Major League Baseball Dataset

Michael Jagdharry and Nicholas Weidner

Data Gathered from:
www.baseball-reference.com

2004-2019 seasons for all 30 teams

N = 480

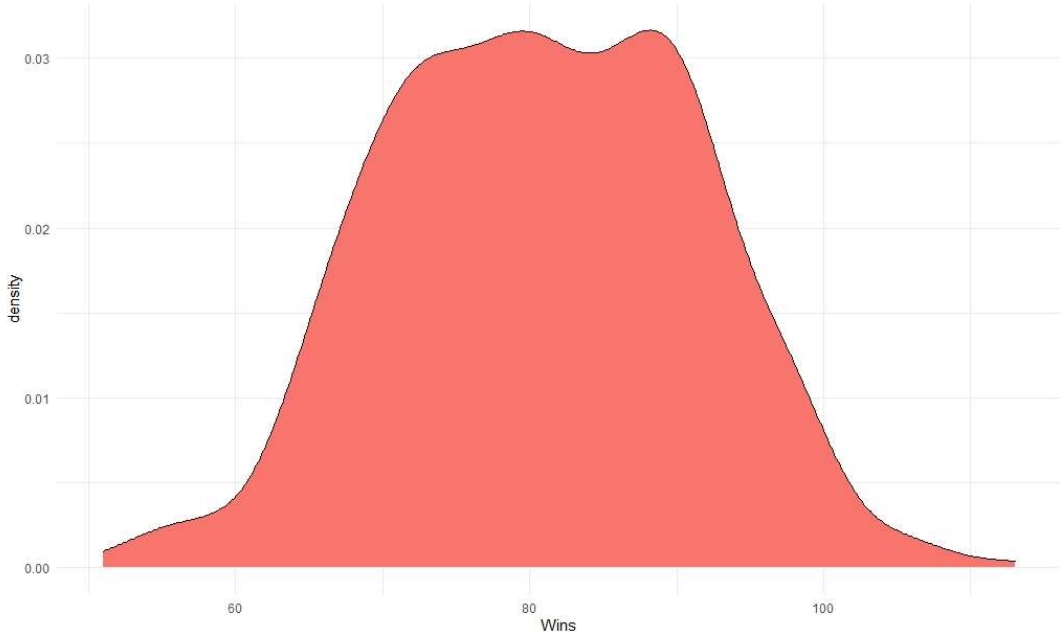
