Comparison of Regression Methods Using Facebook Data

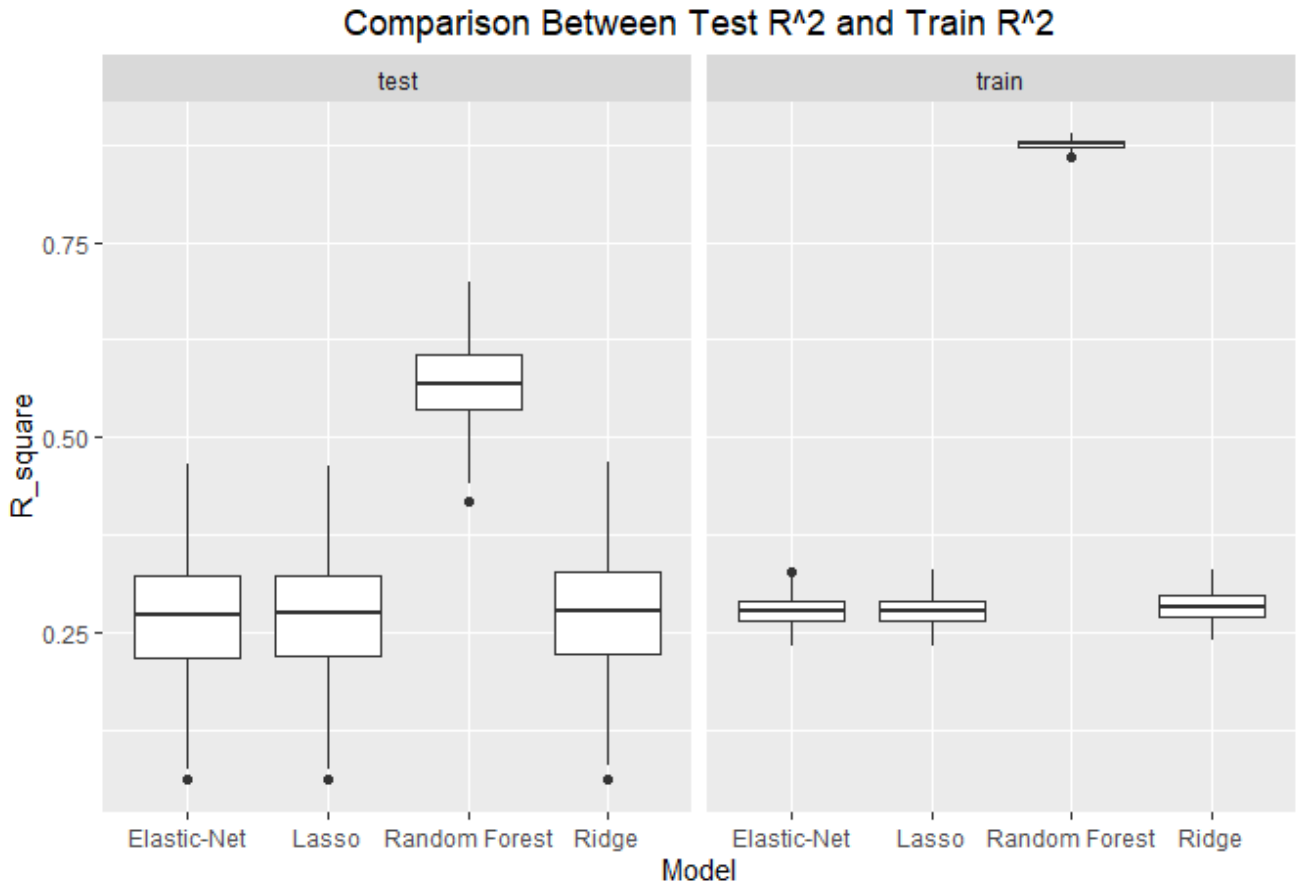
Danny Huang

