Homework3

Fjolle Gjonbalaj

08/04/2021

QUESTION 1.

1. The reason why this would be a difficult task is that the rate of crime could be affected by factors other than the number of cops in a city. That is, it is possible that just having a terror alert system in place, regardless of whether there were more cops or not, criminal activity might decrease since robbers could be afraid of the elevated terror attacks. Moreover, the terror alert system might also cause many tourists to hold back and not visit the city where the alert is present, so there are less victims in the streets. For those reasons, it is difficult to simply look at the data and run regressions of Crime on Police.
2. They checked for ridership levels on the metro system to check whether or not the number of victims on High terror days was changer or not. After testing this hypothesis they found no statistical difference between the control Metro ridership group and the treatment group and that the number of victims actually remained largely unchanged. Again, this is to conclude that it is not so easy to establish causation. However, this thought process could eventually lead us in the right direction of establishing true causal relationship between crime level and number of cops in a city.
3. They controlled for Metro ridership in order to establish whether or not there was lower criminal activity in the streets after the High terror alert due to the fact that there were less tourists, and hence less victims in the streets during that day.
4. This table shows the results of regressing crimes by district on a dummy variable set to 1 on high alert days in district one and another dummy variable set to 1 on high alert days in other districts. What the first column in this table shows is that the number or crimes falls by a significant number of approximately 2.6 crimes per day. On the other districts, on the other hand, this reduction is much smaller and not statistically significant, with only approximately 0.6 crime reductions per day. This way, the table shows that there is a significant change (decrease) in criminal activity when a higher police number is available in district 1. However, the same conclusion does not hold for other districts.

QUESTION 2.

First Model: Hand-Built Linear Model.

Second Model: Model Based on Forward Selection

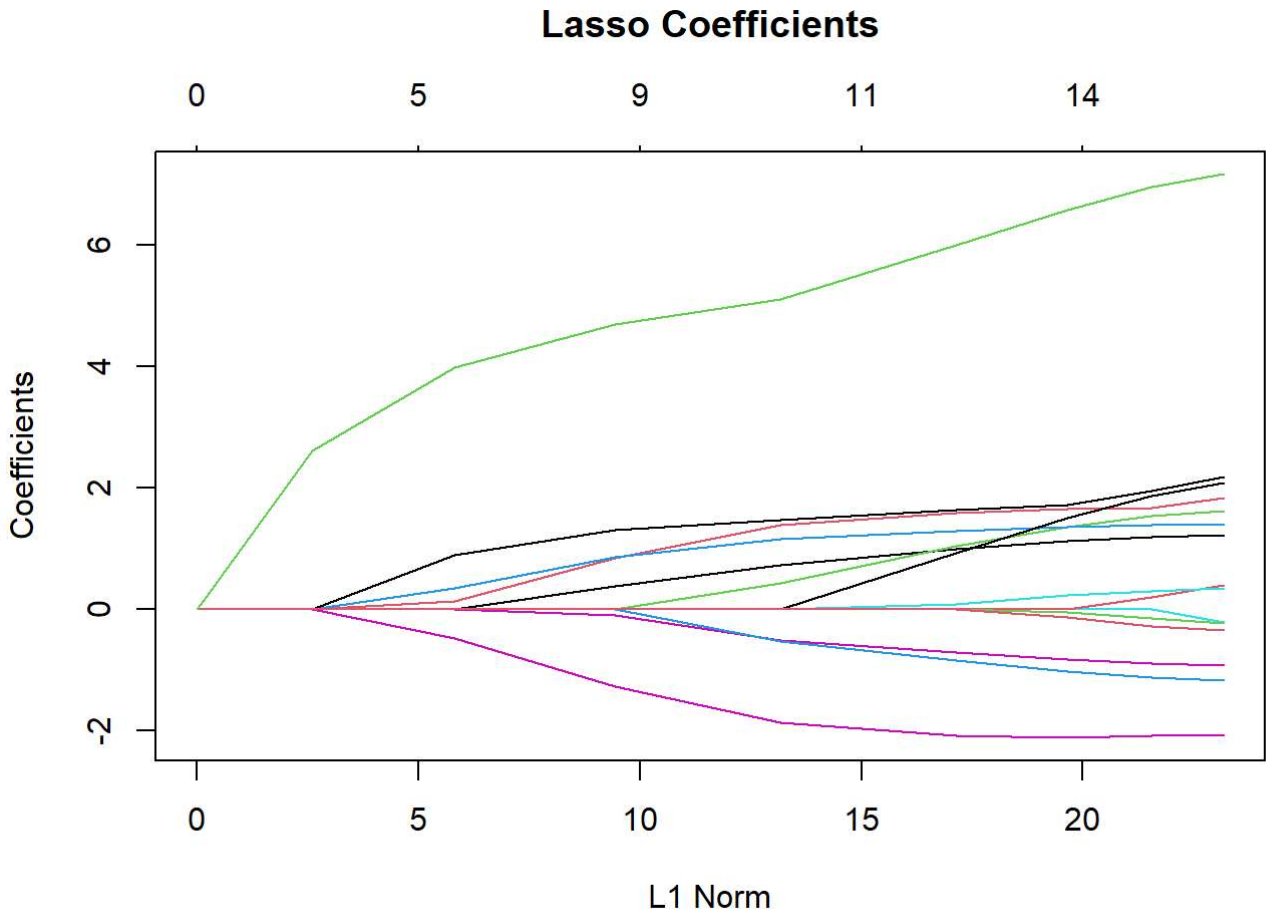
	Hand-Built Linear Model		*Forward Selection Linear Model*	
Predictors	Estimates	p	Estimates	p
(Intercept)	38.71	**<0.001**	63.06	**0.001**
cluster	0.00	**<0.001**	-0.00	**0.020**
size	0.01	**<0.001**	-0.02	**0.024**
empl_gr	0.21	**<0.001**	13.09	**<0.001**
leasing_rate	0.06	**<0.001**	-0.08	**0.034**
stories	-0.08	**<0.001**	-0.30	0.078
cd_total_07	-0.01	**<0.001**	-0.05	**<0.001**

