

# ONLINE NEWS POPULARITY

**FINAL REGRESSION PROJECT**

STA 9890 UWA – STATISTICAL LEARNING FOR DATA MINING

MAY 5, 2021 – PROF. KAMIAR RAHNAMA RAD

**GROUP 4:**

**MICHAEL S. BONETTI**

<https://github.com/mbonetti-nyc/Online-News-Popularity>



# BRIEF DESCRIPTION

ONLINE NEWS POPULARITY | UCI ML REPOSITORY



## About this dataset

**Summarizes heterogenous set of features**

- About published Mashable articles over a 2-year period
  - Datapoints: 39,797 (*stated*), 39,644 (*observed*) ( $n$ )
  - Response (*shares*): the goal is to predict the number of shares in social networks (popularity)



## Attributes

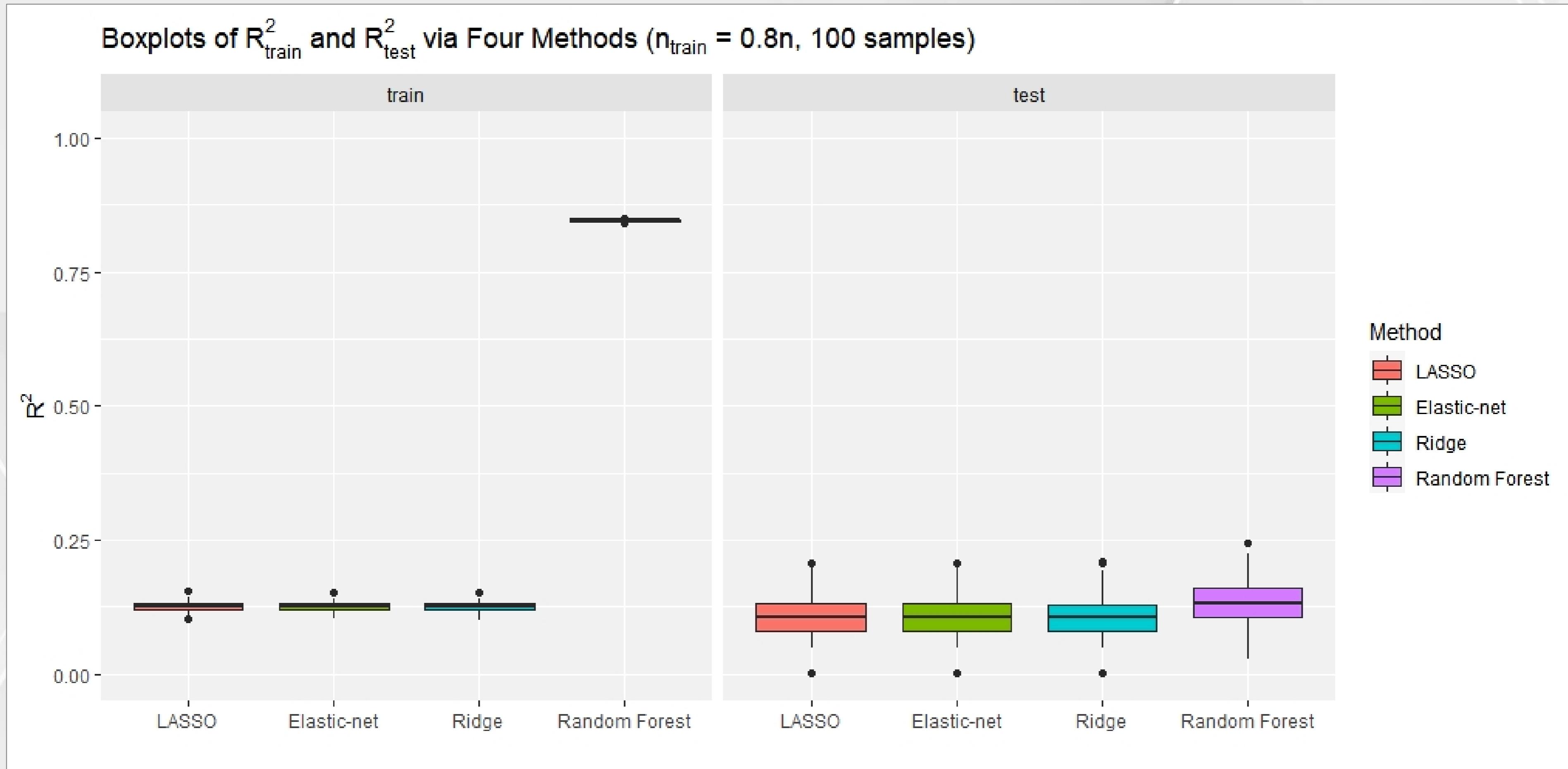
- **61 attributes ( $p$ )**
  - 58 predictive attributes (19 categorial & 39 numerical), 2 non-predictive, 1 target
  - Predictors
    - 6 data channels: lifestyle, entertainment, business, social media, tech, & world
    - Days of the week, and closeness to LDA (*Latent Dirichlet allocation*)
- **No missing values were found upon preliminary analysis**

