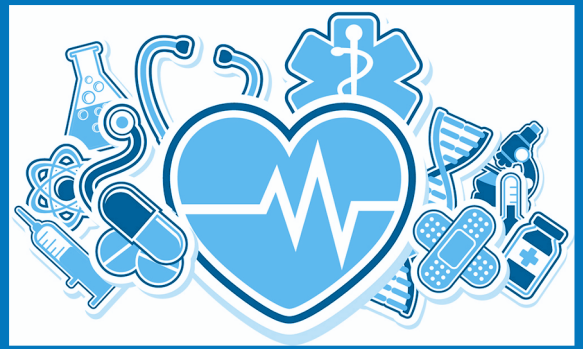
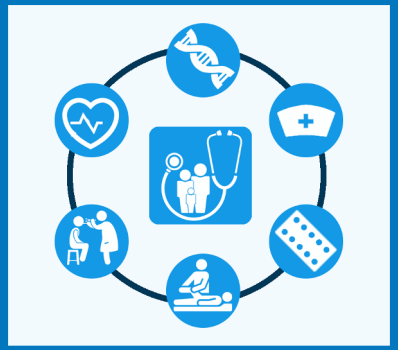
Predicting Premature Death Rate in US Counties

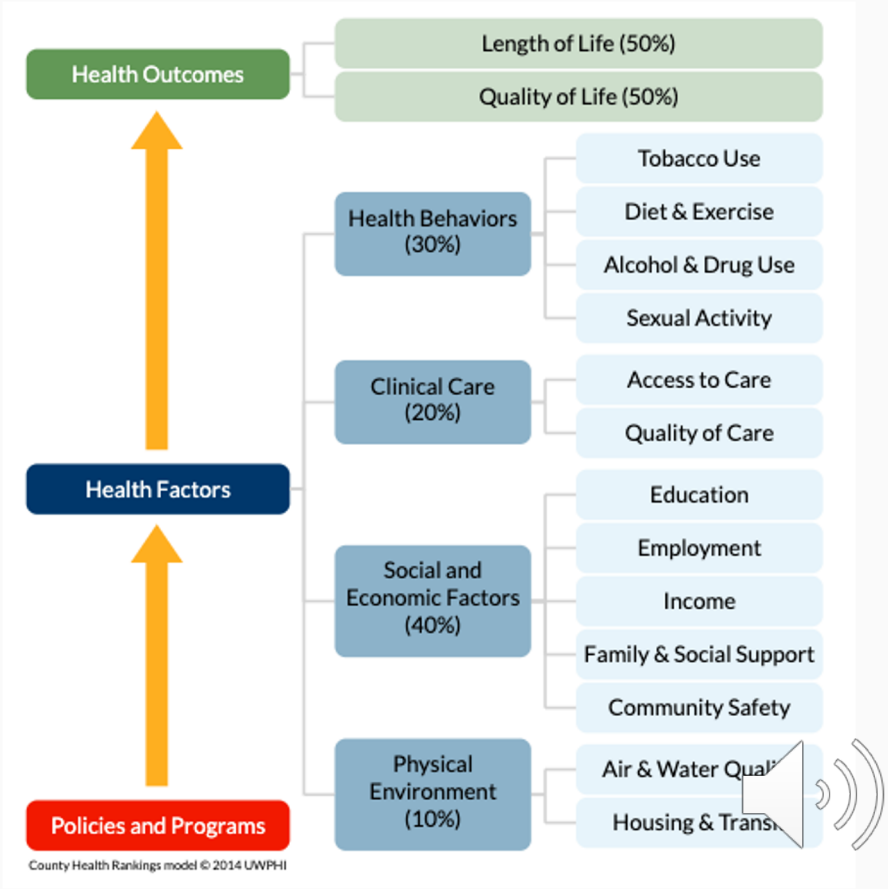
Mai Le - https://github.com/lnm453/STA9890





