STA 9890 Project

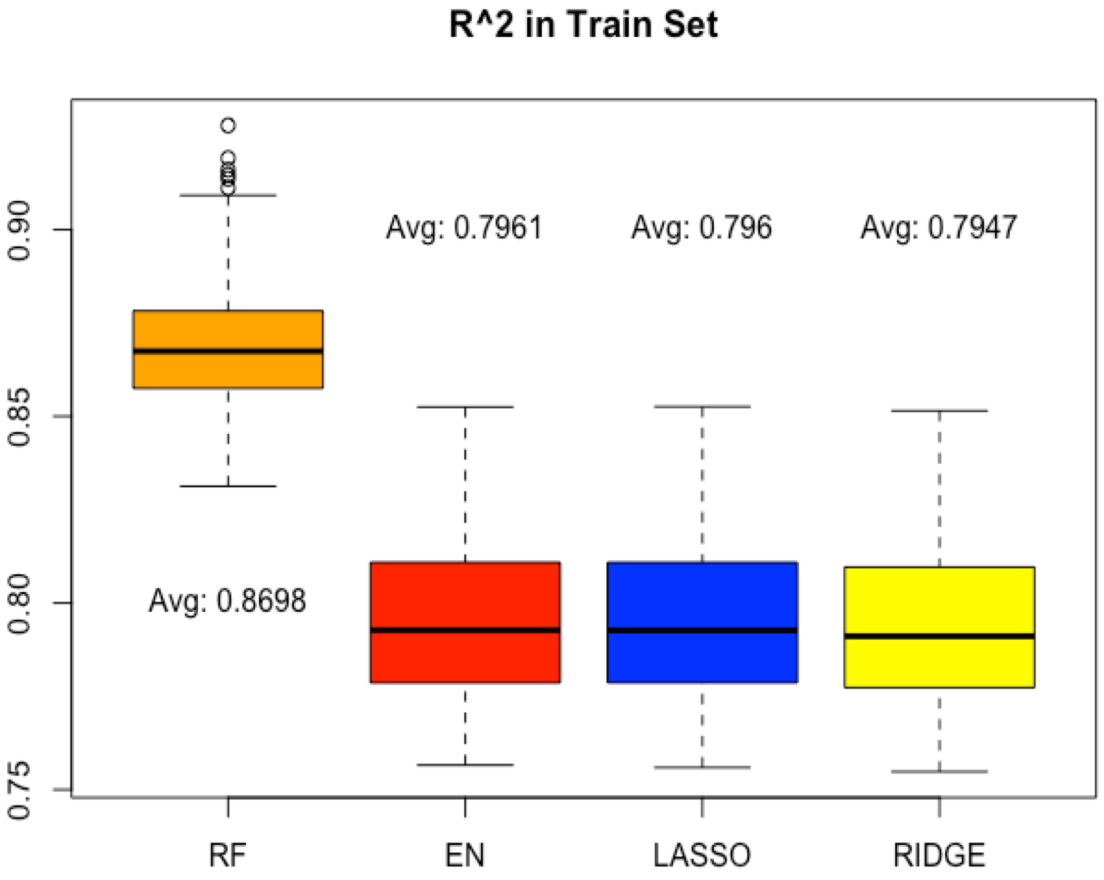
Jungkyng Lim

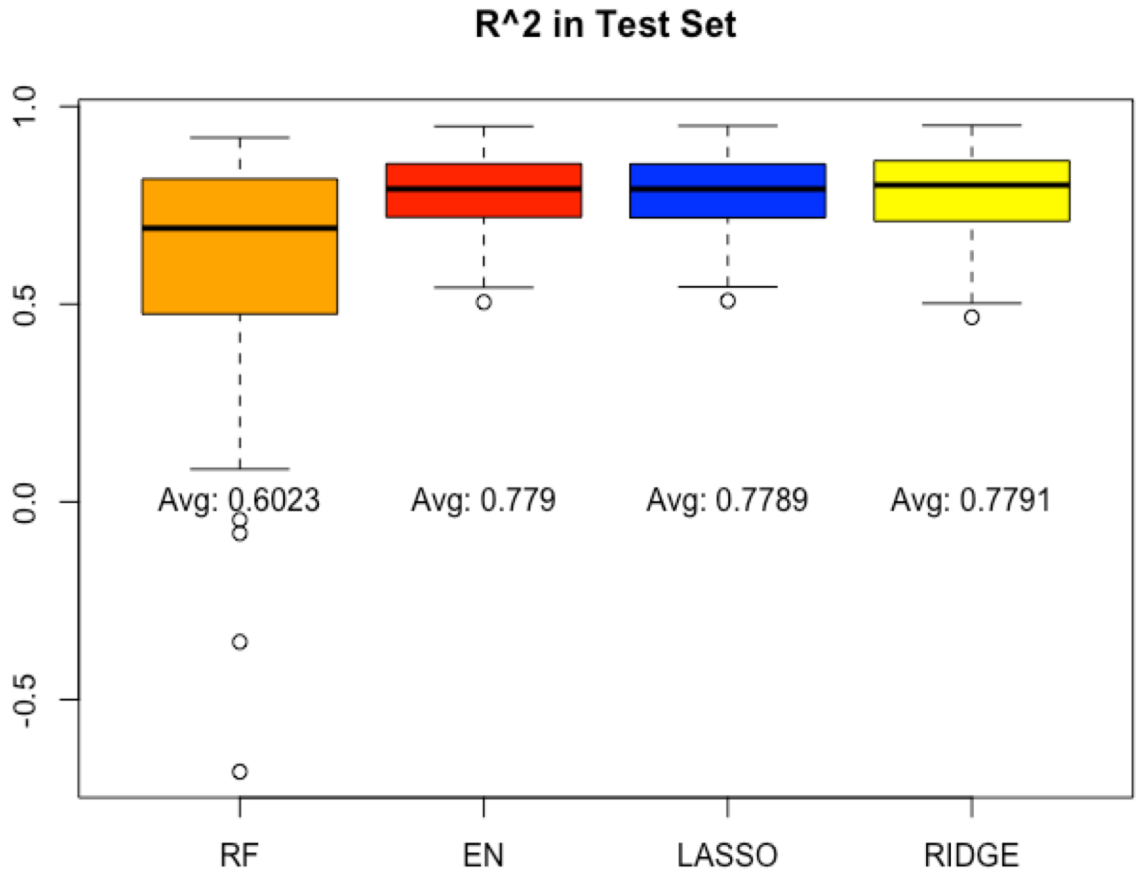
EMPID# 15000106

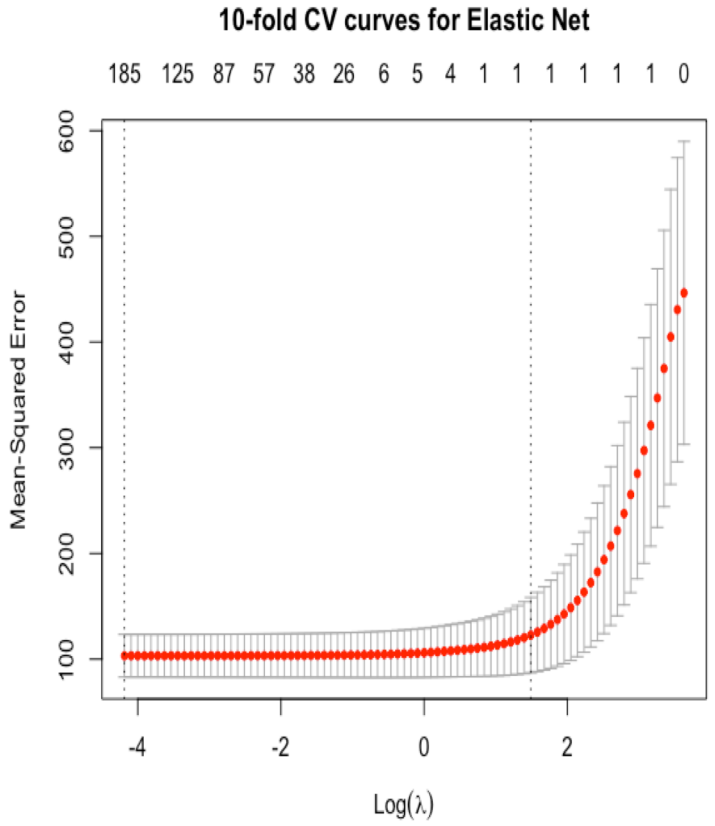
Data

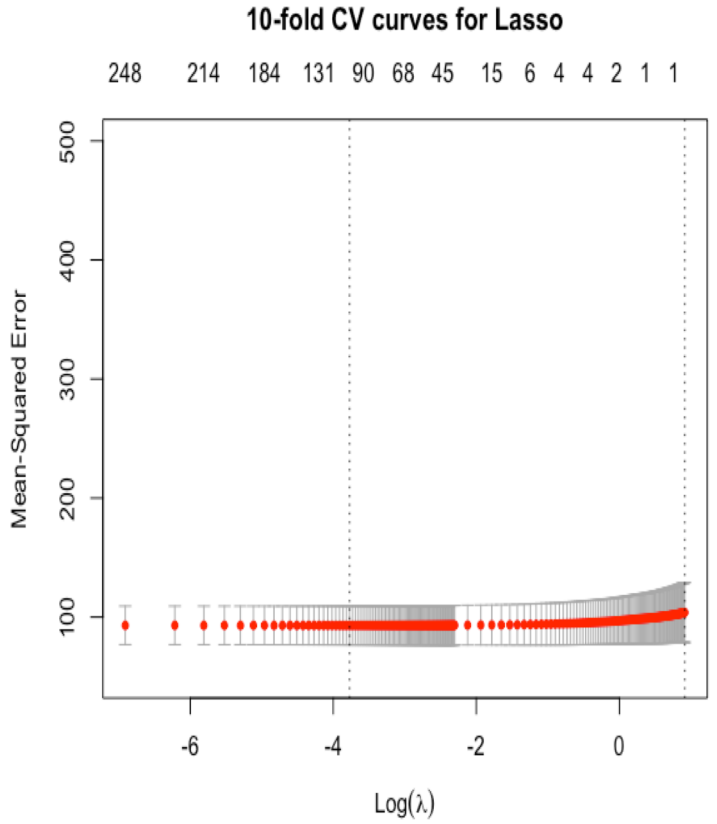
- Blog Feedback Dataset
  - http://archive.ics.uci.edu/ml/datasets/BlogFeedback
- Response Variable : Number of comments in the next 24 hours
- Predictors : 280 predictors
  - Statistics of overall attributes (AVG, SD, Min, Max, Med)
  - Each post's characteristics (Length, frequent words(word bags), Day of post, etc..)
  - Relationship of Parent Blog post, if any
- Shape of Data :
  - N = 52,397
  - P = 280

