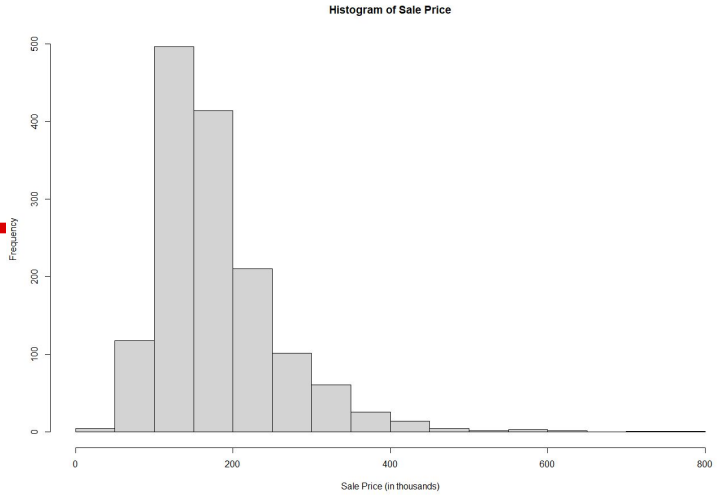
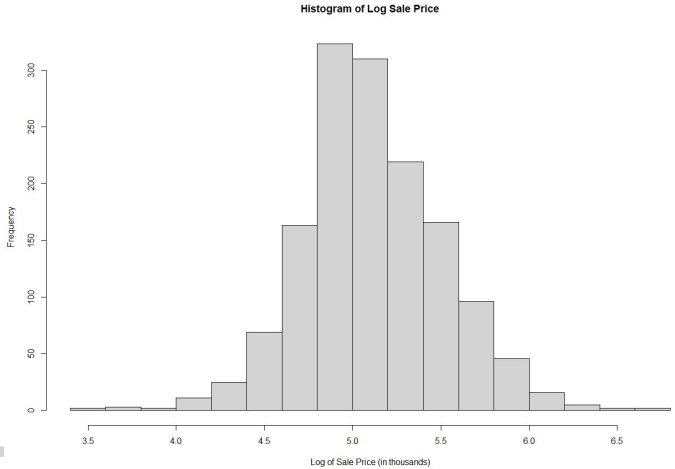
Predicting House Sale Price

