Using World Bank Data to Understand Working of Economies

*Congrui Li a, Nidhi Rastogi a, Pu Zhang b*

*aDepartment of Computer Science, Rensselaer Polytechnic Institute*

*bDepartment of Industrial and Systems Engineering, Rensselaer Polytechnic Institute*

*Author Email:* [*lic10@rpi.edu*](mailto:lic10@rpi.edu)*,* [*raston@rpi.edu*](mailto:raston@rpi.edu)*,* [*zhangp5@rpi.edu*](mailto:zhangp5@rpi.edu)

Gross domestic product (GDP) is a monetary measure of the value of all final goods and services produced in a period (quarterly or yearly). Nominal GDP estimates are commonly used to determine the economic performance of a whole country or region, and to make international comparisons. [1] A lot of factors may effect a country’s GDP, including but not limited to governance, population, labor, health of people, agriculture and so on.

The World Bank has free and open access to data about development in countries around the globe. It is a perfect data source for us to investigate the correlation of a country’s economy and other various development factors. We would also like to find out how of the effect of each factor to the economy.

We chose GDP per capita as our dependent variable. Other 16 development indicators were selected as our independent variables, which are listed below.

Subjective governance factors [2]:

* Control of Corruption
* Government Effectiveness
* Political Stability
* Regulatory Quality
* Rule of Law
* Voice and Accountability

Objective statistical factors [3]:

* Arable land (hectares per person)
* Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate)
* Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)
* Life expectancy at birth, female (years)
* Life expectancy at birth, male (years)
* Population growth (annual %)
* Population, total
* Adolescent fertility rate (births per 1,000 women ages 15-19)
* Proportion of female in national parliaments (%)
* Urban population (% of total)

All of these indicators along with GDP per capita are available at the World Bank website for a certain country at a certain year. We chose these indicators for the following reasons:

1. These indicators cover a wide range of aspects including governance, population, labor, health of people, and agriculture.
2. Some special indicators related to female living conditions are also included, such as adolescent fertility rate, proportion of female in national parliaments.
3. When similar indicators are available, we selected the one with the most complete data. (For example, besides labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate), labor force participation rate, male (% of male population ages 15-64) (modeled ILO estimate) and labor force participation rate, male (% of male population ages 15+) (national estimate) are also available.)

Eventually, we used the data from 200 countries over 10 years, more specifically, from 2005 to 2014, since these are the most complete data available. This gives us approximately 2000 data points to be used. Detailed data preprocessing procedures we went through will be available in the actual paper, including the normalization of each factors, the handling of missing values and so on.

Therefore, our investigation turned into a regression problem with GDP per capita as the dependent variable and the other 16 selected indicators as the independent variables.

Going into the data mining part, we applied 7 different supervised learning models for the regression analysis:

* Linear regression with Lasso (Lasso regression)
* Support vector regression (SVR)
* Regression tree
* Conditional inference tree (Ctree)
* Random forest
* Conditional inference random forest (Cforest)
* eXtreme gradient boosting (XGBoost)

The elaborated parameter selection criteria for each model will be available in the actual paper. Cross validation was also conducted for the models. The adjusted R-squared is used as the metric of a model’s accuracy. Comparison is shown in Figure 1 below.

