TOM'S HARDWARE: PREDICTING POPULARITY OF TECH TOPICS

Nika Kondzhariya & Julia Nebia

Project Description:

- **Predict popularity of the topic**

- **Use 4 techniques:**
  - Lasso
  - Ridge
  - Elastic-Net
  - Random Forest

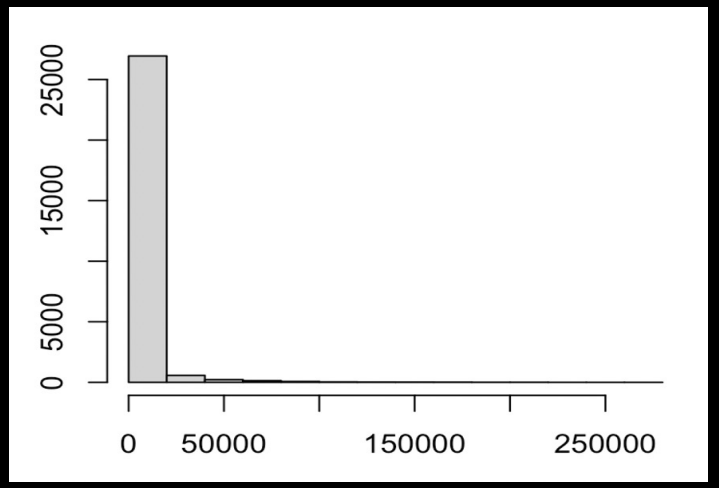
Data Description:

- n = 28,054

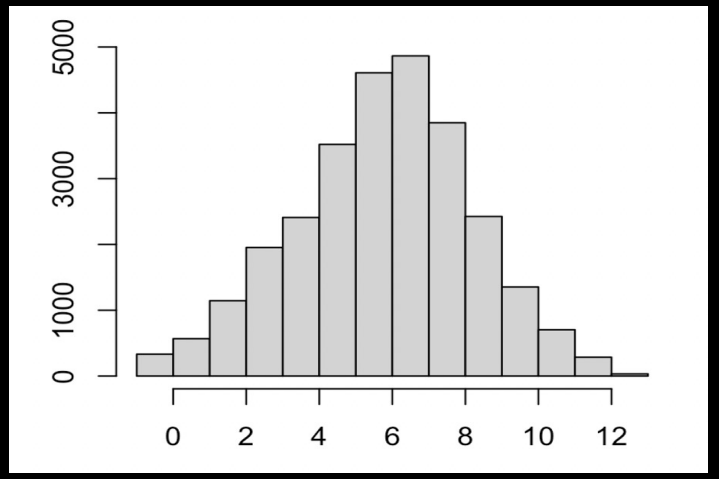
- p = 96 (all numeric, named X1-X96):
  - Number of discussions, posts, readers

- Response Variable (numeric, named Y):
  -average number of displays of the posts

