GDP growth prediction

Matej Sebenik

2022-08-07

1. Introduction:
The goal of this paper is the creation of a predictive algorithm, based on the economic data form the World Bank, that will predict if the economy of any given country any given year is in a downturn (defined as having negative annual GDP growth). The metric used to evaluate the algorithm is accuracy. In absence of official accuracy targets, a success rate of at least 75% will be considered fine, and a success rate of 90% or more will be considered excellent.

The data set is available at **this link**.

Key steps in the calculation of the predictive algorithm are as follows:
1. Data set download, preparation and data cleaning.
2. Data visualizations.
3. Algorithm training and verification.

The data set, after all the relevant data manipulation, is previewed below:

```
##   country year GDP_growth_annual.x economic_downturn unemployment_rate.x
## 1 Albania 1996            9.099999                 0              13.931
## 2 Albania 1997          -10.919984                 1              16.876
## 3 Albania 1998            8.829424                 0              20.047
## 4 Albania 2012            1.417243                 0              13.380
## 5 Albania 2013            1.002018                 0              15.870
## 6 Albania 2014            1.774449                 0              18.050
##   unemployment_rate.y GDP_growth_annual.y inflation_GDP_deflator.x
## 1              14.611           13.322333                38.172058
## 2              13.931            9.099999                11.239644
## 3              16.876          -10.919984                 6.730860
## 4              13.480            2.545406                 1.042715
## 5              13.380            1.417243                 0.288746
## 6              15.870            1.002018                 1.549917
##   inflation_GDP_deflator.y inflation.x inflation.y govt_debt_share_of_gdp.x
## 1                 9.970663   12.725478    7.793219                 37.48106
## 2                38.172058   33.180274   12.725478                 53.10782
## 3                11.239644   20.642859   33.180274                 55.56570
## 4                 2.314744    2.031593    3.429123                 63.66915
## 5                 1.042715    1.937621    2.031593                 70.58077
## 6                 0.288746    1.625865    1.937621                 73.32023
##   govt_debt_share_of_gdp.y gross_savings_share_of_gdp.x
## 1                 35.75690                    16.132506
## 2                 37.48106                     5.572394
## 3                 53.10782                    17.337436
## 4                 69.63767                    19.625104
## 5                 63.66915                    17.741294
## 6                 70.58077                    15.929387
##   gross_savings_share_of_gdp.y total_reserves.x total_reserves.y
## 1                    16.357511        323376839        265298058
## 2                    16.132506        342425753        323376839
## 3                     5.572394        417921457        342425753
## 4                    20.545456       2599863597       2471402725
## 5                    19.625104       2773278107       2599863597
## 6                    17.741294       2665215805       2773278107
##   current_account_balance.x current_account_balance.y external_debt_stocks.x
## 1                -107300000                 -11500000              491994449
## 2                -272232500                -107300000              516476961
## 3                 -65070000                -272232500              623592792
## 4               -1256644800               -1667175109             7384500627
## 5               -1184891052               -1256644800             8647043654
## 6               -1425385680               -1184891052             8512452310
##   external_debt_stocks.y country_number
## 1              458862738             78
## 2              491994449             78
## 3              516476961             78
## 4             6484194296             78
## 5             7384500627             78
## 6             8647043654             78
```

