Process of building the model:

* Retrieve the data from IMF / Data Bank.
* To handle the missing data, the blank slot will be estimated by calculating the average number of 10 slots ahead.
* Real GDP is calculated by GDP/GDP Deflator.
* For missing value in the field of total debt and short-term debt (HK, SGP & KR), used 0 to replace the null value. (This might affect the accuracy for these 2 column)

Code (Steps):

* Reading data
* Split the dataset
* Run the model
* Tune the parameter & Cross validation
* Output Result

Current Result:

* The machine learning model provides an unsatisfactory result using the annual data from IMF & Data Bank.
* The data is not granular enough and the quality is bad and fragmented.
* Lack of data causes the model couldn't be trained accurately.
* 3 sets of results (Method):

1. Feature scaling per country
2. Feature scaling all countries together
3. Exclude most of the rows that included missing values

* Overall, the accuracy is high but recall is too low, which means guessing every row is false can achieve a high rate of accuracy.

|  | Method 1 | Method 2 | Method 3 |
| --- | --- | --- | --- |
| Best Performance (Recall) | SVM | N/A | KNN |

* Based on the idea of more “Hit” is better than accuracy, method 3 with the KNN model is more suitable for us.

Output:

| Model | Feature scaling by country | Feature scaling all in a single file |  |
| --- | --- | --- | --- |
| SVM |  |  |  |
| Decision Tree |  |  |  |
| NB |  |  |  |
| MLP |  |  |  |
| KNN |  |  |  |
| Adaboots |  |  |  |

Future Development

* Output the graph to visualize the result
* Maybe the real GDP can be replaced by real GDP per capita to increase the accuracy.
* Apply the undersampling technique to duplicate the “Hit” record so it has more balanced data to train the model, thus increasing the hit rate.
* Develop an API tool to retrieve all the data from the web, and increase the number of countries to be tested.
* In the future, we can use models like LADtree and DNDT to test.