Final Project

Communities and Crime dataset

STA 9890 – Statistical Learning for Data Mining

Spring 2020 – May 20, 2020

Ariel Chajet

Data Description

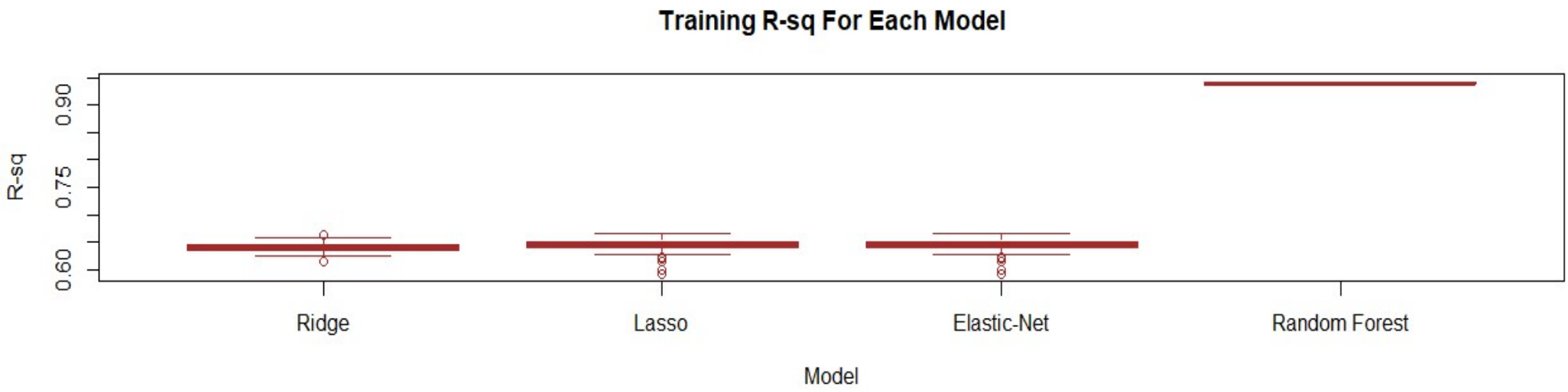
- The data compiles socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR. (**Source**: https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized)

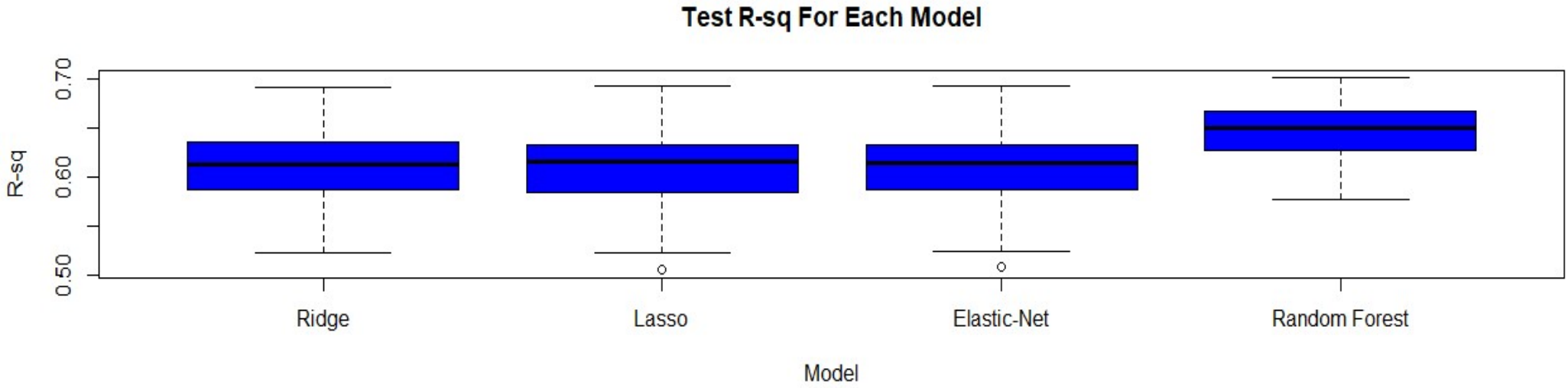
- Each record in the data represents an entire town/city and all related census and crime attributes pertaining to that city.