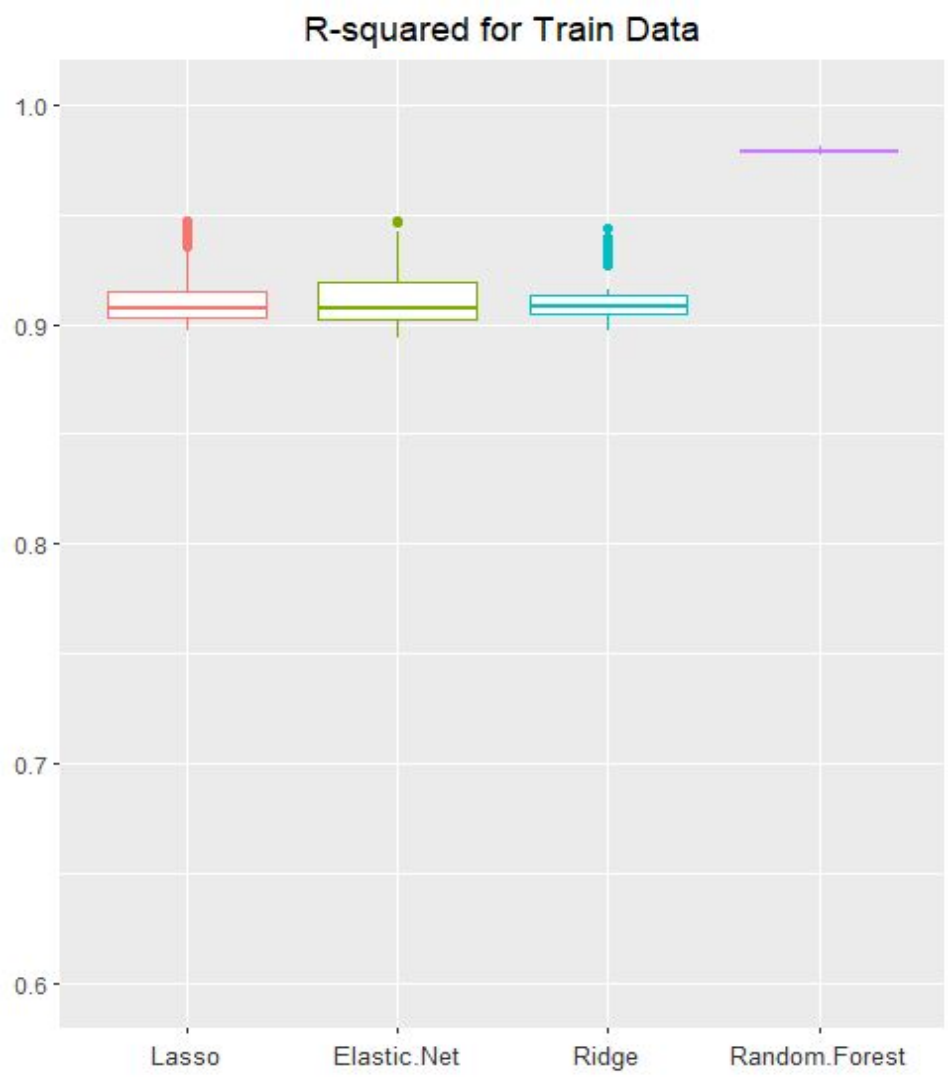
Jacob Bayer and Wei Bin Li

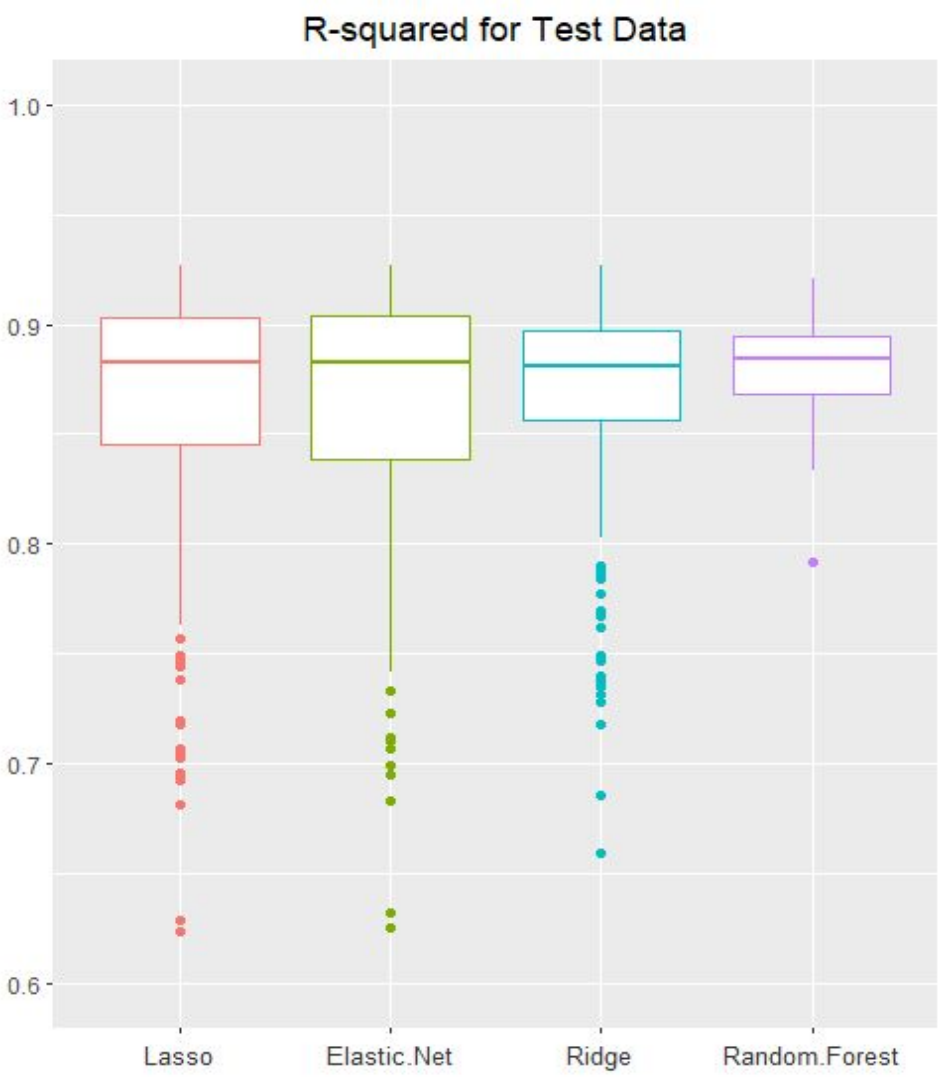
About the data

- Dependent Variable: sale price of a house
- Dimensions: 1460 x 79
- 18 numerical & 61 categorical predictors
