

STA 9890 HOTEL SCORE PREDICT PROJECT

ZILING WAN

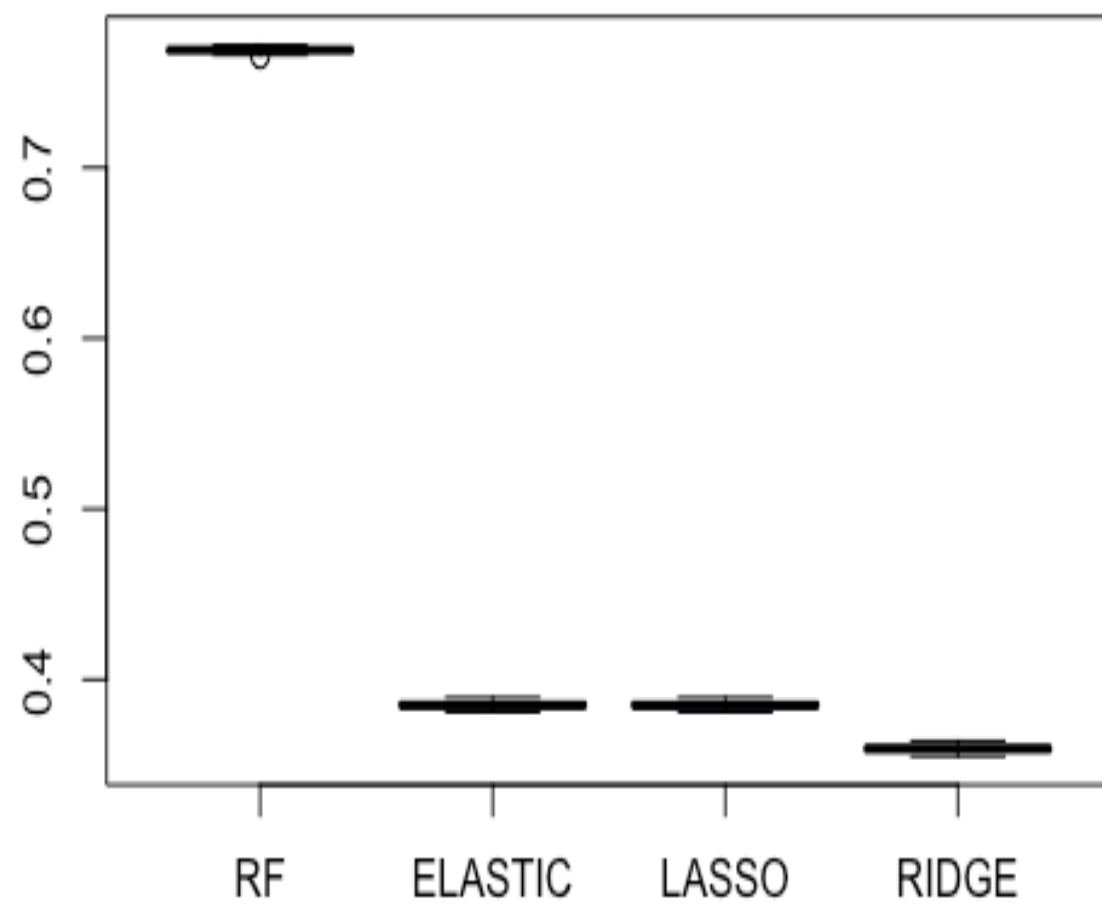
Github: <https://github.com/ZWan7/ZILING-WAN---STA9890>