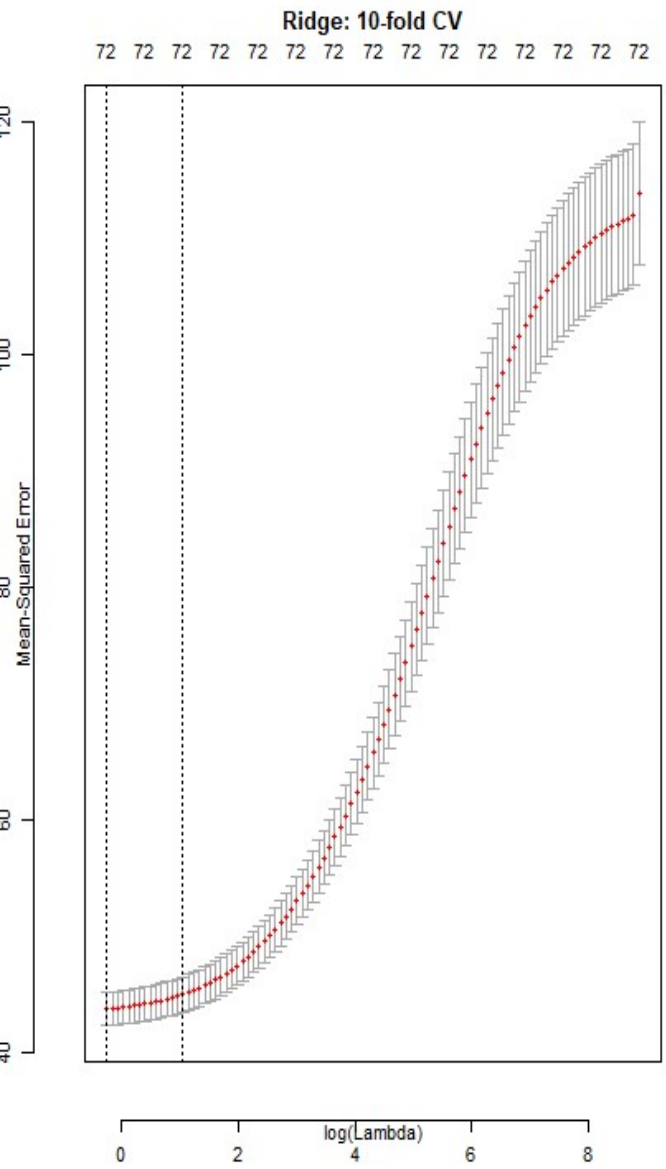
- The selected response variable represents the rate of violent crimes per 100K population, which is positively skewed. The square root of the rate of violent crimes was ultimately used to help center the distribution of the response.

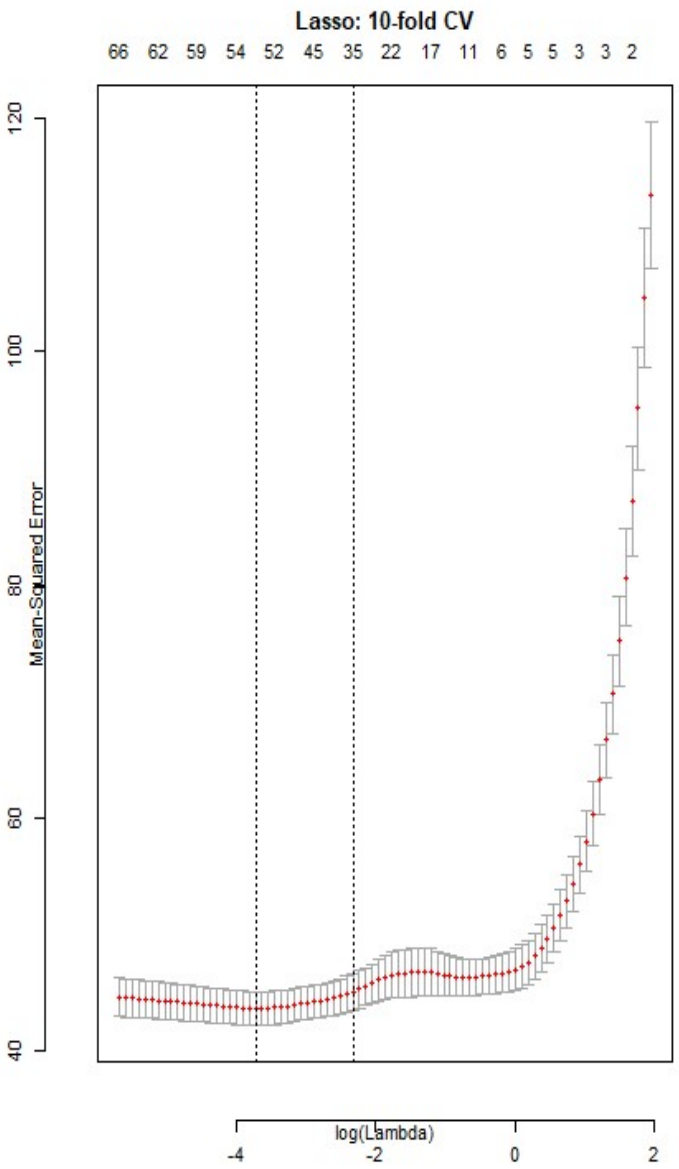
- The raw data contains 2,215 observations with 147 features (5 non-predictive features, 125 predictive, and 18 potential response variables). All predictor and response variables are quantitative. Ultimately, the working data set was reduced to 72 predictors for modeling.