Distribution of Y

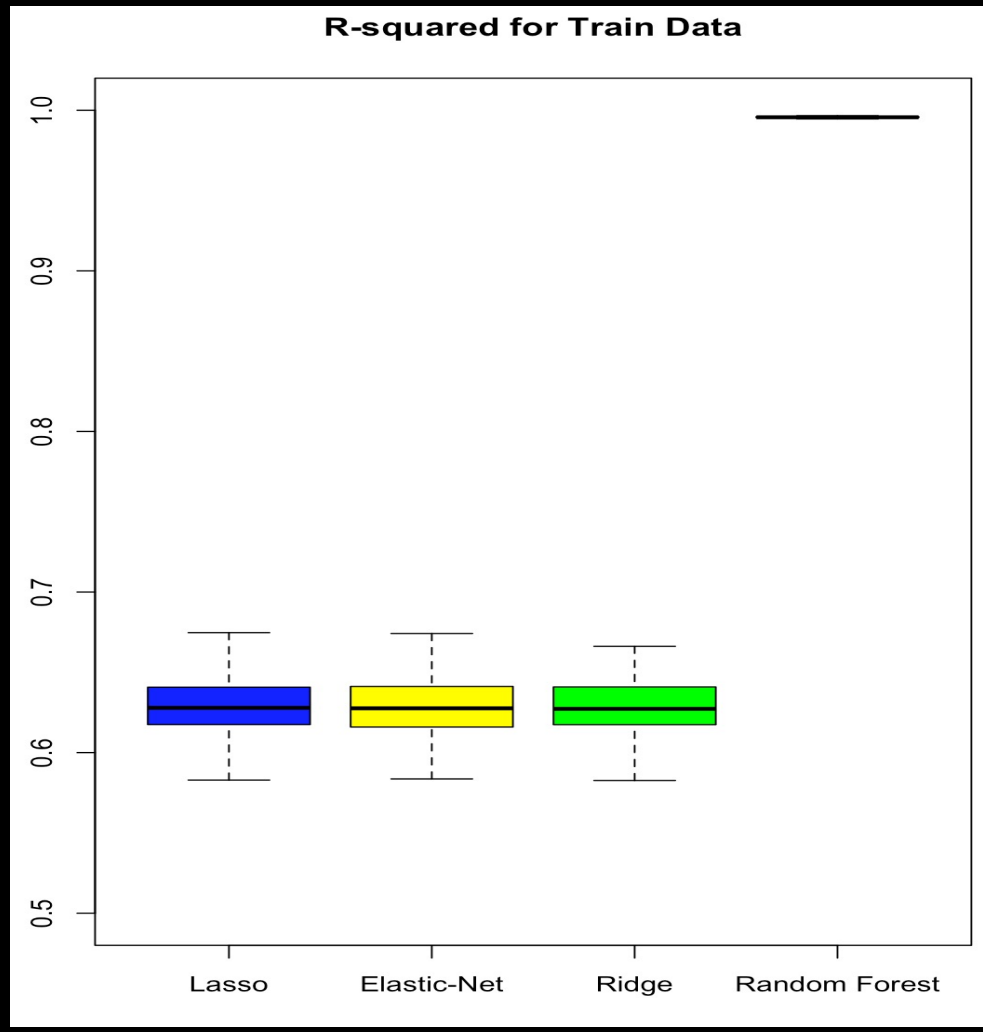


Distribution of log(Y)