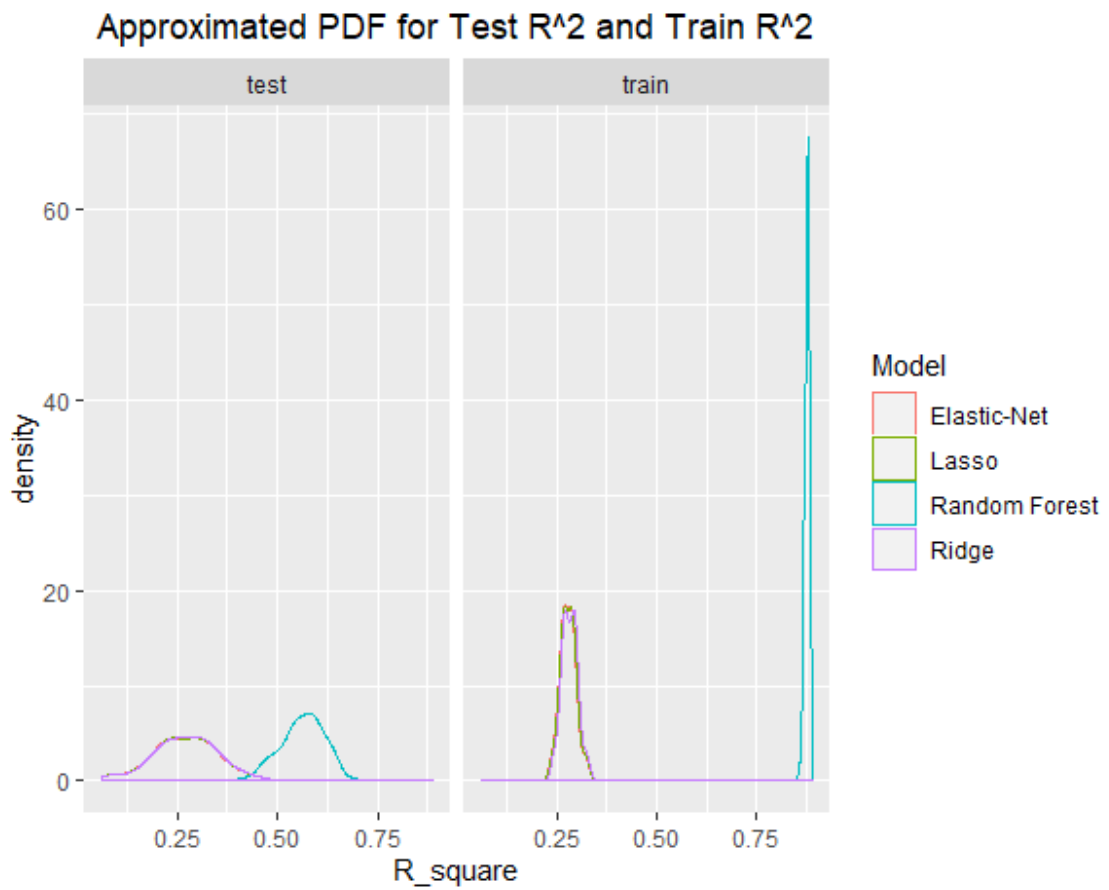
Methods Used: Ridge, Lasso, Elastic Net, Random Forest (15-trees)