## For this project

- **25% random sample taken**
  - 9,911 observations ( $n$ ): 25% of 39,644
  - 57 predictive attributes ( $p$ ): 3 attributes removed (*url, timedelta, is\_weekend*)



## BOXPLOTS OF $R^2$



# CROSS-VALIDATION CURVES

VIA 10-fold | LASSO, ELASTIC-NET (EN), RIDGE

**LASSO Runtime 1:** 1.98 secs.

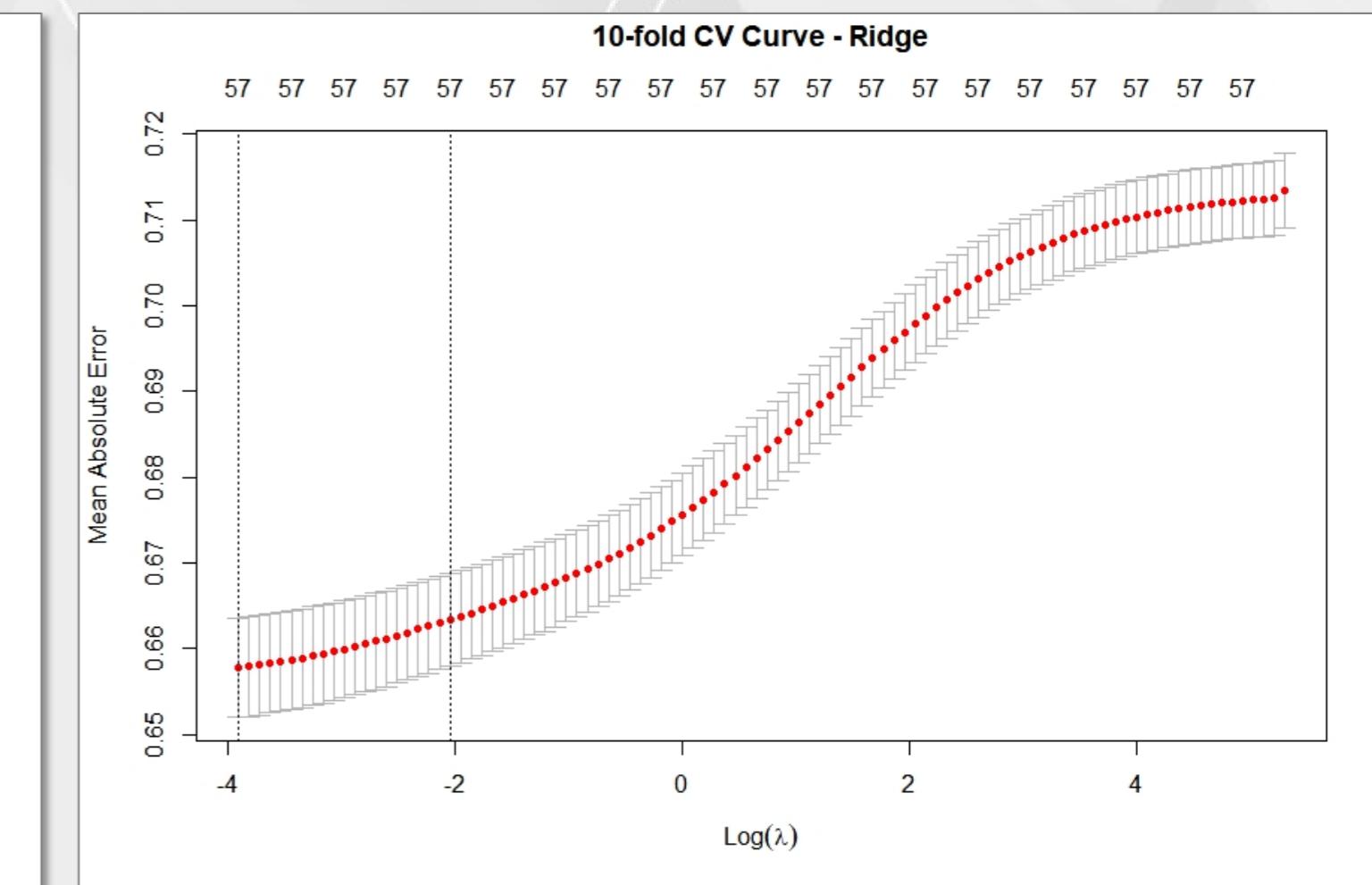
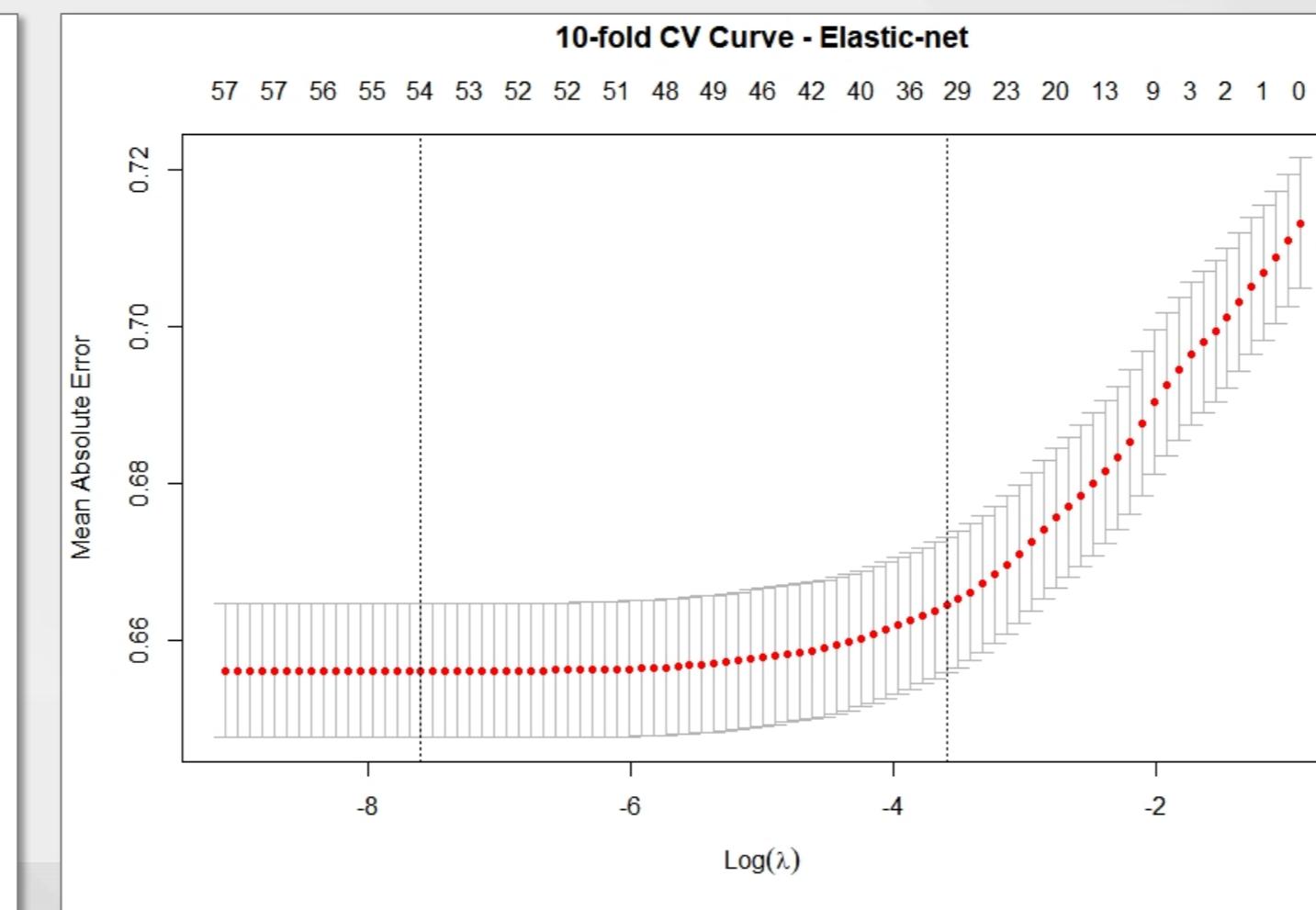
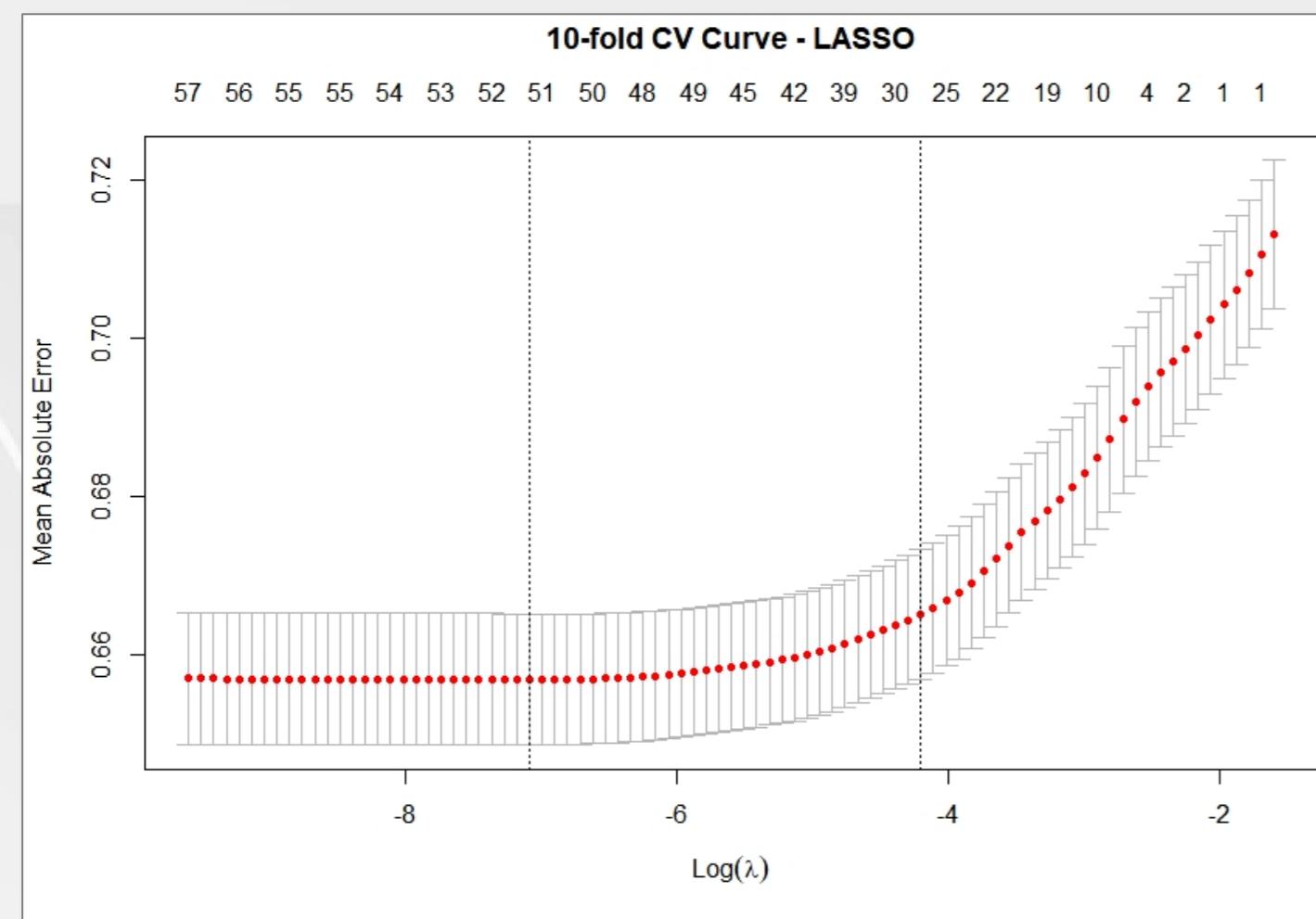
**LASSO Runtime 2:** 2.31 secs.