- P = 188 after creating dummy variables



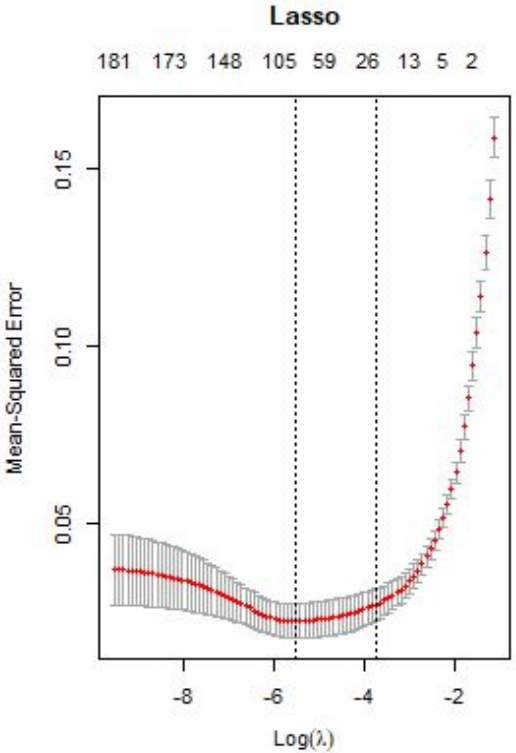


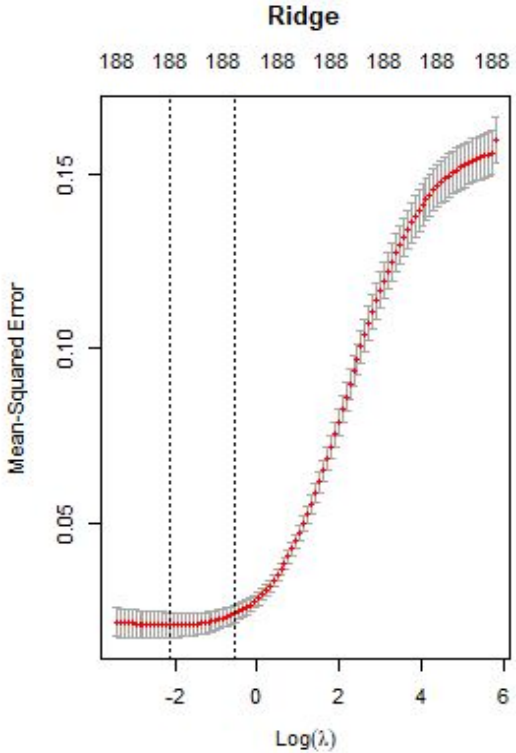

R-squared for Train Data
1.0
0.9
0.8
0.7
0.6
Lasso
Elastic.Net
Ridge
Random.Forest