DESCRIPTION OF DATA

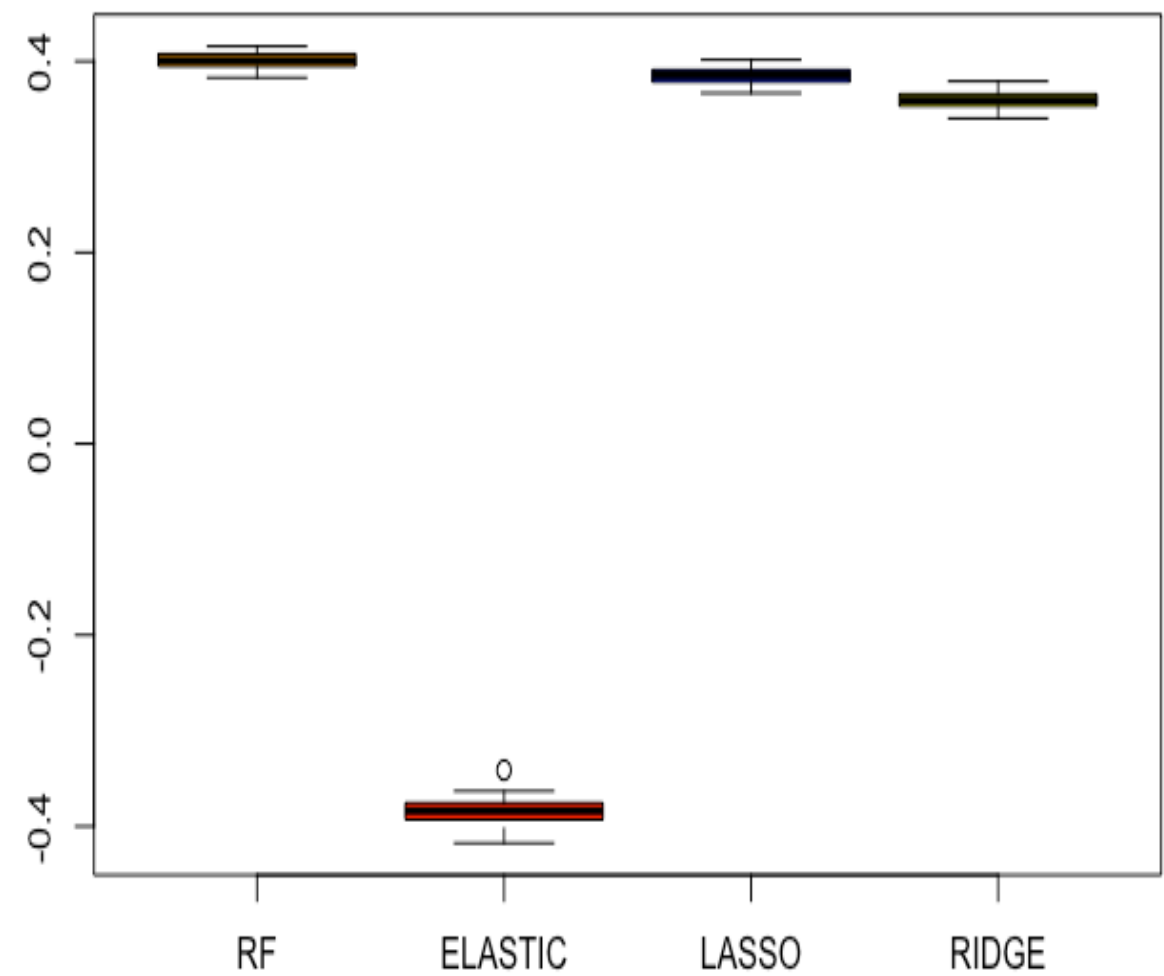
- ▶ Source of Dataset
 - ▶ <https://www.kaggle.com/ycalisar/hotel-reviews-dataset-enriched/data#>
- ▶ Observations: 65534
- ▶ Features: 41 (10 after processing)
- ▶ Response : Hotel Score
- ▶ Predictors: # of positive words, Average score, Review positivity rate, etc.
- ▶ Regression Methods: Elastic Net, Ridge, Lasso, Random Forest

