Prediction of Life Expectancy in Developing vs Developed Countries

Haozhen Ni, Huimin Zhang, Yanhua Lin, Yingzhen Zhao, Yujing Wen, Ziqian Liao

Introduction

Here we have a dataset from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO). The dataset records the life expectancy for each country in the world from 2000 to 2015 annually. Besides, the dataset includes several variables in different aspects with 22 columns, including immunization factors, mortality factors, economic factors and social factors.

Objectives

- Identify key variables that influence the values of life expectancy between countries with different status: developing or developed.
- Model the relationship between key factors and the values of life expectancy.
- Use findings to propose insights into real-life factors that affect the world's wellbeings.

Data Cleaning and Analysis

There are several missing values in the original dataset, which may be due to the lack of the data from certain small countries. Also, from the correlation plot, we can see that there is a certain degree of severely high correlation between variables. As a result, we remove several variables and only contain those we think have the potential to predict the response variables.

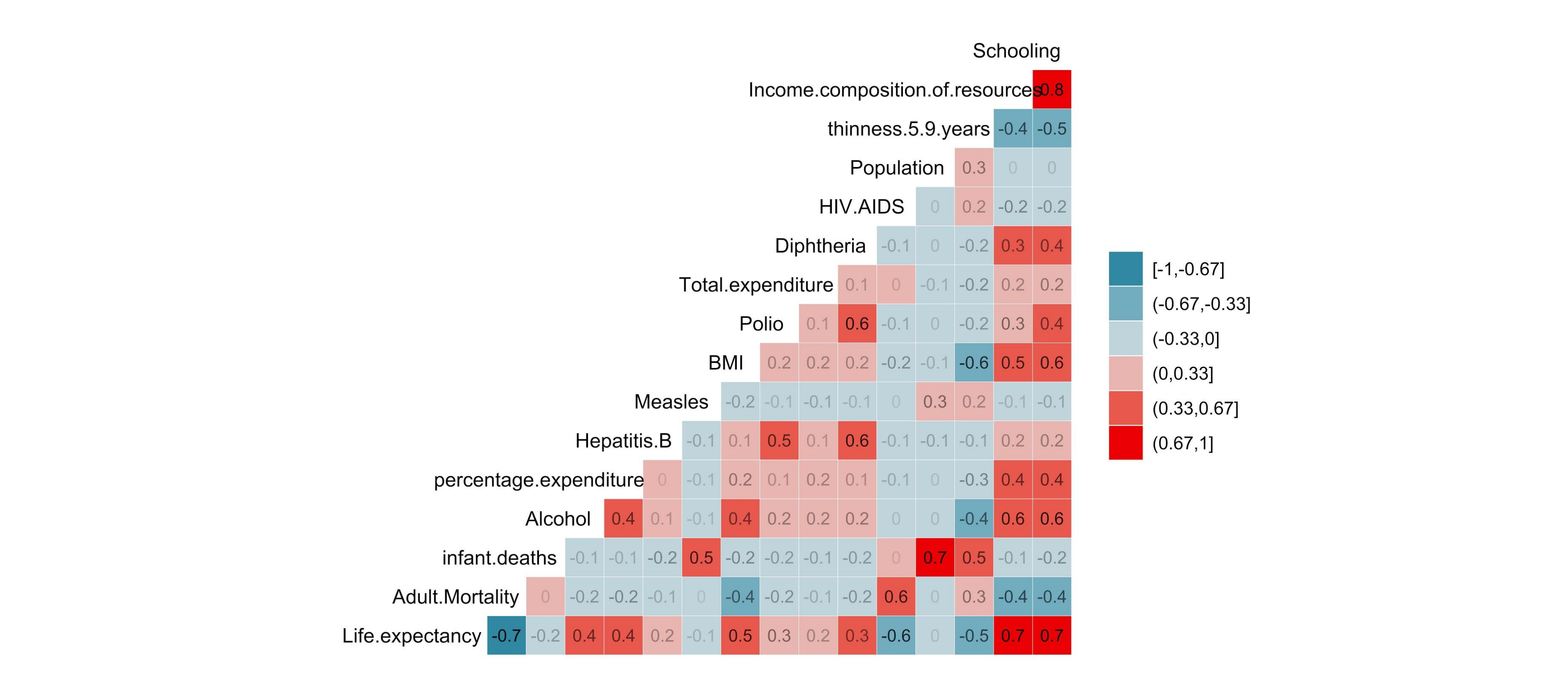

Figure1. *Correlation plot after removing highly correlated variables*

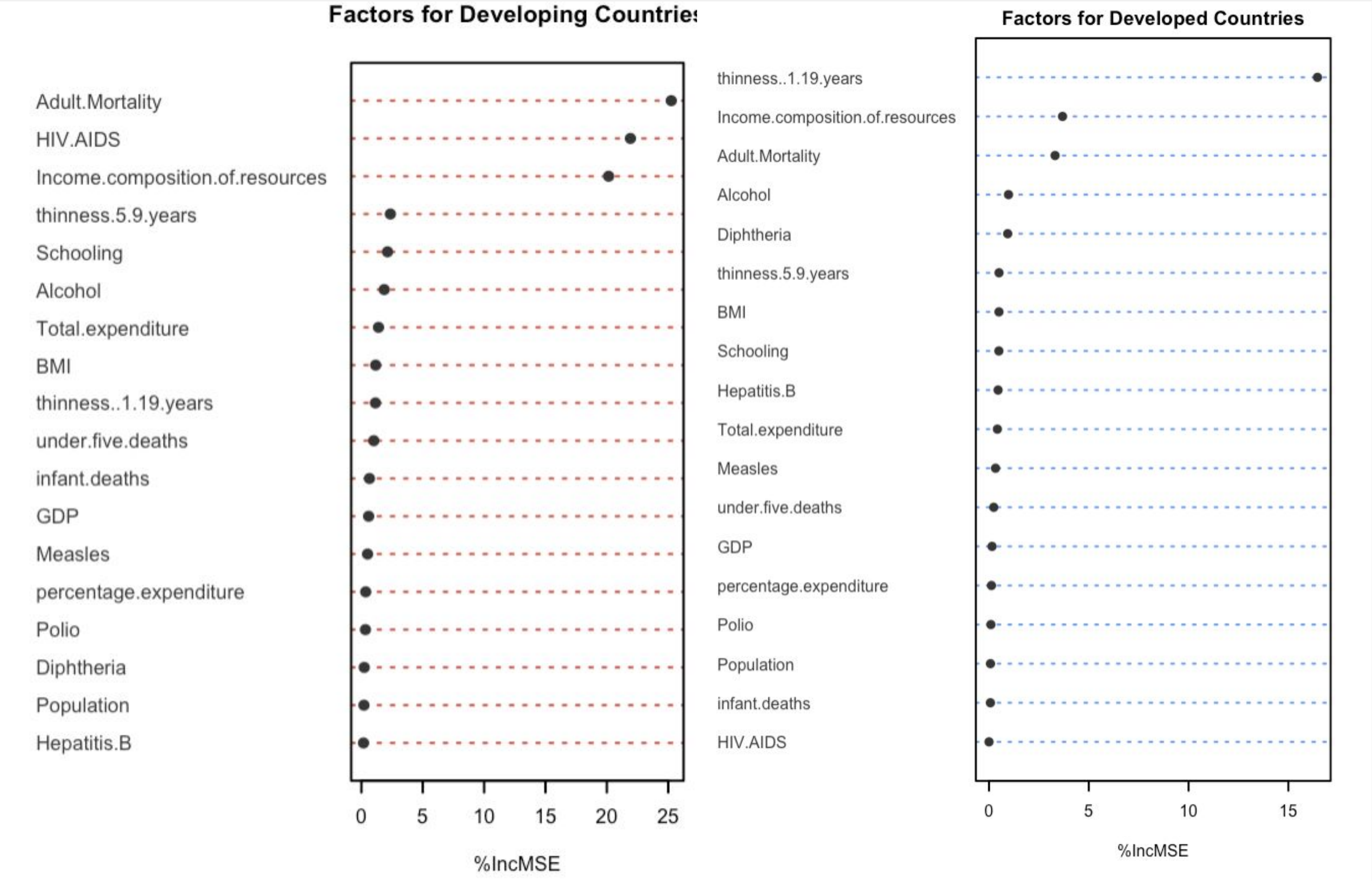

Figure2. *Factors sorted by significance in predicting life expectancy within the model based on random forest*

Modeling

Random Forest

To investigate factors that have an impact on the life expectancy of developed countries and developing countries, respectively, we applied the random forest algorithm to build a regression model. We found through cross validation that the recommended number of predictors chosen at each node (mtry) was 18. By tuning the model, we finally got an RMSE of 0.489 for predictions of life expectancy of developing countries, and an RMSE of 0.573 for that of developed countries. Such a result suggests that our regression model is effective in predicting the desired response.

Discussion