P = 43

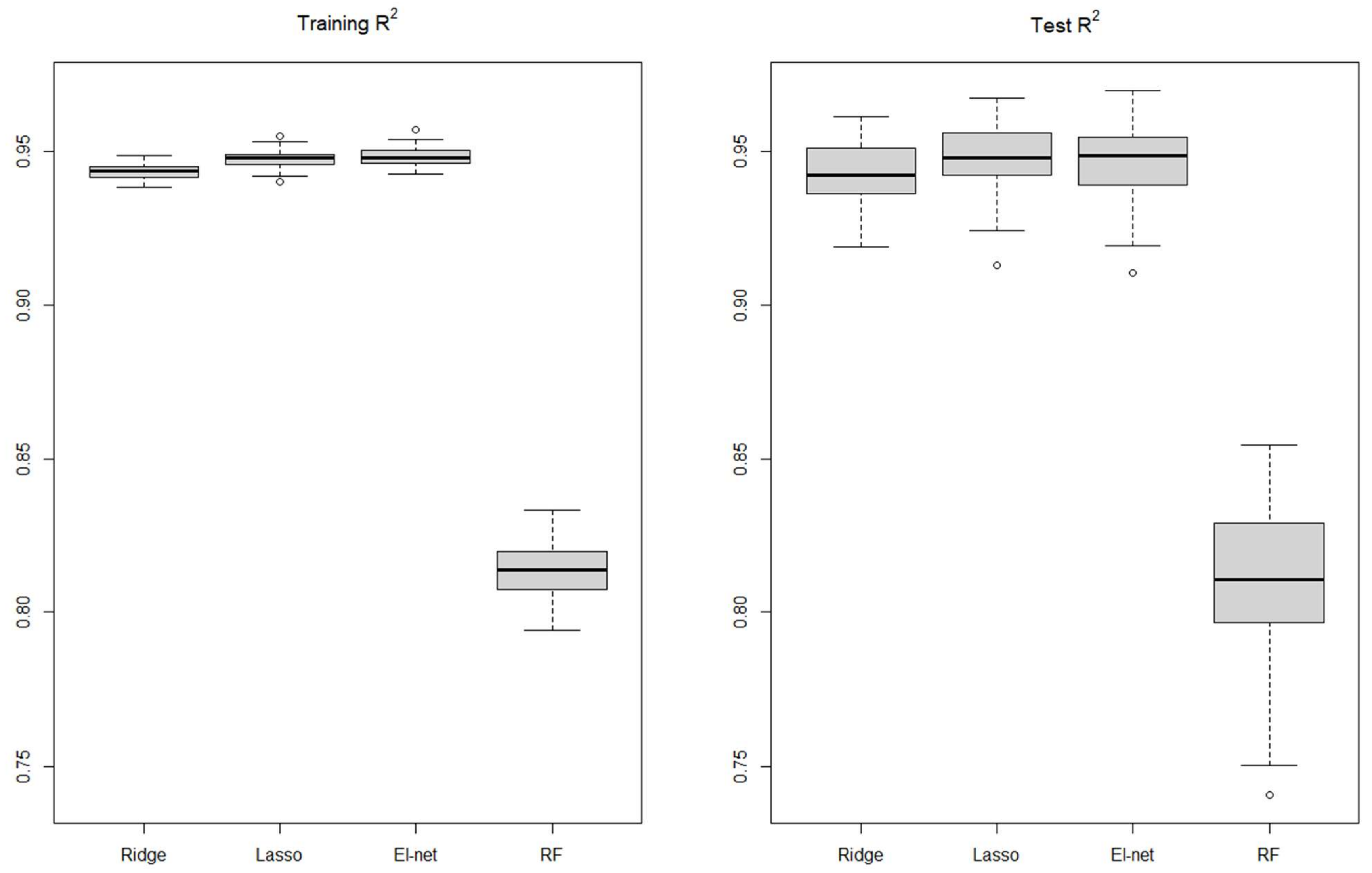
Response Variable: Wins per season

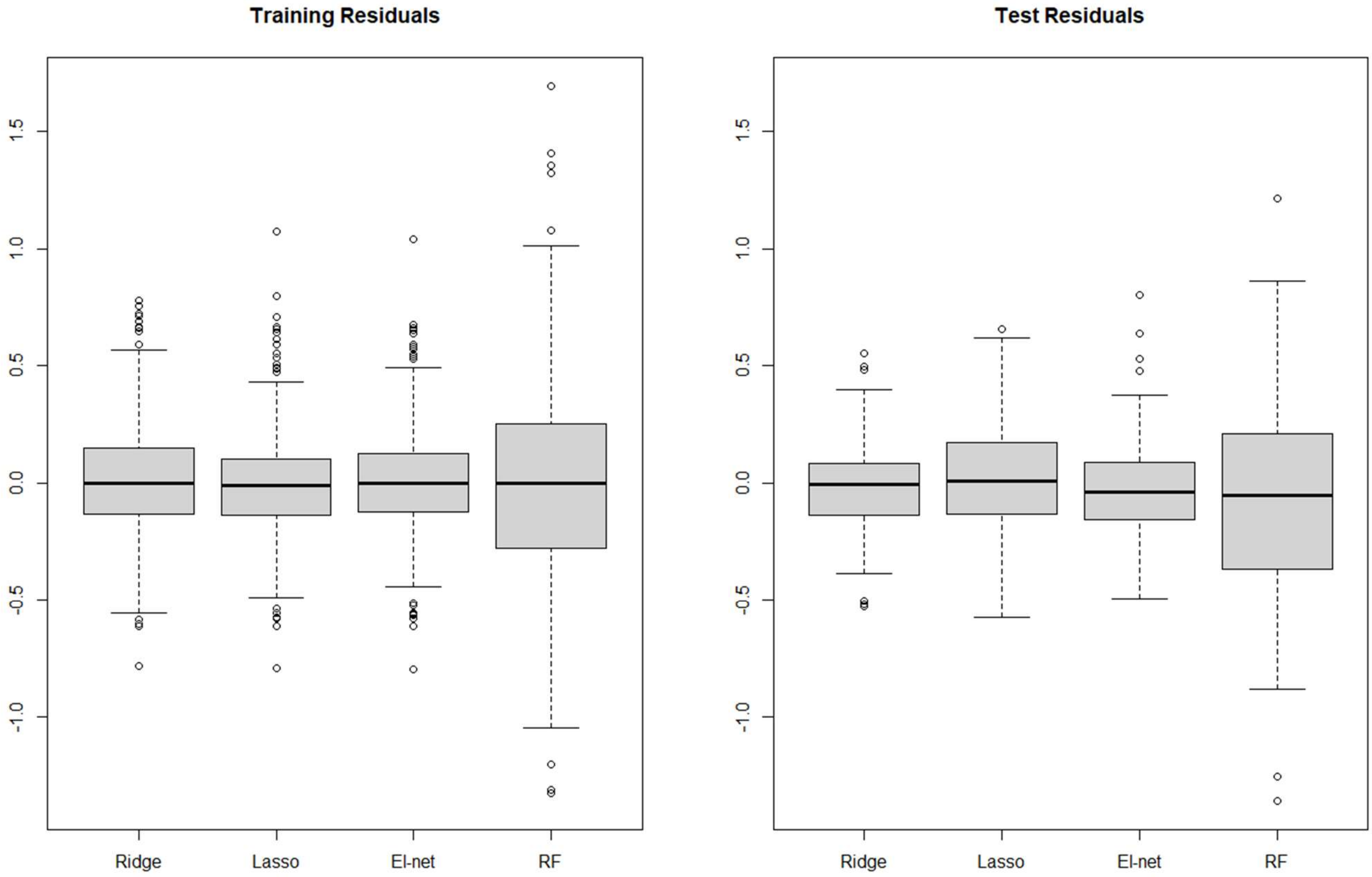
Min:	51 Wins
Max:	113 Wins
Mean:	80.89 Wins

