Project Report

JyunYu Cheng; Yiji Gao

5/8/2022

Abstract

In this project, we focus on the question of how to classify countries. We use the methods of model selection, PCA and cluster. Bagging model is the best model to make prediction under this circumstance. Through PCA and cluster, all countries are divided into 4 categories. They are very similar to the classification in the real world. So, our conclusion is that our variables are chosen meaningfully.

Introduction

At reality, countries are divided into several categories, including developed countries, countries to be consider developed and developing countries. Different institutions make use of different standards to make the judgement. For example, UNDP adopt the data of "human development" index while the World Bank focus on the income. Different standard may cause confusion. As a result, we collect the data from all the countries (few of them are excluded due to the missing of data) in 2018, including all aspects of a citizen life. Ranging from macroeconomics to microeconomics, GDP index, food and education are all considered.

In conclusion, we will compare between the 2018 data and the categories in the reality. We will focus on the difference and analyze them.

Methods

At first, we get the data from the UN websites. There are 14 parameters describing each country. Almost every country is included, except for some that lacks too many data.

Secondly, we use some basic data verbs like summarize, mutate, select to wrangle the data. During the process, we can achieve many basic understandings of the data. For example, mean and standard deviation contributes to the rough situation in general.

Furthermore, we take advantage of different models to make perdition.

Finally, PCA and cluster are put into practice. We use the collected data to divide all countries into 4 categories. Comparison are made between various times using our index and the reality. Maps are made to make it more precisely.

Part 1 Wrangle the data and create a treemap with labels

During the process in this part, we discover that the standard deviation of GDP is very high, which means that GDP is very different between countries. The second picture indicates that even in the same category according to the reality, the GDP amount differs a lot. It also shows that the GDP is not the only incidator to determine if the country is a developed or developing country.

Part 2 : Predictive Model Building

In the real world, there are too many factors that may affect one country's GDP, such as population, monetary policy and industry structure. As a result, we want to use the data we collect and analyze it, then try to predict the GDP by other factors, then compare the predict result with the real situation. At this part, we want to use the variables in the data to predict the GDP of each country,for example, food production index, foreign direct investment and unemployment rate etc. We will use four models : Random forest model, Bagging model, CART model and Boosting model to do the prediction.Then we will assess which model is better by the RMSE of each model. After deciding the model which we want to use, we will use it to do the prediction and plot the figure to compare the real situation and the predictive situation. In the beginning, we have to clean to data, delete some wrong values and some information we don't need. Secondly, we will separate the data to training data and testing data, because the data we collected can't all be used to train, we have to remain some data to test, so that we can assess the performance of the model. At this step, we decide to let 80% of the data be the training set, and the remaining 20% be the testing set. Step description: Step1. Build four models Step2. Model selection(compare the RMSE of every model and choose the best model) Step3. Conclusion

Step 1 : Build the model

```
## Distribution not specified, assuming gaussian ...
```