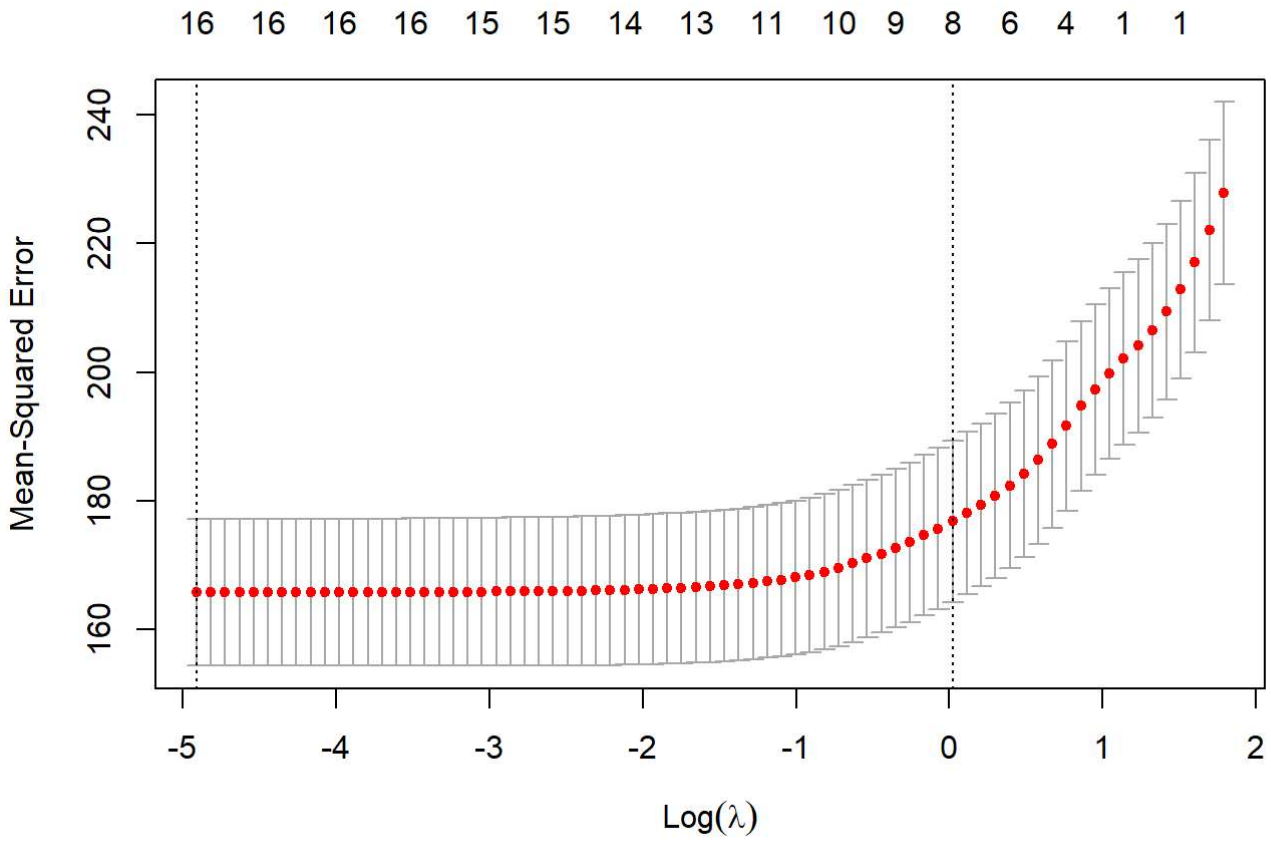
Electricity_Costs	-787.36	**<0.001**	1096.56	**<0.001**
age	-0.03	**<0.001**	-0.09	0.066
renovated	-1.10	**0.001**	-0.95	0.164
class_a	4.45	**<0.001**	29.93	**<0.001**
class_b	1.44	**0.001**	31.31	**<0.001**
green_rating	-0.61	0.218		
net	-4.85	**<0.001**	-3.63	0.080
amenities	0.76	**0.017**	-4.94	**0.013**
Gas_Costs	779.47	**<0.001**	769.30	0.586
hd_total07	-0.01	**<0.001**	-0.02	**<0.001**
size * stories	-0.00	0.863	-0.00	**0.002**
cd_total_07 * Electricity_Costs	0.13	**<0.001**		
Electricity_Costs * hd_total07	0.44	**<0.001**	0.40	**<0.001**
Electricity_Costs * cd_total_07			0.72	**<0.001**
Electricity_Costs * empl_gr			-326.61	**<0.001**
Electricity_Costs * class_a			-390.15	**<0.001**
Electricity_Costs * size			1.44	**<0.001**
hd_total07 * size			0.00	**<0.001**
class_a * hd_total07			-0.00	**<0.001**
hd_total07 * empl_gr			-0.00	**<0.001**
cd_total_07 * empl_gr			-0.00	**0.012**
size * cluster			0.00	**<0.001**
Electricity_Costs * cluster			0.16	**<0.001**
Electricity_Costs * Gas_Costs			-233151.34	**<0.001**
Electricity_Costs * age			7.09	**<0.001**
hd_total07 * age			0.00	**<0.001**