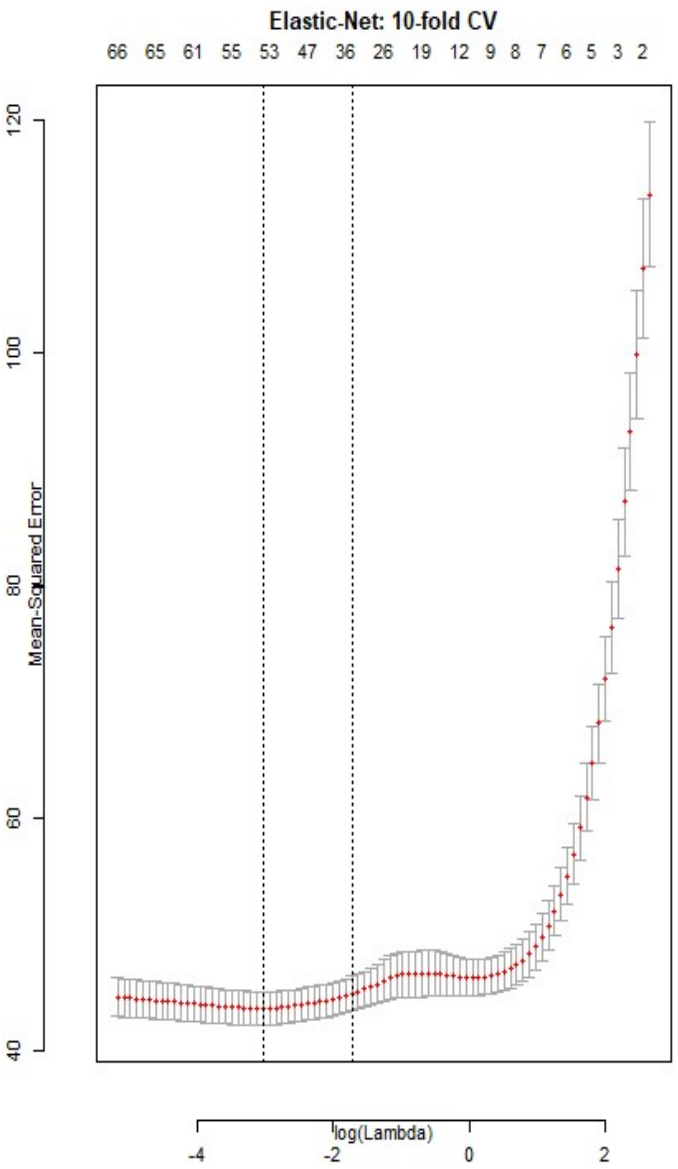
- The data was downloaded from the URL and saved locally as a CSV file. Values of "?" were converted to null values in excel before importing data into R. The column names were also added as headers to the CSV file.


