inflation_crises
year	1.000000	0.421927	0.441790	0.189083	0.531304	0.401267	0.000000	0.812338	0.327599	0.379579
systemic_crisis	0.421927	1.000000	0.392061	0.170234	0.372682	0.048303	0.059151	0.217482	0.071946	0.255574
exch_usd	0.441790	0.392061	1.000000	0.152233	0.586554	0.000000	0.000000	0.164333	0.000000	0.058890
domestic_debt	0.189083	0.170234	0.152233	1.000000	0.658024	0.000000	0.103045	0.154396	0.142807	0.331542
external_debt	0.531304	0.372682	0.586554	0.658024	1.000000	0.326063	0.000000	0.343229	0.120118	0.281289
gdp_weighted	0.401267	0.048303	0.000000	0.000000	0.326063	1.000000	0.000000	0.055706	0.175819	0.128280
inflation_annual_cpi	0.000000	0.059151	0.000000	0.103045	0.000000	0.000000	1.000000	0.000000	0.041555	0.022690
independence	0.812338	0.217482	0.164333	0.154396	0.343229	0.055706	0.000000	1.000000	0.045114	0.000000
currency_crises	0.327599	0.071946	0.000000	0.142807	0.120118	0.175819	0.041555	0.045114	1.000000	0.243260
inflation_crises	0.379579	0.255574	0.058890	0.331542	0.281289	0.128280	0.022690	0.000000	0.243260	1.000000

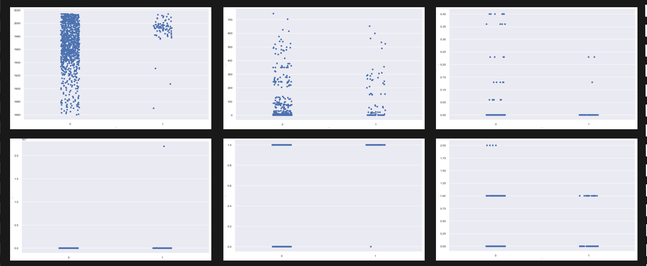
PHI-K CORRELATION

Phi K correlation is a non-parametric measure of association that can be used with categorical data. It is similar to Pearson's correlation, but it does not require the assumption that the data is normally distributed.

	year	systemic_crisis	exch_usd	domestic_debt	external_debt	gdp_weighted	inflation_annual_cpi	independence	currency_crises	inflation_crises
year	1.000000	0.421927	0.441790	0.189083	0.531304	0.401267	0.000000	0.812338	0.327599	0.379579
systemic_crisis	0.421927	1.000000	0.392061	0.170234	0.372682	0.048303	0.059151	0.217482	0.071946	0.255574
exch_usd	0.441790	0.392061	1.000000	0.152233	0.586554	0.000000	0.000000	0.164333	0.000000	0.058890
domestic_debt	0.189083	0.170234	0.152233	1.000000	0.658024	0.000000	0.103045	0.154396	0.142807	0.331542
external_debt	0.531304	0.372682	0.586554	0.658024	1.000000	0.326063	0.000000	0.343229	0.120118	0.281289
gdp_weighted	0.401267	0.048303	0.000000	0.000000	0.326063	1.000000	0.000000	0.055706	0.175819	0.128280
inflation_annual_cpi	0.000000	0.059151	0.000000	0.103045	0.000000	0.000000	1.000000	0.000000	0.041555	0.022690
independence	0.812338	0.217482	0.164333	0.154396	0.343229	0.055706	0.000000	1.000000	0.045114	0.000000
currency_crises	0.327599	0.071946	0.000000	0.142807	0.120118	0.175819	0.041555	0.045114	1.000000	0.243260
inflation_crises	0.379579	0.255574	0.058890	0.331542	0.281289	0.128280	0.022690	0.000000	0.243260	1.000000

BIVARIATE ANALYSES

Bivariate analyses are statistical techniques that examine the relationship between two variables. They can be used to determine whether there is a significant association between two variables, and the strength of the association.