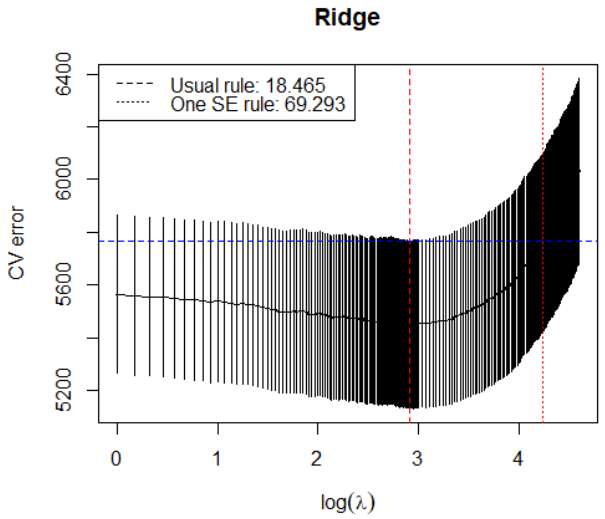
Number of Observation: 22498

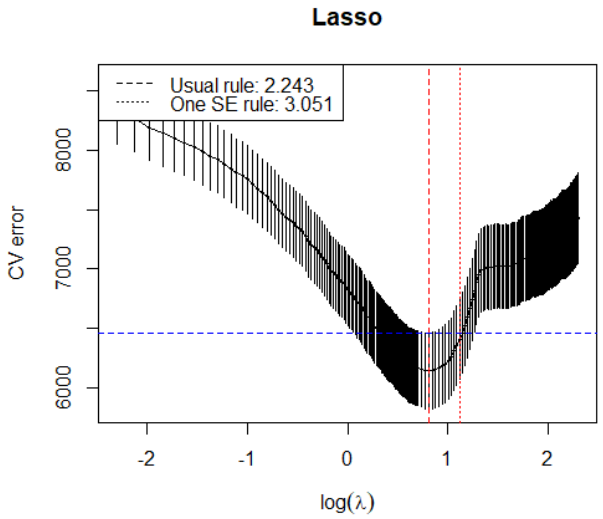
Number of Features: 52

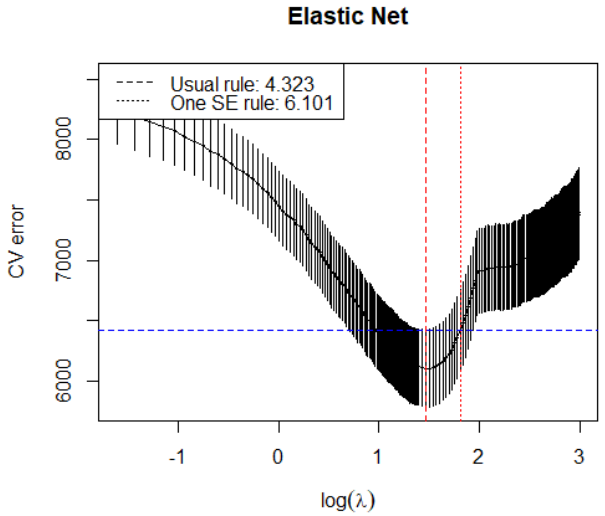
Response Variable: The Number of Comments in a Facebook Post

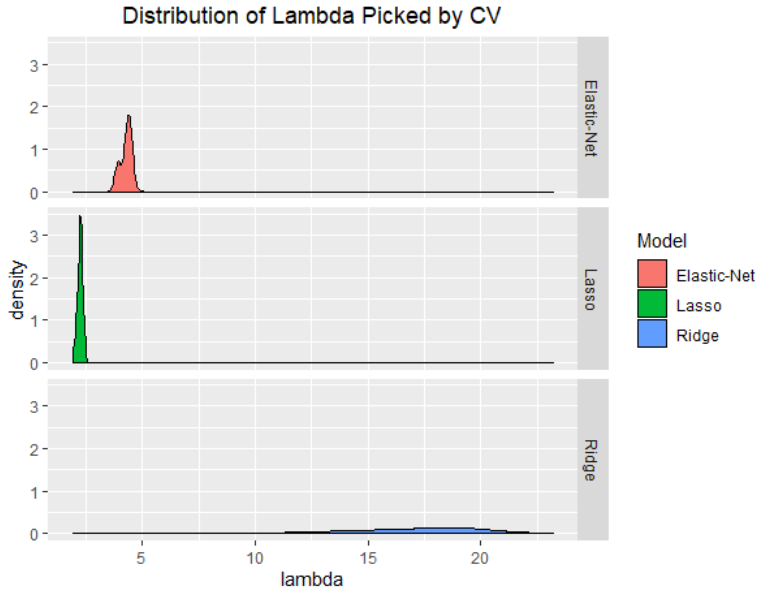

Comparison Between Test R^2 and Train R^2
test
train
0.75
0.50
0.25
R_square
Elastic-Net
Lasso
Random Forest
Ridge
Model


Approximated PDF for Test R^2 and Train R^2
test
train
60
40
20
0
density
0.25
0.50
0.75
R_square
Model
Elastic-Net
Lasso
Random Forest
Ridge