**EN Runtime 1:** 1.32 secs.

**EN Runtime 2:** 1.74 secs.

**Ridge Runtime 1:** 1.03 secs.

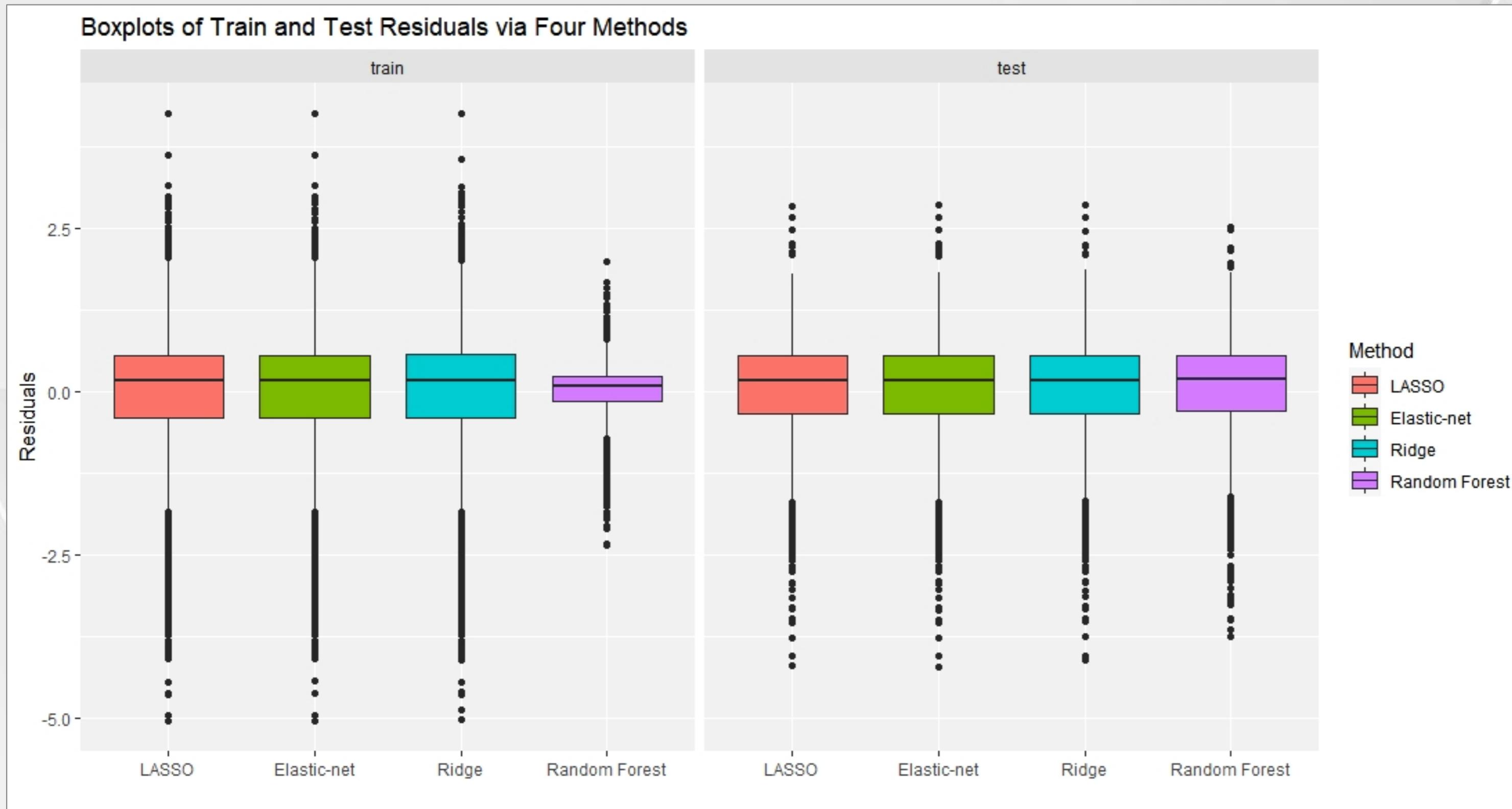
**Ridge Runtime 2:** ~ 26.61 hrs.