The plots show that for developed countries, society and economy are the main factors affecting Life Expectancy, while for developing countries, immunity and mortality or medical level are the main factors affecting Life Expectancy.

For developed countries, due to a longer period of economic prosperity, the gap between the rich and the poor is also more severe. Poor people in developed countries have relatively low incomes, and therefore cannot receive high-level education due to the inability to afford high tuition. They have less knowledge of healthy diets and more shallow understanding of the effects of various factors on their health, which is not conducive to the development of healthy living habits. At the same time, it is difficult to buy healthy foods at relatively high prices for a long time with low income, and there is no way to maintain a reasonable dietary structure and good eating habits, which affect the life expectancy of the country.

For developing countries, due to the underdeveloped economy, the quality of living environment and the level of medical and health services will be relatively low. The poor quality of living environment does not guarantee daily clean water or less polluted environment, which will breed many bacteria and diseases. At the same time, due to the low level of medical and health services, many infectious diseases cannot be well controlled, and may form a vicious circle, which affects the life span of the population.

Conclusion

Based on the analysis, for developed countries, the most important factor affecting Life Expectancy is thinness 1-19 years, while for developing countries, the most important factor is Adult Mortality.

According to the above figure, it can be seen that the main factors affecting Life Expectancy in different countries may include: 'Income composition of resources', 'Adult Mortality', 'thinness 1-19 years / thinness 5-9 years' and 'Alcohol'. At the same time, 'population', 'GDP' and 'percentage expenditure' have little impact on Life Expectancy.

Reference

Dataset link:
https://www.kaggle.com/kumarajarshi/life-expectancy-who