Machine Learning Applications In Predicting Life Expectancy And A Nation's Development Status

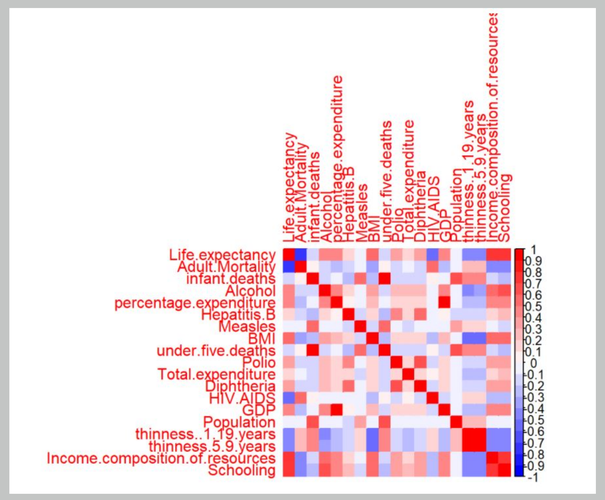
Tristan Luu
Fall 2021
Department of Data Analytics - Denison University

Abstract

- This research is divided into two parts. Firstly, I use linear regression, with year as the fixed effect, to predict the life expectancy of different nations. In the second part, I use two machine learning methods, decision tree and logistic regression, to classify whether a country is a developed or developing nation.
- After comparing the results, I conclude that decision tree performs slightly better than logistic regression in this scenario.

Explanatory Data Analysis

- The Life Expectancy dataset from WHO consists of 22 columns and 2938 rows with information related to health, GDP, mortality rate, etc in 193 countries from 2000 to 2015. There are 20 predicting variables, which can be divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.
- Correlation matrix between the variables
-



- In order to build the models, I separate the dataset into training and testing sets (80/20 ratio).
- In the following theorem, we prove that this is actually a stronger condition than is needed to ensure cycle consistency.

Predicting Life Expectancy - Process

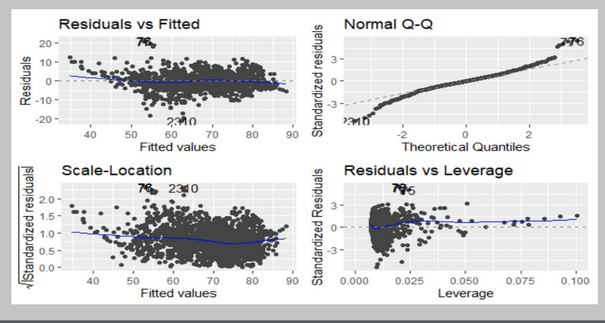
- Since this is a time-series dataset, my approach is to use *linear regression* with the *year* variable as the fixed effect to measure life expectancy of different nations.
- I first start with using all of the variables in the dataset, then I drop out some of insignificant variables with p-values larger than 0.1.
- My final model uses *adult mortality*, *percentage expenditure*, *BMI*, *deaths under 5 year-old*, *diphtheria*, *HIV AIDS*, *income composition of resources*, *schooling*, and *year* to predict life expectancy.
-

```
lm(formula = Life.expectancy ~ Year + Adult.Mortality + percentage.expenditure +
    BMI + under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources +
    Schooling, data = train1)
```