cd_total_07 * hd_total07	0.00	**<0.001**
cd_total_07 * Gas_Costs	2.63	**<0.001**
hd_total07 * Gas_Costs	0.70	**<0.001**
size * Gas_Costs	-0.79	0.162
class_a * leasing_rate	0.04	**0.041**
Gas_Costs * amenities	598.59	**<0.001**
class_a * amenities	-2.34	**0.002**
age * Gas_Costs	-17.72	**<0.001**
age * amenities	-0.01	0.340
cd_total_07 * age	0.00	0.102
empl_gr * Gas_Costs	-302.98	**0.010**
class_a * empl_gr	0.10	**0.020**
class_a * Gas_Costs	-140.09	0.452
leasing_rate * size	0.00	**0.006**
cd_total_07 * net	0.00	**0.001**
age * class_b	-0.06	**<0.001**
class_a * age	-0.03	0.105
Electricity_Costs * class_b	-532.37	**<0.001**
hd_total07 * class_b	-0.00	**<0.001**
cd_total_07 * cluster	-0.00	**0.004**
empl_gr * cluster	0.00	**0.001**
age * size	-0.00	**<0.001**
class_a * size	-0.02	**<0.001**
size * class_b	-0.02	**<0.001**
size * amenities	0.01	**0.001**
empl_gr * size	0.00	0.186
hd_total07 * renovated	0.00	**0.021**
hd_total07 * stories	0.00	**<0.001**
class_b * stories	0.27	**0.007**
renovated * stories	-0.20	**<0.001**

size * renovated	0.01	**<0.001**
hd_total07 * amenities	-0.00	0.057
Electricity_Costs * leasing_rate	1.94	**0.019**
cd_total_07 * class_b	-0.00	**<0.001**
class_a * cd_total_07	-0.00	**<0.001**
cd_total_07 * stories	0.00	**<0.001**
age * stories	0.00	**0.007**
class_a * stories	0.25	**0.019**
cluster * class_b	-0.00	0.110
net * Gas_Costs	-326.14	0.099
Gas_Costs * stories	-30.76	**0.009**
cd_total_07 * size	-0.00	**0.025**
leasing_rate * Gas_Costs	4.50	0.107
Observations	7820	7820
R^2 / R^2 adjusted	0.397 / 0.396	0.484 / 0.480

Third Model: Lasso





Lasso Model Predictor Estimates

Variable	Estimate

Variable	Estimate
(Intercept)	28.4209220
cluster	1.2173177
size	1.8381229
empl_gr	1.6201395
leasing_rate	1.4038241
stories	-0.2082164
age	0.0000000
renovated	-0.9241263
class_a	2.1812377
class_b	0.3908465
green_rating	-0.2311674
net	-1.1683912
amenities	0.3469201
cd_total_07	-2.0597298
hd_total07	2.0855071
Gas_Costs	-0.3470659
Electricity_Costs	7.1581711

The Lasso model gives us the variance-bias trade off. When we observe lambda to be high, the variance decreases but the bias increases. To select the lambda parameter I use cross-validation and compute the leave-one-out cross validation error for each value.

In the graph I show the lambda with the least CV Mean Squared Error. The range of the cv MSE is rather wide and gives us results which suggest that theere is a wide range of values for lambda that give us similar errors. Since lambda is essentially zero, the results from the Lasso model will be close to the least squares model, and hence result in a high variance but low bias.

In the table created we see that all coefficient estimates are nonzero.

For the linear model

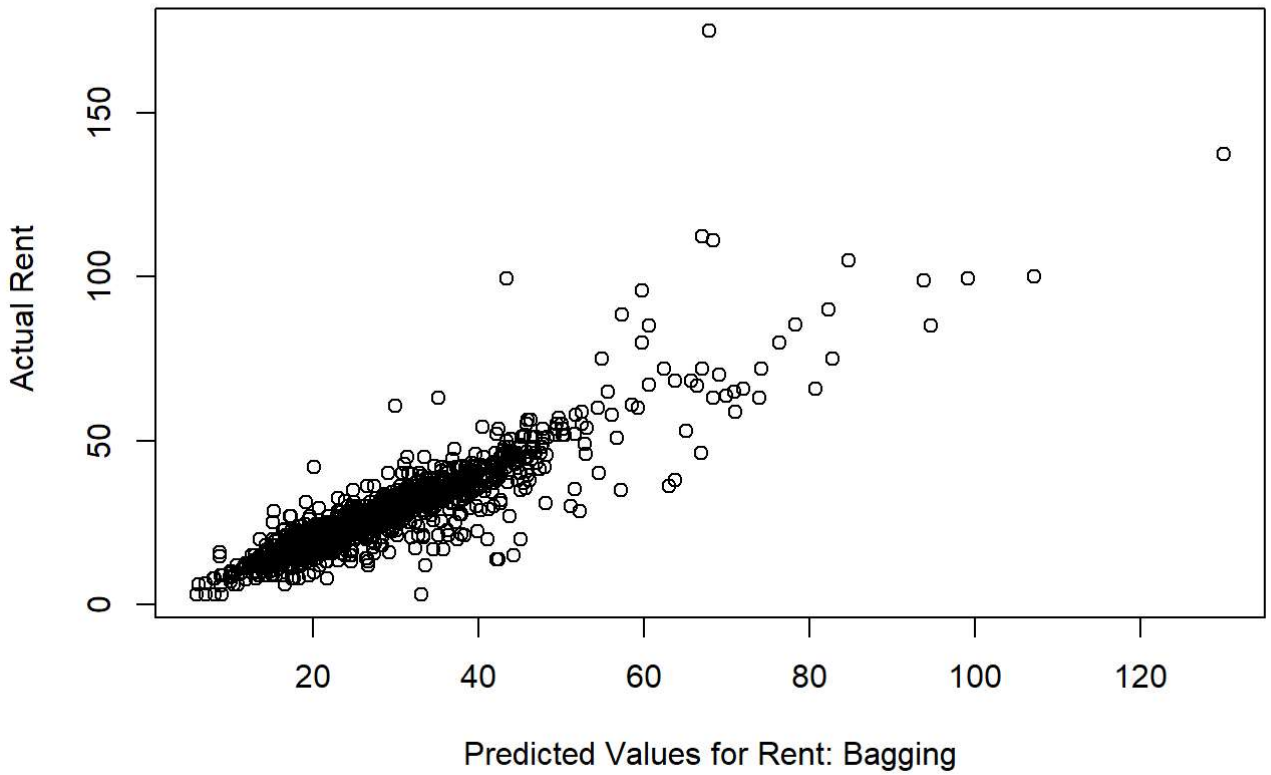
	Mean RMSE
Hand-Built Linear Model	11.77410
Forward Selection Linear Model	10.98140