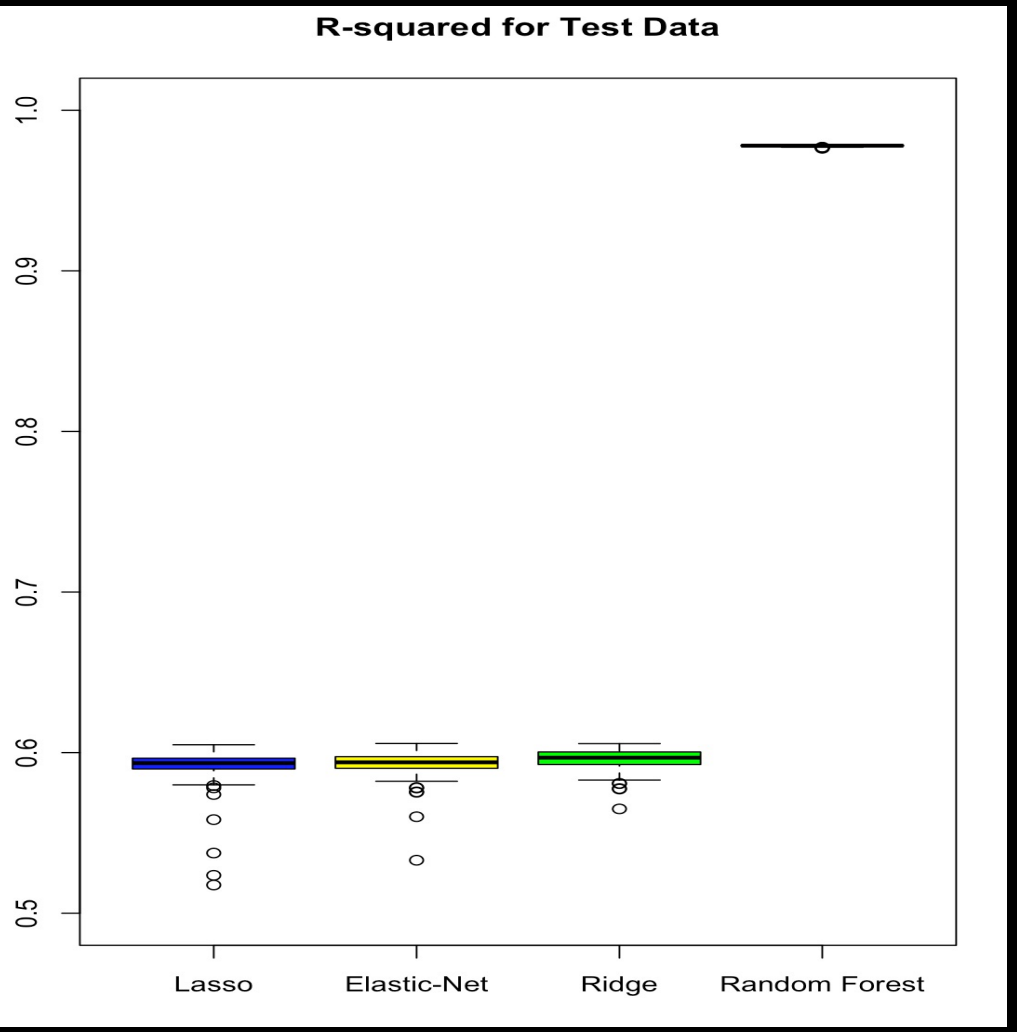


BOXPLOTS OF R-SQUARED

TRAIN DATA (n=1000)

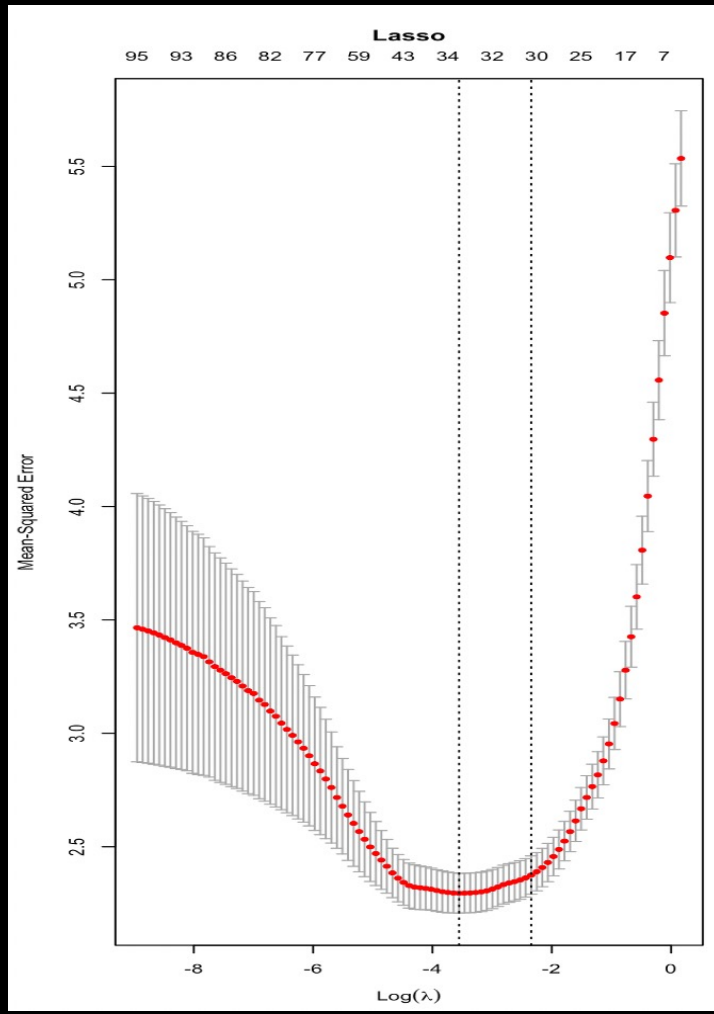


TEST DATA (n=27,054)