BOX PLOT OF R SQUARE

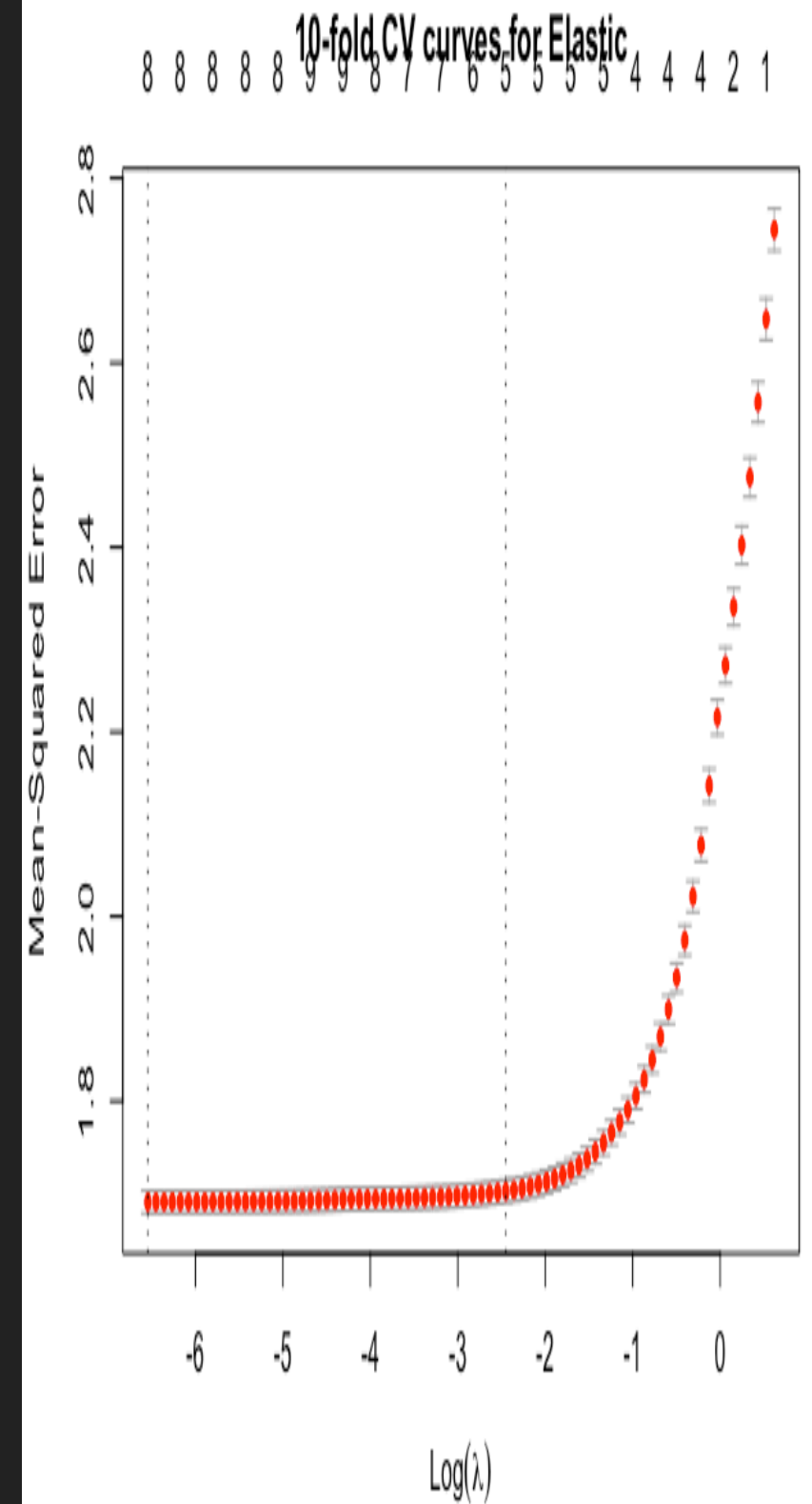