	Mean RMSE
Lasso	12.15874

Here we can see that the lasso regression is not superior in predicting in comparison to the OLS, based on RMSE. Both Lasso and OLS give us similar results in that they both have a high variance but a low bias.

Forth model: Tree-Based Models: Bagging

```
##
## Call:
##  randomForest(formula = Rent ~ cluster + empl_gr + stories + age +      renovated + leasin
g_rate + class_a + class_b + amenities +      green_rating + net + cd_total_07 + hd_total07 +
Precipitation +      Gas_Costs + size + Electricity_Costs, data = greenbuildings_train,
mtry = 17, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 17
##
##           Mean of squared residuals: 49.71979
##                     % Var explained: 79.02
```

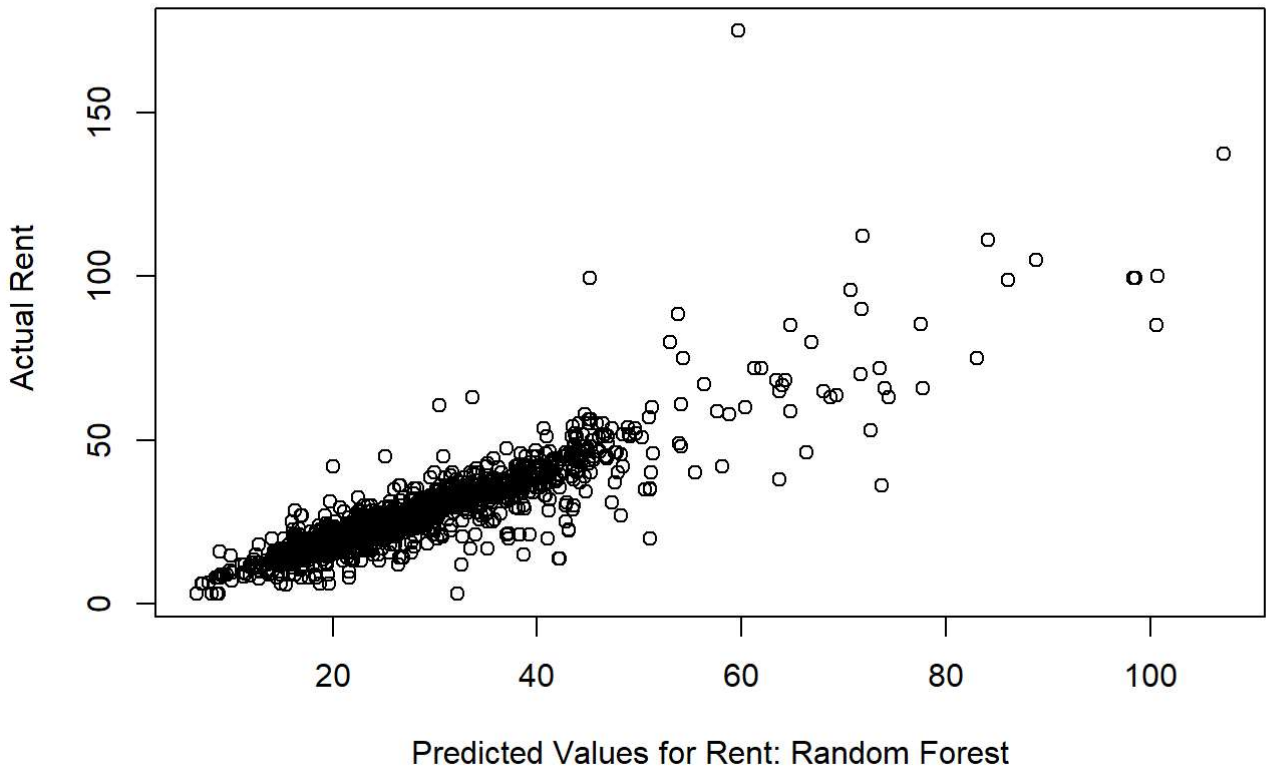


Trees that are fitted individually tend to result in a high variance. That is, a decision tree of one part of the data can result in outcomes very different from another part of the data. This is the reason why the bagging procedure is useful in this case. The latter uses bootstrapping by taking repeated samples from the training set and averages the predictions. This method gives us a reduced variance and provides us with a more accurate prediction. With bagging we've created 500 trees. Although each individual tree has high variance and low bias, averaging them out will give us a much smaller variance.

