Data Mining – Exercise 3

Yangxi & Chen & Shuyan

Problem 1

1.Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime?

The reason is simultaneity. It's hard to tease out the difference between more policy leading to crime or more crime leading to more police. We can assume that when a city has more police, there would be lower crime rate, but a high crime rate area, it's common to hire more cops. As a result, the issue of endogeneity arises. Hence, the coefficient would be biased if we directly make regression of 'Crime' on 'Police'.

2.How were the researchers from UPenn able to isolate this effect?

What the researchers at UPenn did was to find a natural experiment. They were able to collect data on crime in DC, where the number of police is unrelated to crime. By law, because Washington, D.C., is likely to be a terrorism target, and there exists a terrorism alert system. When the terror alert level goes to orange, then extra police are put on the Mall and other parts of Washington to protect against terrorists. It has nothing to do with street crime or things like that, but when you have the extra police there for terrorism-related reasons, they're on the streets, they make the streets safer, and things like murder, robbery, assault go down. Besides, they chose ridership as a control variable, and then checked the hypothesis that tourists were less likely to visit Washington or to go out by looking at ridership levels on the Metro system.

As a results, from table 2 we see that controlling for ridership in the METRO, days with a high alert (which was a dummy variable) have lower crime as the coefficient is negative for sure.

3.Why did they have to control for Metro ridership? What was that trying to capture?

If people were not out and about during the high alert days there would be fewer opportunities for crime and hence less crime (not due to more police). Controlling for Metro would make sure the citizens outdoor are unchanged, which means the crime rate does not decrease due to less outdoor activities causing by alert. The goal is to clearly isolate other interrupting factors and makes the coefficient completely reflect the relation between the number of cops and crime rate. The results from the table tells us that holding ridership fix more police has a negative impact on crime.

4.Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

In table 4 they just refined the analysis a little further to check whether or not the effect of high alert days on crime was the same in all areas of town. Using interactions between location and high alert days they found that the effect is only clear in district 1. Again, this makes a lot of sense as most of the potential terrorists targets in DC are in District 1 and that’s where more cops are most likely deployed to.

Problem 2

Model Selection (Tree/Randomforest/Boosting)

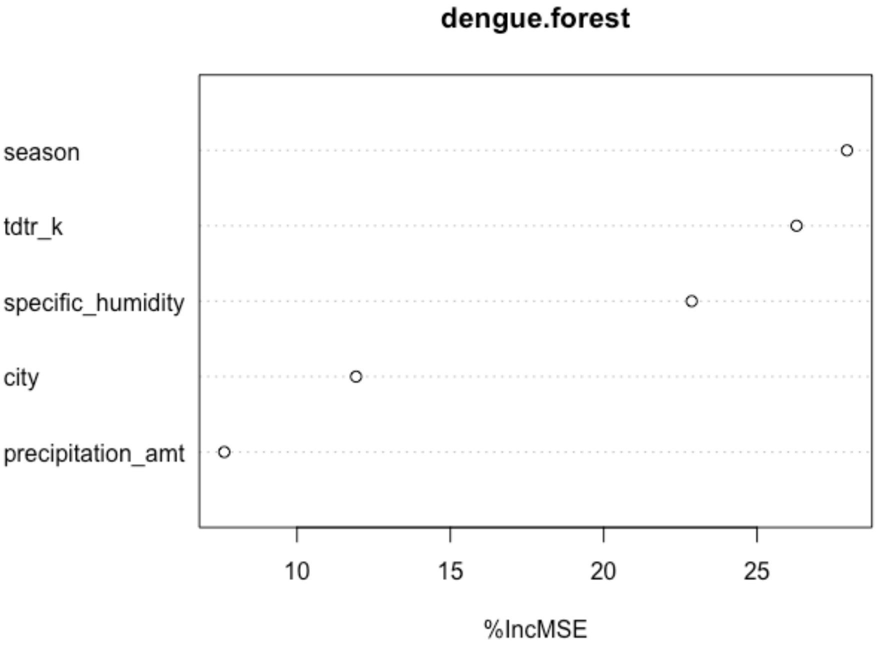
Using tree, randomforest and boosting to predict in the training set with cross validation in default.