**Runtime 1:** Using 25% random sample dataset

**Runtime 2:** Using complete dataset (all 39,644 obs.)

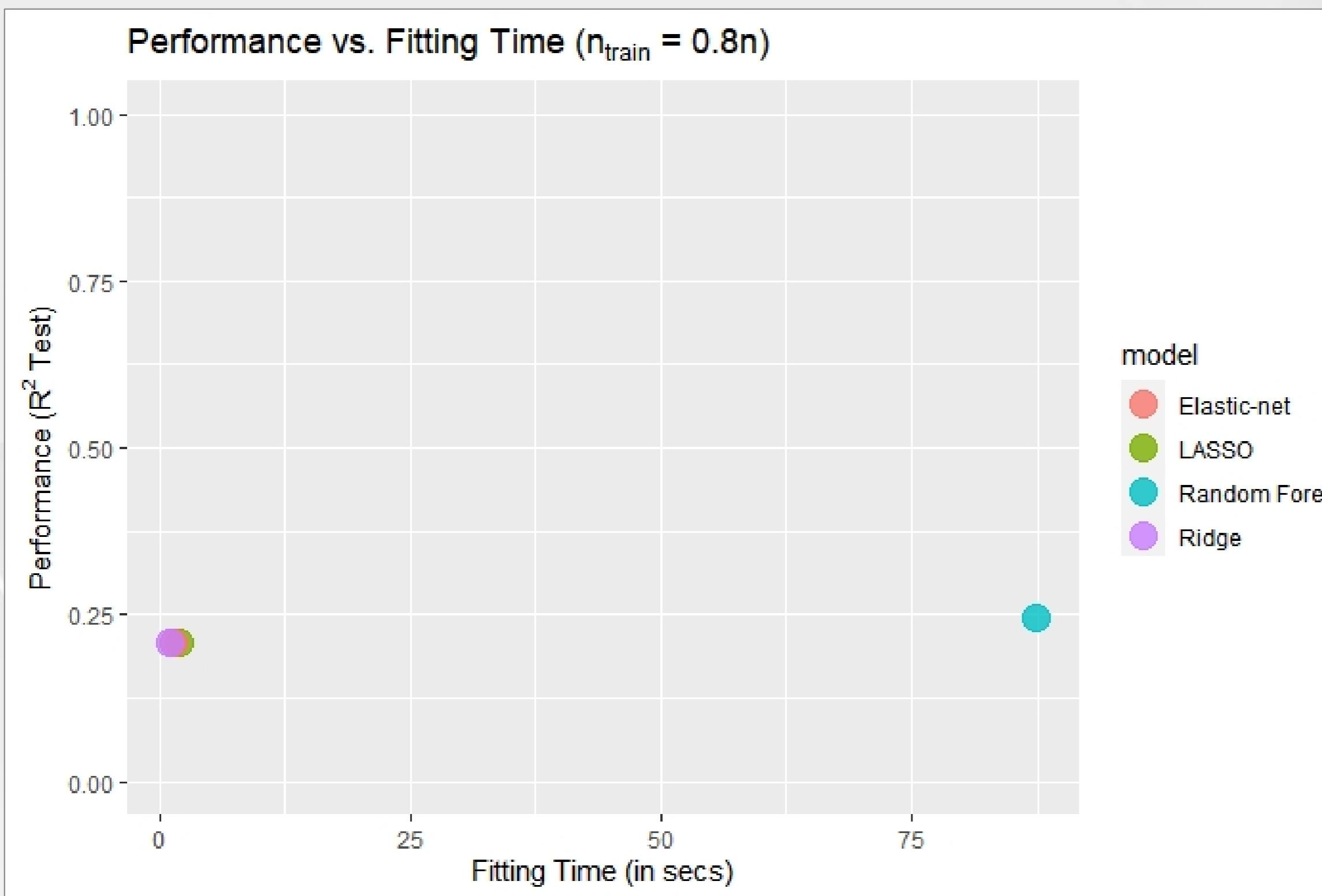
# BOXPLOTS OF RESIDUALS



## Observations:

- Residual means near zero
- LASSO, EN, & Ridge have similar residual variance in training and testing sets
- **Random Forest (RF)**
  - Has smaller residual variance when compared against LASSO, EN, and Ridge in training set
  - RF training variance is smaller than its testing variance
  - Has similar variance against other methods in testing set