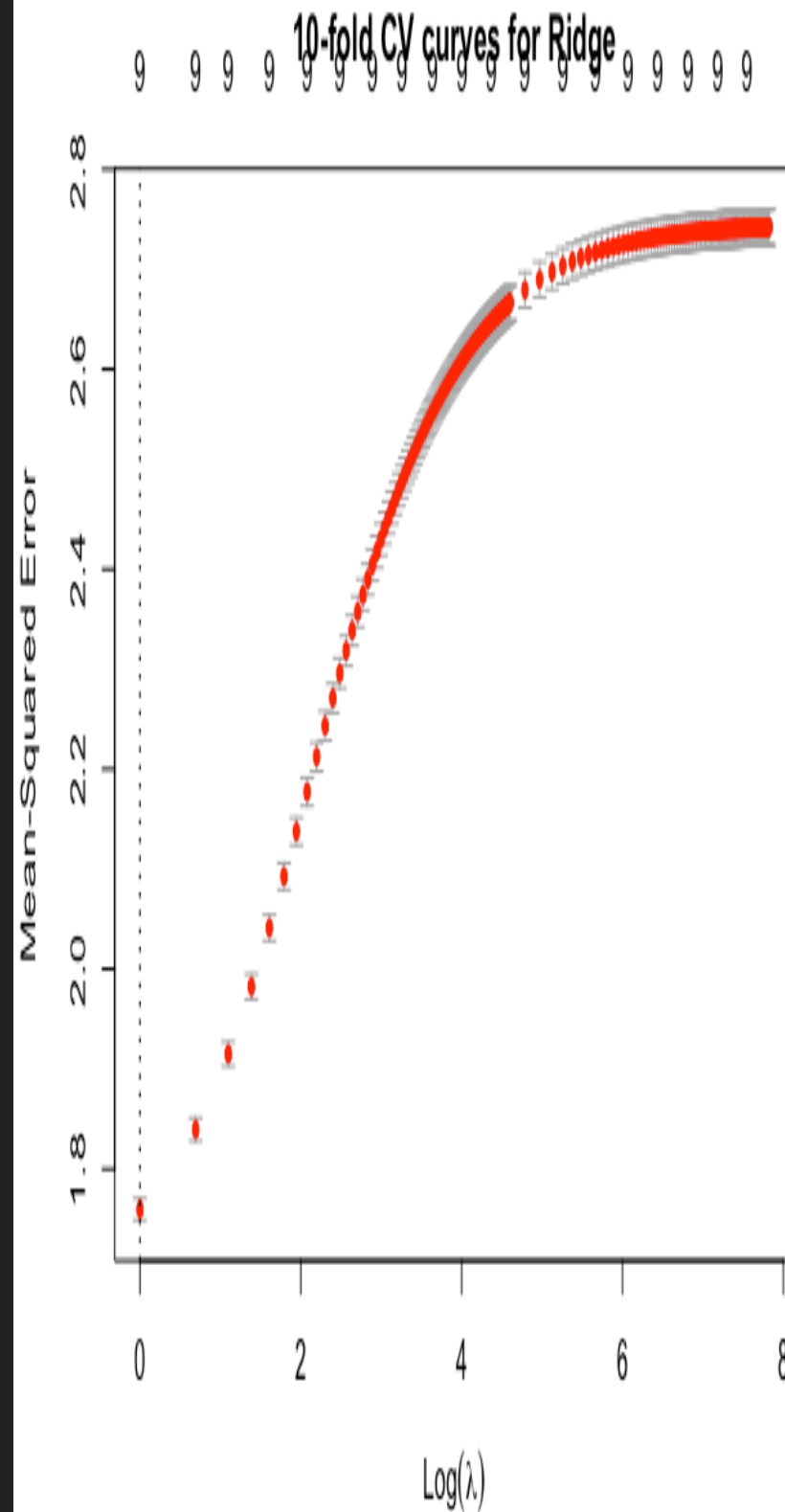