Step2: Calculate the RMES of each model

RMSE_Random_forest

```
## [1] 0.8911728
```

RMSE_Bagging

```
## [1] 0.2261014
```

RMSE_CART

```
## [1] 0.9252854
```

RMSE_Boosting

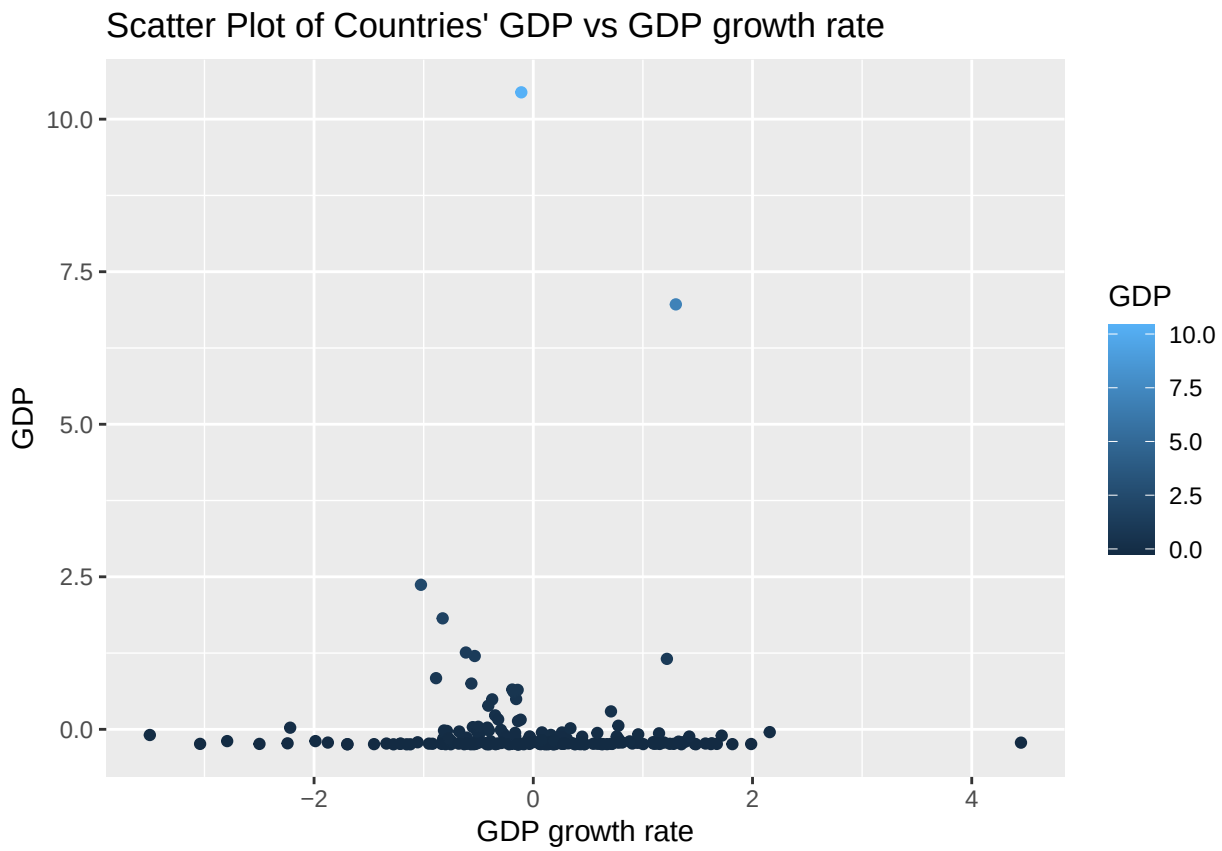
```
## [1] 1.074493
```

From the above numbers, we can see that Bagging model has the smallest RMSE of each models,so we decide to choose Bagging model as the best predictive model and use it to do the prediction.