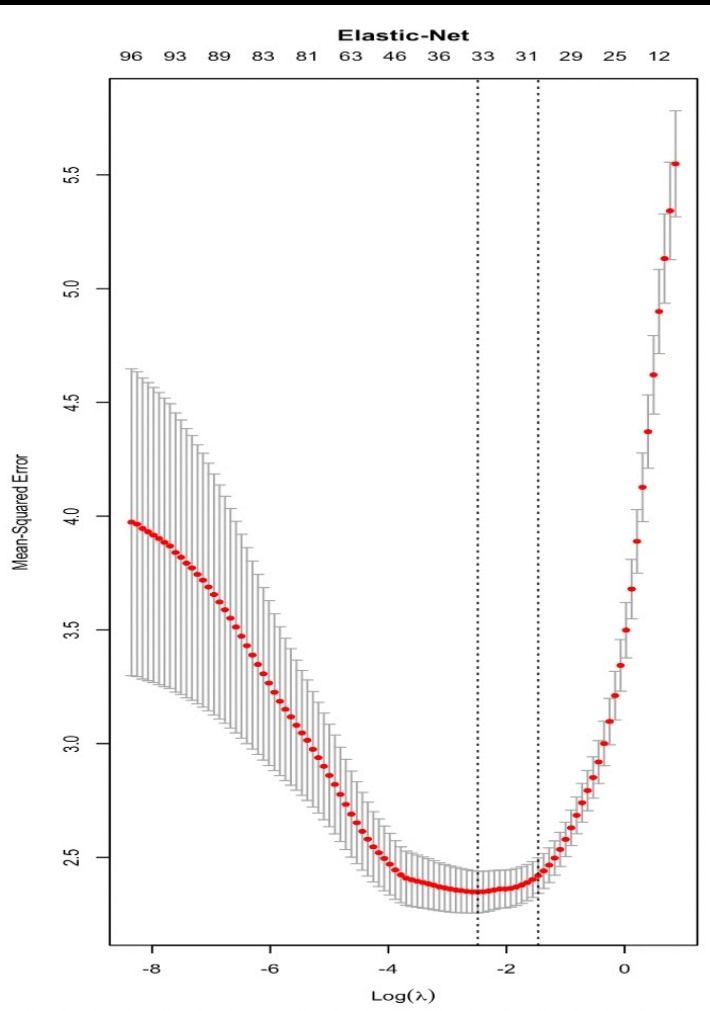
10-FOLD CV CURVES

0.404 sec



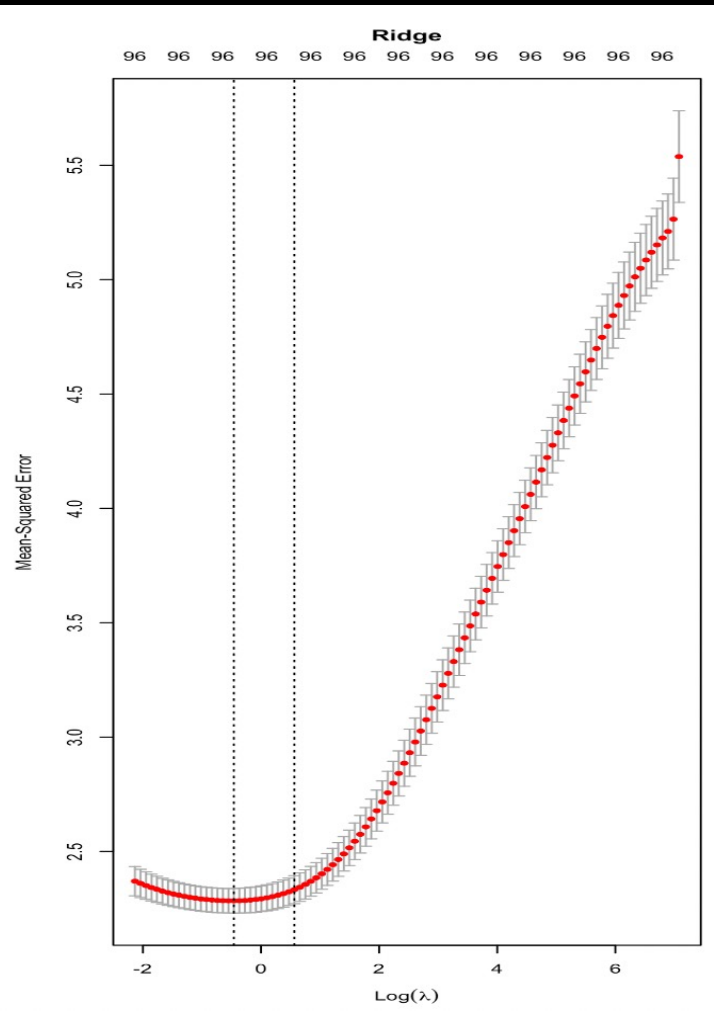
LASSO

0.425 sec



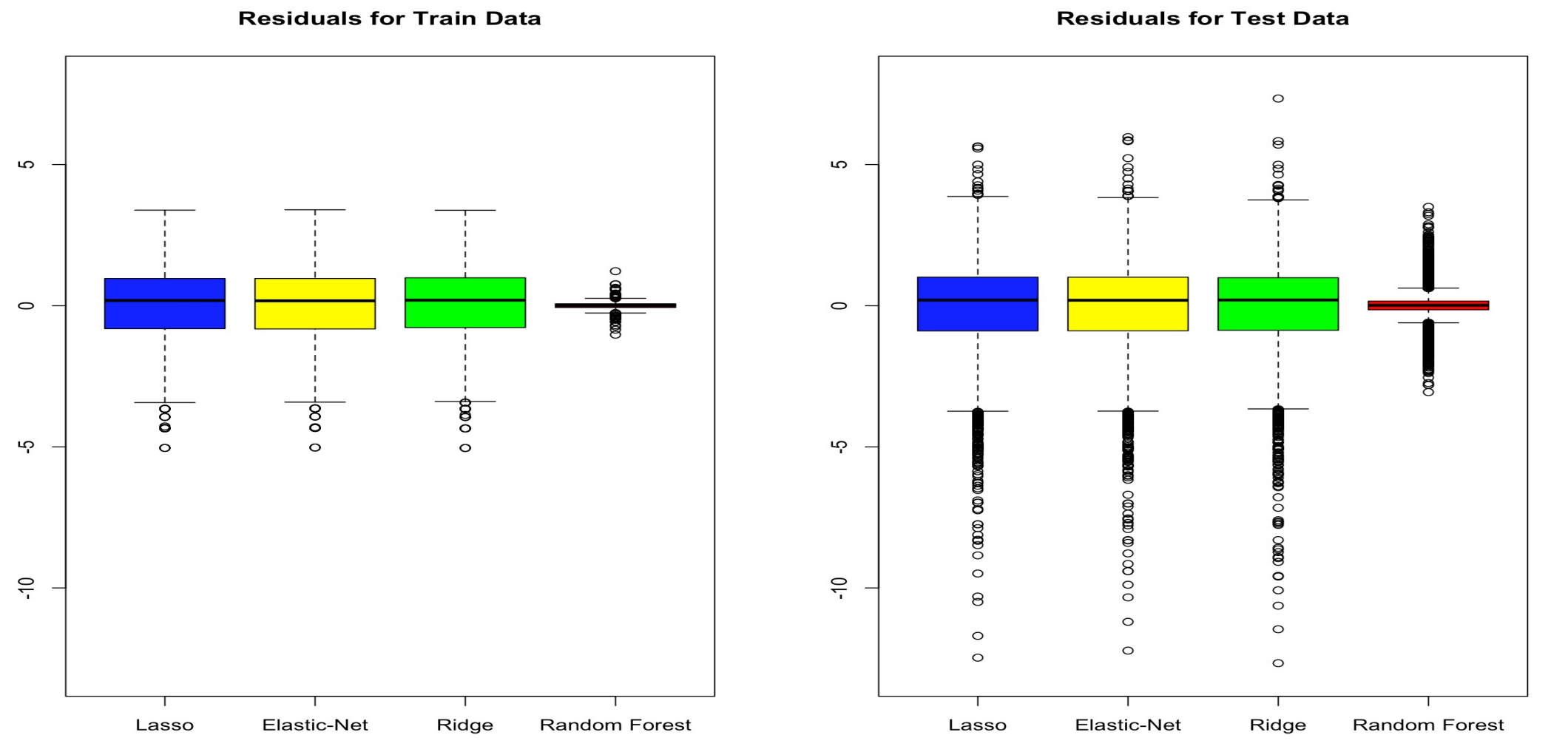
ELASTIC NET

0.204 sec

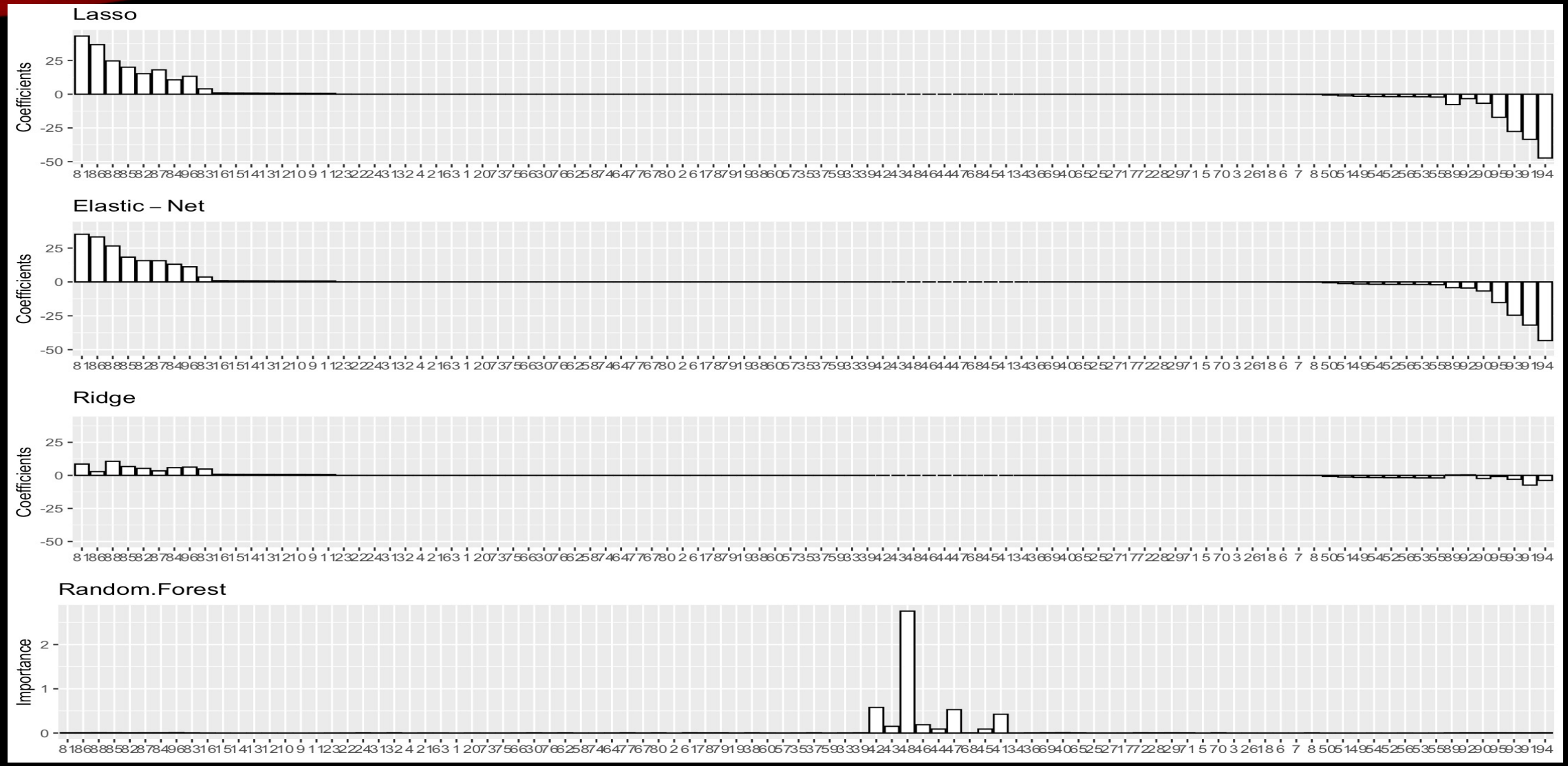


RIDGE