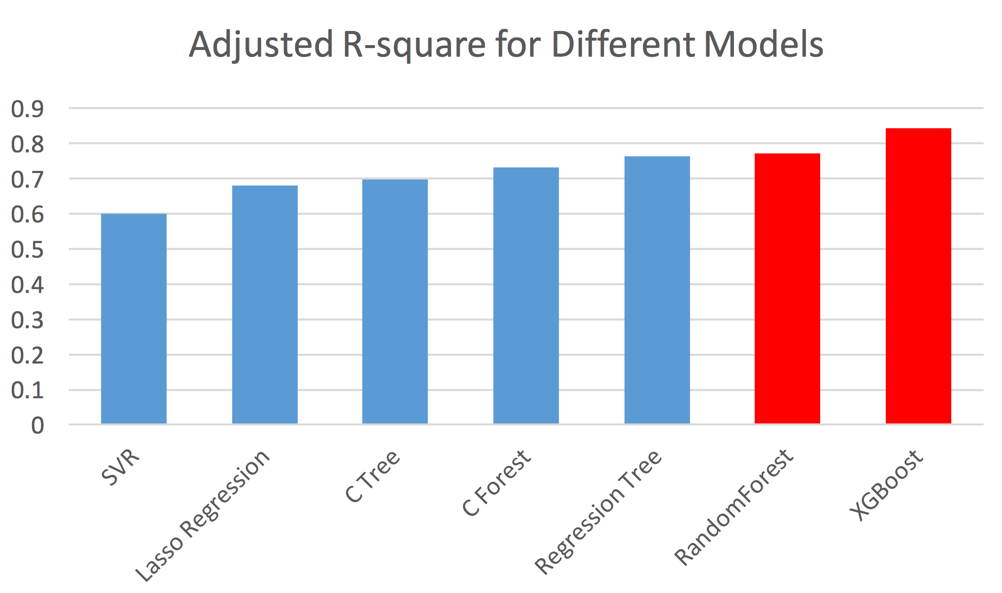


Figure 1: Comparison of adjusted R-squared among different models

For each model, we measures of importance of each factors. The overall importance of the factors are aggregated by the importance given by different models, weighted with the accuracy of each model. Comparison chart of the importance score for each factor is shown in Figure 2.

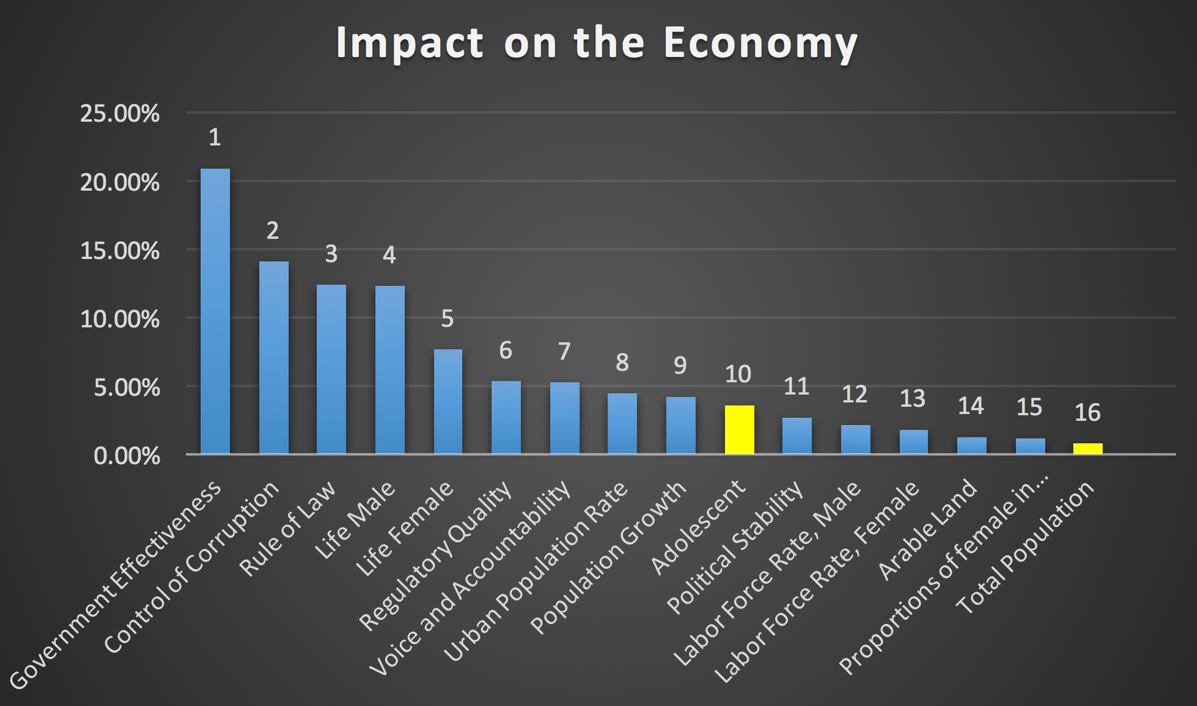


Figure 2: Ranking of the importance score for each factor

In Figure 2, the blue colored factors are the positively correlated factors while the yellow ones are negatively correlated. The Y-axis means how many percent of each factors contribute to the overall correlation.