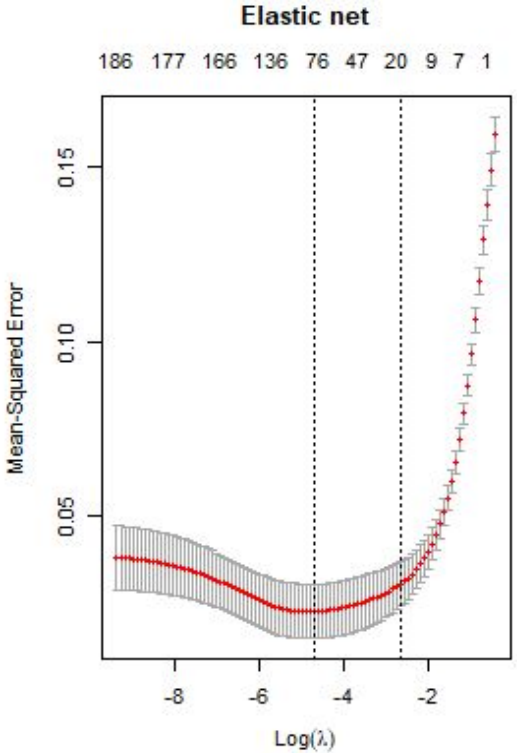

R-squared for Test Data
1.0
0.9
0.8
0.7
0.6
Lasso
Elastic.Net
Ridge
Random.Forest