BOXPLOTS OF RESIDUALS



ESTIMATED COEFFICIENTS



MORE ON COEFFICIENTS

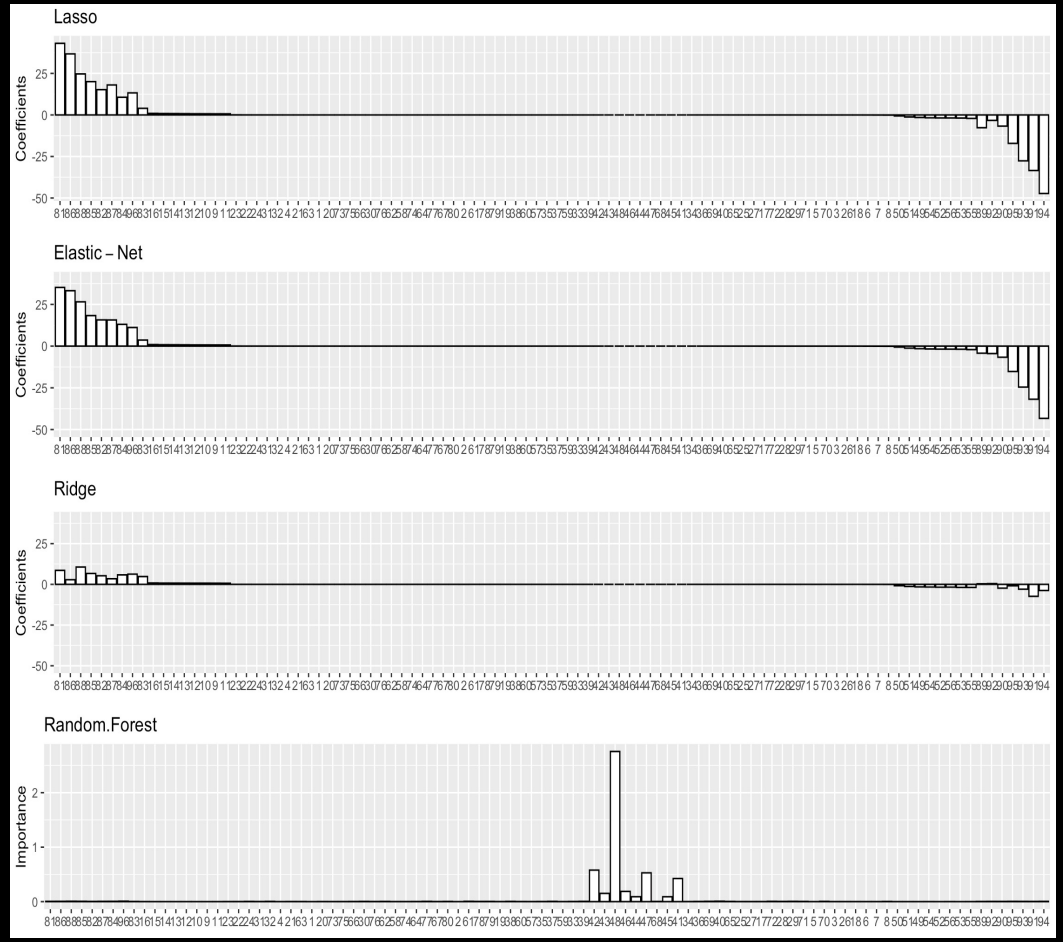
Lasso/Elastic-Net/Ridge:

- **Positive:**
  - The number of discussions
  - The average number of authors

- **Negative:**
  - The average length of a discussion
  - Measure of burstiness level for a topic

Random Forest:

- Number of new authors



RESULTS

	90% Interval for test R-square	Time to perform (entire dataset)
LASSO	[0.5778227, 0.6015839]	3.241 secs
Elastic-Net	[0.5781852, 0.6024422]	3.615 secs
Ridge	[0.5828034, 0.6027388]	3.813 secs
Random Forest	[0.9770214, 0.9785912]	16.247 mins

Trade-off: The better performance – the more time required to build

Thank you for attention! Do you have any questions?