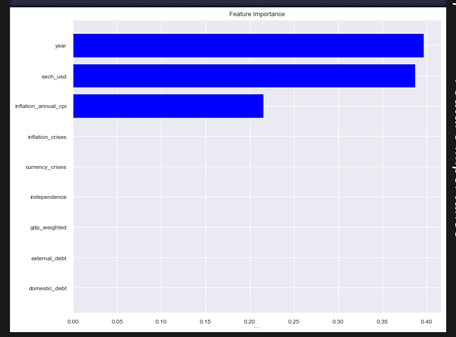
results

Overfitting - is a phenomenon where a model learns the training data too well, resulting in poor performance on new data. This happens when the model learns the noise and irrelevant details in the training data, rather than the underlying patterns. As a result, the model is not able to generalize to new data that it has not seen before. So, we have to ignore methods that have extremely high score on Train Accuracy test (when it is more than 0.99)

CLASSIFICATION METHODS

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	F1_Train	F1_Test
1	Tuned Decision Tree	0.944669	0.927673	0.735849	0.620690	0.590909	0.600000	0.655462	0.610169
4	Bagging Classifier	0.995951	0.937107	0.943396	0.482759	1.000000	0.736842	0.970874	0.583333
3	Tuned Random Forest	0.994602	0.933962	0.924528	0.448276	1.000000	0.722222	0.960784	0.553191
0	Decision Tree	1.000000	0.924528	1.000000	0.448276	1.000000	0.619048	1.000000	0.520000
5	Tuned Bagging Classifier	0.995951	0.930818	0.943396	0.379310	1.000000	0.733333	0.970874	0.500000
2	Random Forest	1.000000	0.930818	1.000000	0.344828	1.000000	0.769231	1.000000	0.476190

Models comparison



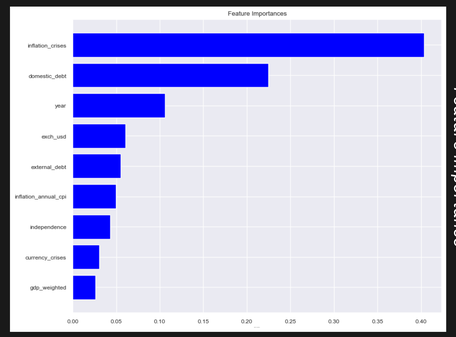
Feature importance

Based on F1 test, the best not overfitting method is **Tuned Decision Tree** scoring 0.61

BOOSTING METHODS

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1	Test_F1
4	XGBoost	1.00	0.95	1.00	0.55	1.00	0.84	1.00	0.67
5	XGBoost Tuned	0.99	0.93	1.00	0.59	0.93	0.63	0.96	0.61
1	AdaBoost Tuned	0.94	0.93	0.30	0.34	0.76	0.71	0.43	0.47
2	Gradient Boost	0.99	0.93	0.85	0.31	1.00	0.82	0.92	0.45
3	Gradient Boost Tuned	1.00	0.93	0.98	0.31	1.00	0.75	0.99	0.44
0	AdaBoost	0.95	0.92	0.38	0.24	0.77	0.70	0.51	0.36

Models comparison



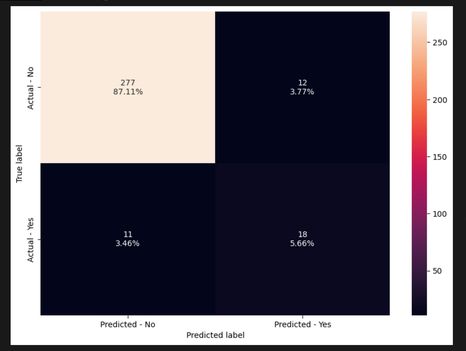
Feature importance

Based on F1 test, the best not overfitting method is **Tuned AdaBoost** scoring 0.47

conclusion

On one hand false positive prediction could lead to panic among citizens, taking their money from the banks and only increasing the chances of crisis. On the other hand, false negative prediction leads to lack of preparation of people and companies with harsher crisis consequences for everyone. So, I want to make balanced model and **F1 test will be the most accurate in this case**.

If we ignore over fitting methods (that scored >0.99 on Train Accuracy test), **the best method for predicting crisis in this dataset is Tuned Decision Tree**, scoring 0.61 on F1 test.



Tuned Decision Tree confusion matrix

POSSIBLE IMPROVEMENTS

While the tuned tree classifier was the best model in this study, there are still some areas that could be improved to make predictions even more accurate

- A **larger dataset** with more countries included would likely allow the model to learn more about the different factors that contribute to crises and would therefore improve its accuracy.
- Another area that could be improved is the **selection of features**. There are other features that could be used to improve the accuracy of the model. For example, features related to political stability or social unrest could be included in the model.
- **More classification methods** could also significantly benefit my result since, as we could see, even the best methods weren't using all of the features while some of them could have been correlating with the outcome in a more complicated way.

references

- https://scikit-learn.org/stable/index.html
- https://www.kaggle.com/datasets/chirin/africa-economic-banking-and-systemic-crisis-data
- https://spcs-programs.instructure.com
- https://github.com/msNPS/ML-Crysis-Prediction