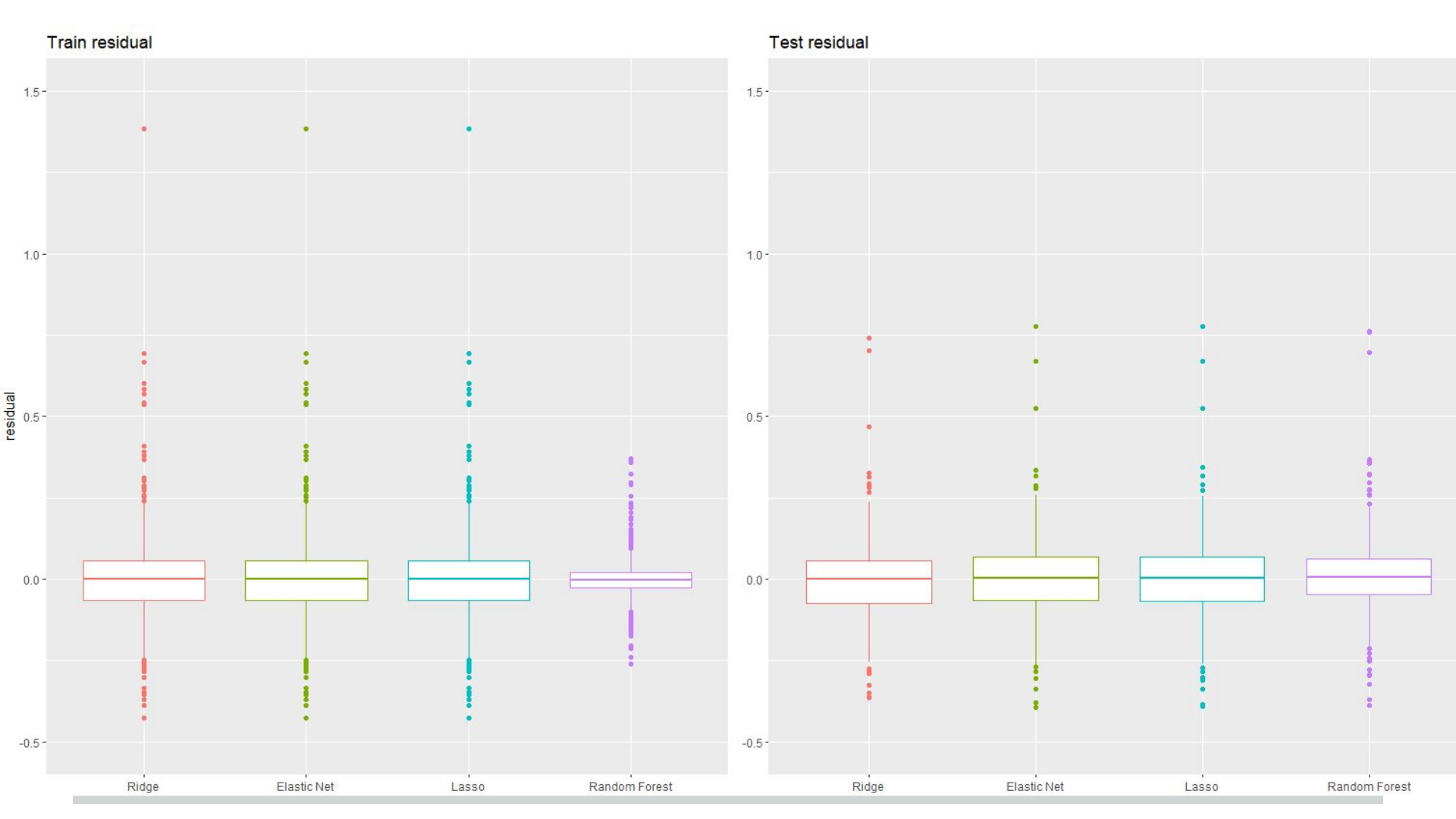
CV plots