Training R-sq For Each Model
R-sq
0.60
0.75
0.90
Ridge
Lasso
Elastic-Net
Random Forest
Model

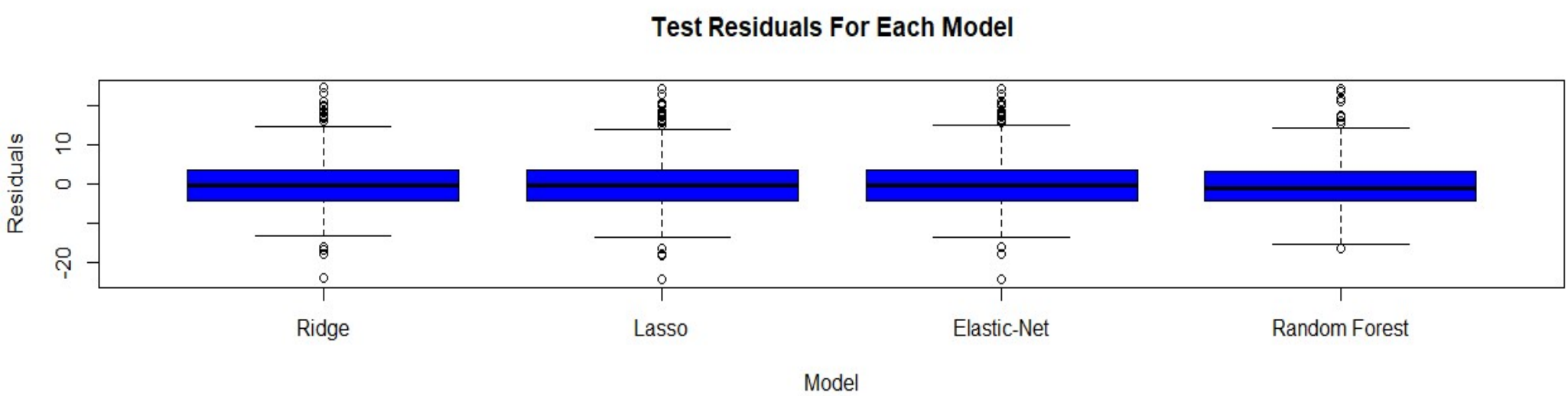

Test R-sq For Each Model
R-sq
0.50
0.60
0.70
Ridge
Lasso
Elastic-Net
Random Forest
Model

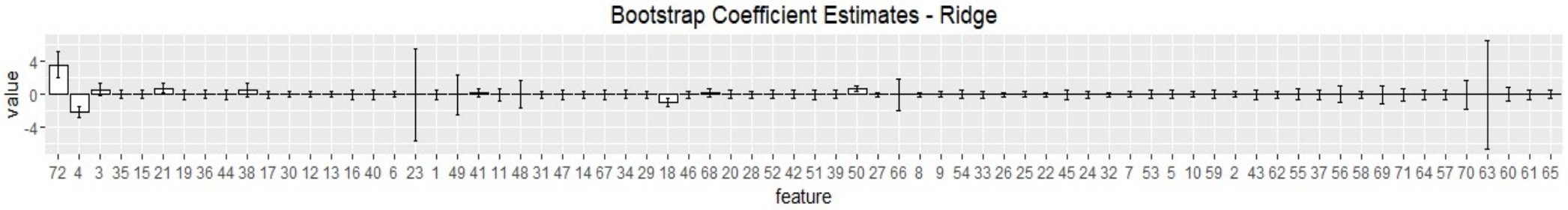

Ridge: 10-fold CV
72 72 72 72 72 72 72 72 72 72 72 72 72 72 72
Mean-Squared Error
40
60
80
100
120
log(Lambda)
0
2
4
6
8

