English Premier League Statistics

- Data source: https://www.kaggle.com/idoyo92/epl-stats-20192020
- Each row is a summary of an EPL match from one team's perspective
- 576 Records & 44 features
- Response variable: goals scored in one match for one team
- With one hot encoding the categorical variables, I used p = 42 predictors.
- Models: Random Forest, Ridge Regression,

  Lasso Regression, Elastic-Net Regression

R-Square Plots for Four Models

- Randomly split the dataset into two mutually exclusive datasets:
  - 80% Training data: 460
  - 20% Testing data: 116
- Use training data to fit lasso, elastic-net, ridge, and random forest. Tune the hyper-parameters using 10-fold CV.
- For each model, calculate R-square with following formula (similar with R^2train) and repeat the process for 100 times.

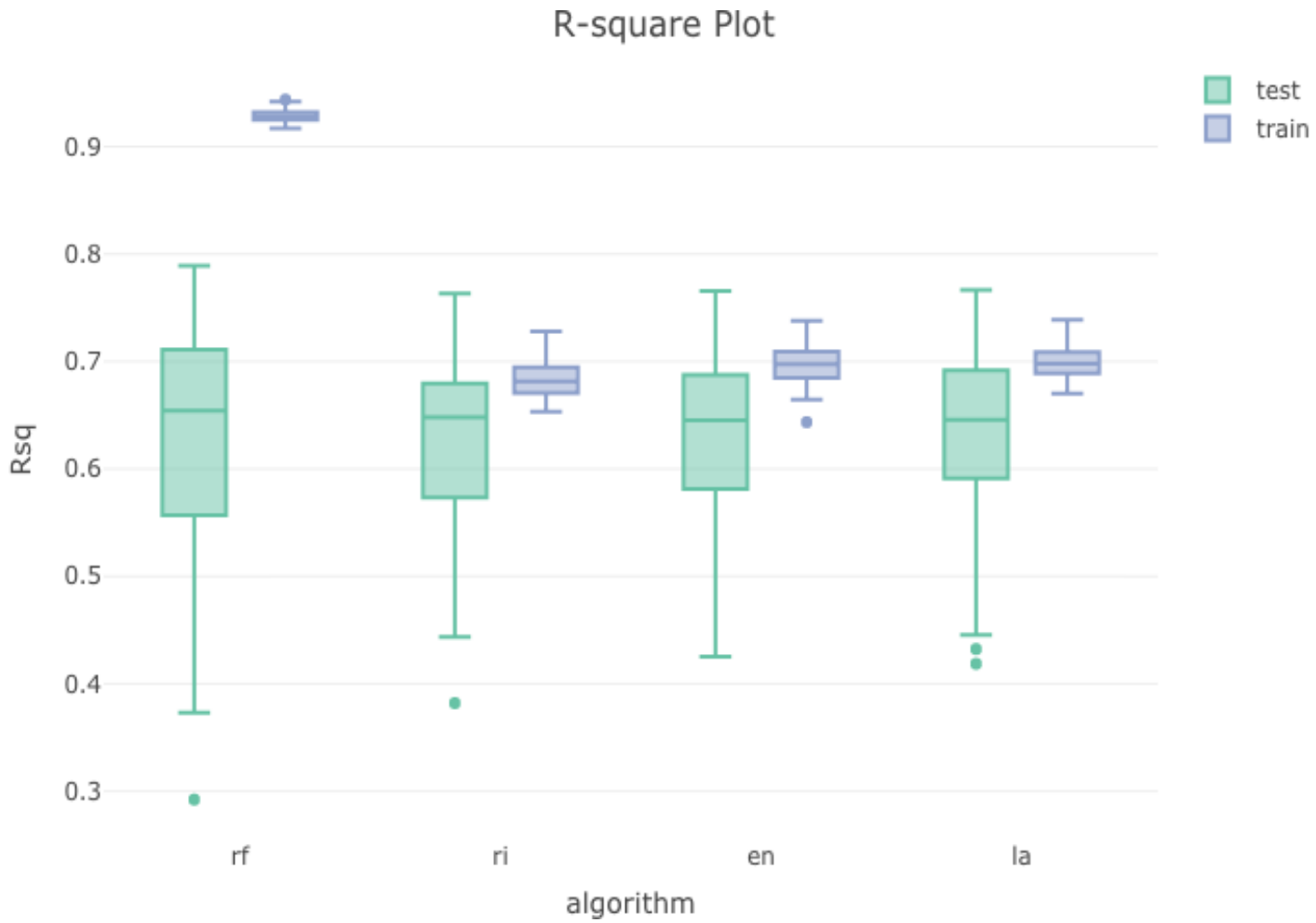
$$R^2_{test} = 1 - \frac{\frac{1}{n_{test}} \sum_{i \in D_{test}} (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Test Mean