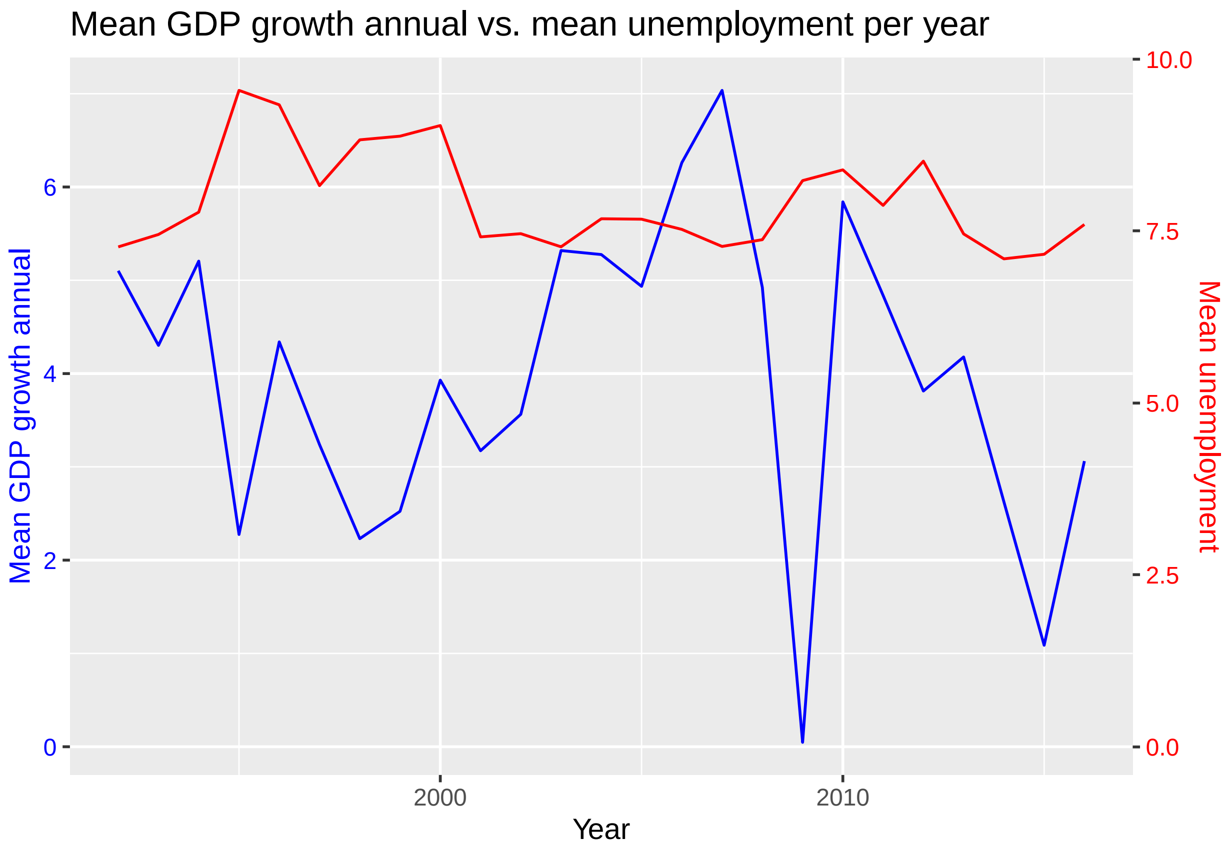
2. Data description:

The data set is divided into 22 columns. These contain:
1. The name of the country.
2. The year of the value.
3. GDP_growth_annual.x contains the annual GDP growth for that year, in percent.
4. Economic_downturn is the previous column, recoded so that any negative annual GDP growth returns 1, and stagnation or positive growth returns 0. This column is necessary for the classification part of the algorithm development.
5. Unemployment_rate.x returns that year's unemployment rate, in percent.
6. Unemployment_rate.y returns the previous year's unemployment rate, in percent.
7. GDP_growth_annual.y returns the previous year's annual GDP growth, in percent. This value is not actually used anywhere, due to very high correlation with the primary annual GDP growth.
8. Inflation_GDP_deflator.x returns the inflation for the entire economy, for that year, in percent.
9. Inflation_GDP_deflator.y returns the inflation for the entire economy, for previous year, in percent.
10. Inflation.x returns the inflation for a fixed basket of goods, for that year, in percent.
11. Inflation.y returns the inflation for a fixed basket of goods, for previous year, in percent.
12. Govt_debt_share_of_gdp.x returns the size of government's debt, for that year, in percent.
13. Govt_debt_share_of_gdp.y returns the size of government's debt, for previous year, in percent.
14. Gross_savings_share_of_gdp.x returns the size of savings in a country, for that year, in percent.
15. Gross_savings_share_of_gdp.y returns the size of savings in a country, for previous year, in percent.
16. Total_reserves.x returns the size of government's reserves, for that year, in USD.
17. Total_reserves.y returns the size of government's reserves, for previous year, in USD.
18. Current_account_balance.x returns the balance of a country's cash outflows/inflows, for that year, in USD.
19. Current_account_balance.y returns the balance of a country's cash outflows/inflows, for previous year, in USD.
20. External_debt_stocks.x returns the amount owned to foreign subjects, for that year, in USD.
21. External_debt_stocks.y returns the amount owned to foreign subjects, for previous year, in USD.
22. Country_number is a unique numeric identifier of any given country.

The data set has 507 entries/rows.