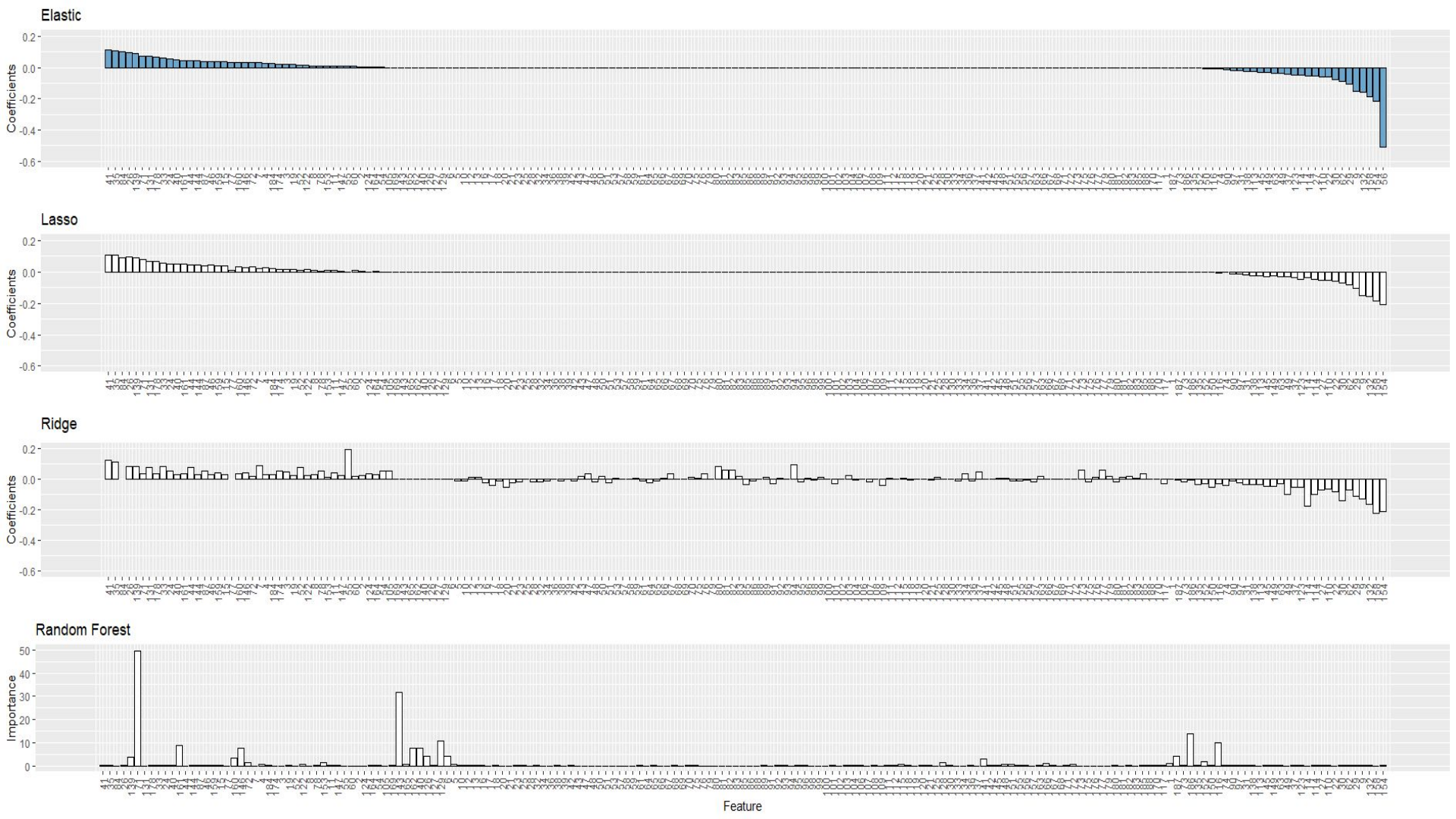
Ridge	0.44s
Lasso	0.62s
Elastic Net	0.61s