Rsq.test.rf = 0.629, Rsq.test.ri = 0.625,
Rsq.test.en= 0.630, Rsq.test.la= 0.633

Train Mean

Rsq.train.rf = 0.929, Rsq.train.ri= 0.684,
Rsq.train.en= 0.698, Rsq.train.la= 0.699



From the boxplot we could see:

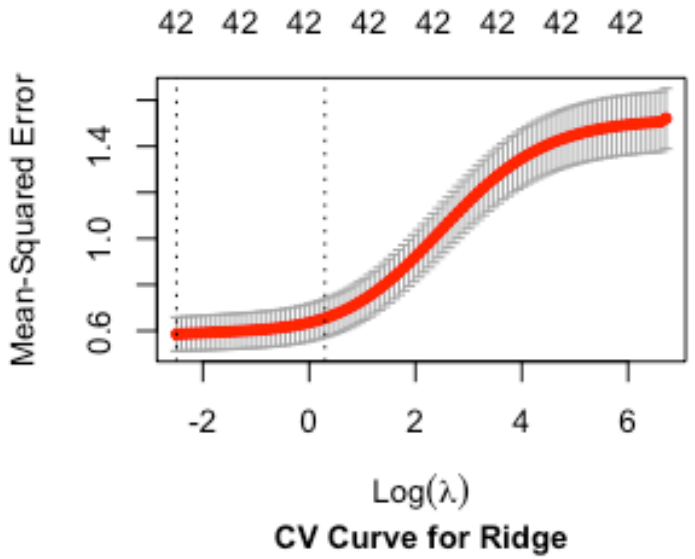
- Average testing R-square < average training R-square
- Random forest: Overfitting problem
- Elastic-Net & Lasso & Ridge: similar