In this case we use a regression random forest type and apply it to all the variables in the data set. We get a Mean of squared residuals equal to 48.92116 and % Var explained in the model equal to 78.87.

Since Lasso has a higher RMSE than the linear regression model, I perform two tree decision models using bagging and random forest. Although no model is superior to the other in all possible aspects, the random forest model improves upon the bagging procedure in that the former bootstraps on the training samples also. However, this model does not consider all variables in each split but rather selects a set of these variables for the tree split. I use a square root of the number of original variables for the sample of predictors to be considered. This gives us around 4 variables to be considered at each split. This method improves on the bagging method in that by only considering 4 variables instead of all 17 of them we reduce the highly correlated predictions. Hence, random forest usually gives us a lower variance.

Fifth model: Random Forest



The plot shows the random forest model prediction accuracy.

Comparison of all 5 Predictive Models:

```
## [1] 11.76975
```

```
## [1] 11.0412
```

```
## [1] 12.15172
```

LOOCV RMSE per Model

	LOOCV RMSE

	LOOCV RMSE
Hand-Built Model	11.769746
Forward Selection Model	11.041201
Lasso Model	12.151720
Bagging Model	7.154923
RandomForest Model	7.378792

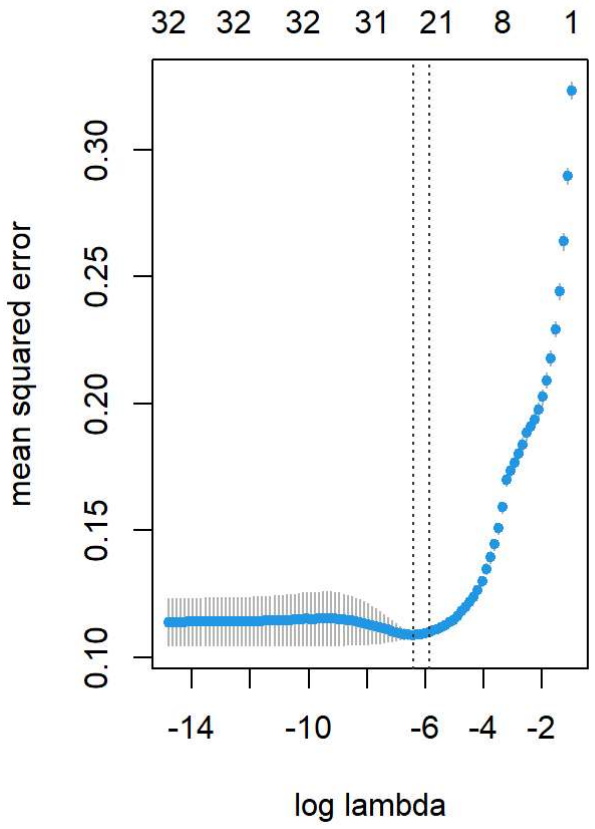
In order to compare all the models considered above we can use the leave one out cross validation. The list shows the RMSEs from cross validation.

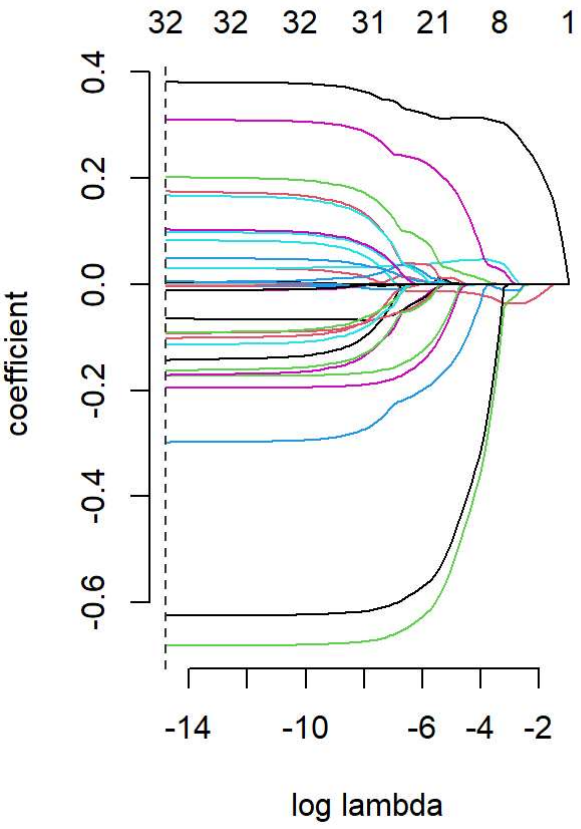
With a lower RMSE the tree decision models perform a lot better than the other three models. I use Random Forest as a predictive model for rent since it is quite similar to the RMSE for bagging but simultaneously produces improved results to those of bagging.

PROBLEM 3.