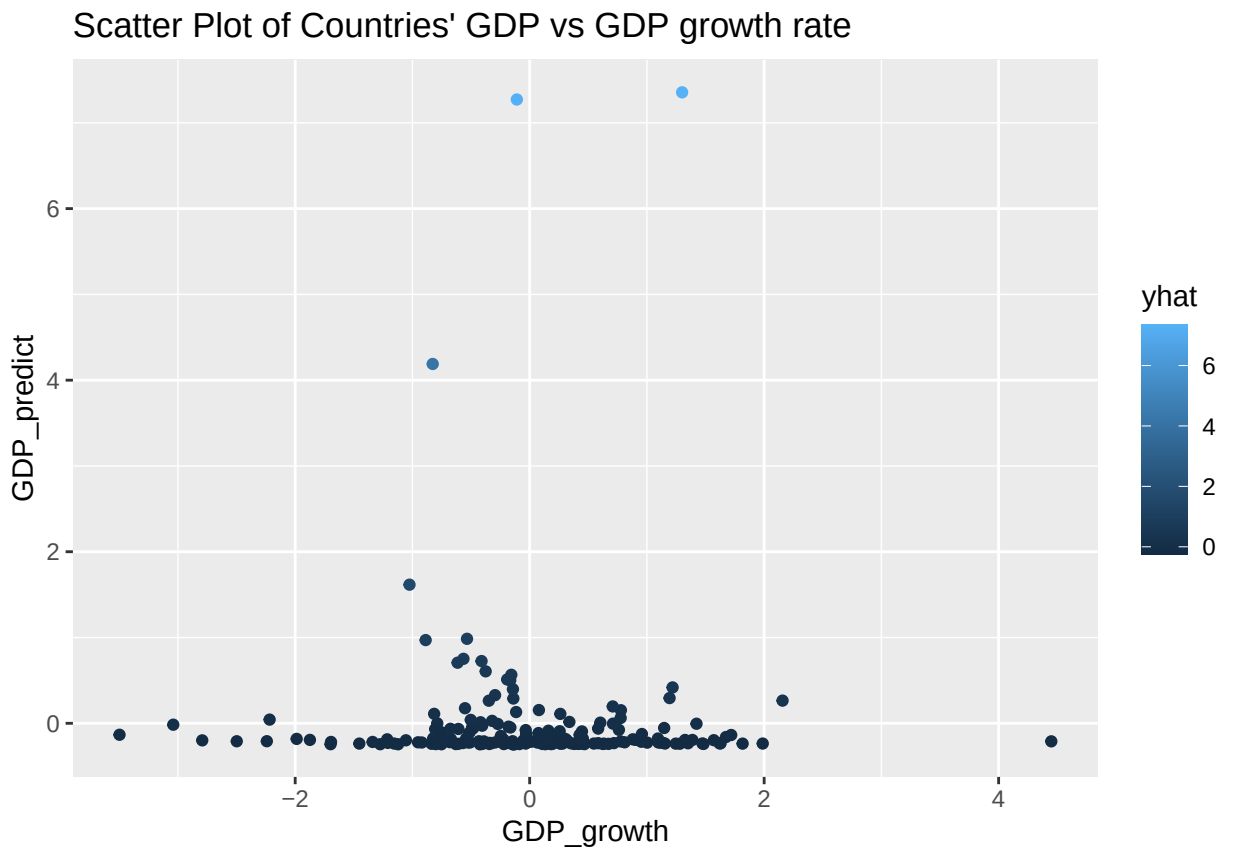
Step3: Plot the pictures

Now, we want to see the relationship between GDP growth rate and GDP of each countries. Firstly, we will see the scatter plot of the original data.Then we will see the plot of the relationship between the model's prediction of the GDP and GDP growth.

This is the scatter plot of Real GDP growth rate and GDP.



This is the scatter plot of the relationship between the model's prediction of the GDP and GDP growth.



Conclusion in Part 2

From the above figure, we can see that the distribution of predicted value are very similar to the real values. As a result, we can say that our Bagging model has good ability of predict the GDP by the other variables. If policy maker or financial institution want to know the GDP of one country, they can consider to use our Bagging model to do the prediction and decide the government policy and investment policy.