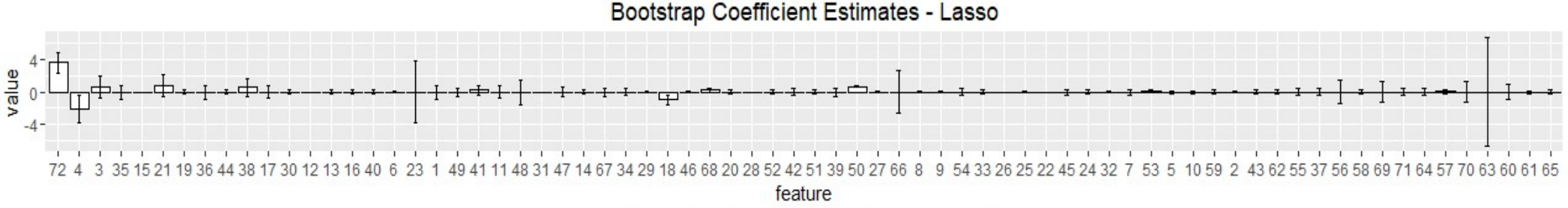

Lasso: 10-fold CV
66 62 59 54 52 45 35 22 17 11 6 5 5 3 3 2
Mean-Squared Error
40
60
80
100
120
log(Lambda)
-4
-2
0
2

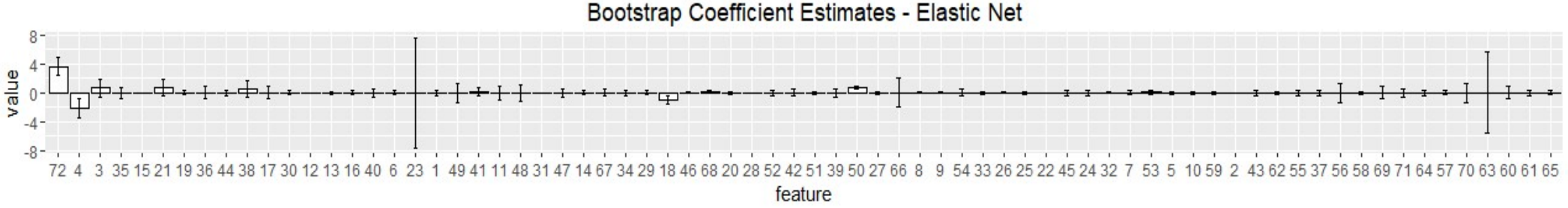

Elastic-Net: 10-fold CV
66 65 61 55 53 47 36 26 19 12 9 8 7 6 5 3 2
Mean-Squared Error
40
60
80
100
120
log(Lambda)
-4
-2
0
2

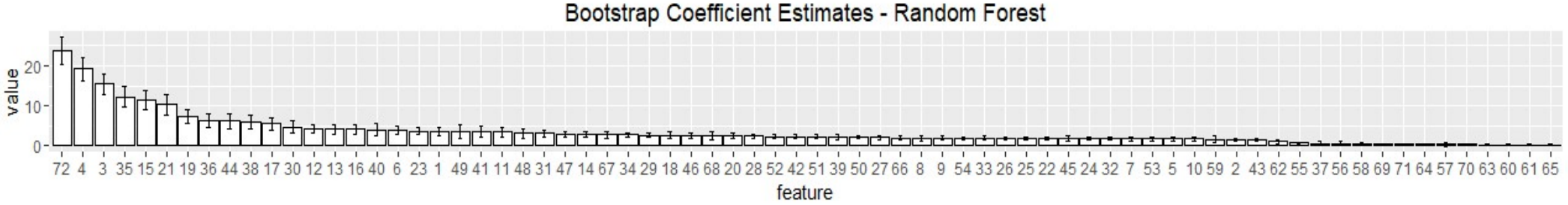

Training Residuals For Each Model
Residuals
20
0
-20
Ridge
Lasso
Elastic-Net
Random Forest
Model


Test Residuals For Each Model
Residuals
10
0
-20
Ridge
Lasso
Elastic-Net
Random Forest
Model


Bootstrap Coefficient Estimates - Ridge
value
4
0
-4
72 4 3 35 15 21 19 36 44 38 17 30 12 13 16 40 6 23 1 49 41 11 48 31 47 14 67 34 29 18 46 68 20 28 52 42 51 39 50 27 66 8 9 54 33 26 25 22 45 24 32 7 53 5 10 59 2 43 62 55 37 56 58 69 71 64 57 70 63 60 61 65
feature