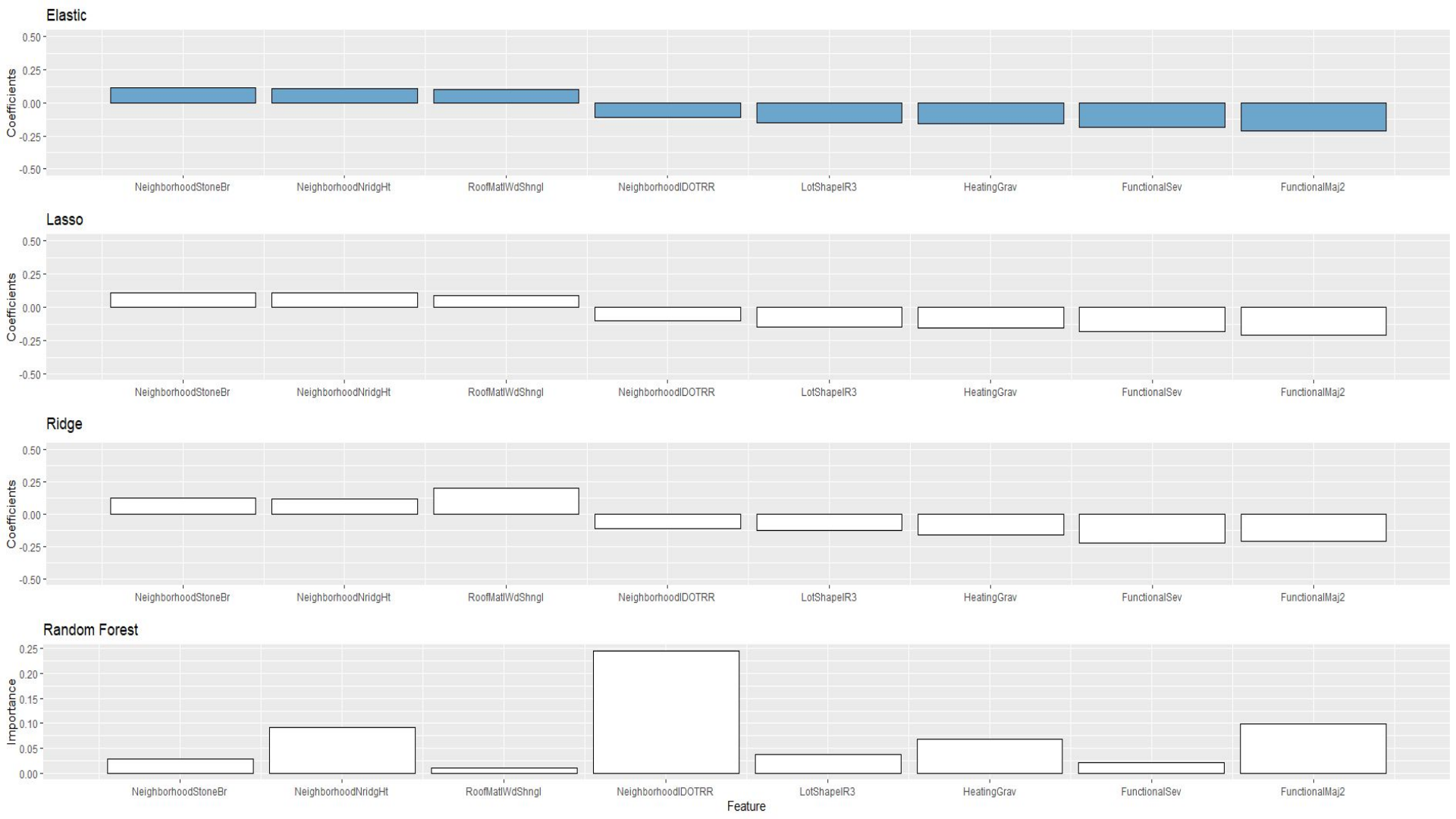


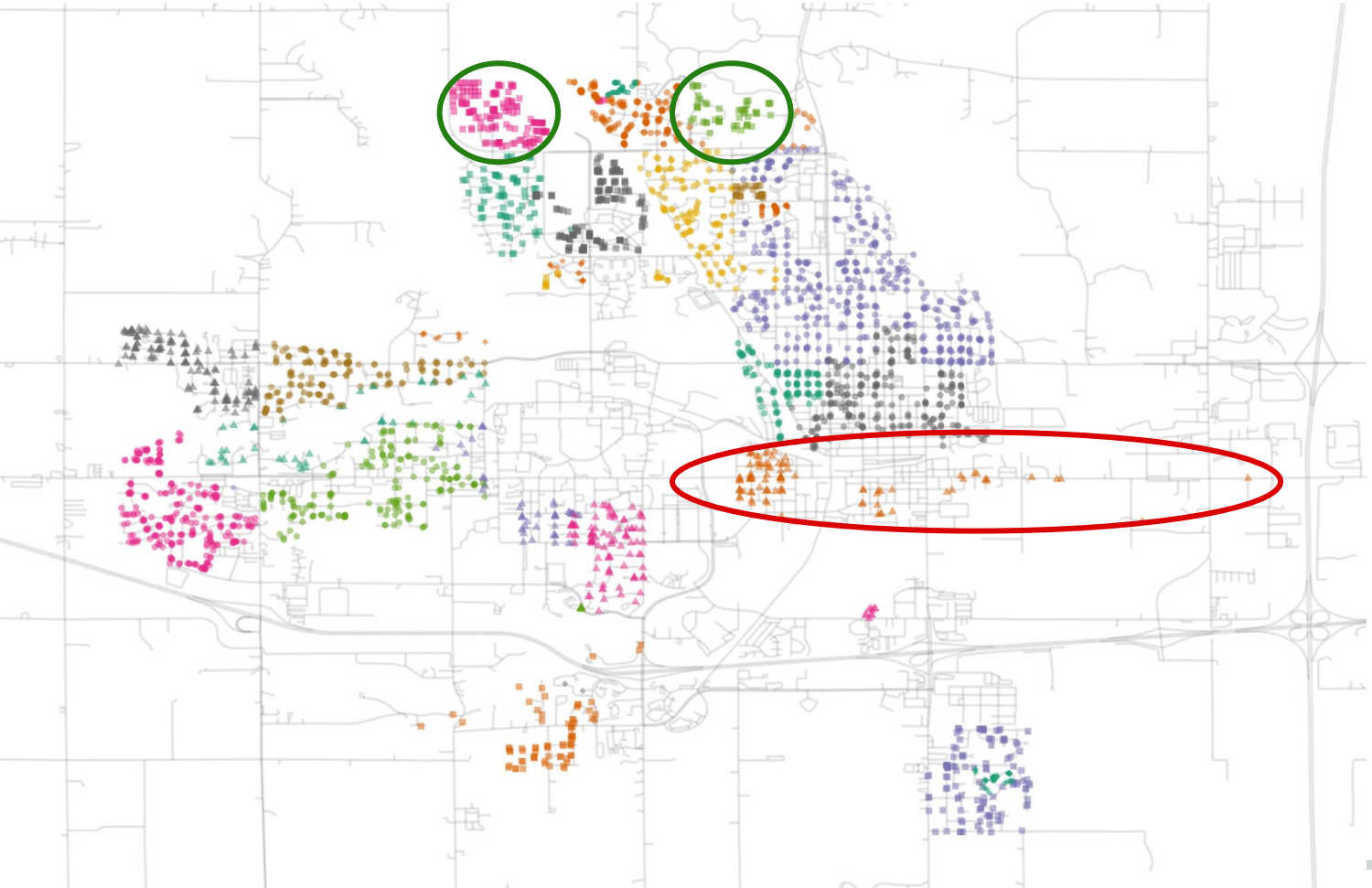





Train residual
Test residual
residual
1.5
1.0
0.5
0.0
-0.5
Ridge
Elastic Net
Lasso
Random Forest


Elastic
Coefficients
0.2
0.0
-0.2
-0.4
-0.6
Lasso
Coefficients
0.2
0.0
-0.2
-0.4
-0.6
Ridge
Coefficients
0.2
0.0
-0.2
-0.4
-0.6
Random Forest
Importance
50
40
30
20
10
0
Feature


Elastic
Coefficients
0.50
0.25
0.00
-0.25
-0.50
NeighborhoodStoneBr
NeighborhoodNridgHt
RoofMatlWdShngl
NeighborhoodIDOTRR
LotShapeIR3
HeatingGrav
FunctionalSev
FunctionalMaj2
Lasso
Coefficients
0.50
0.25
0.00
-0.25
-0.50
NeighborhoodStoneBr
NeighborhoodNridgHt
RoofMatlWdShngl
NeighborhoodIDOTRR
LotShapeIR3
HeatingGrav
FunctionalSev
FunctionalMaj2
Ridge
Coefficients
0.50
0.25
0.00
-0.25
-0.50
NeighborhoodStoneBr
NeighborhoodNridgHt
RoofMatlWdShngl
NeighborhoodIDOTRR
LotShapeIR3
HeatingGrav
FunctionalSev
FunctionalMaj2
Random Forest
Importance
0.25
0.20
0.15
0.10
0.05
0.00