First Model: Hand-Built Linear Model

Second Model: Lasso





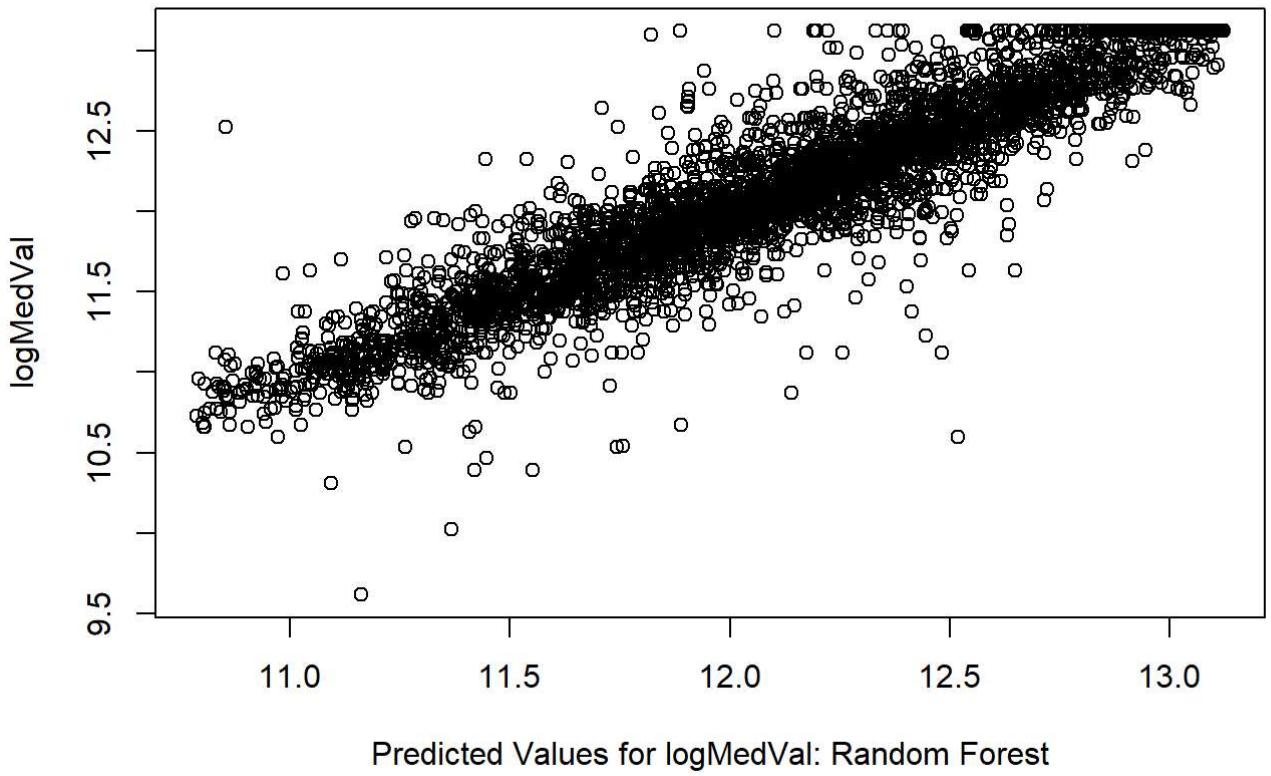
```
## 32 x 1 sparse Matrix of class "dgCMatrix"
##                                    seg36
## intercept                          12.07
## longitude                          -0.56
## latitude                           -0.62
## housingMedianAge                    0.04
## population                         -0.19
## households                          0.23
## medianIncome                        0.32
## AveBedrooms                         0.01
## AveRooms                           -0.02
## AveOccupancy                        .
## longitude:latitude                  .
## longitude:housingMedianAge         -0.12
## longitude:population                .
## longitude:households               -0.02
## longitude:medianIncome              0.02
## longitude:AveBedrooms               .
## longitude:AveRooms                  0.04
## longitude:AveOccupancy              .
## latitude:housingMedianAge          -0.14
## latitude:population                -0.03
## latitude:households                 .
## latitude:medianIncome               0.00
## latitude:AveBedrooms               -0.03
## latitude:AveRooms                   0.08
## latitude:AveOccupancy               0.00
## longitude:latitude:housingMedianAge .
## longitude:latitude:population       .
## longitude:latitude:households       .
## longitude:latitude:medianIncome    -0.01
## longitude:latitude:AveBedrooms      0.00
## longitude:latitude:AveRooms         .
## longitude:latitude:AveOccupancy     .
```

The Lasso model gives us the variance-bias trade off. When we observe lambda to be high, the variance decreases but the bias increases.

In the graph we observe a wide range of MSE of values for lambda that give us similar errors. Log lambda represents the penalizing factor for the sum of absolute values of coefficients. We obtain the optimal log lambda value by repeating the cross-validation.