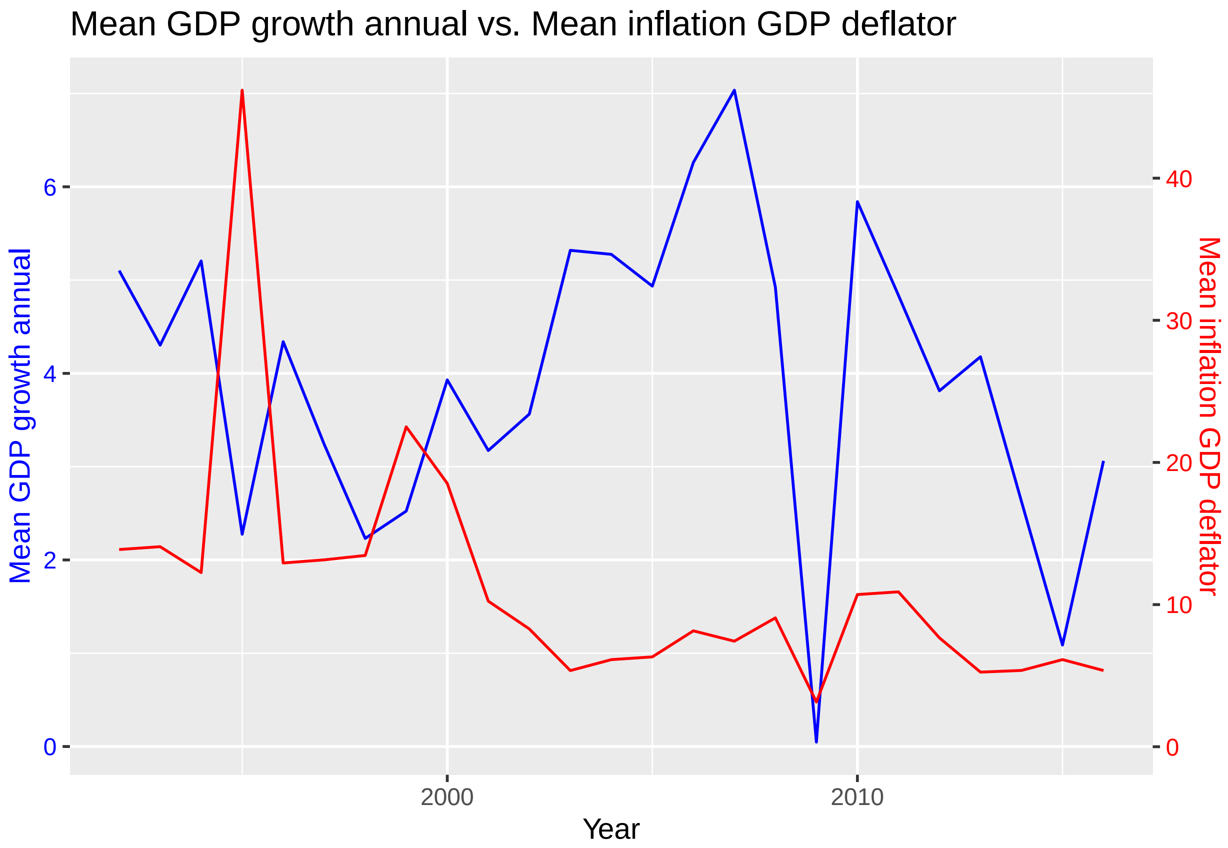
3. Data visualizations:
Several data visualizations were prepared, with the aim of getting acquainted with the data. These visualizations are provided below, along with the relevant descriptions and insights.

1. GDP growth annual vs. unemployment rate:



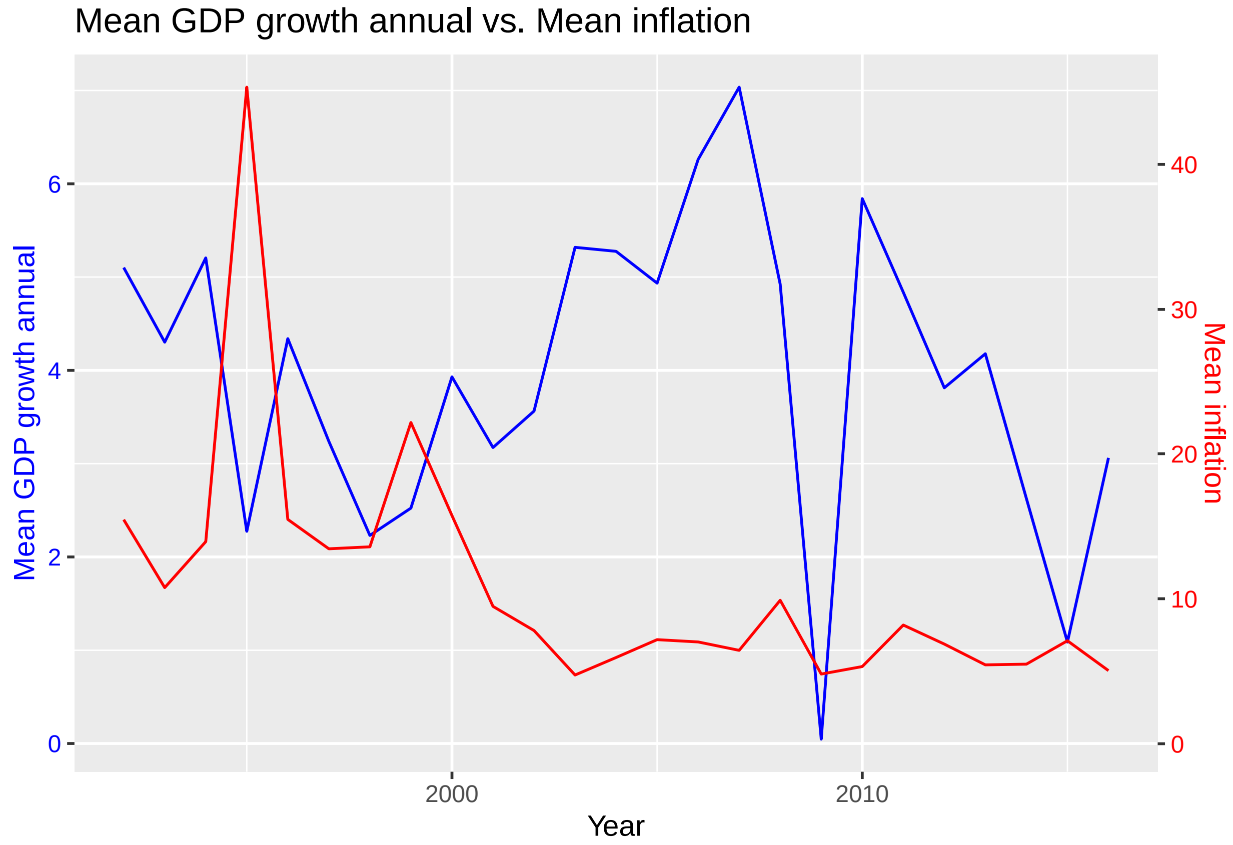
This graph shows the value of annual GDP growth vs. unemployment rate. There appears a relevant correlation between the two values, with the unemployment rising when GDP growth is low or negative, though this trend is somewhat spoiled from about 2003 on, where the values appear to move in tandem.