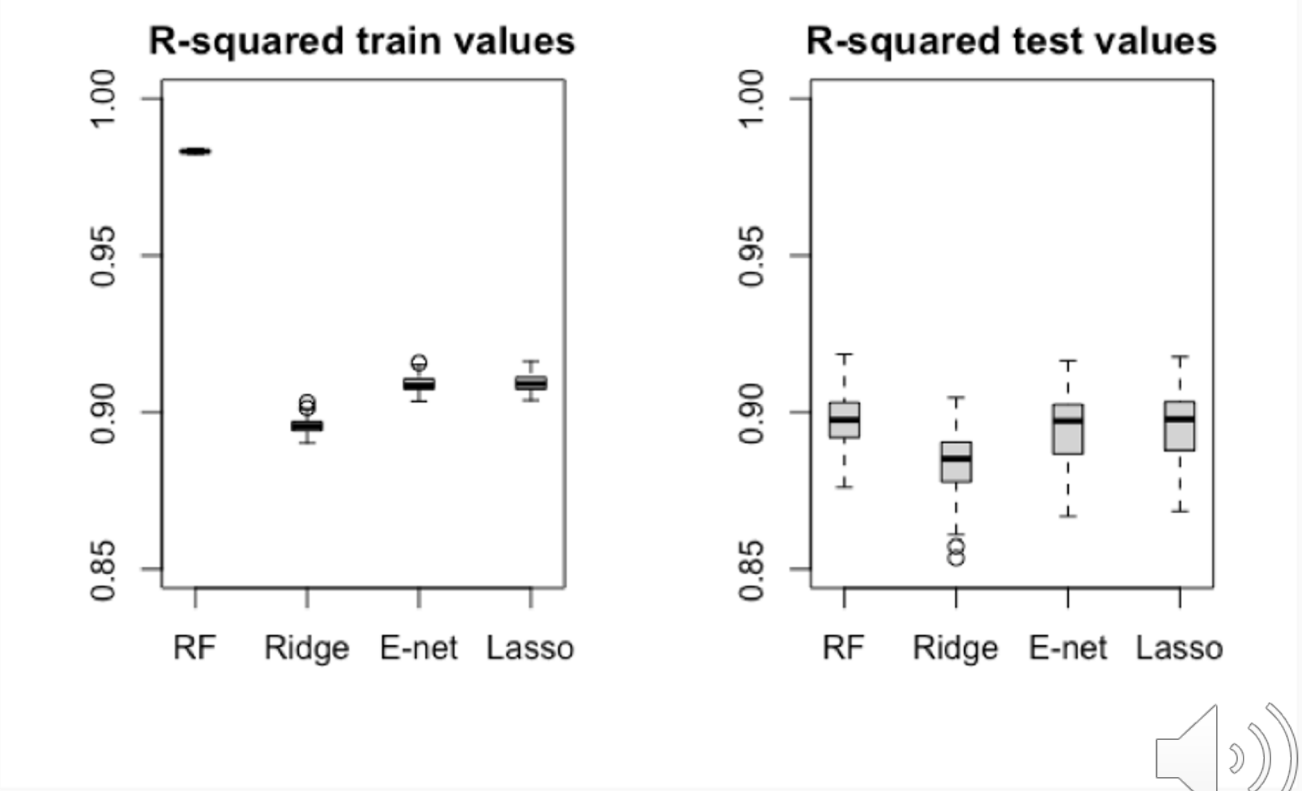
County Health Rankings Data - Overview

- Social, economic and health conditions of all US counties (2016-2018)
- 3144 instances (n)
- 176 predictor variables (p)
- y: ‘Years of potential life lost rate’ (YPLL)