Part 3 : Clustering and PCA

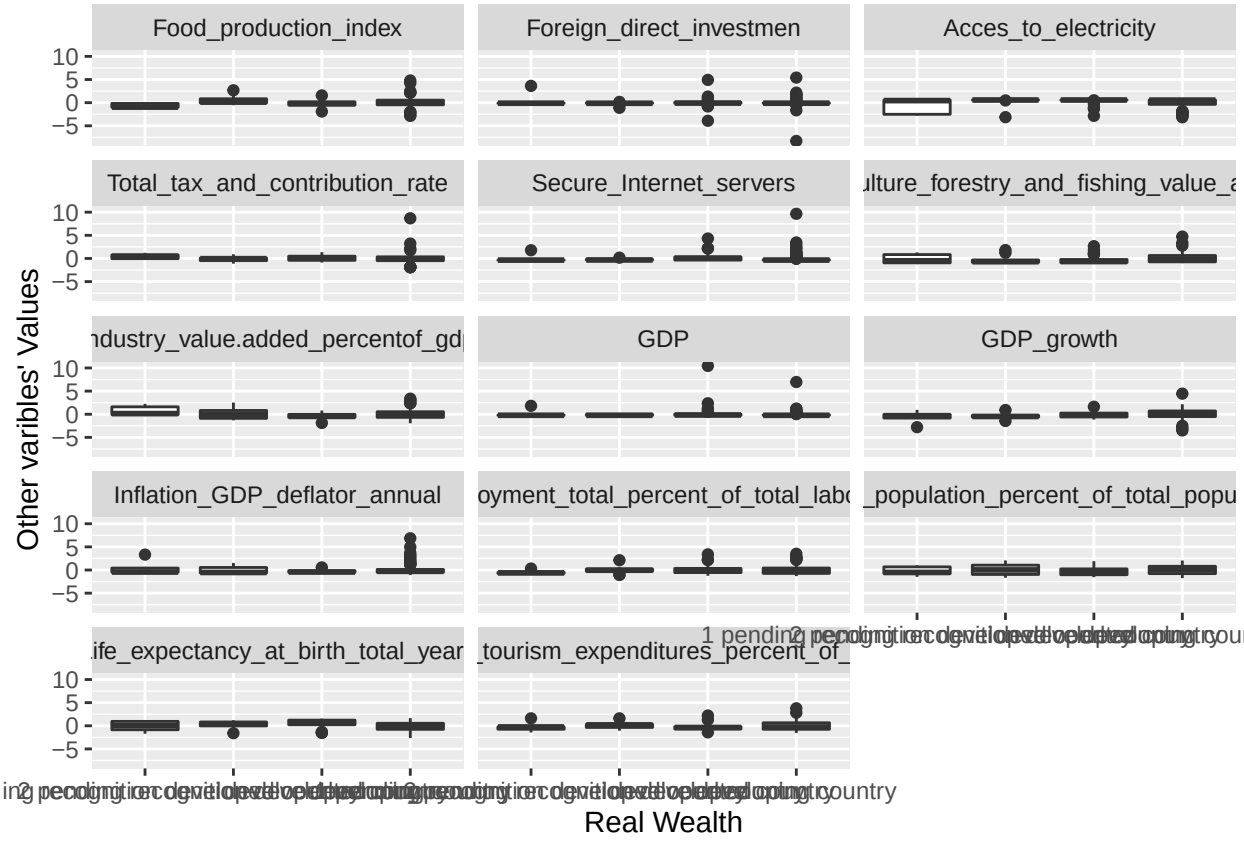
At this part, we want to do the unsupervised learning. In the real world, we can see that there are developing difference between the countries. There are five criteria that can be used for determine the country's developing level: 1. Human Development Index(HDI), 2. High-income economies, 3. Development Assistance Committee, 4. IMF advanced economies, 5. Paris Club members. According to these 5 criteria,which divide the countries to 4 categories: The first one is developed country, which means that it satisfies all 5 criteria; the second one is 2 pending recognition developed country, which means that it sill need to satisfy 2 more criteria to be recognized as developed country; the third one is 1 pending recognition developed country, which means that it sill need to satisfy 1 more criteria to be recognized as developed country, so the 1 pending recognition developed country is better than 2 pending recognition developed country, and the lastest one is developing country, which means that it is still developing so that it only satisfied up to 2 criteria. At this part, we want to give the computer the unsupervised information and let the computer learn to divide the countries into 4 categories which are not defined in advance, and compare it with the original data, then we can assess the performance. We use K-means++ clustering and PCA to do the unsupervised learning.There are 2 steps to do it, 1. run the K-means++ and PCA, 2. assess which method is better.