2. GDP growth annual vs. inflation GDP deflator:



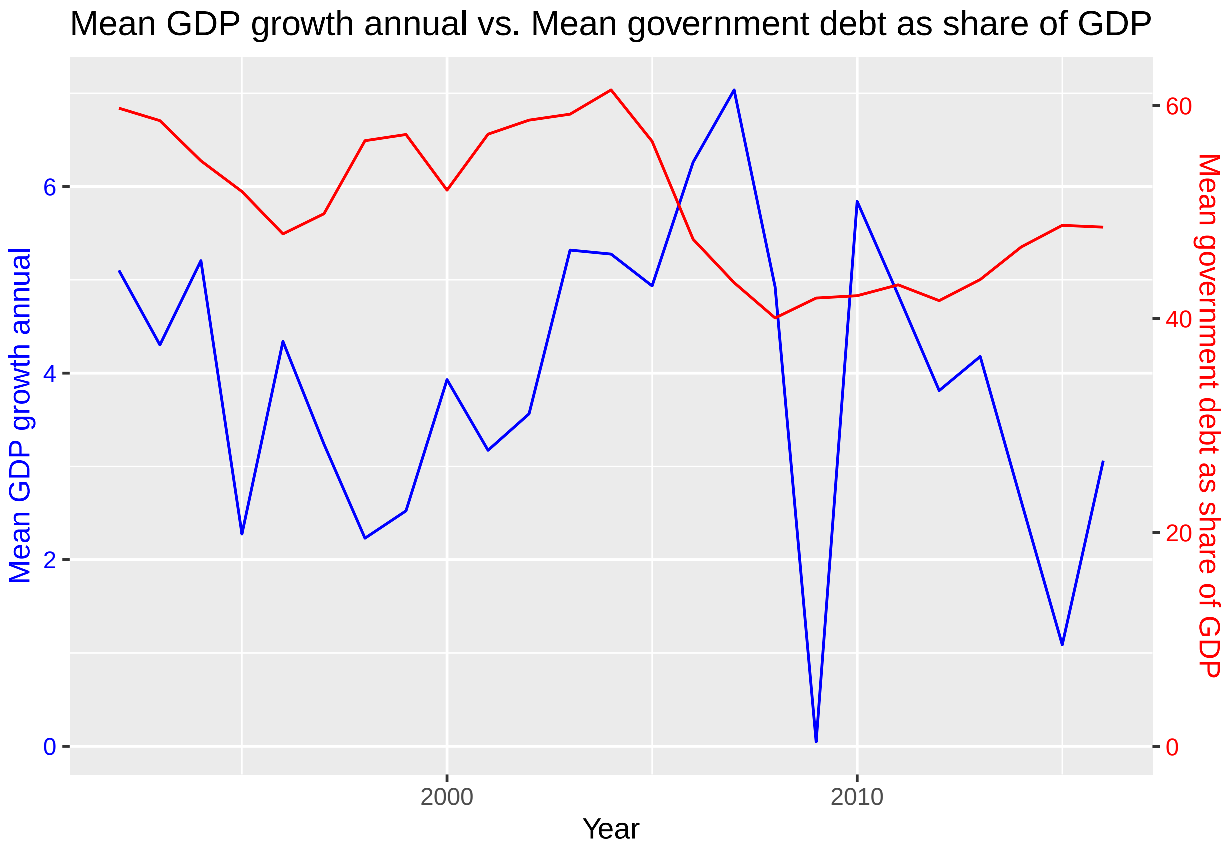
This graph shows the value of annual GDP growth vs. inflation GDP deflator. The values again appear to move opposite to one another, with high inflation GDP deflator associated with low GDP growth. As with the previous graph, this relationship appears to break down with the economic crisis of the late 2010s, and is only reestablished with the last data points, for 2016.