Includes Hitting and Fielding statistics such as: Batting Average, fielding percentage, runs, home runs and errors.

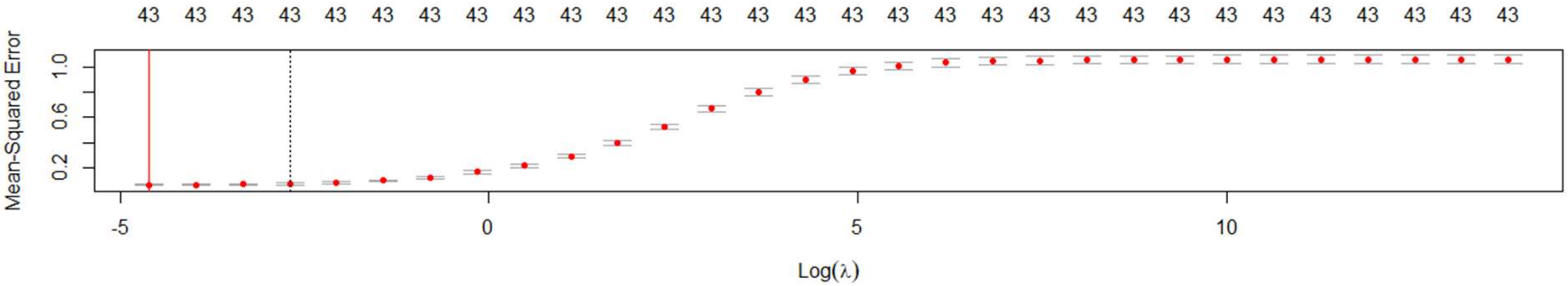
Wins calculated from Wins Above Average



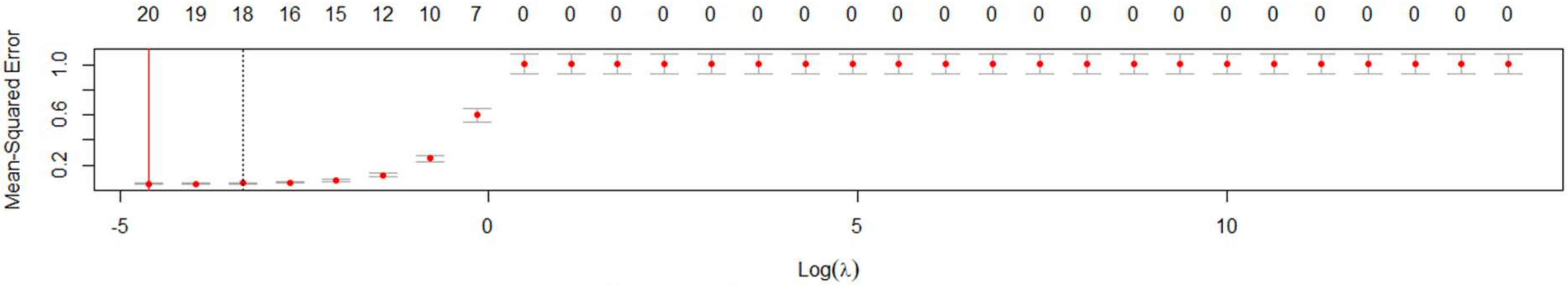

Training R^2
Test R^2
0.95
0.90
0.85
0.80
0.75
0.95
0.90
0.85
0.80
0.75
Ridge
Lasso
El-net
RF
Ridge
Lasso
El-net
RF


