Life Expectancy Prediction

Jingyi Zhu

STA9890 Spring 2020

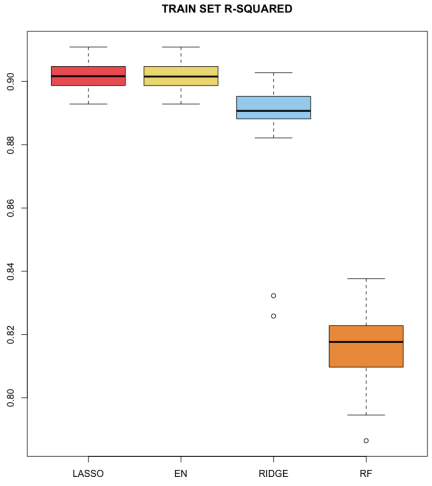
May 18, 2020

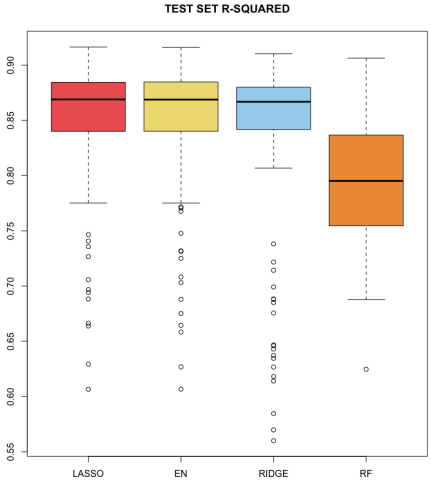
Data Description

For this project, we would like to use related factors to predict **Life Expectancy**. We extracted World Development Indicators of twenty four countries from World Bank Databank.

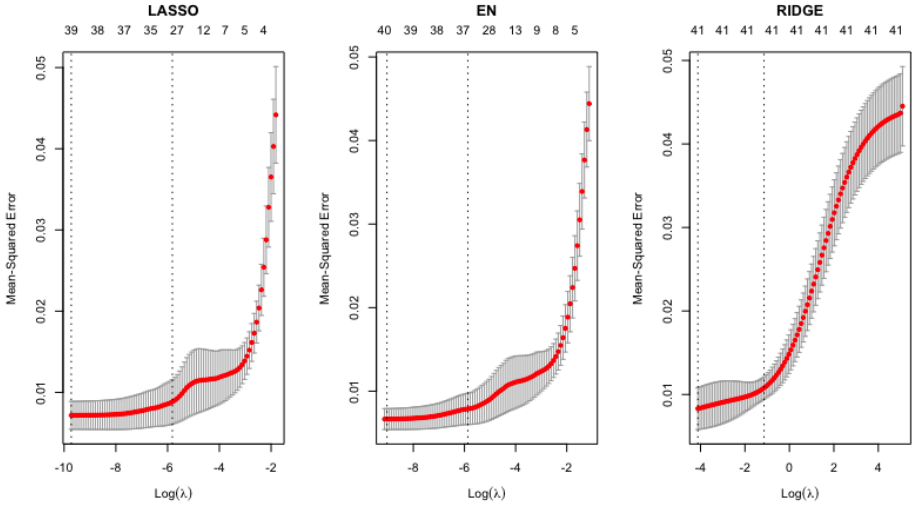
- Data Source: World Bank Databank
- Response Variable: Life Expectancy at birth
- The number of features p: 41
- The sample size n: 480
- Predictors: GDP growth (annual %), GDP per capita (current US$), Physicians (per 1,000 people), Hospital beds (per 1,000 people), Fertility rate, total (births per woman), GINI index (World Bank estimate), and Smoking prevalence, total (ages 15+), etc.

Side-by-side Boxplots of R^2_{test} and R^2_{train}

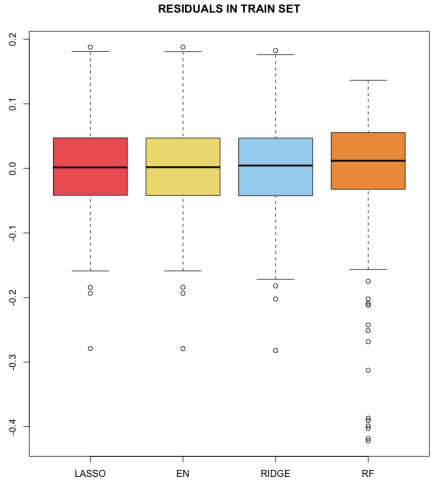


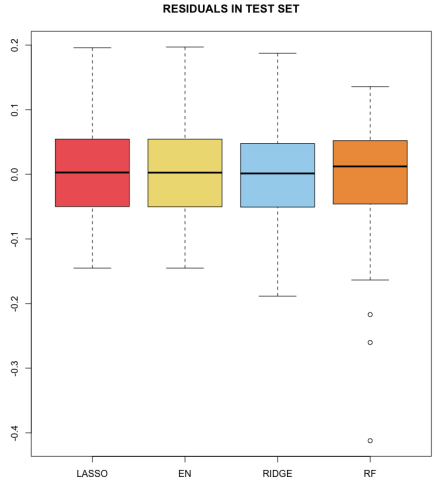


10-fold CV Curves

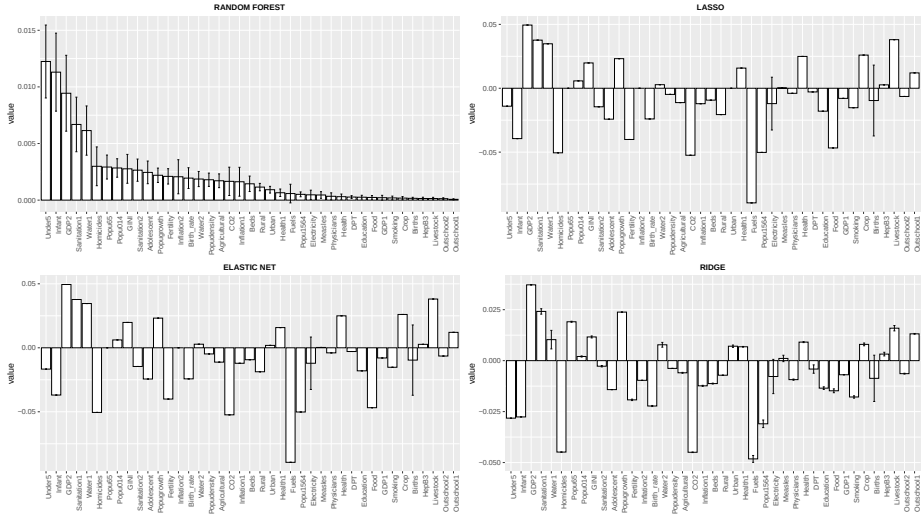


Side-by-side Boxplots of Train and Test Residuals





Bar-plots with Bootstrapped Error Bars



Summary

- No obvious overfitting issue among all 4 methods
- Lasso has the best performance in terms of R-Squared on Test set
- Trade-off between model accuracy and processing time

	MODEL	PERFORMANCE	TIME
1	LASSO	0.8462	0.0986 secs
2	ELASTIC NET	0.8454	0.1449 secs
3	RIDGE	0.8278	0.1034 secs
4	RANDOM FOREST	0.7932	0.0865 secs