3. GDP growth annual vs. inflation:



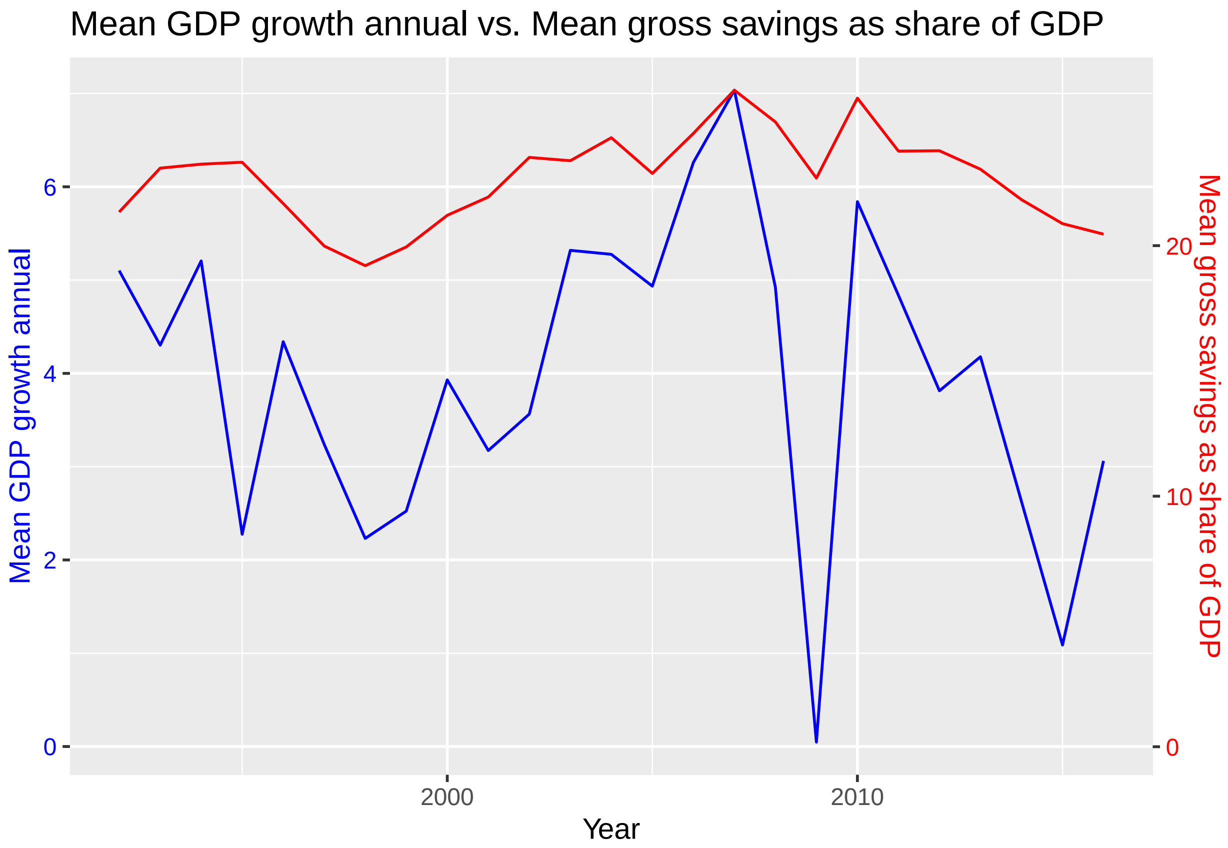
This graph shows the value of annual GDP growth vs. inflation. This graph is very similar to the previous one, since the inflation and inflation GDP deflator are closely related to one another. Again we see the inverse movement of the two values up to 2009, with the inversion dissapearing for the following several years.