Bootstrap Coefficient Estimates - Lasso
value
4
0
-4
72 4 3 35 15 21 19 36 44 38 17 30 12 13 16 40 6 23 1 49 41 11 48 31 47 14 67 34 29 18 46 68 20 28 52 42 51 39 50 27 66 8 9 54 33 26 25 22 45 24 32 7 53 5 10 59 2 43 62 55 37 56 58 69 71 64 57 70 63 60 61 65
feature


Bootstrap Coefficient Estimates - Elastic Net
value
8
4
0
-4
-8
72 4 3 35 15 21 19 36 44 38 17 30 12 13 16 40 6 23 1 49 41 11 48 31 47 14 67 34 29 18 46 68 20 28 52 42 51 39 50 27 66 8 9 54 33 26 25 22 45 24 32 7 53 5 10 59 2 43 62 55 37 56 58 69 71 64 57 70 63 60 61 65
feature


Bootstrap Coefficient Estimates - Random Forest
value
20
10
0
72 4 3 35 15 21 19 36 44 38 17 30 12 13 16 40 6 23 1 49 41 11 48 31 47 14 67 34 29 18 46 68 20 28 52 42 51 39 50 27 66 8 9 54 33 26 25 22 45 24 32 7 53 5 10 59 2 43 62 55 37 56 58 69 71 64 57 70 63 60 61 65
feature

Predictor Index Legend

1	population	**19**	PctUnemployed	**37**	MedNumBR	**55**	PolicReqPerOffic
2	householdsize	**20**	PctEmploy	**38**	HousVacant	**56**	PolicPerPop
3	racepctblack	**21**	TotalPctDiv	**39**	PctHousOccup	**57**	RacialMatchCommPol
4	racePctWhite	**22**	PersPerFam	**40**	PctHousOwnOcc	**58**	PctPolicWhite
5	racePctAsian	**23**	NumImmig	**41**	PctVacantBoarded	**59**	PctPolicBlack
6	racePctHisp	**24**	PctRecentImmig	**42**	PctVacMore6Mos	**60**	PctPolicHisp
7	agePct12t21	**25**	PctRecImmig5	**43**	MedYrHousBuilt	**61**	PctPolicAsian
8	agePct12t29	**26**	PctRecImmig8	**44**	PctHousNoPhone	**62**	PctPolicMinor
9	agePct16t24	**27**	PctRecImmig10	**45**	PctWOFullPlumb	**63**	OfficAssgnDrugUnits
10	agePct65up	**28**	PctSpeakEnglOnly	**46**	MedRent	**64**	NumKindsDrugsSeiz
11	numbUrban	**29**	PctNotSpeakEnglWell	**47**	MedRentPctHousInc	**65**	PolicAveOTWorked
12	medIncome	**30**	PctLargHouseFam	**48**	NumInShelters	**66**	LandArea
13	medFamInc	**31**	PctLargHouseOccup	**49**	NumStreet	**67**	PopDens
14	perCapInc	**32**	PersPerOccupHous	**50**	PctForeignBorn	**68**	PctUsePubTrans
15	NumUnderPov	**33**	PersPerOwnOccHous	**51**	PctBornSameState	**69**	PolicCars
16	PctLess9thGrade	**34**	PersPerRentOccHous	**52**	PctSameHouse85	**70**	PolicOperBudg
17	PctNotHSGrad	**35**	PctPersDenseHous	**53**	PctSameCity85	**71**	PolicBudgPerPop
18	PctBSorMore	**36**	PctHousLess3BR	**54**	PctSameState85	**72**	nonViolPerPop

Conclusions

- The regularized regression methods experienced similar results across training and testing performance.
  - Lasso: reduced to 35 variables
  - Elastic-Net: reduced to 36 variables
- Random Forest showed signs of overfitting in training but only marginally better than the regularized regression methods in the model testing phase.
- Random Forest took approx. 15 times longer to train.
- All models were generally consistent in selecting their variables of importance

Model	Approx time to train
Ridge	**3 min**
Lasso	**3 min**
Elastic-Net	**3 min**
Random Forest	**45 min**

Important Variables for all models:		Other important variables in Reg. Reg. Methods:	
72	nonViolPerPop	**21**	TotalPctDiv
4	racePctWhite	**38**	HousVacant
3	racepctblack	**23**	NumImmig
35	PctPersDenseHous	**18**	PctBSorMore
15	NumUnderPov	**50**	PctForeignBorn