# PERFORMANCE vs. RUNTIME

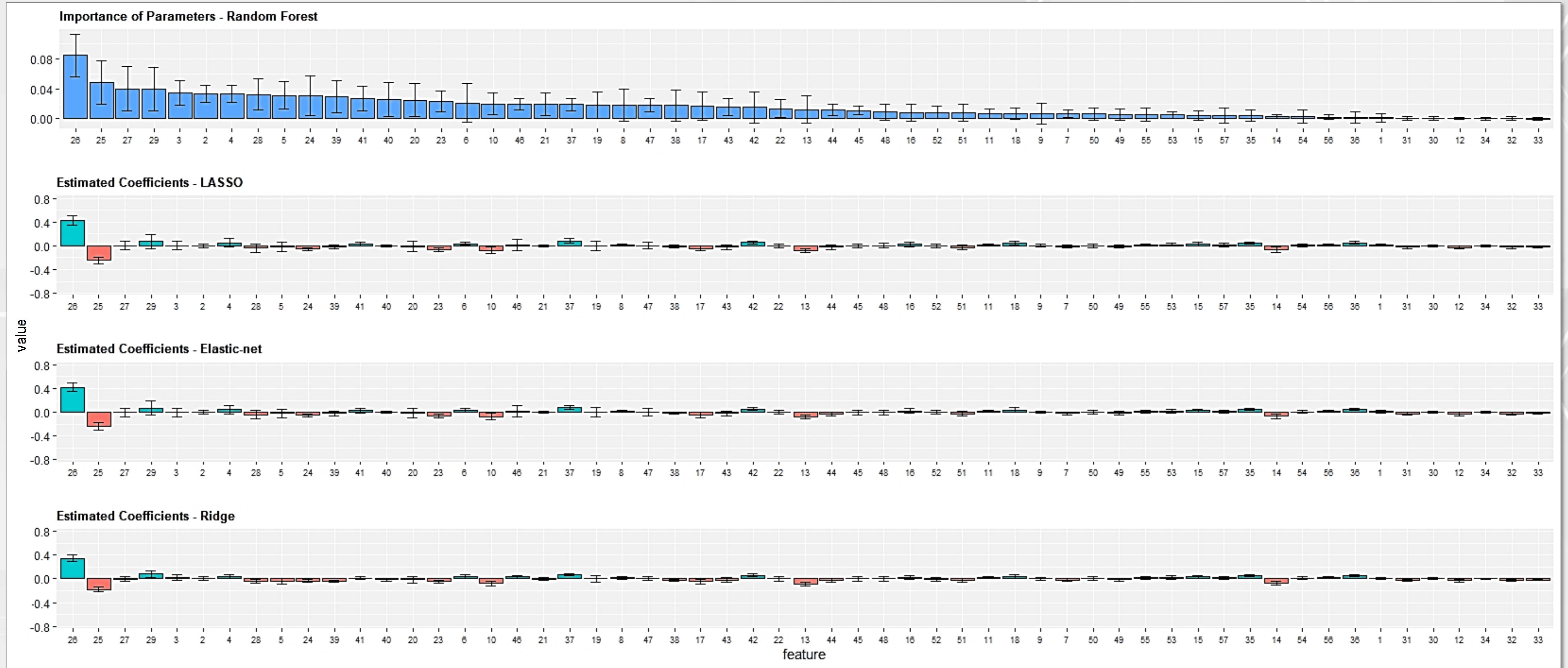


Method	Performance ( $R^2_{test}$ )	Runtime (in secs)	
		CV	Full 25%
LASSO	0.2066	1.98	1.01
Elastic-net	0.2068	1.32	0.93
Ridge	0.2091	1.03	1.17
Random Forest	0.2453	87.4	136.8

## Observations:

- Avg. Performance ( $R^2$ ): 0.217 (21.7%)
- Trade-off? *Slightly.*
  - RF consistently takes longer, but has improved performance
  - However, the additional runtime provides minimal performance improvement

# VARIABLE IMPORTANCE



# CLOSING REMARKS

## ONLINE NEWS POPULARITY | UCI ML REPOSITORY

### Variable Importance (Top 3)

- **Positive influence ↗**
  - 26 - *kw\_avg\_avg* (Avg. keyword (avg. shares))
  - 29 - *self\_reference\_avg\_shares*  
(Avg. shares of referenced articles in Mashable)
  - 37 - *LDA\_00* (Closeness to LDA topic 0)
- **Negative influence ↙**
  - 25 - *kw\_max\_avg* (Avg. keyword (max. shares))
  - 10 - *average\_token\_length* (Avg. length of words in content)
  - 13 - *data\_channel\_is\_entertainment* (Is data channel 'Entertainment'?)

Table 5: Ranking of features according to their importance in the RF model.

Feature	Rank (#)	Feature	Rank (#)
Avg. keyword (avg. shares)	0.0456 (1)	Closeness to top 1 LDA topic	0.0287 (11)
Avg. keyword (max. shares)	0.0389 (2)	Rate of unique non-stop words	0.0274 (12)
Closeness to top 3 LDA topic	0.0323 (3)	Article text subjectivity	0.0271 (13)
Article category (Mashable data channel)	0.0304 (4)	Rate of unique tokens words	0.0271 (14)
Min. shares of Mashable links	0.0297 (5)	Average token length	0.0271 (15)