NeighborhoodStoneBr
NeighborhoodNridgHt
RoofMatlWdShngl
NeighborhoodIDOTRR
LotShapeIR3
HeatingGrav
FunctionalSev
FunctionalMaj2
Feature


North_Ames
College_Creek
Old_Town
Edwards
Somerset
Northridge_Heights
Gilbert
Sawyer
Northwest_Ames
Sawyer_West
Mitchell
Brookside
Clear_Creek
Meadow_Village
Briardale
Bloomington_Heights
Veenker
Northpark_Villa
Blueste
Greens
Green_Hills
Landmark
Crawford
Iowa_DOT_and_Rail_Road
Timberland
Northridge
Stone_Brook
South_and_West_of_Iowa_State_University

Model Performance

MODEL	90% CI for R^2	MEAN R^2 PERFORMANCE	TIME (Seconds)
LASSO	(0.70,0.92)	0.85	0.71
ELASTIC NET	(0.70,0.92)	0.84	0.94
RIDGE	(0.74,0.92)	0.86	0.52
RANDOM FOREST	(0.85,0.91)	0.88	15.11

Conclusion

- Random forest is the best model, but it takes the longest to calculate
- The most important factor in predicting house price is location (neighborhood)
- The second most important factor appears to be whether or not the house is damaged