There are several reasons we did not use principal component analysis (PCA) before applying our models to the data:

1. PCA could lead to loss of interpretability.
2. Our models already have the feature selection functionality since they all provide measurements for the importance of each feature.
3. PCA might be inappropriate for features in different scales, especially when there are both subjective scores and objective data in different units.

After the data analysis, we selected the following 12 countries as: Australia, Brazil, China, Egypt, Ethiopia, Germany, India, Japan, Nigeria, Russia, Sweden, and United States for empirical study. These 12 countries roughly cover 40% of the area of all countries, produce 60% of the globe GDP, and have 60% of the global population. They also show the diversity regarding economic development, geographical location among themselves. In other words, these 12 countries would be a very good representation of the world.

According to our analysis, the top three important subjective governance factors are: 1. Government Effectiveness; 2. Control of Corruption; and 3. Rule of Law. The The top three important objective statistical factors are: 1. Life expectancy at birth, male (years); 2. Life expectancy at birth, female (years); 3. Urban population (% of total).

To show the correlation, motion chart visualizations of dependent variable versus the top three important subjective governance factors as well as dependent variable versus the top three important objective statistical factors were created. Once again, the data used are for the selected 12 countries over 10 years (from 2005 to 2014). Visualizations are available through the following URL:

<https://goo.gl/yLSmmX>

<https://goo.gl/3vn9KU>

Multiple conclusions could be drawn from the motion charts. For example, if we set Government Effectiveness as the X-axis and GDP per capita as Y-axis, not matter how the 12 data points (which stand for 12 countries) move on the chart over the 10 years, they will always stay in two clusters, which are high government effectiveness, high GDP per capita cluster as well as low government effectiveness, low GDP per capita cluster. This means government effectiveness does have strong positive correlation with GDP per capita, which approves our previous analysis. More detailed conclusions will be shown in the paper.

As some side work, we also used ARIMA (0, 2, 2) model for time series analysis. Based on the data in the past 34 years (from 1981 to 2014), we predicted the total GDP and GDP per capita for the 12 countries until the years of 2025. The motion chart visualization of the analysis result is available through this URL: <https://goo.gl/BOsZsD>.

The outline presentation of our work is also available through this URL:

Conclusions are listed below:

* Overall, governance factors have big correlation than other statistical factors on a country’s economy.
* Population growth has positive correlation with GDP per capita; while total population has negative correlation with GDP per capita.
* XGBoost and Random Forest have the best fit on the data among all the models we used.

Our future work includes the followings:

* Using other economic indicators along with GDP per capita as a better representation of a country’s economy to serve as our dependent variable.
* Include more development indicators as independent variables in our investigation, such as indicators related to environment, science & technology, and maybe even refugee.
* Refine the regression models by using different parameters and/or parameter values to get better accuracy; also consider other kinds of regression models for this purpose.
* Improvement the metric of model accuracy; also improvement the metric of importance of each factor.
* Discover more interesting stories using the huge amount of data in the World Bank.
* Conduct unsupervised learning analysis for the data, such as K-means clustering, to discover new findings and stories.
* Use other techniques and tools to do better visualizations of the data.

### References

[1] <https://en.wikipedia.org/wiki/Gross_domestic_product>

[2] data available through this URL: [http://databank.worldbank.org/data/reports.aspx?source=worldwide-governance-indicators#](http://databank.worldbank.org/data/reports.aspx?source=worldwide-governance-indicators)

[3] data available through this URL: <http://data.worldbank.org/indicator/all>