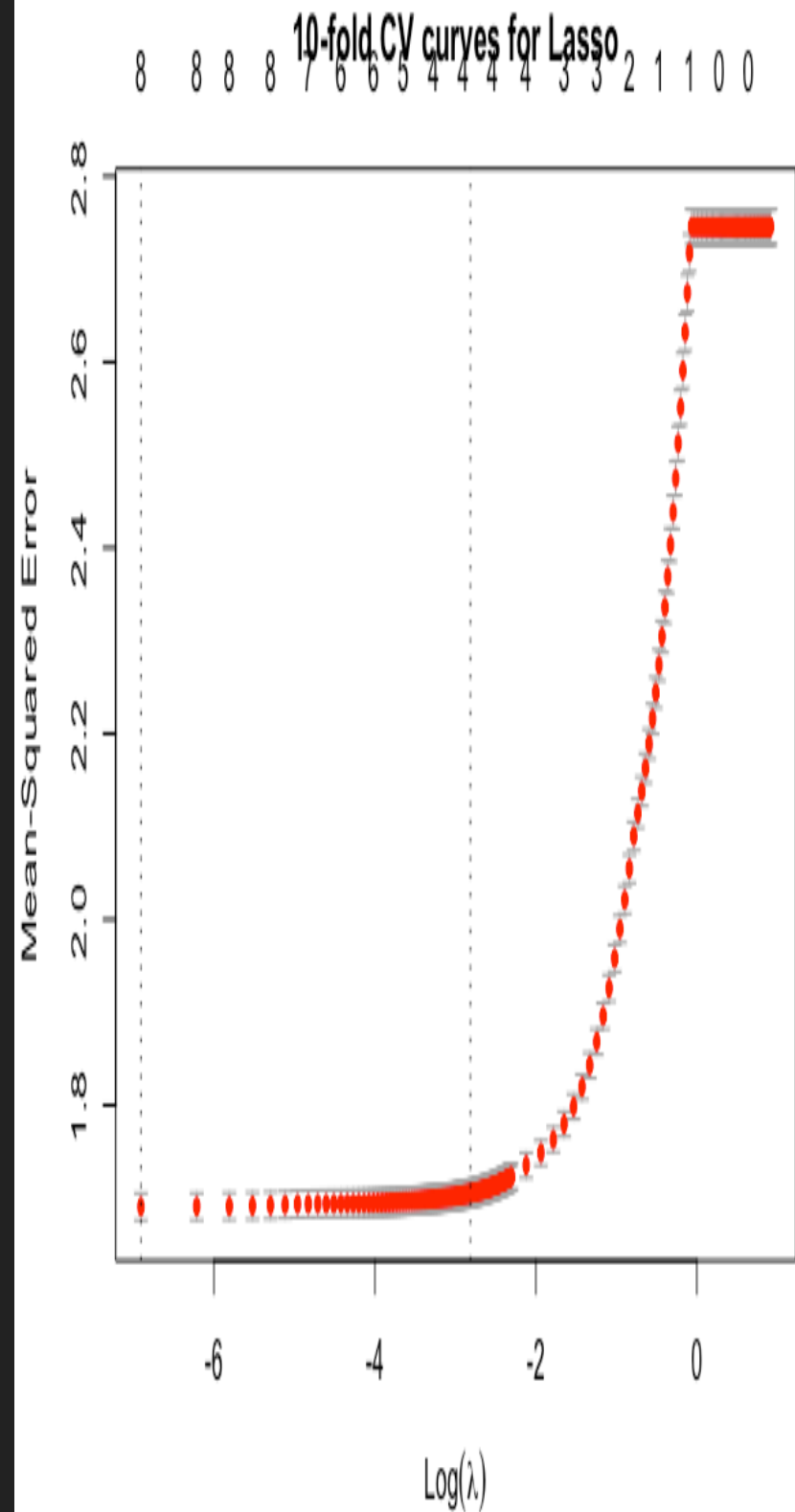
Boxplot for R Square in Train Data