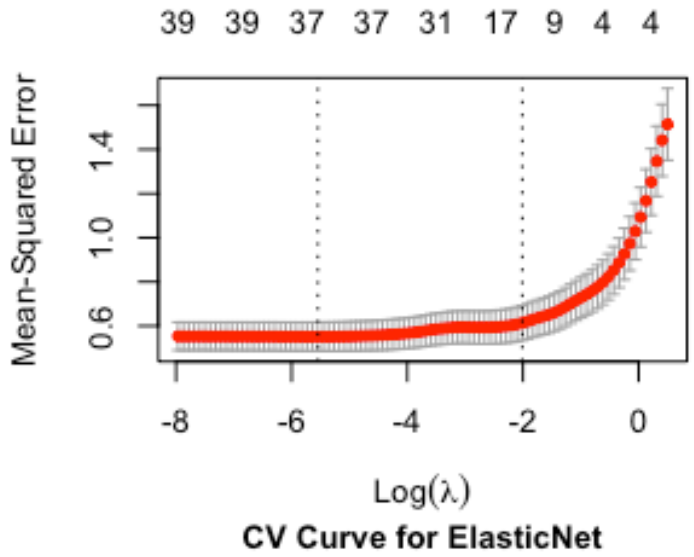
10–Fold CV Curve

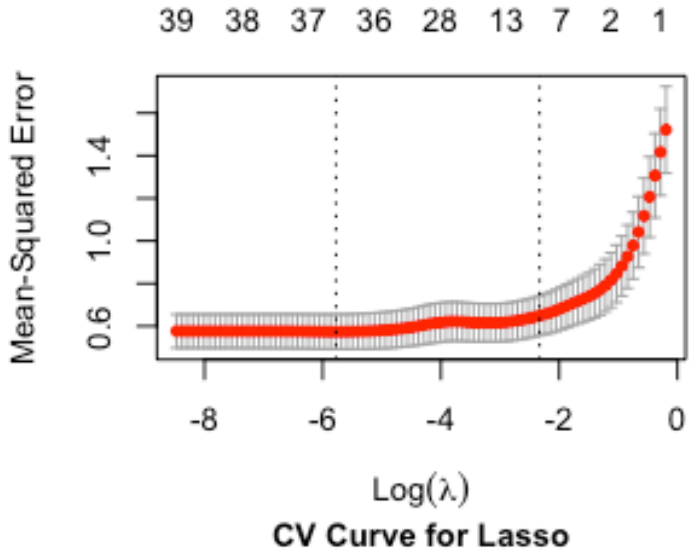
- Use usual rule to choose Lambda

- cv.fit.ri$lambda.min = 0.08246107
- cv.fit.en$lambda.min = 0.003899566
- cv.fit.la$lambda.min = 0.003104606

- Ridge uses all 43 features, ElasticNet keeps 37 features, Lasso also keeps 37 feature.