Training Residuals
Test Residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
Ridge
Lasso
El-net
RF

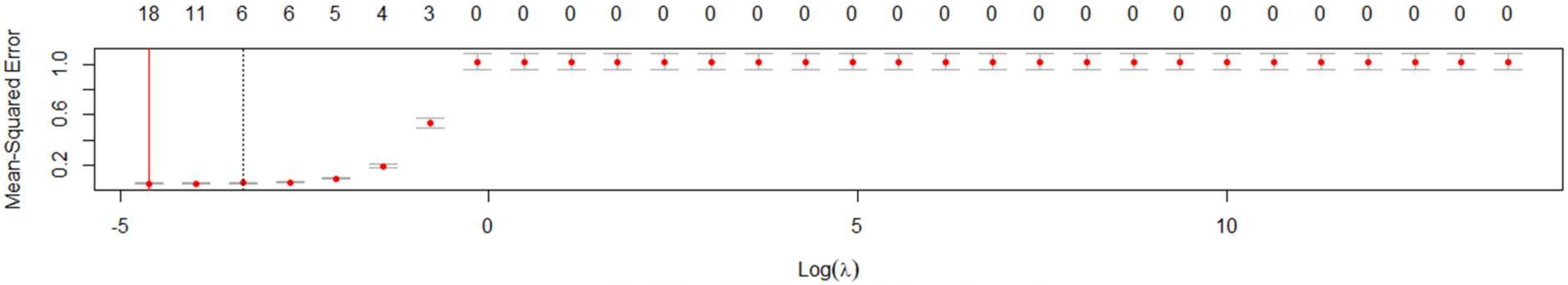
CV Curves



10-Fold Cross Validation Curve for Ridge Regression



10-Fold Cross Validation Curve for El-net Regression

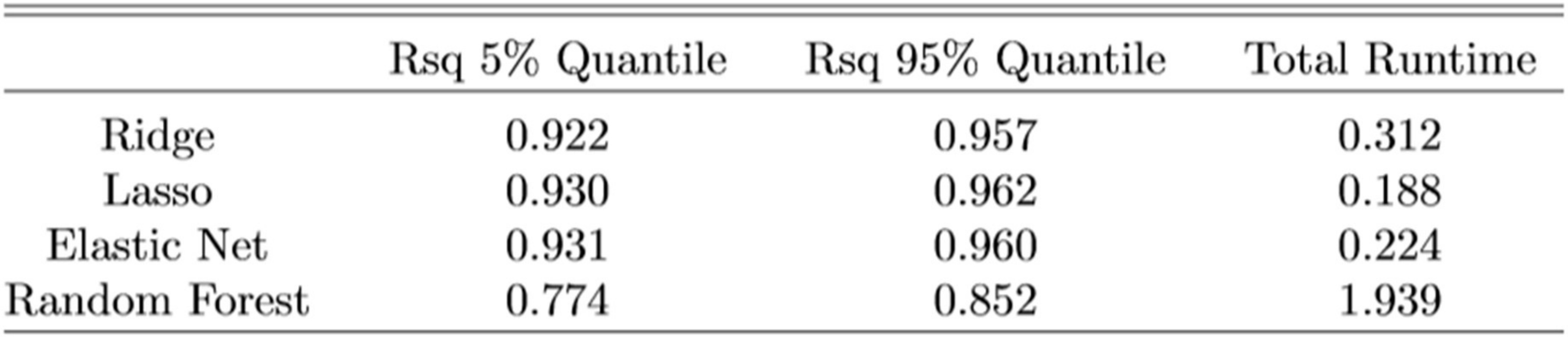


10-Fold Cross Validation Curve for Lasso Regression

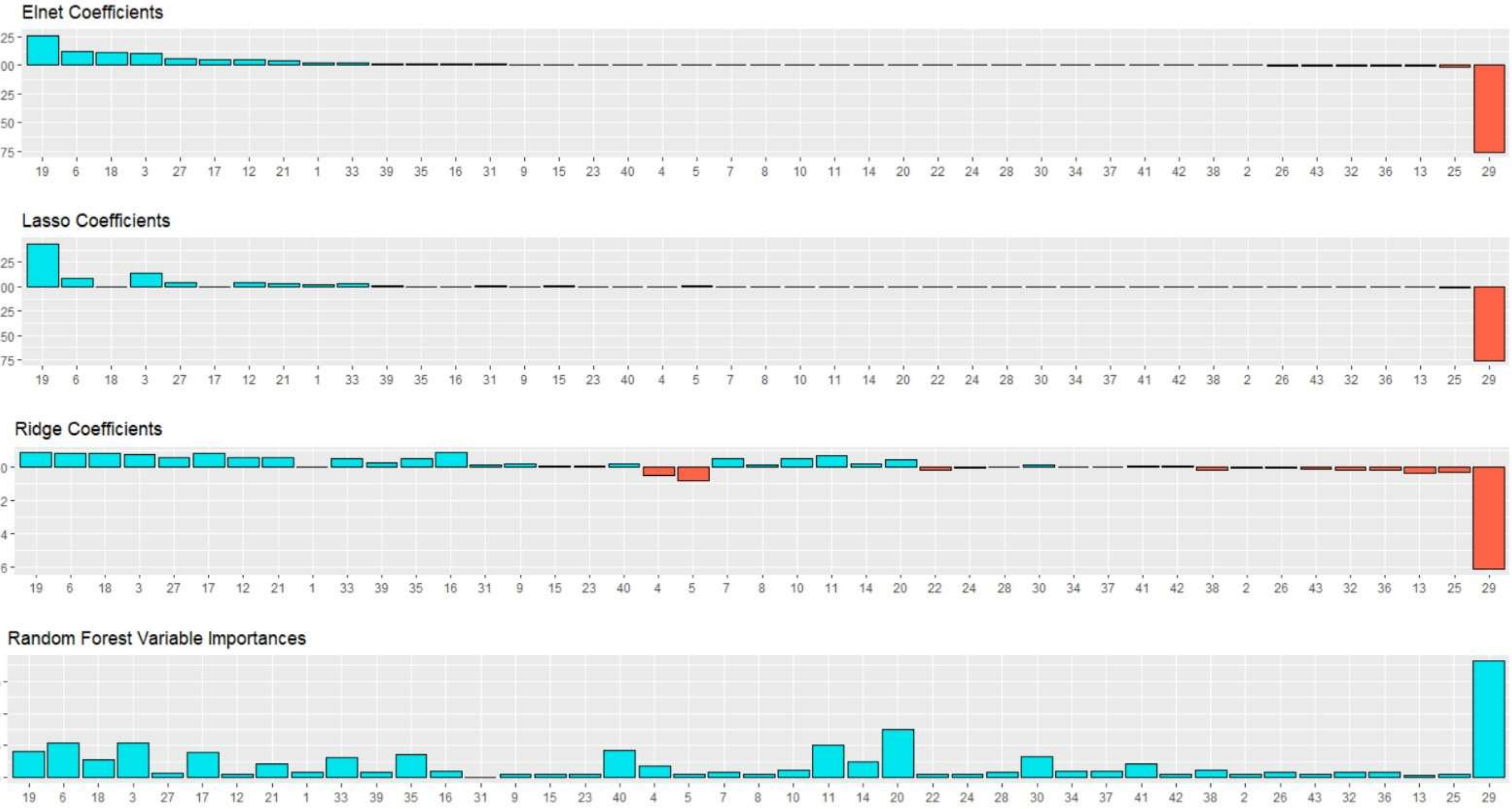
Table 1: CV Times (secs)

Ridge	0.098
Lasso	0.074
Elnet	0.079

90% Test Intervals and Runtimes



	Rsq 5% Quantile	Rsq 95% Quantile	Total Runtime
Ridge	0.922	0.957	0.312
Lasso	0.930	0.962	0.188
Elastic Net	0.931	0.960	0.224
Random Forest	0.774	0.852	1.939



Elnet Coefficients

Lasso Coefficients

Ridge Coefficients

Random Forest Variable Importances

Concluding Remarks

- For this data set regularization is preferable to random forest for the accuracy and time issues previously noted

- Maximizing team wins is a combination of minimizing opponents runs and maximizing runs for your team

- Further analysis could look into how individual players contribute to each of these variables which have been identified as important for maximizing wins in order to select a team which is more likely to win