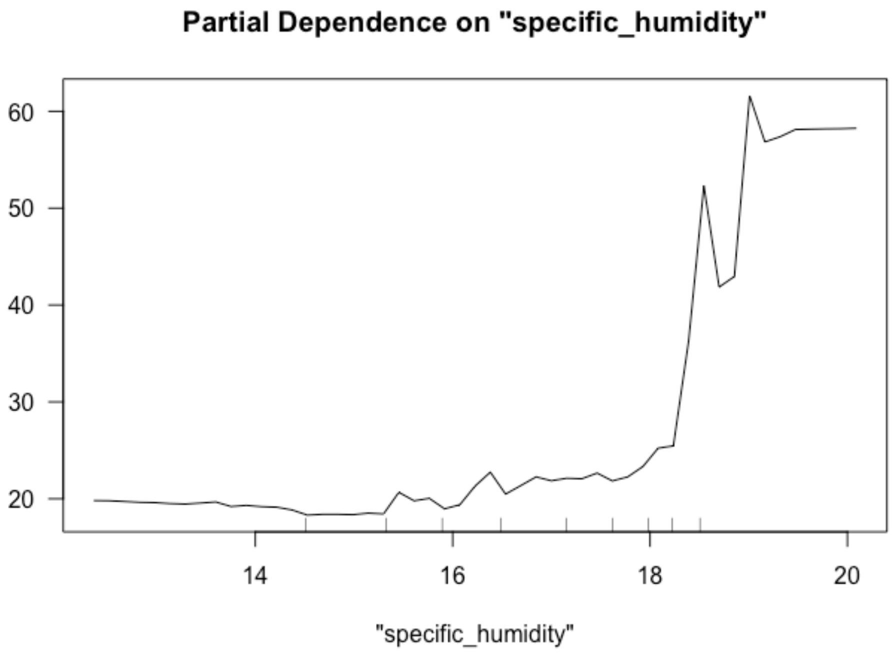
The RMSE among tree/randomforest/boosting within testing set are as follows.

Table Model RMSE

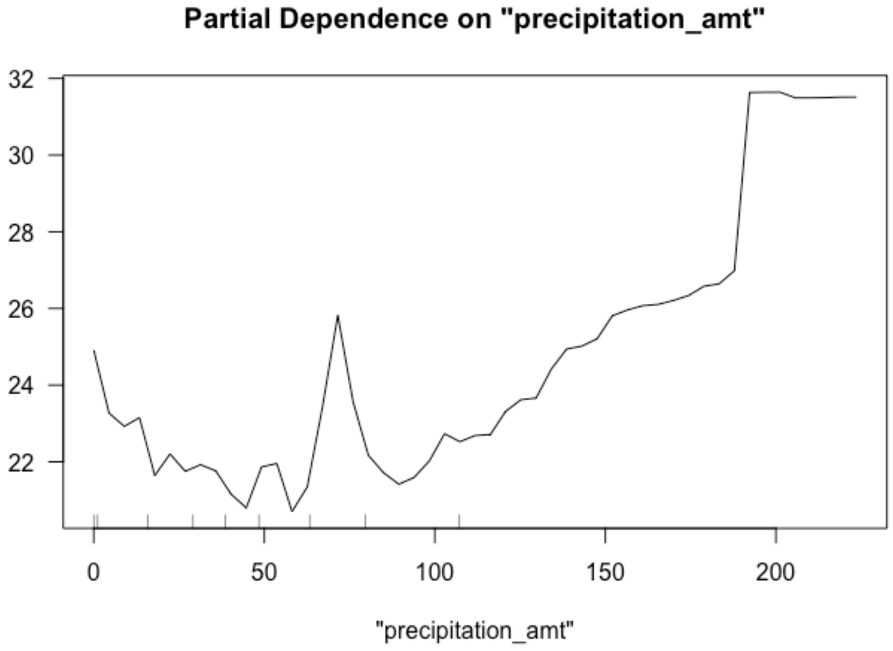
Model	RMSE
RMSE_tree	29.50763
RMSE_random_forest	27.37638
RMSE_gbm	27.58587

Here are the graph of partial dependence among **specific_humidity**, **precipitation_amt** and **season**.