Run the PCA and Clustering

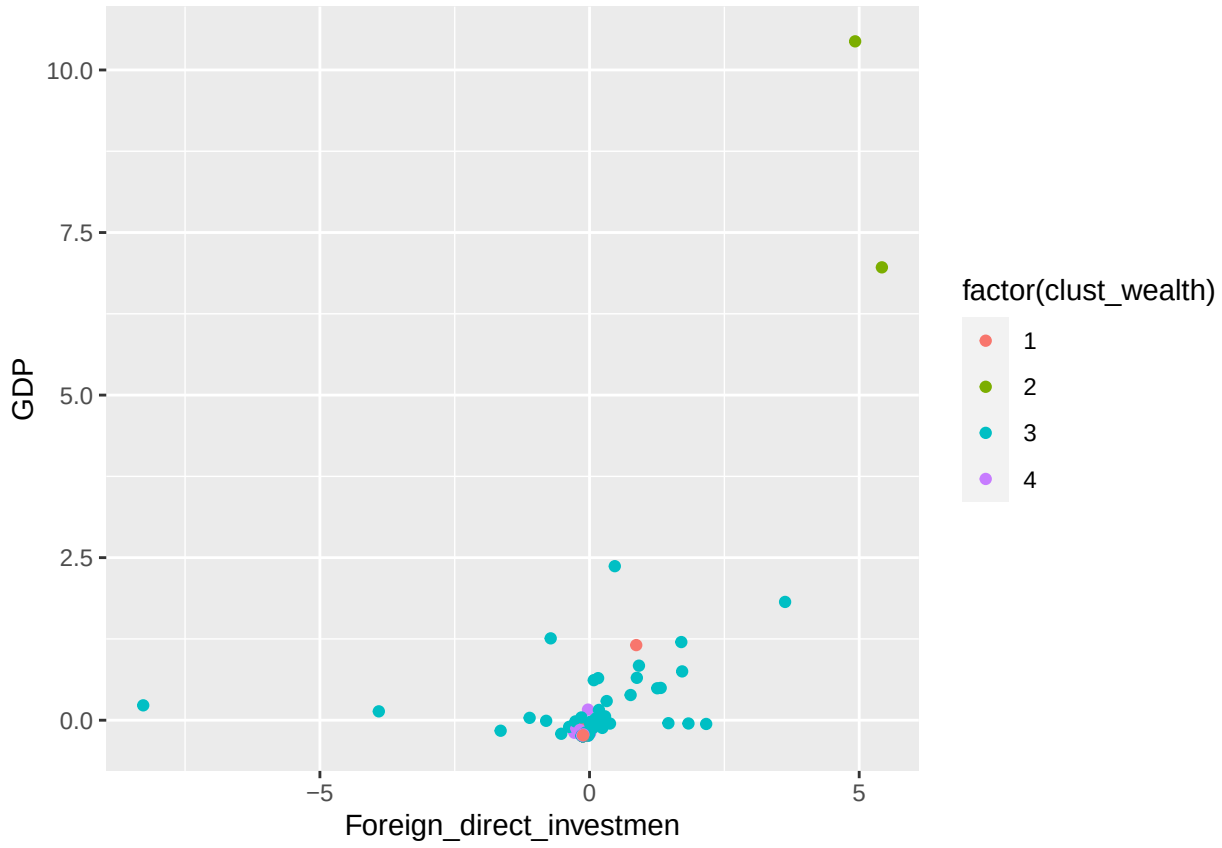
Clustering part

Assess : Clustering

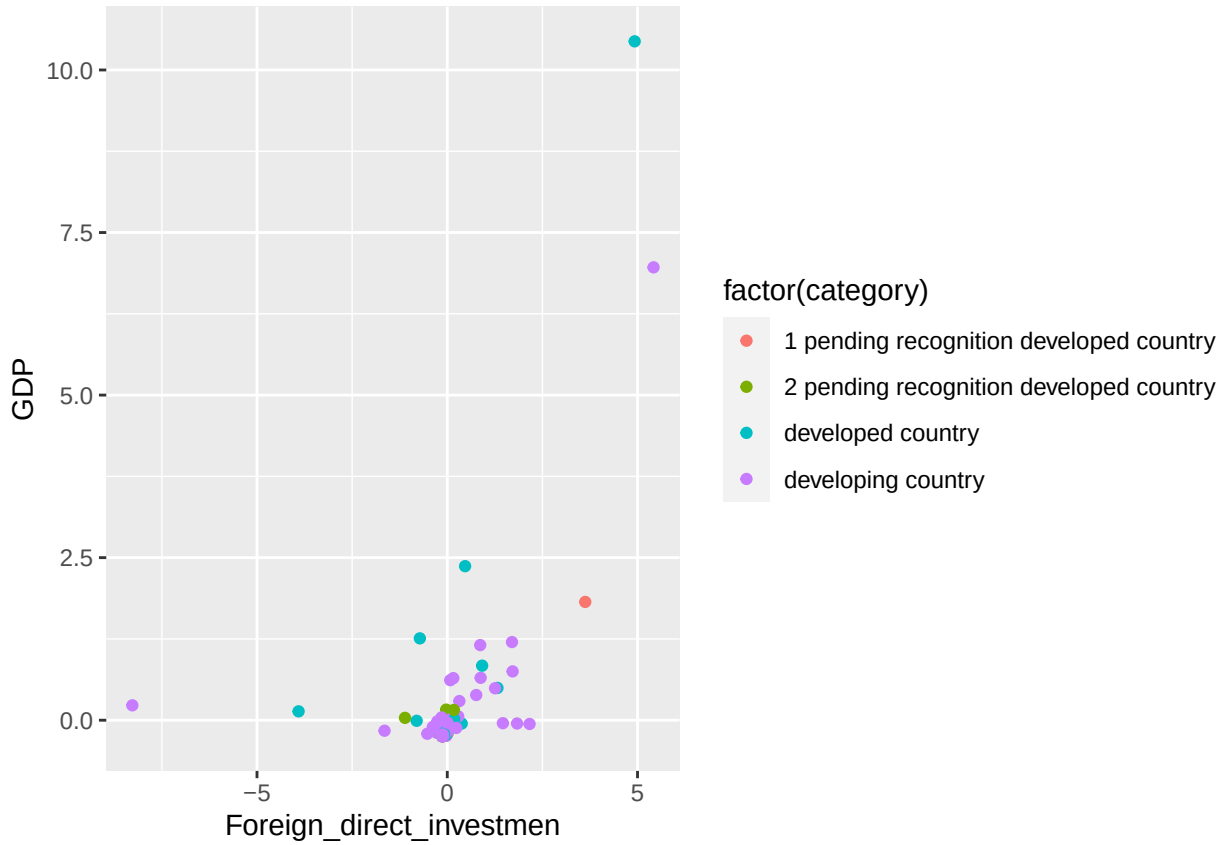
Category: Assess clustering results



According to this figure, we can see that we may distinguish countries by Foreign_direct_investmen and GDP_growth. Labels emerged naturally from clustering.