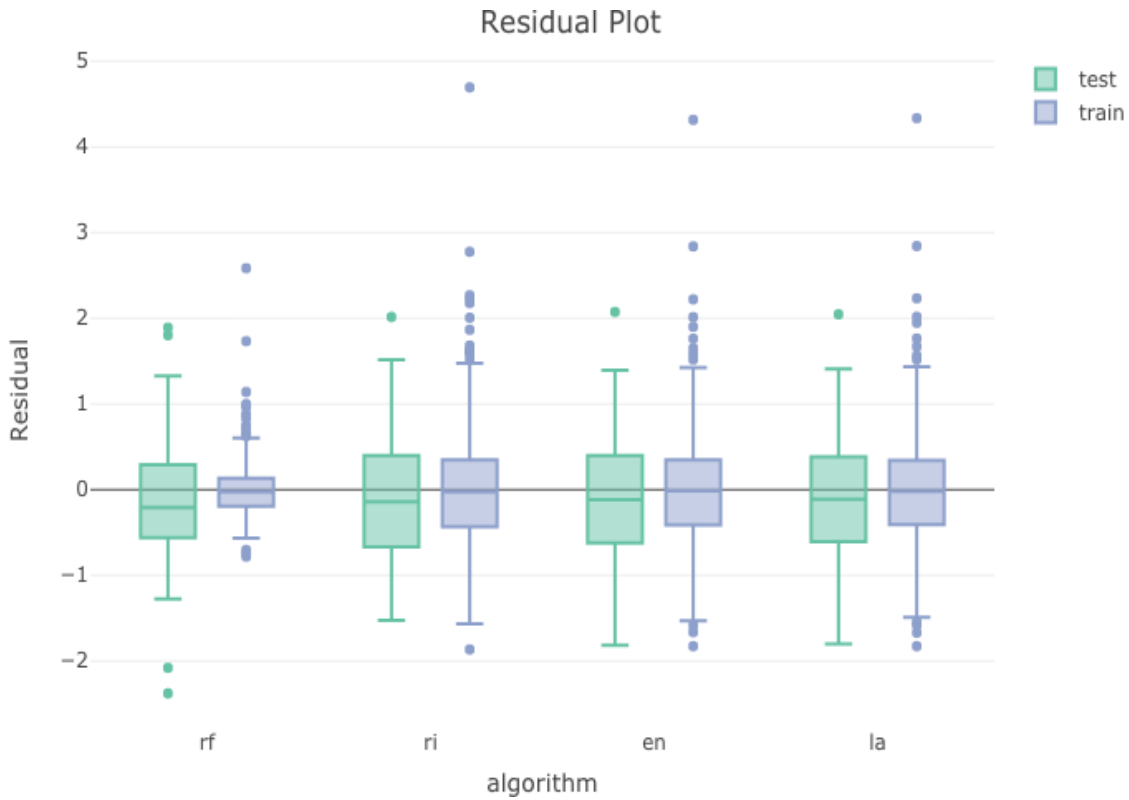


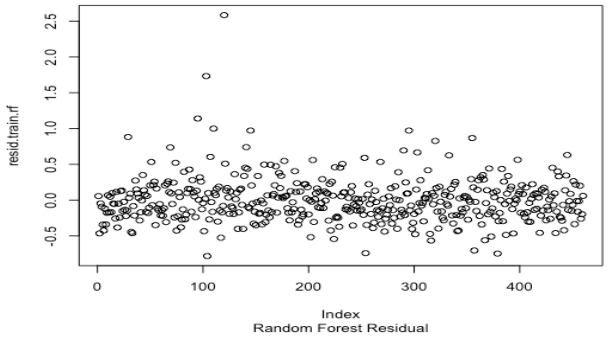


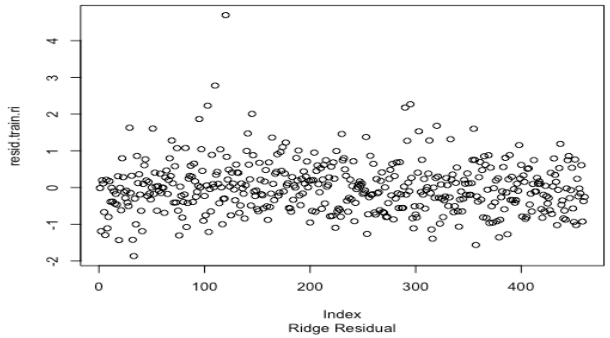


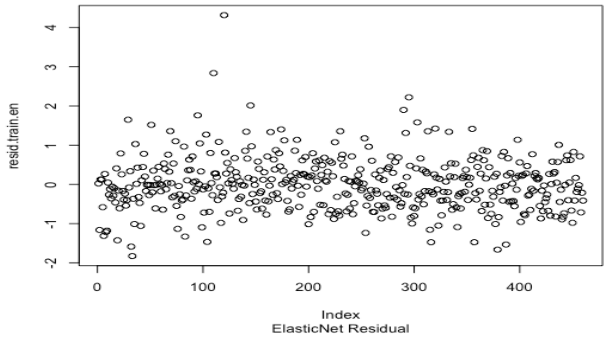
Residuals

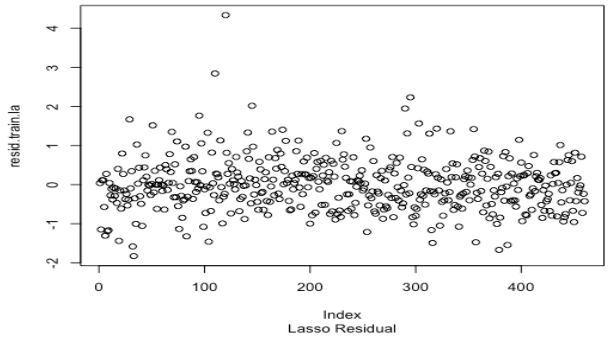
- e = y - ŷ
- Mostly center around 0
- Randomly dispersed -- linear model is appropriate