Third model: Tree-Based Models Bagging

Random Forrest

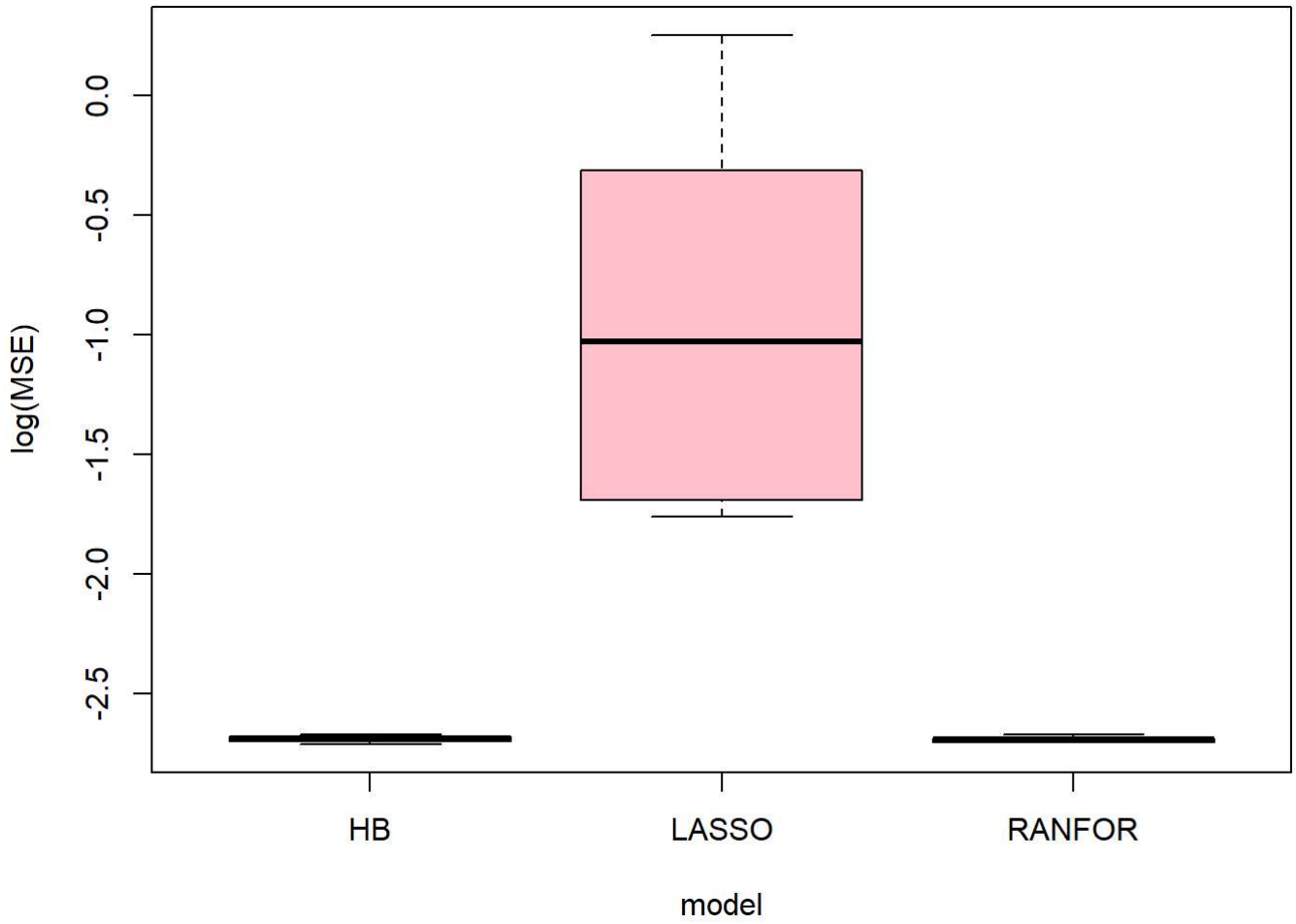


The plot shows the random forest model prediction accuracy.

Comparison of the 3 Predictive Models: Hand-Built Linear Model, Forward Selection, Lasso, Bagging and Random Forest

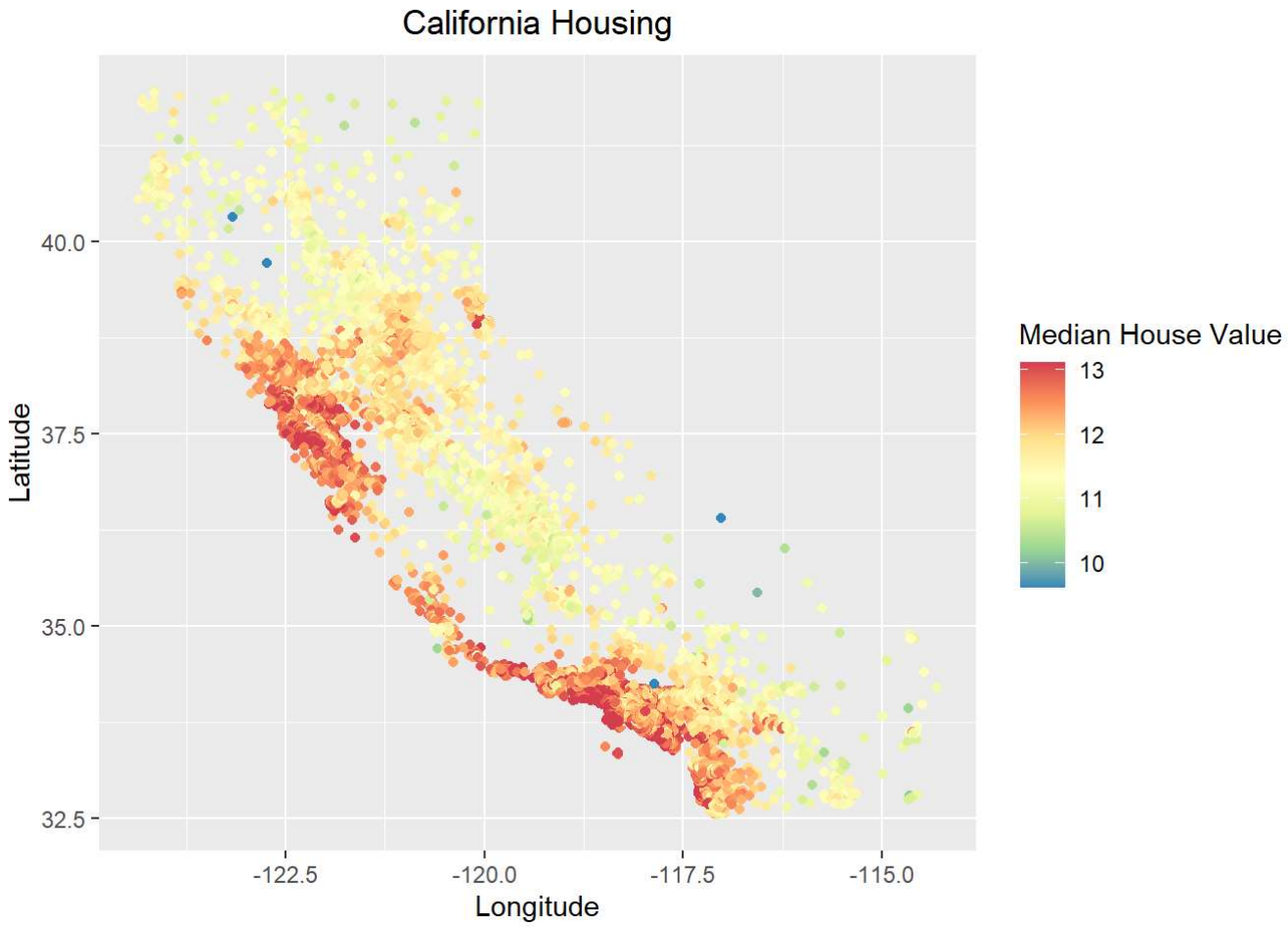
Out of sample prediction

```
## 12345678910
```