Predicting Life Expectancy - Result

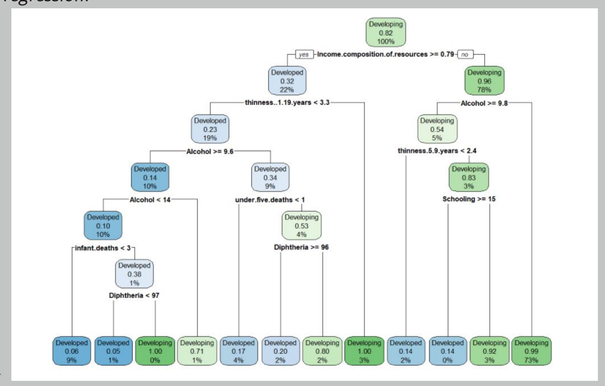
```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.229e+01  5.802e-01  90.134  < 2e-16 ***
Year2001                        -8.745e-01  4.944e-01  -1.769 0.077053 .
Year2002                        -1.414e+00  4.790e-01  -2.952 0.003191 **
Year2003                        -1.321e+00  4.786e-01  -2.761 0.005820 **
Year2004                        -8.811e-01  4.848e-01  -1.817 0.069292 .
Year2005                        -1.201e+00  4.841e-01  -2.481 0.013177 *
Year2006                        -1.382e+00  4.754e-01  -2.907 0.003689 **
Year2007                        -1.448e+00  4.851e-01  -2.985 0.002870 **
Year2008                        -1.280e+00  4.824e-01  -2.654 0.008011 **
Year2009                        -1.302e+00  4.808e-01  -2.707 0.006837 **
Year2010                        -1.323e+00  4.906e-01  -2.696 0.007074 **
Year2011                        -1.707e+00  4.921e-01  -3.469 0.000532 ***
Year2012                        -1.834e+00  4.871e-01  -3.766 0.000170 ***
Year2013                        -1.493e+00  4.865e-01  -3.068 0.002182 **
Year2014                        -1.377e+00  4.859e-01  -2.834 0.004637 **
Year2015                        -1.106e+00  4.938e-01  -2.240 0.025219 *
Adult.Mortality                 -1.755e-02  9.096e-04 -19.294  < 2e-16 ***
percentage.expenditure           3.415e-04  4.668e-05   7.316 3.60e-13 ***
BMI                              3.408e-02  5.324e-03   6.402 1.88e-10 ***
under.five.deaths               -2.058e-03  5.485e-04  -3.752 0.000180 ***
Diphtheria                       4.378e-02  4.106e-03  10.663  < 2e-16 ***
HIV.AIDS                        -5.041e-01  1.972e-02 -25.557  < 2e-16 ***
Income.composition.of.resources  9.236e+00  7.198e-01  12.831  < 2e-16 ***
Schooling                        9.237e-01  4.743e-02  19.474  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.94 on 2151 degrees of freedom
  (175 observations deleted due to missingness)
Multiple R-squared:  0.8207,    Adjusted R-squared:  0.8188
F-statistic: 428.1 on 23 and 2151 DF,  p-value: < 2.2e-16
```



Classify a nation's status

- I want to investigate the performance of a parametric and a non-parametric algorithm when classifying a nation's status (whether it is a developed or developing country). Since the dependent variable is categorical, my chosen methods are *decision tree* and *logistic regression*.



- *Decision tree* results in an accuracy of 96.2 percent and an area under the curve of 90.7.
- *Logistic Regression* has an accuracy of 93.3 percent and an area under the curve of 83.1.

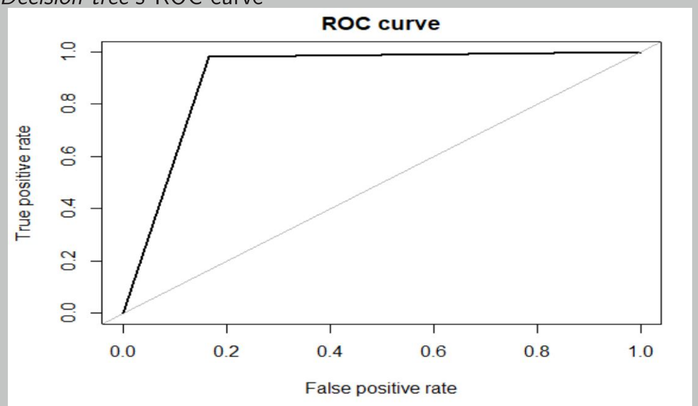
Logistic Regression

- Due to variables being highly correlated, I also add an interaction term: thinness = thinness..1.19.years x thinness.5.9.years.
-

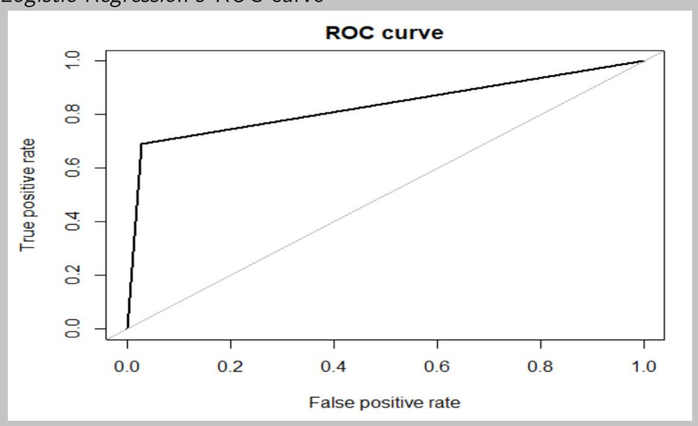
```
glm(formula = Status ~ Alcohol + Hepatitis.B + under.five.deaths +
    thinness + Income.composition.of.resources, family = binomial,
    data = train2)
```

ROC curves

Decision tree's ROC curve



Logistic Regression's ROC curve



In conclusion, *Decision tree* performs slightly better than *Logistic Regression* when classifying status of a nation.

Reference

- Rajarshi, Kumar. "Life Expectancy (WHO)." Kaggle, 10 Feb. 2018, https://www.kaggle.com/kumarajarshi/life-expectancy-who.