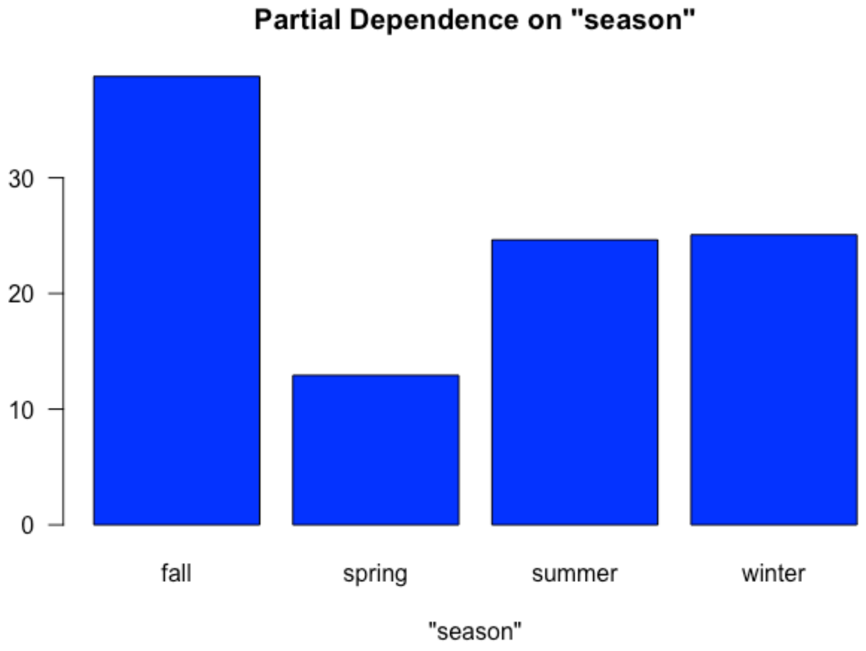




Partial Dependence on "specific_humidity"



Partial Dependence on "precipitation_amt"


Partial Dependence on "season"
30
20
10
0
fall
spring
summer
winter
"season"

Problem 3

Introduction

Considering the data set on green buildings in *greenbuildings.csv*, which contains data on 7894 commercial rental properties from across the United States, and 685 of them have been awarded green buildings. The goal of this report is to build the best predictive model possible.

Model Selection

At the beginning, because rent income is decided by rent and leasing rate, so we create a column – **rent_income=rent*leasing rate**. In addition, there are two green certifications: **LEED and Energystar**, and the column **'*green rating*'** can identify these two features, so we maker rename it as **'*green_certified*'**.

1. Forward Selection Linear Model

Firstly, we use forward selection to build a linear model. Then, we get the best linear model:

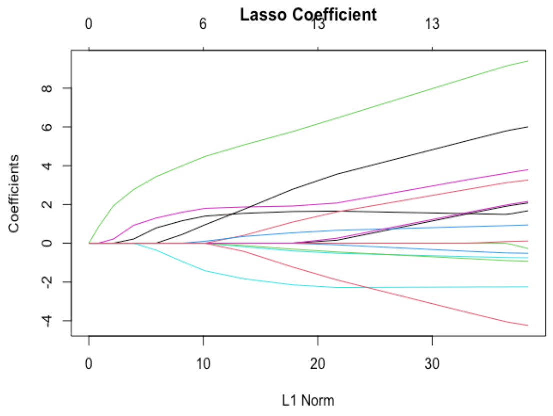
rent_income = Electricity_Costs +size + cd_total_07 + Precipitation + class_a + class_b + age + amenities + hd_total07 + net + Electricity_Costs:size +Electricity_Costs:cd_total_07 + Electricity_Costs:Precipitation + cd_total_07:Precipitation + Electricity_Costs:age + size:Precipitation + Precipitation:class_a + cd_total_07:age + size:cd_total_07 + Precipitation:hd_total07 + class_a:amenities + size:hd_total07 + age:hd_total07 + Electricity_Costs:hd_total07

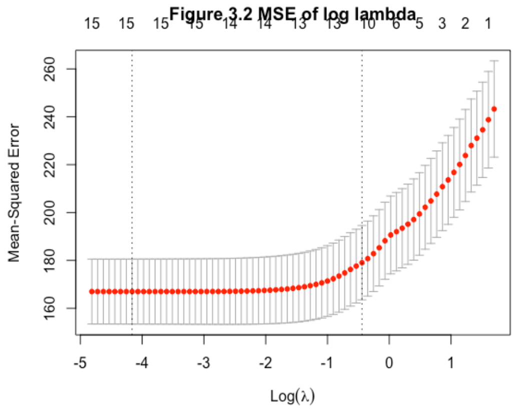
Because the forward model includes many variables, and nonzero choices of beta j also has cost(the variance), which is hidden in the maximum likelihood estimation. By using a shrinkage method–*Lasso*, we continue by optimality and concentrate on stabilizing the system.