4. GDP growth annual vs. government debt as share of GDP:



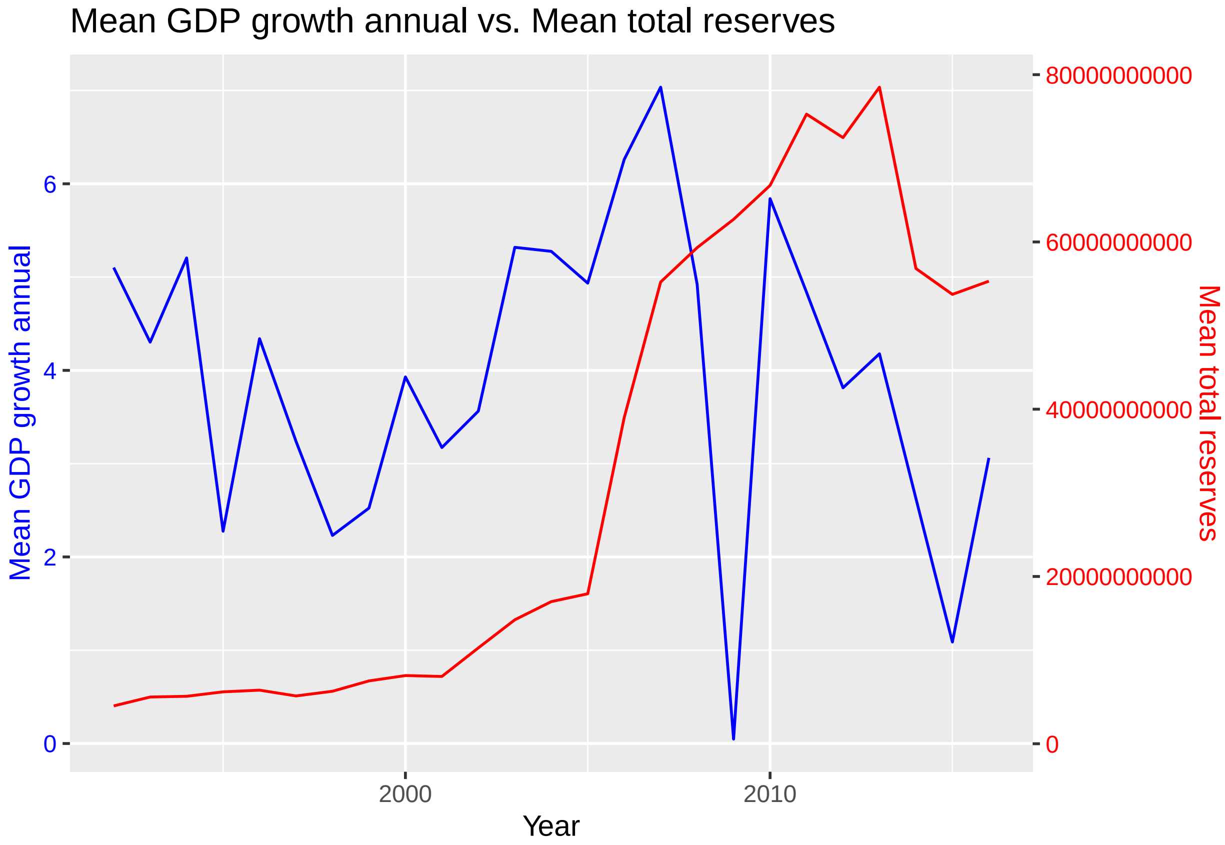
This graph shows the value of annual GDP growth vs. government debt as share of GDP. The two values on display generally mirror each other, which is odd (low GDP growth should cause grater levels of government debt?). Just before the crisis the two lines diverge, which makes sense. Oddly, during the crisis, debt increases only marginally, and then slowly rises through the uneven economic recovery following.

5. GDP growth annual vs. savings as share of GDP:



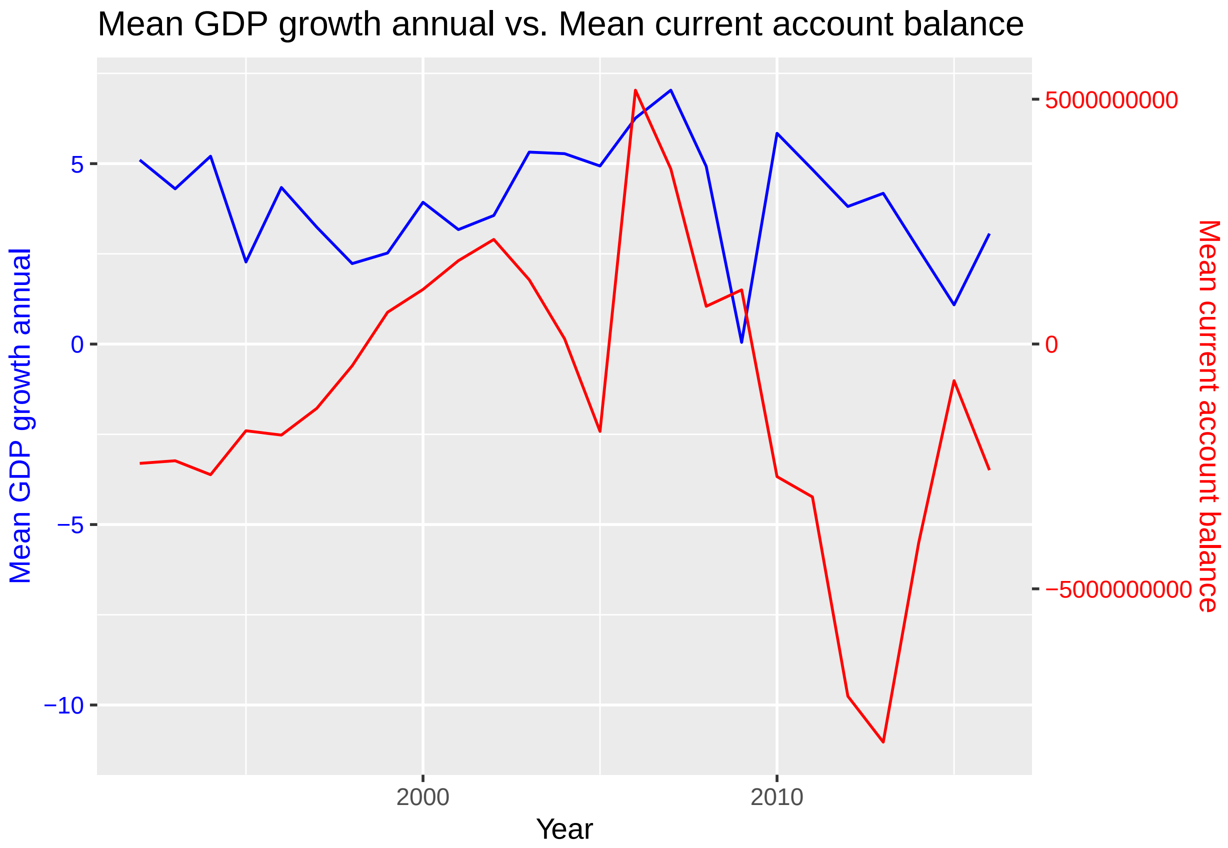
This graph shows the value of annual GDP growth vs. savings as share of GDP. The two values neatly follow each other. When GDP grows, so do the savings in a society, and vice versa.