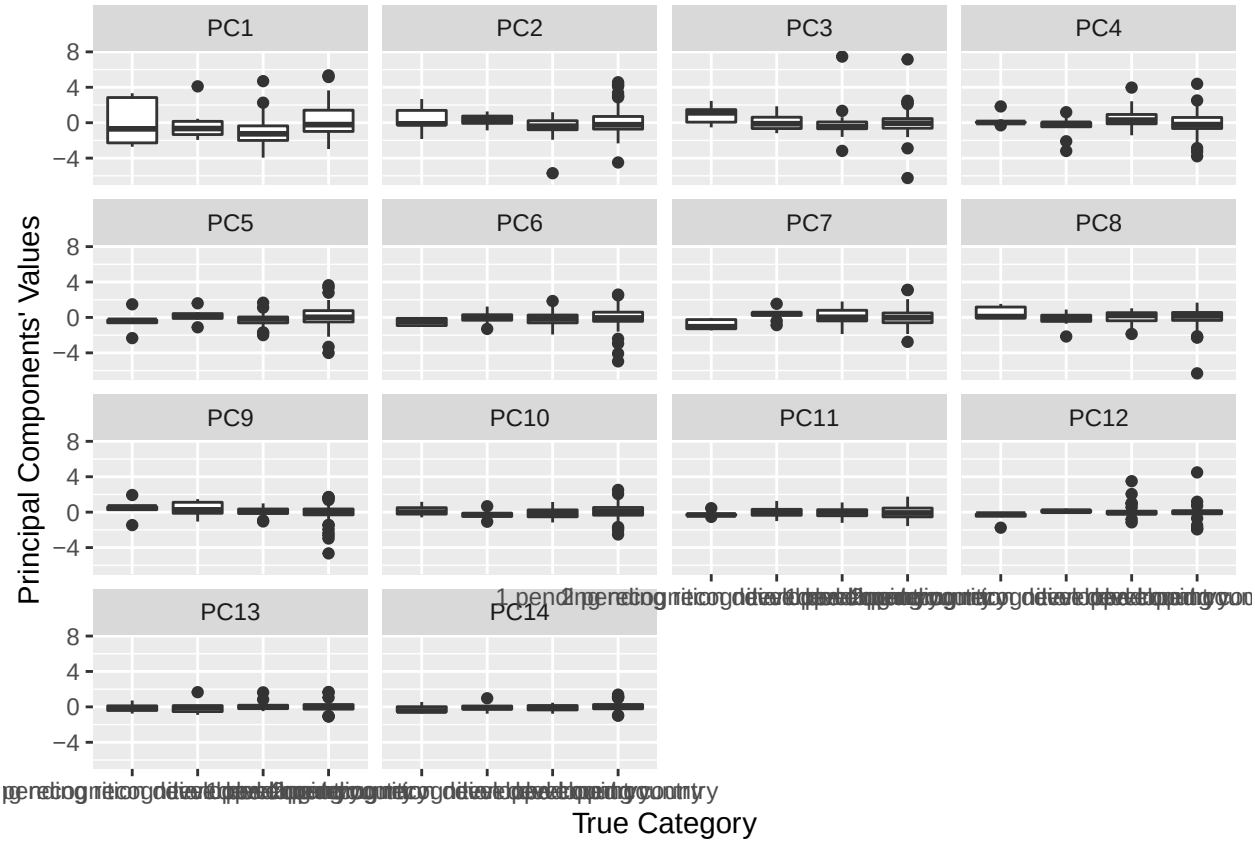
By clustering, We can see that green group is concentrated in the 0 and the purple ones are on the right, and the blue group is on the bottom of the figure.