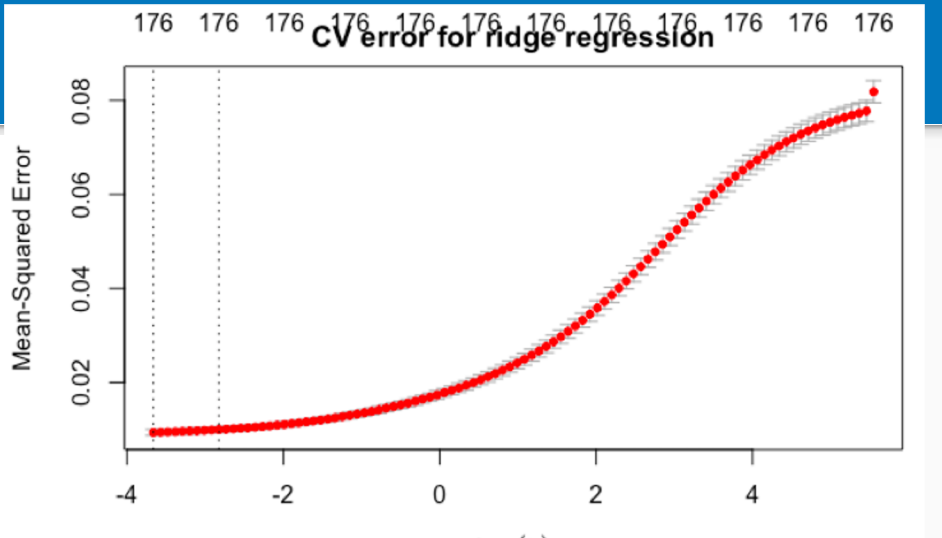
Model performance measured by R-squared values

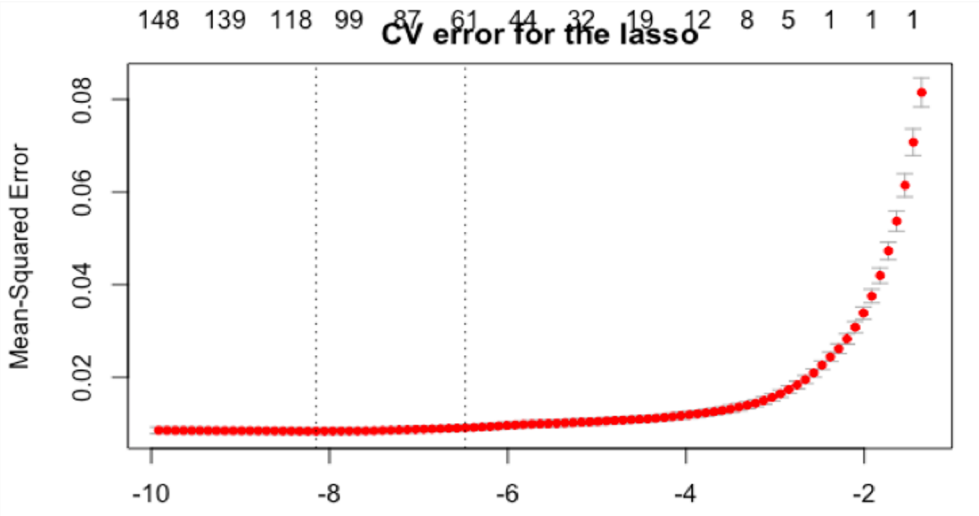
- All four models seem to be overtraining
- Random forest - highest level of overtraining
- Test values have higher SD

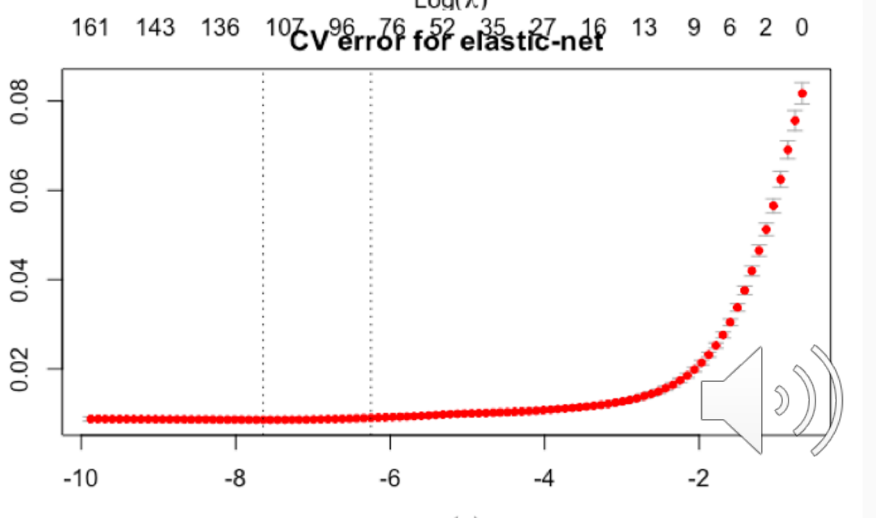


10-fold Cross Validation curves

- Lasso: 61 - 118 features
- Elastic-net: 76 - 107 features
- Ridge regression: 176 features