Best keyword (avg. shares)	0.0294 (6)	Number of words	0.0263 (16)
Avg. shares of Mashable links	0.0294 (7)	Day of the week	0.0260 (18)
Closeness to top 2 LDA topic	0.0293 (8)	Number of words in the title	0.0161 (31)
Worst keyword (avg. shares)	0.0292 (9)	Number of images	0.0142 (34)
Closeness to top 5 LDA topic	0.0288 (10)	Number of videos	0.0082 (44)

### 25% Random Sample vs. Full Dataset Comparison

- **Results were generally the same**
  - Performance decreased, and Ridge / RF caused runtimes to substantially increase
  - CV curves had same shapes, while boxplot variances shrank

### Improvements can be made...

- RF (still the best performer), AdaBoost, SVM, kNN, NB
- ... but human behavior is unpredictable!
  - Therefore,  $R^2$  between 10 – 20% for social sciences is acceptable



# THANK YOU!

---

## Q&A



## **REFERENCES**

---

Fernandes, K., Vinagre, P., Cortez, P., & Sernadel, P., . (2013). *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*. Porto, Portugal; Braga, Portugal; Alveiro, Portugal: INESC TEC Porto/Universidade do Porto; ALGORITMI Research Centre, Universidade do Minho; Universidade de Aveiro.

Ozili, P. K. (2016, September 2). *What is The Acceptable R-Squared Value?*

Retrieved from ResearchGate: [https://www.researchgate.net/post/what\\_is\\_the\\_acceptable\\_r-squared\\_value](https://www.researchgate.net/post/what_is_the_acceptable_r-squared_value)

UCI Center for Machine Learning and Intelligent Systems. (2015, May 31). *Online News Popularity Data Set*.

Retrieved from UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/online+news+popularity#>