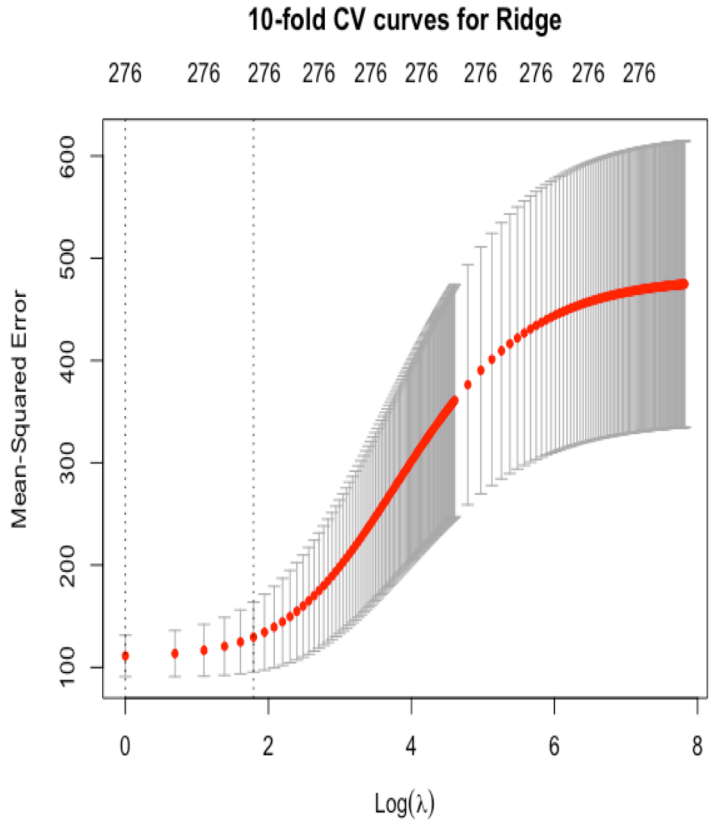
Boxplots of R^2_{test}, R^2_{train}







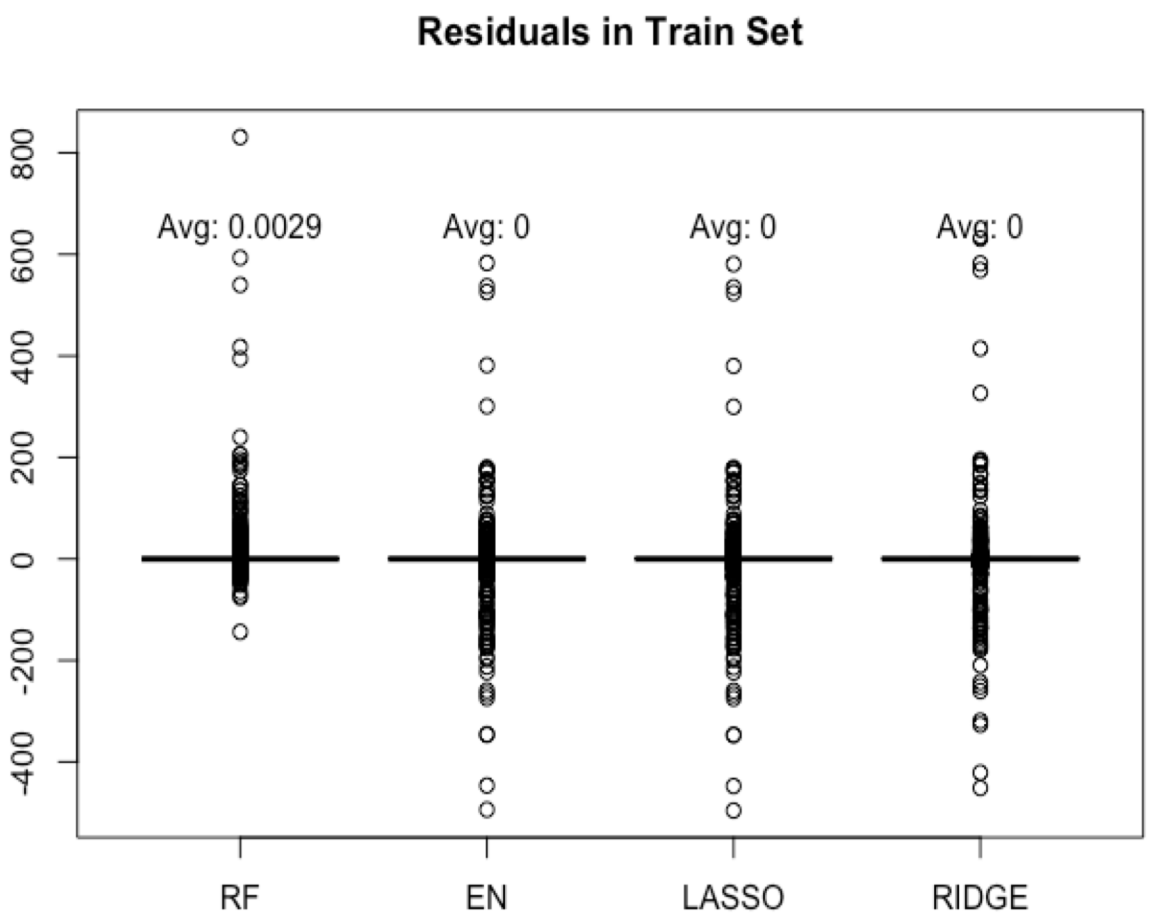


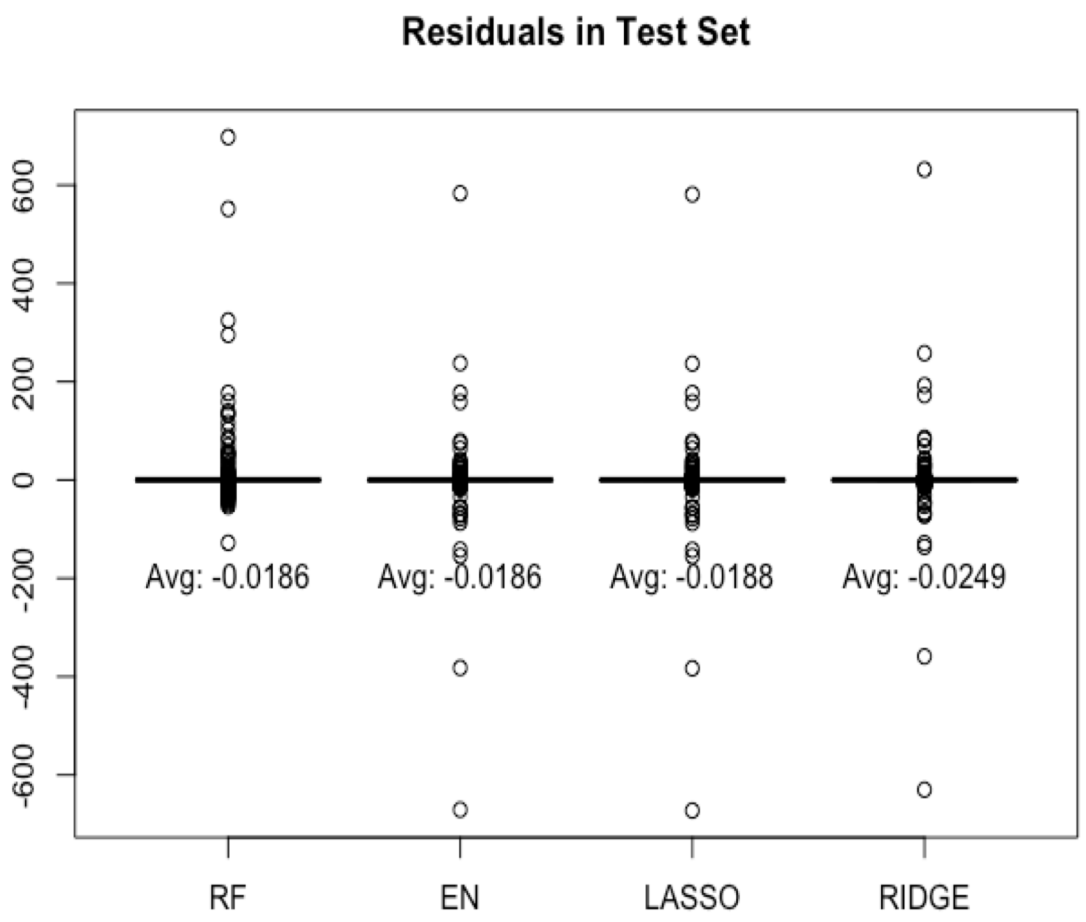


10-fold CV curves for lasso, elastic-net, ridge

- 10-fold CV MSE for models. First dotted vertical line in each plot represents the λ with the smallest MSE and the second represents the λ with an MSE within one standard error of the minimum MSE.