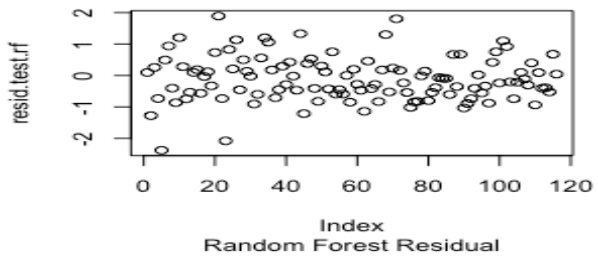


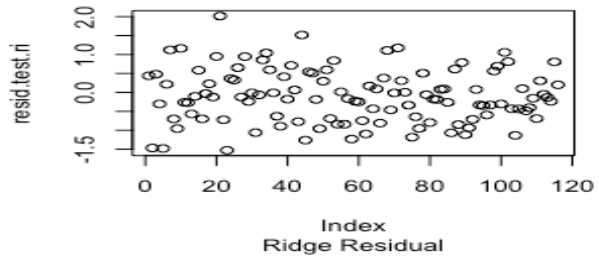


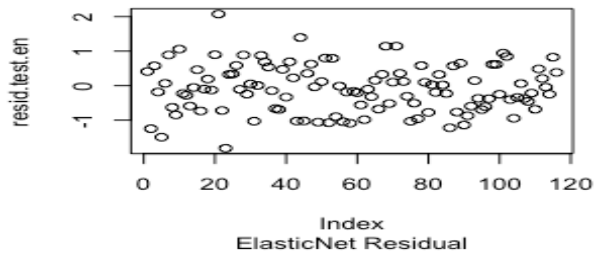


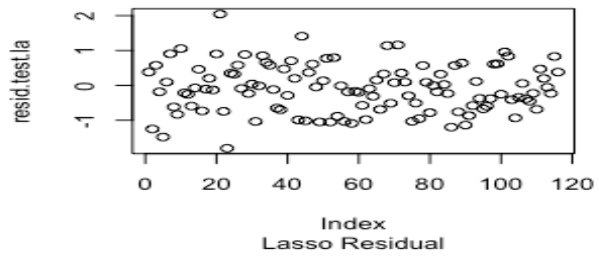


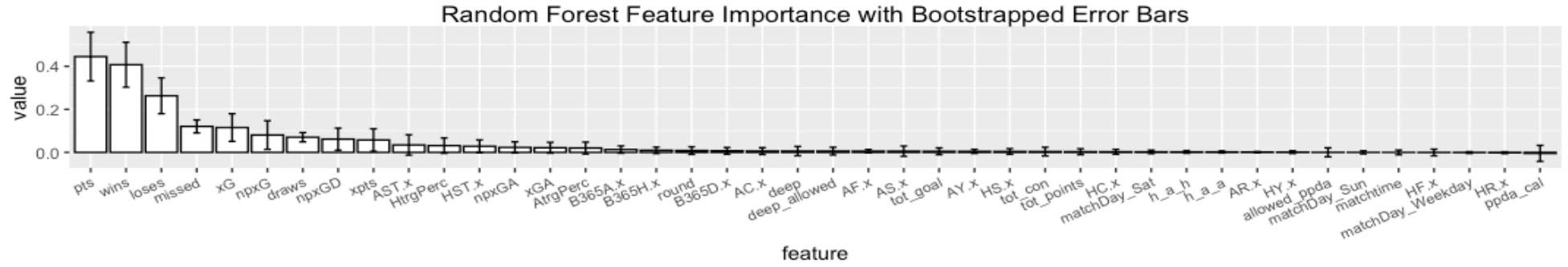


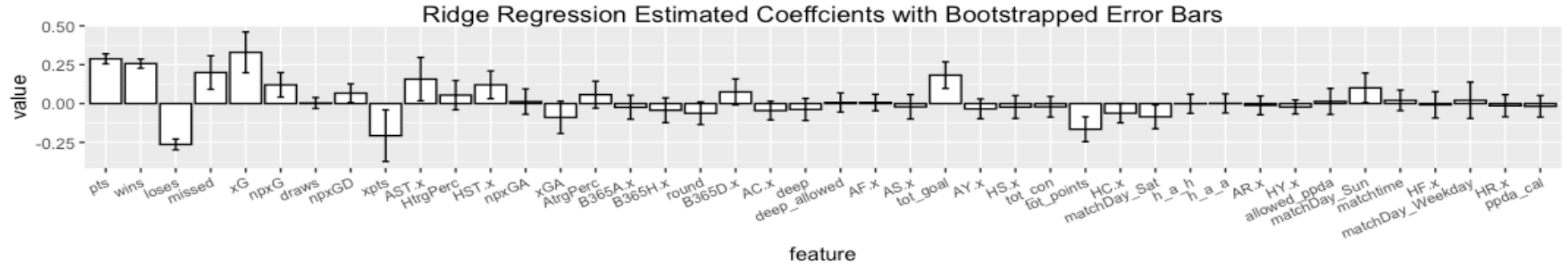


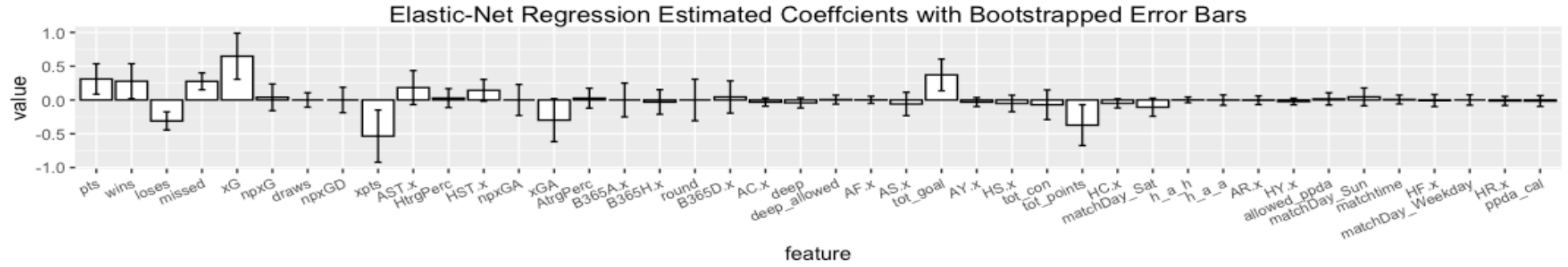


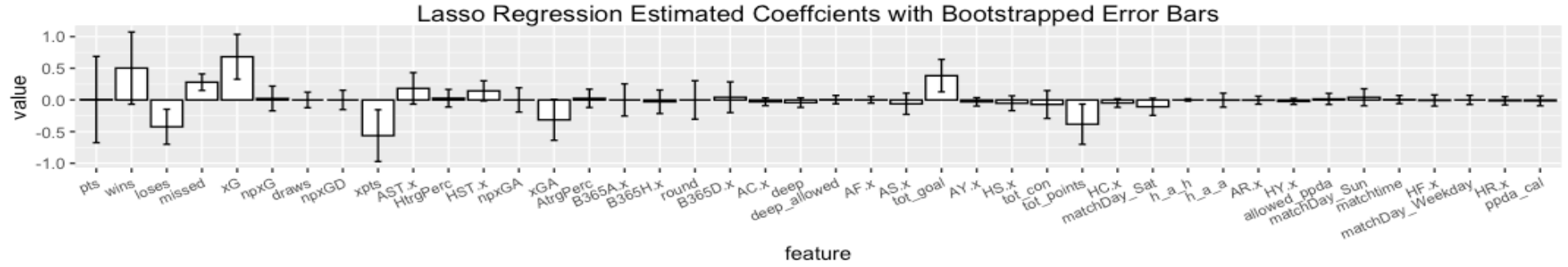












Feature Importance & Estimated Coefficients

- pts, wins, loses
  – actual win/lose
- xG index
  - expected goals
- xGA
  - xG index for opposite team
- tot_goal
  - total goals team has scored so far

Summary

	Performance	Training time
Random Forest	Overfitting	1.51s
Ridge Regression	Good	0.13s
Elastic-Net Regression	Best	0.15s
Lasso Regression	Best	0.15s