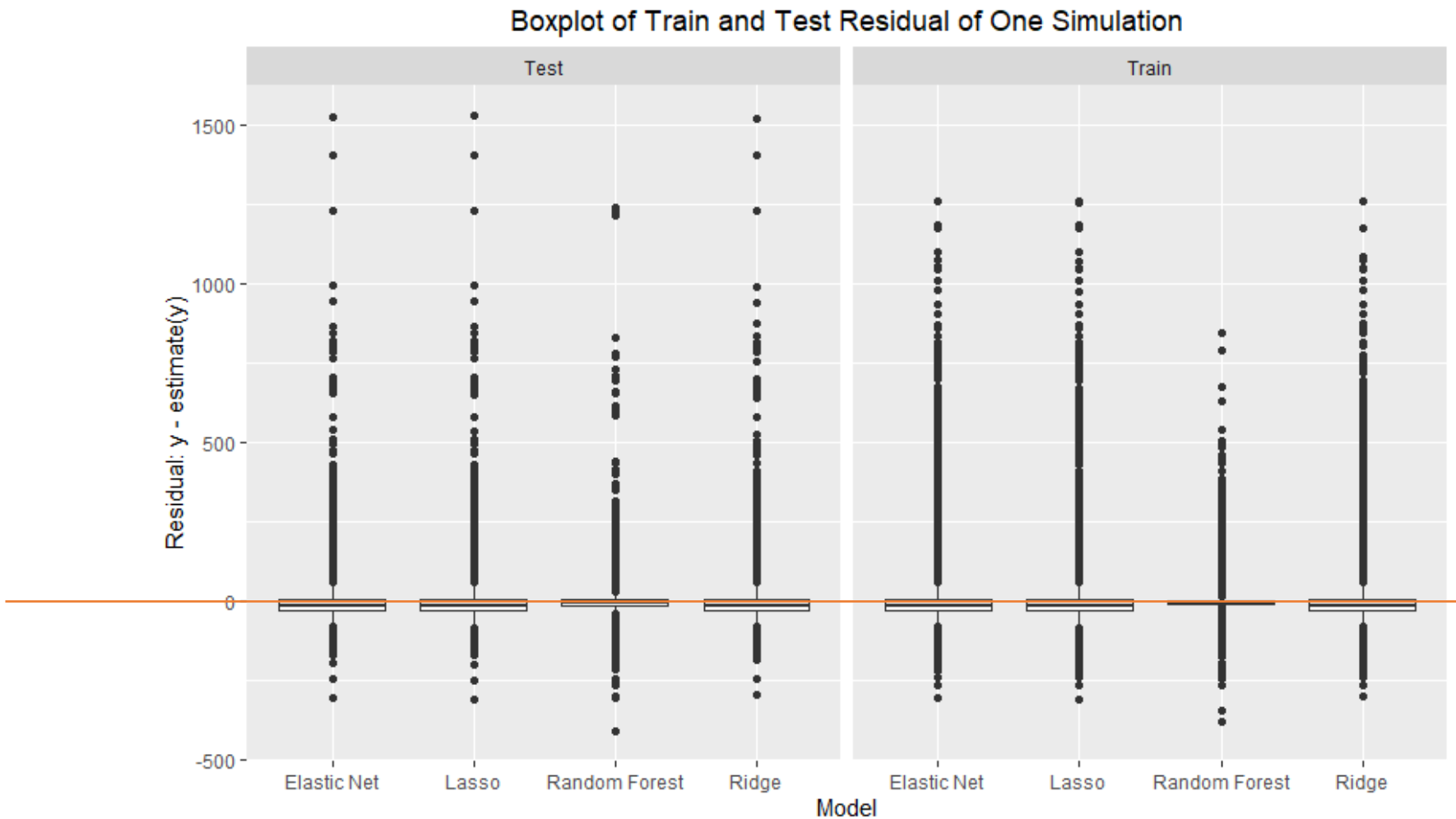


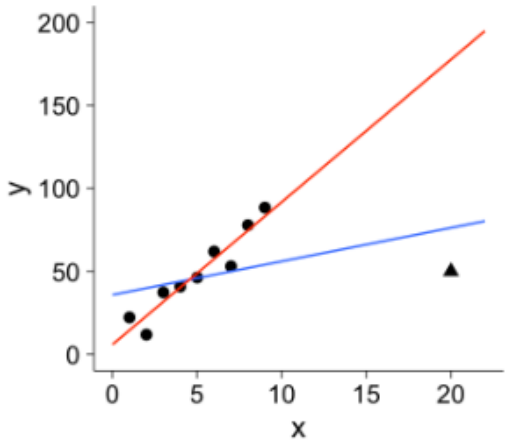