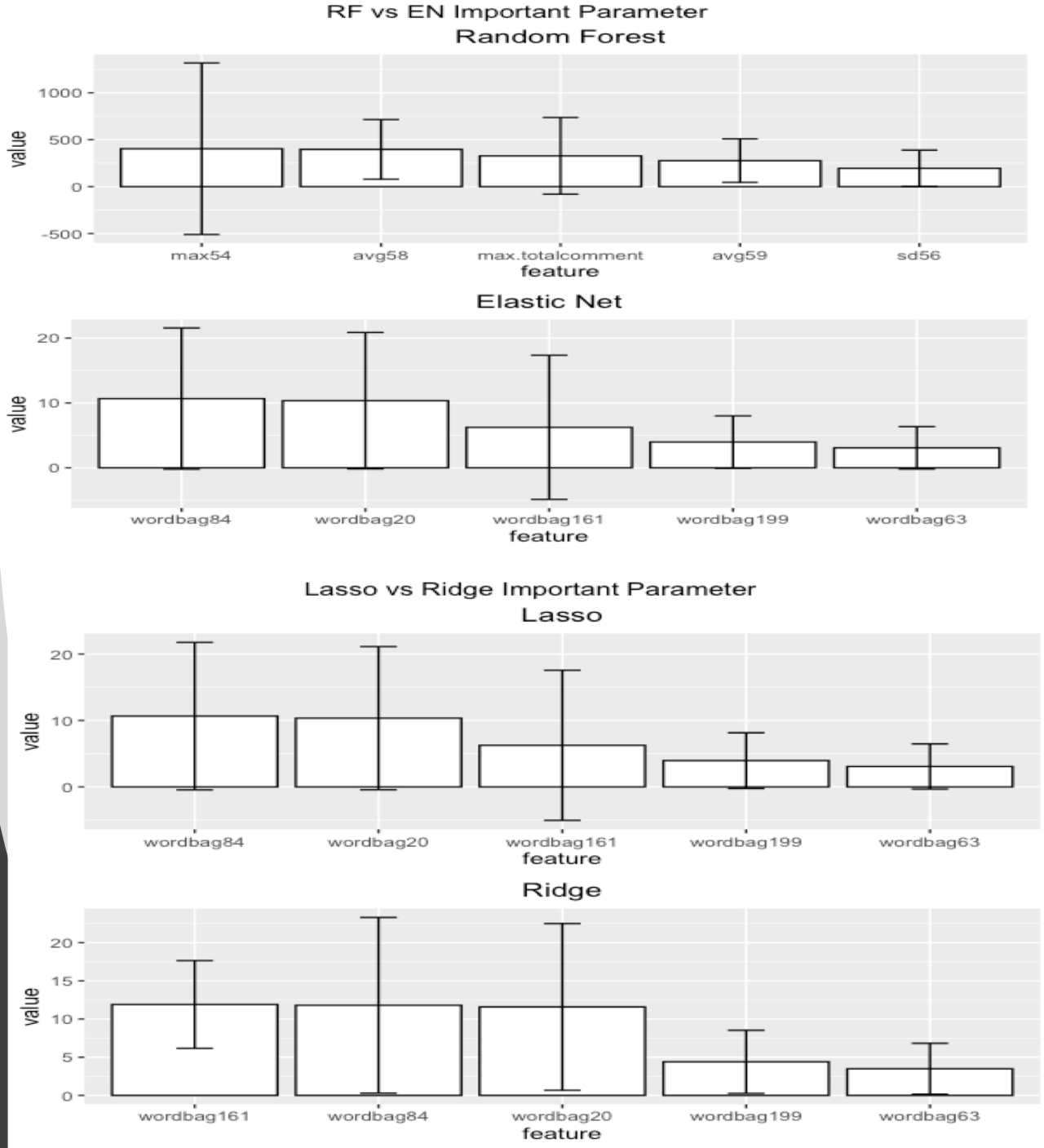
Residuals


Residuals in Train Set
800
600
400
200
0
-200
-400
Avg: 0.0029
Avg: 0
Avg: 0
Avg: 0
RF
EN
LASSO
RIDGE


Residuals in Test Set
600
400
200
0
-200
-400
-600
Avg: -0.0186
Avg: -0.0186
Avg: -0.0188
Avg: -0.0249
RF
EN
LASSO
RIDGE

Estimated coefficients and parameter

- Statistical parameters are important in Random Forest
- Wordbags – what kind of word is contained in the article is important
- Relationship with Parent pages is low



Summary

- Which one performance the best?
  - Random Forest is overfitting.
  - Elastic Net performs the best in R^2 and similar to lasso and ridge
- Time

Model Type	Time for Training for 1 times	Time for Training 100 times
Ridge	31.091 sec	52.57 min
Lasso	28.860 sec	47.23 min
Elastic-Net	25.967 sec	52.25 min
Random Forest	1726.375 sec	589.16 min

* Random Forest is without CV and ntree with 100

- Next Step – What will be in the Wordbag