

Project Presentation: Which variables influence the life expectancy?

Chenyue Zhu

12/4/2019

Problem Statement and Background

- ▶ In the past: the demographic factors such as income distribution and natural growth rate.
- ▶ Recently: the development index, immunization factors, and education level

So I added the new indexes together with the traditional ones to analyze their relationship with life expectancy.

Methods/approaches

- ▶ Data wrangling
- ▶ Data visualization
- ▶ Regression model

Methods/approaches used

Data wrangling

- ▶ tidyverse: clean up the raw data set, select useful variables, combine different predictors and outcome in one data set.

Data visualization

- ▶ ggplot: visualize the distribution of variables to observe its patterns.

Regression model

- ▶ Linear regression
- ▶ KNN model
- ▶ CART
- ▶ Random Forest I can choose from the three different models and see which one can explain the data set best by comparing the MAE, RMSE, Rsquared.

```
## Parsed with column specification:
```

```
## cols(
```

Preliminary Results and Conclusions

- Import the morality rate raw data set.

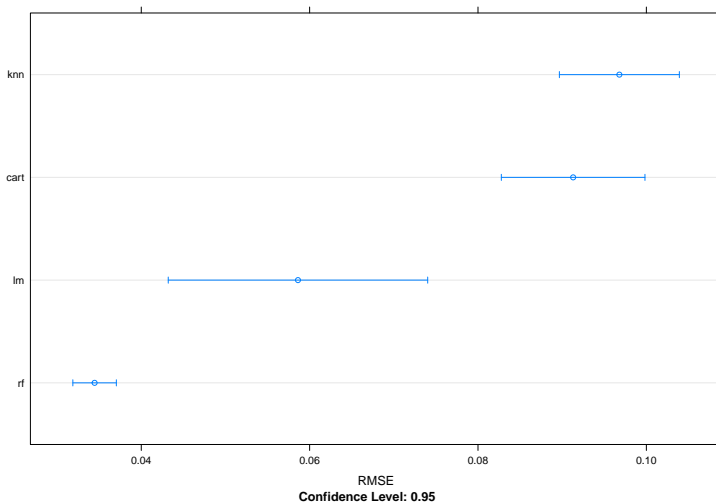
```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_character(),
##   X5 = col_character()
## )
```

- Clean the raw data set.

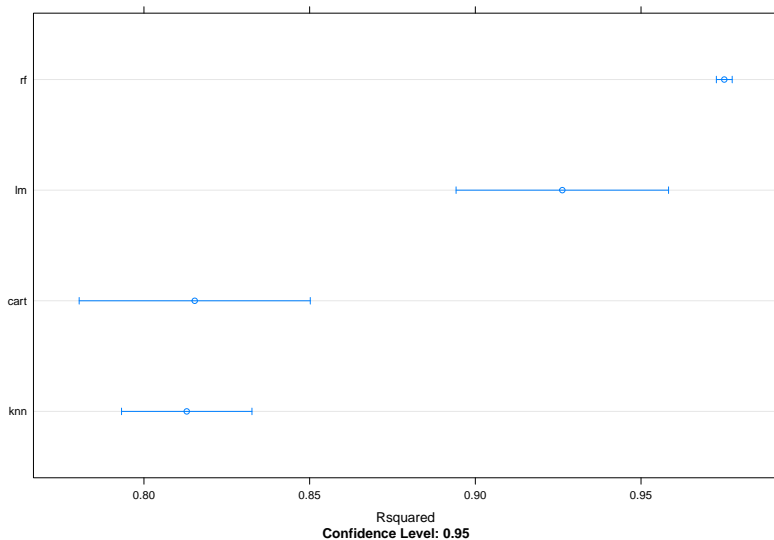
```
## # A tibble: 6 x 3
## # Groups:   country, year [6]
##   country      year morality_rate
##   <chr>        <chr> <chr>
## 1 Afghanistan 2016    245
## 2 Afghanistan 2015    233
```

Model Comparison

► RMSE



Examine the fit across each of the models.



► Test the Predictive Accuracy of the Best Model