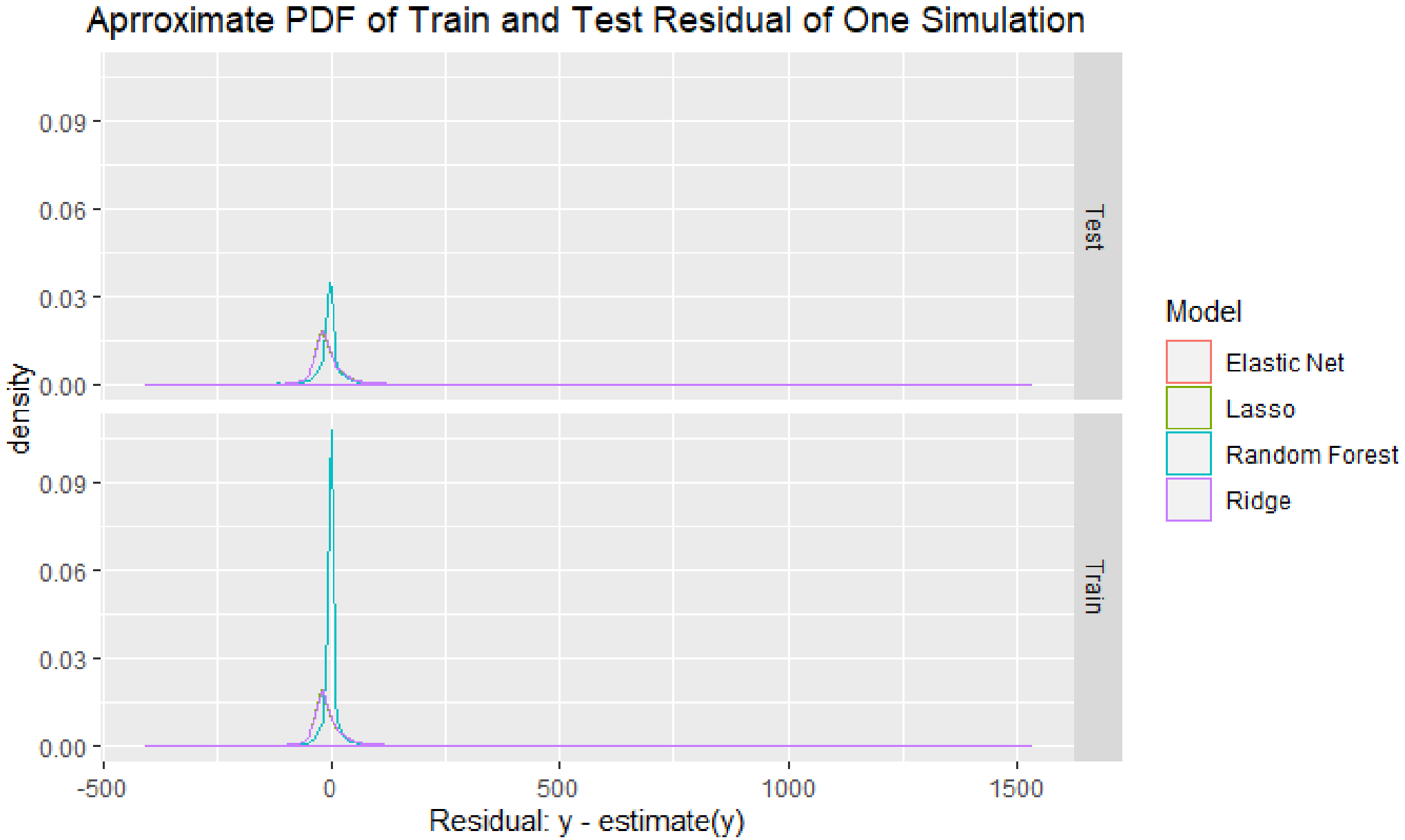
Average Lambda Chosen By 10-fold CV:

Ridge: 16.8
Lasso: 2.25
Elastic-Net: 4.29

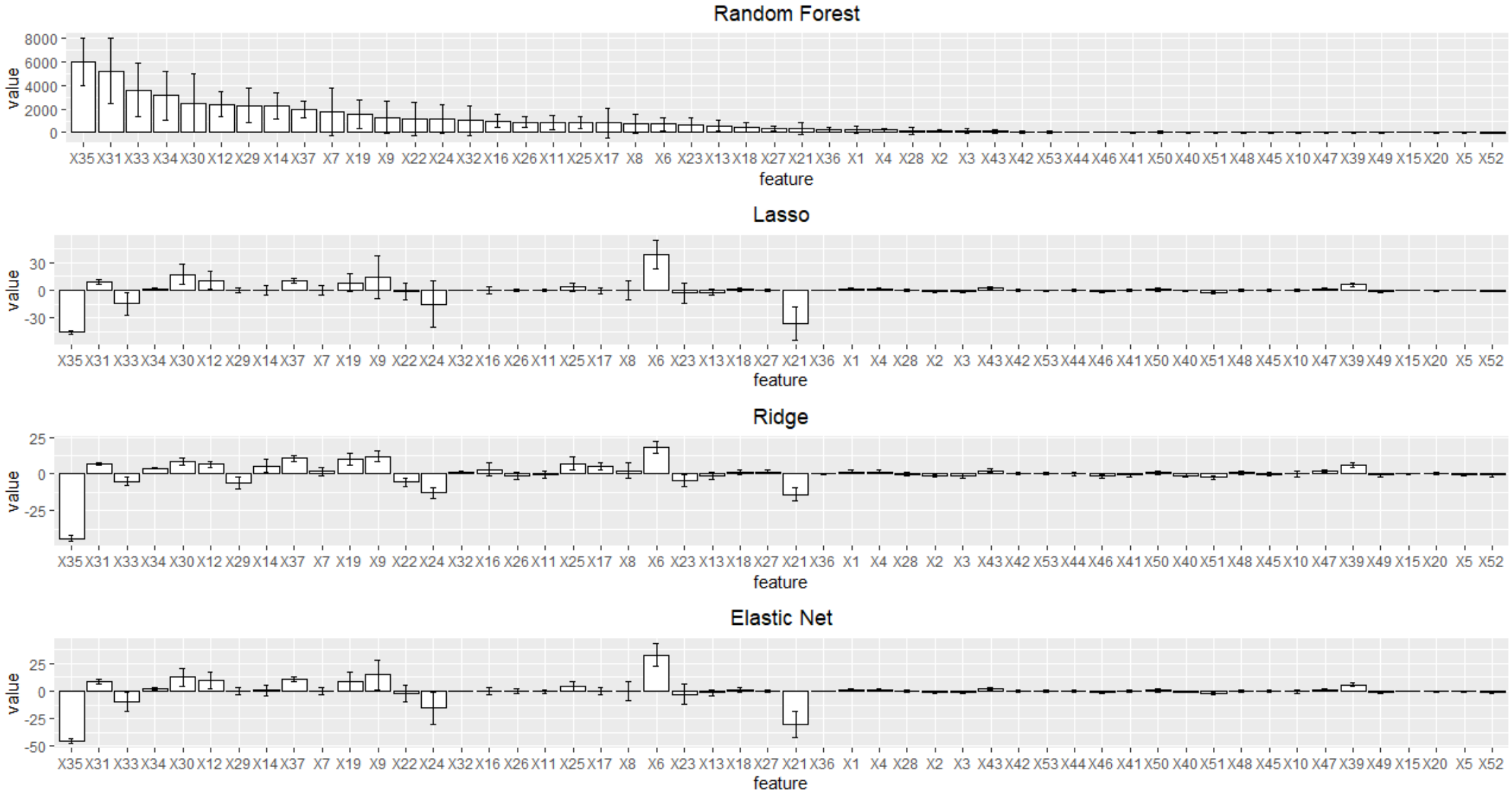




Except Random Forest, Other models seem to have a mean for residuals below 0


Aprroximate PDF of Train and Test Residual of One Simulation
density
0.09
0.06
0.03
0.00
Test
Train
-500
0
500
1000
1500
Residual: y - estimate(y)
Model
Elastic Net
Lasso
Random Forest
Ridge

The Coefficient Estimates with Bootstrapped Error bar



Average Time per Simulation with 22980 observations

Ridge: 0.510 seconds in 10-fold CV

Lasso: 0.501 seconds in 10-fold CV

Elastic-Net: 0.592 seconds in 10-fold CV

Random Forest (15-trees) = 12 seconds

For random forest, the computation is significantly larger than Ridge, Lasso, and Elastic Net regressions.