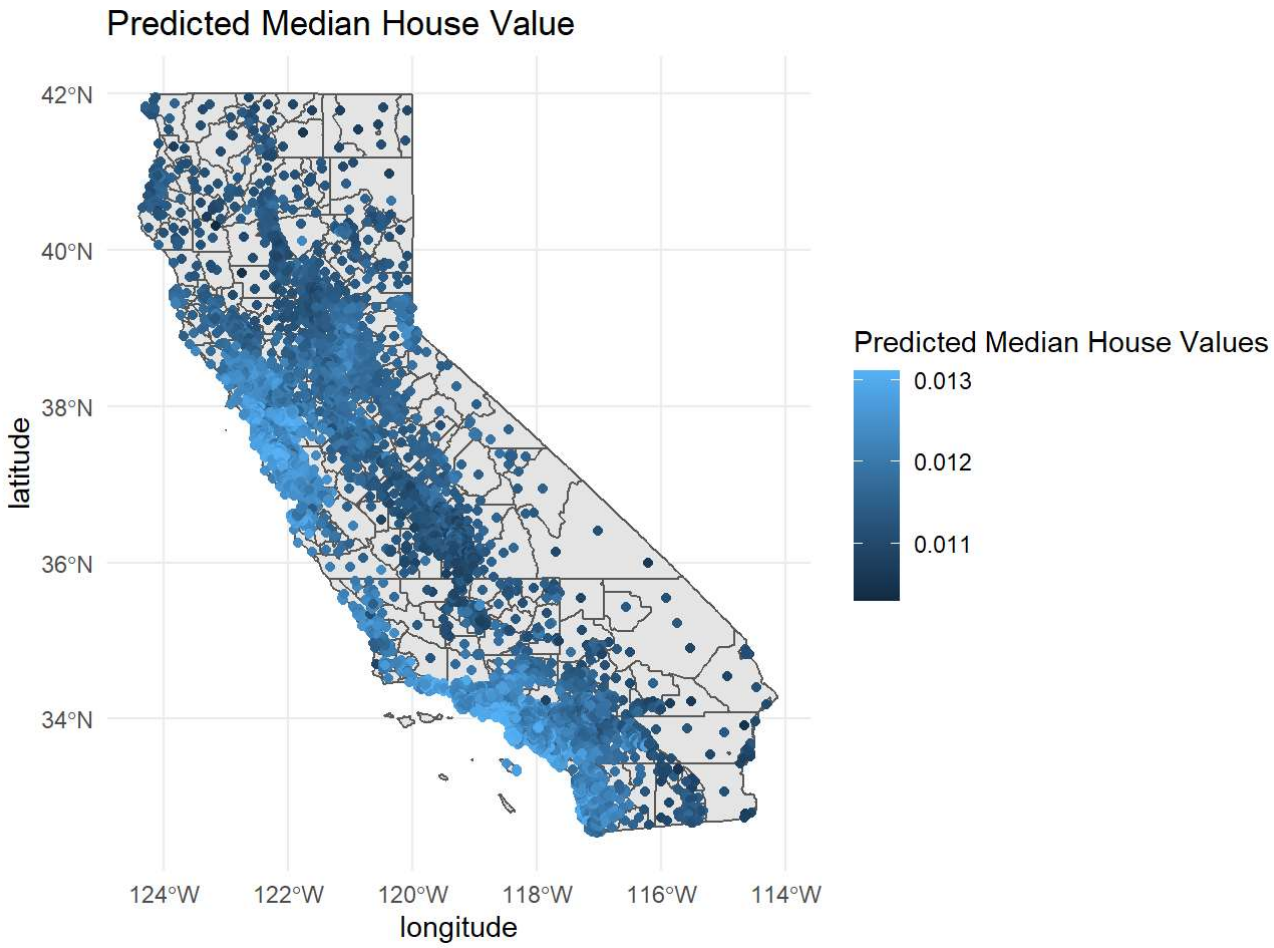
Although each time I run the code I get a different result of the boxplots due to different training and testing data sets, the most common result is that the Random Forest Model gives us the lowest log(MSE). The random forest model bootstraps on the training samples. This model does not consider all variables in each split but rather selects a set of these variables for the tree split.

First plot: Original Data




```
## Using FIPS code '06' for state 'CA'
```

Second Plot: Model's predictions



Third Plot: Model's error's/residuals