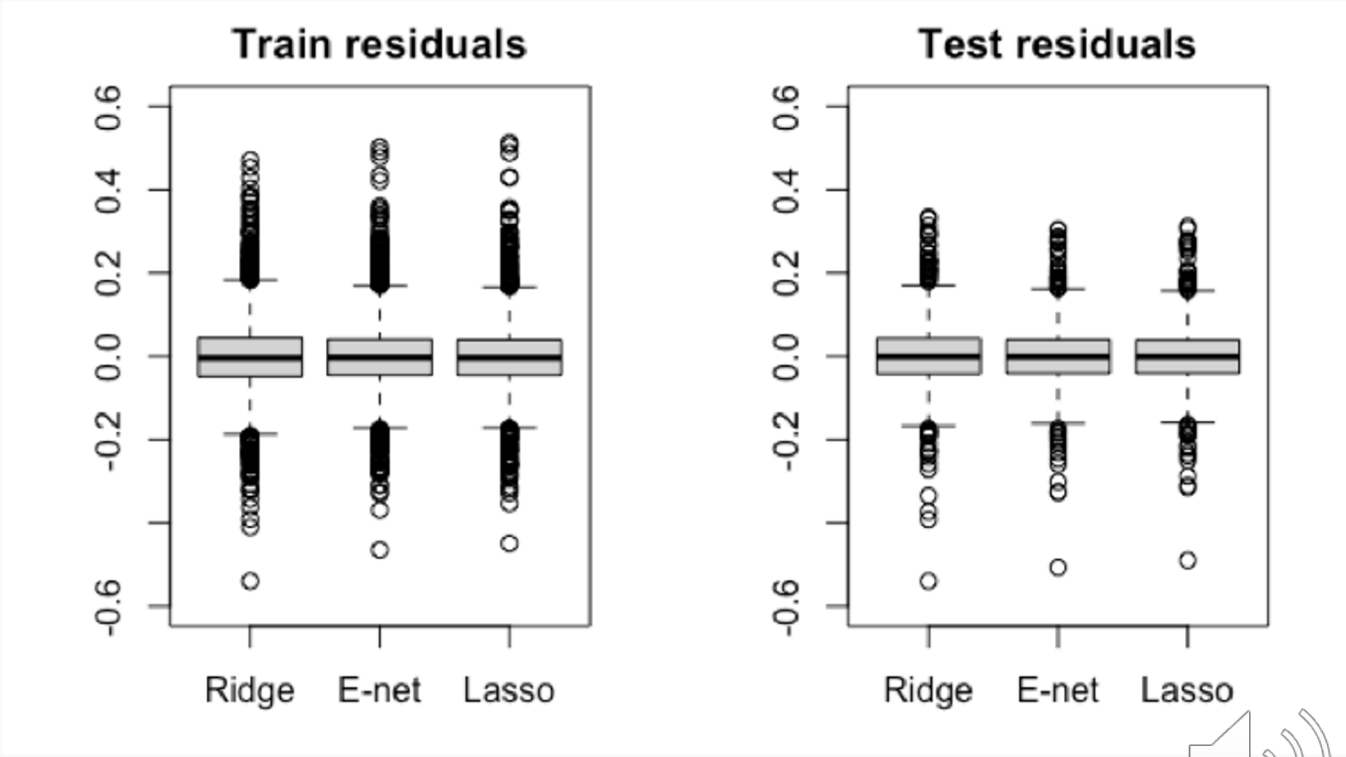






10-fold Cross Validation: train and test residuals

- Mean: 0
- SD: low
- Test residuals have lower variability than train residuals



Importance of parameters (RF) and Estimated Coefficients (Rid, E-net, Lasso)