Boxplot for R Square in Test Data

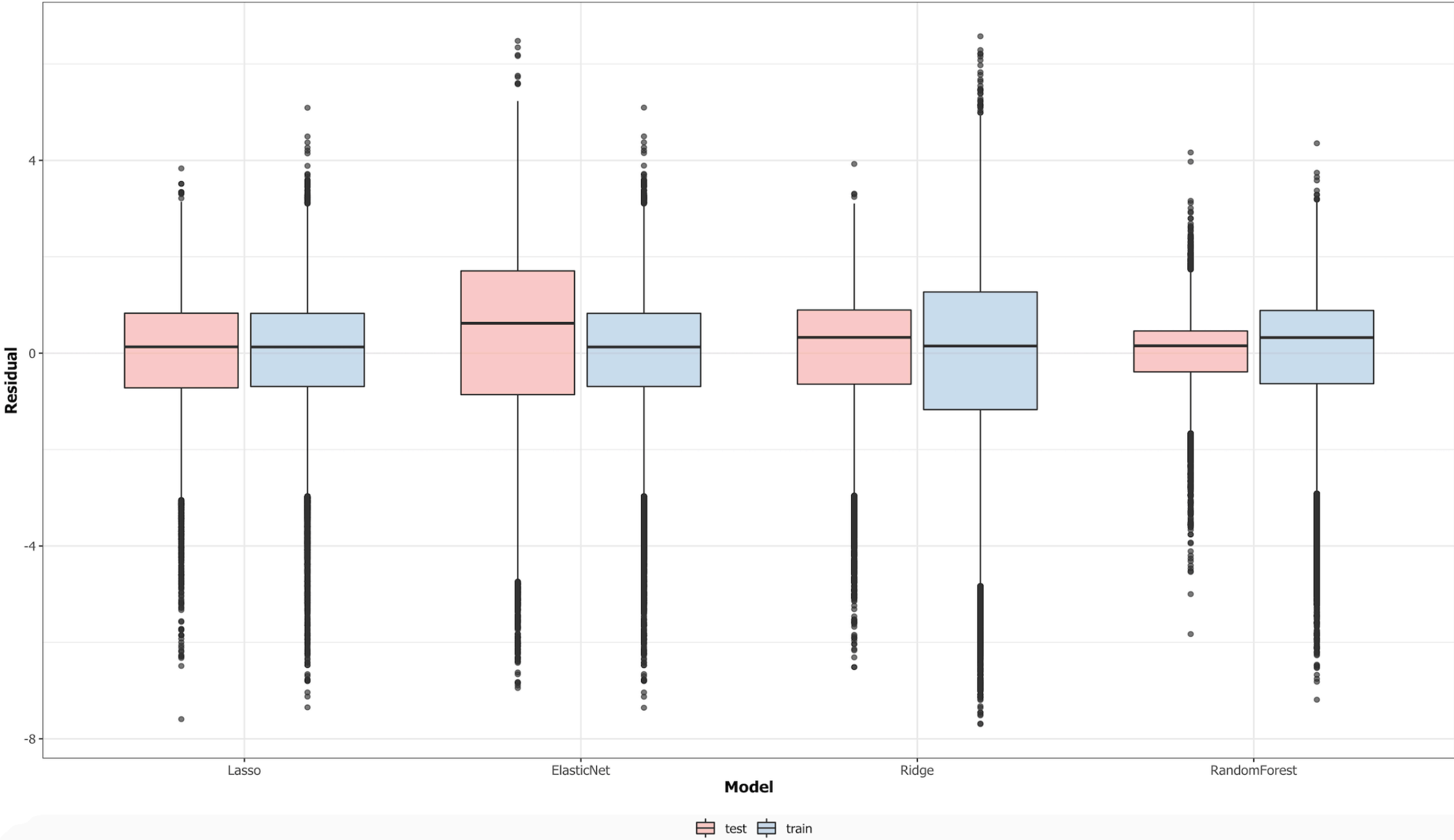


10-FOLD CV CURVES FOR THREE MODELS

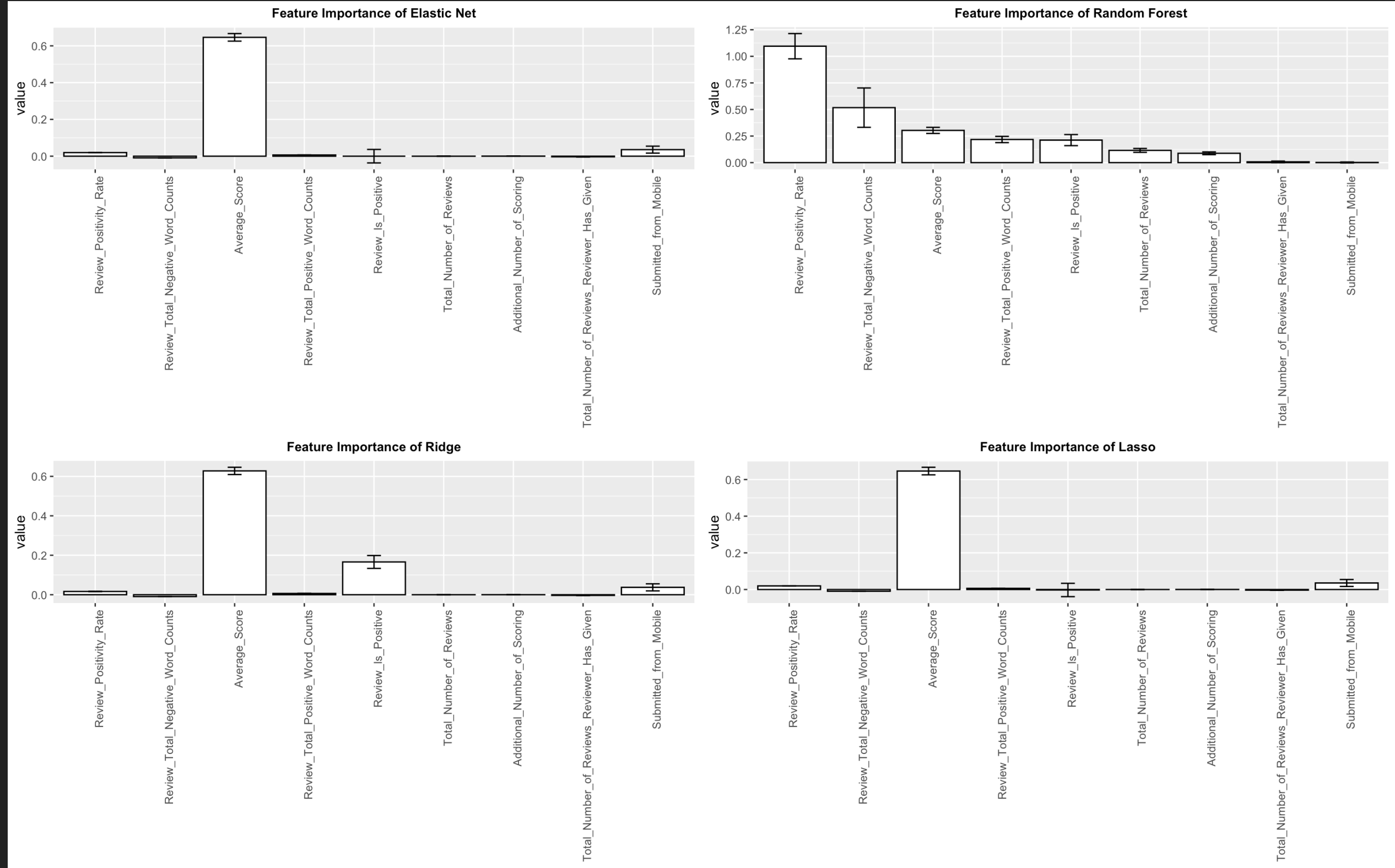


SIDE BY SIDE BOX PLOT OF RESIDUALS

Boxplot of Residual_train and Residual_test for Each Model



FEATURE IMPORTANCE IN FOUR MODELS



SUMMARY

Model	Performance	Time
Lasso	0.385	1.44
Elastic Net	0.384	1.34
Ridge	0.359	1.52
RandomForest	0.401	38.98