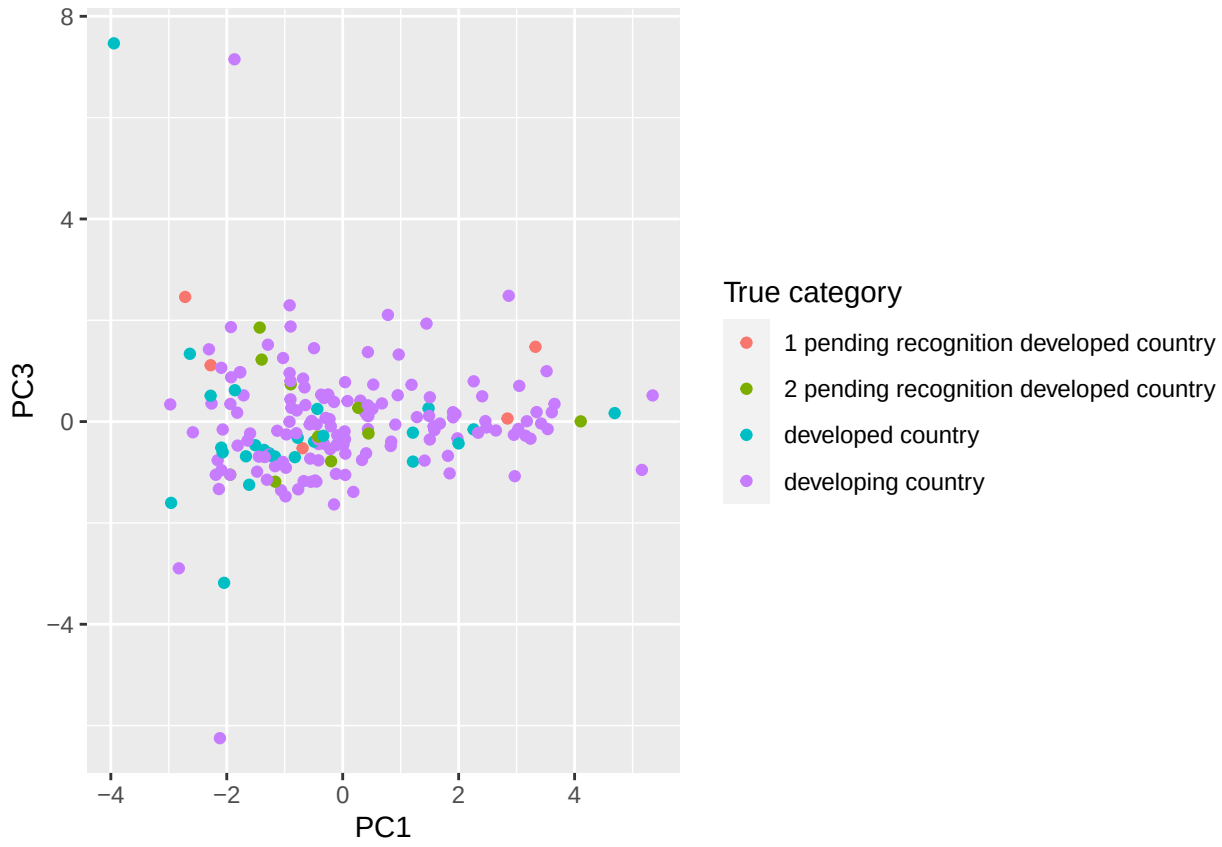
This is the true category group of country, compare to the above figure(clustering), we can see that clustering can help us distinguish the wealth of countries.

PCA part

Find the PC which can help us to distinguish



From this figure, we can see that PC1 and PC3 may have some trends, so we may use it to distinguish category of countries. Then we plot it into one figure to see their abilities of distinguish.



We only can see that the developed countries usually be in the left side,but PCA still can't distinguish the category of the country successfully.

Conclusion in Part3

From this part, we can see that Clustering (K-Means++) is a better way for us to distinguish the category of the country.However, although it clustering is better, it still can't distinguish the category very accurately. As a result, we may have more information to do the unsupervised learning successfully next time. But we still have some useful information that can be used for policy makers, we can find that Foreign Direct Investment can affect the country's developing level, as a result, if the Foreign Direct Investment can be higher, the country has more probabilities to become the developed country.

Conclusion

From part 2 we do above, we can see that the distribution of predicted value are very similar to the real values. As a result, we can say that our Bagging model has good ability of predict the GDP by the other variables. In part 3, we can conclude that clustering (K-Means++) is a better way for us to distinguish the category of the country.

Appendix

Pictures and tables in Part 1

```
##        avg_gdp       sd_gdp   q05_gdp    q95_gdp
## 1 473308864382 1.927853e+12 922917439 1.7295e+12
```