Order of Importance for RF:

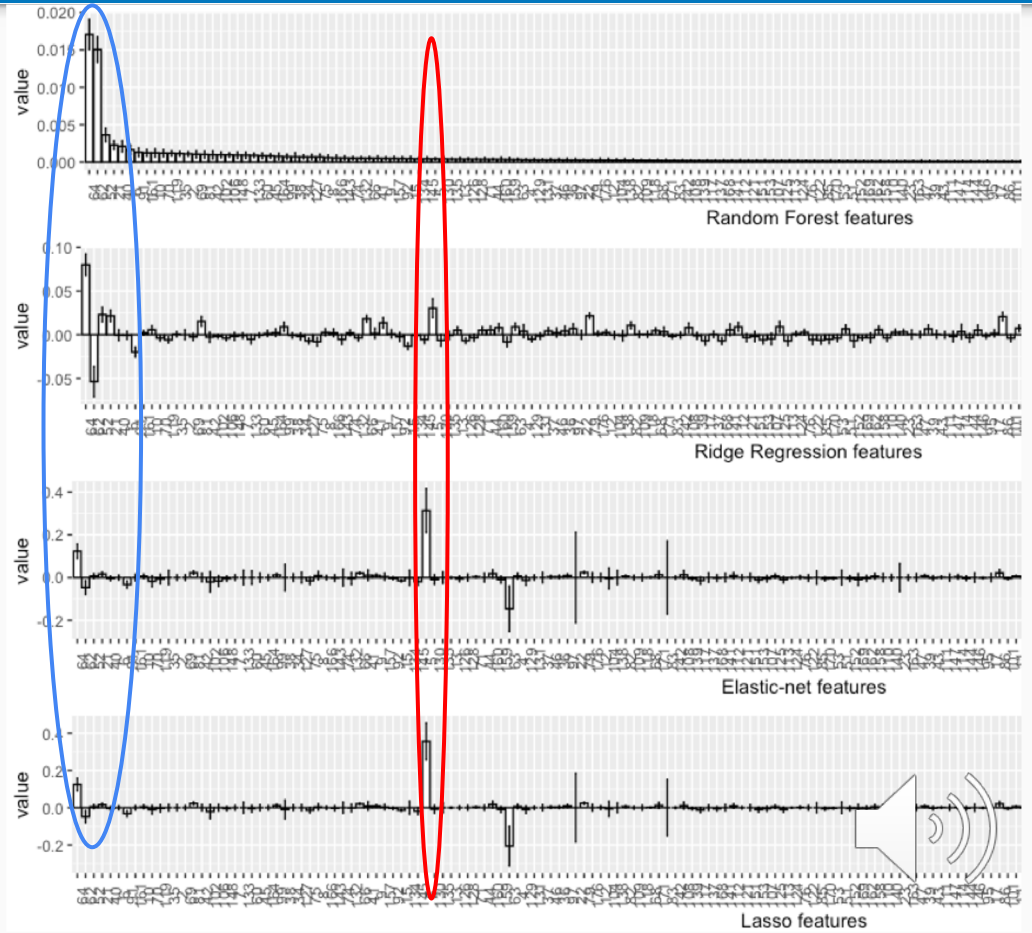
1. **Age adjusted death rate**
2. **Life expectancy**
3. Injury death rate
4. Teen birth rate

Est. Coefficient Size for Ridge Regression:

1. **Age adjusted death rate**
2. **Life expectancy**
3. Number of deaths
4. Injury death rate

Est. Coefficient Size for E-net and Lasso:

1. Number of deaths
2. **Age adjusted death rate**
3. **Life expectancy**



In closing

	Random Forest	Ridge Regression	Elastic-net	Lasso
Rsq values	Test < Train	Test≈ Train	Test≈ Train	Test≈ Train
Residuals	Low residuals Small variability	Low residuals Small variability	Low residuals Small variability	Low residuals Small variability
Model complexity	More complex	More complex	More sparse	More sparse
Training Time (s)	117.843	1.423	1.820	1.725