6. GDP growth annual vs. total reserves:



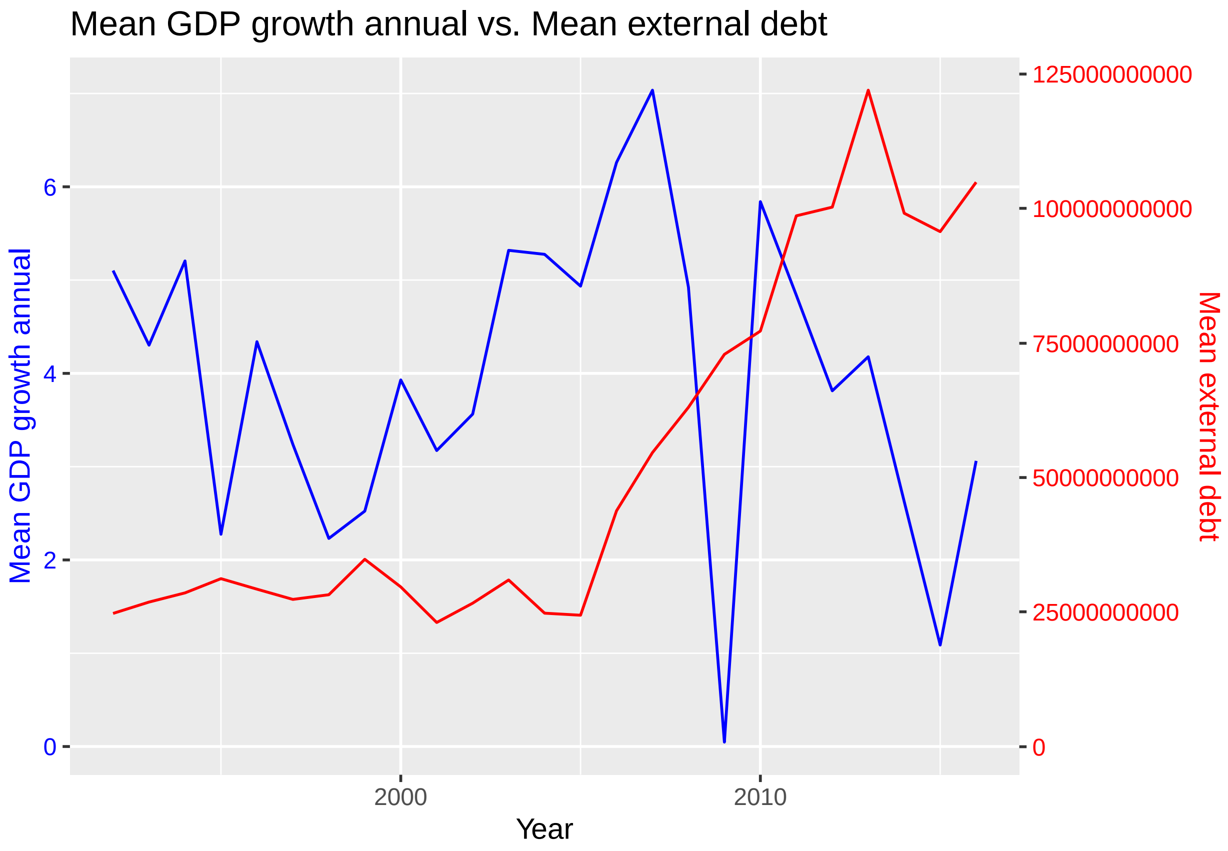
This graph shows the value of annual GDP growth vs. total reserves. The values in this graph are confusing, as they indicate a massive jump in the absolute value of reserves just before the crisis. In fact, this is due to the fact that the values for several large economies (Russian Federation, Indonesia, Brasil) only appear during this period, massively increasing the mean values displayed in the graph. After the crisis, the values behave more logically, falling with the reduced economic activity.

7. GDP growth annual vs. current account balance:



This graph shows the value of annual GDP growth vs. current account balance. Due to the nature of the current account balance, a correlation between the two values is not to be expected. Current account balance is a measure of the flow of money into or out of a given country. If we had the perfect data for the entire planet, the mean value of the current account balance would always be zero, as each dollar that goes out of one country enters another country. Given that we do not actually have perfect data, an in depth analysis shows that the appearance of the Russian Federation hugely boosts the value before the crisis (the country being a massive exporter of oil and gas). Afterwards, the large importers Brasil and India drag the value down.

8. GDP growth annual vs. external debt:



This graph shows the value of annual GDP growth vs. external debt. Similarly to previous entries, this graph reflects the inclusion of the Russian Federation, Brasil and Indonesia in the mid 2010s. Even without these three outliers, the debt still increases with years.

4. Modelling approach and results:

4.1. Modelling approach:

The goal of the algorithm is to predict if the economy of a country in a year is in a downturn, defined as a negative annual GDP growth. The metric used to evaluate regression models, RMSE, will be transformed into accuracy, so that a comparison with the classification methods can be implemented.

The most popular regression types were tested, these being the general linear model, localized regression and the k nearest neighbor method. Additionally, from the family of classifications, lda, decision tree and random forest algorithms were tested. Qda algorithm was discarded due to various errors. All the results were imputed into a single overview table.

The data set was not further divided into training and test sets. Instead, cross validation was used for initial model evaluation. Spans and degrees were not regulated.

The column GDP_growth_annual.y (which means previous year's GDP annual growth) was not used in algorithms, since it is too closely related to current year's GDP annual growth.

4.2. Results:

```
##                  algorithm_name  initial_metric test_value validation_value
## glm_results      "glm"           "RMSE"         3.40895    3.09823
## knn_results      "knn"           "RMSE"         3.68924    3.28105
## gamLoess_results "gamLoess"      "RMSE"         2.75386    2.88922
## lda_results      "lda"           "accuracy"     0.90449    0.88235
## rpart_results    "decision tree" "accuracy"     "0.89696"  "0.84314"
## rf_results       "random forest" "accuracy"     "0.91011"  "0.84314"
##                  validation_accuracy
## glm_results      0.84314
## knn_results      0.82353
## gamLoess_results 0.82353
## lda_results      0.88235
## rpart_results    "0.84314"
## rf_results       "0.84314"
```

The validation (final) accuracy shows that all the algorithms achieve the accuracy above 75 %, and thus can be considered fine. None are better than 90 %, and so none can be considered excellent. The best validation accuracy is returned by the lda method (**88.235 %**).

5. Conclusion: This report presented the development and results of the algorithm used to predict the economic performance of countries based on the World Bank's economic data set. The best results were produced using the lda method of classification. The results are, at a first glance, fine, although they do raise several relevant questions and offer additional improvement possibilities. These are stated below.

1. The final data set is quite small. It begins at **16492**, but after removing all the empty values, ends at **507**. This limits the quality of the algorithms produced, possibility of additional/advanced methods used etc. It also removes most of the developed world from the data set, leaving mostly developing states. Thus, if certain columns currently in use would be dropped or the limitations of the specific models regarding non existent data be otherwise circumvented, more data could be made available, improving the models.
2. Referring to point 1 above, hyperparameters could be used to refine individual models. Additionaly, an ensemble of the models could be created, further increasing the accuracy of the overall model.
3. As seen in the data visualizations, certain states only appear in the middle of the data interval used, which can skew the algorithms. These states or the problematic economic indicators should potentionally be excluded.
4. The last three economic indicators (reserves, current account balance and external debt) are all in absolute dollar terms. Perhaps these should be transformed into the share in GDP, to maintain data cohesiveness (other indicators are in shares, not absolute numbers).
5. Only one inflation indicator should be kept?
6. Recode states into developed, medium, developing?
7. Eventually, the data source could be changed, so that more granular data are obtained (recessions are generally based on quarterly data, not yearly data). This would entail massive changes to the foundations of this model. In this eventuality, only data from previous quarter could be used (predicting downturns based on concurrent data is not as useful as being able to predict them in advance).
8. Aesthetically, all the gdp should either be upper or lower case. Also, why are the last two model results (decision tree and random forest) in the table 'results' in quotes, preventing any calculations with them?

Best regard,

Matej Sebenik
Ljubljana
Slovenia
EU