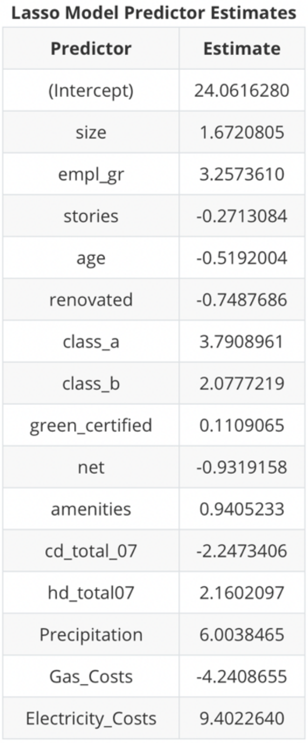
2. Lasso Model

Because the forward model includes many variables, and nonzero choices of *beta j* also has cost (the variance), which is hidden in the maximum likelihood estimation. By using a shrinkage method--Lasso, we continue by optimality and concentrate on stabilizing the system.

The accuracy of lasso model mainly rely on the choice of lambda, below is the coefficient plot. In the lasso model, we need to consider the variance-bias trade off. Large lambda means the high biase and the low variance.





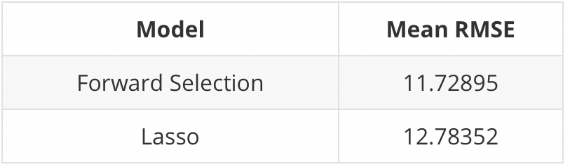


Lasso Model Predictor Estimates

Predictor	Estimate
(Intercept)	24.0616280
size	1.6720805
empl_gr	3.2573610
stories	-0.2713084
age	-0.5192004
renovated	-0.7487686
class_a	3.7908961
class_b	2.0777219
green_certified	0.1109065
net	-0.9319158
amenities	0.9405233
cd_total_07	-2.2473406
hd_total07	2.1602097
Precipitation	6.0038465
Gas_Costs	-4.2408655
Electricity_Costs	9.4022640

we can see that the best lamda is ***0.008***, which is very close to zero, the result of lasso model will be close to the OLS models with high variance but low bias. Table shows when we using ***lambda=0.008***, the coefficient estimates.

- **Compare the Forward Selection and Lasso.**



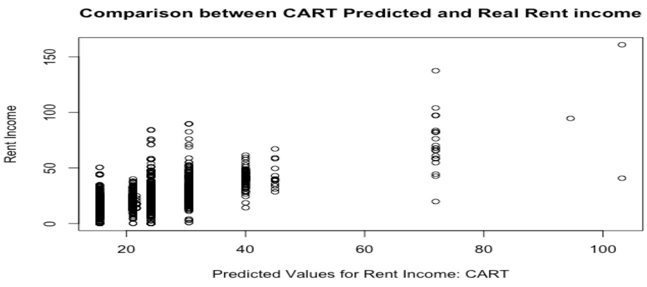
Model	Mean RMSE
Forward Selection	11.72895
Lasso	12.78352

Now we use cross validation to calculate the mean *RMSE* of forward selection model and lasso model. By comparing the average *RMSE*, we conclude that the lasso model does not predict better.

The results above shows that lasso model has higher RMSE than the linear regression, so next we use 4 tree models: *classification and regression tree, random forest, bagging, and boosted tree.*

3. CART

A *Classification And Regression Tree (CART)*, is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

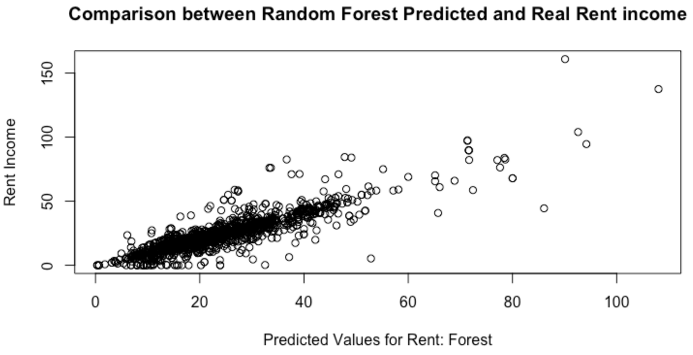


The right plot shows that the classification and regression procedure.

4. Random Forest

Random forest builds decision trees by bootstrapping the training samples, and it only choose a set of variables for the tree split. We use only *4 predictors* instead of all variables so to avoid highly correlated predictions.

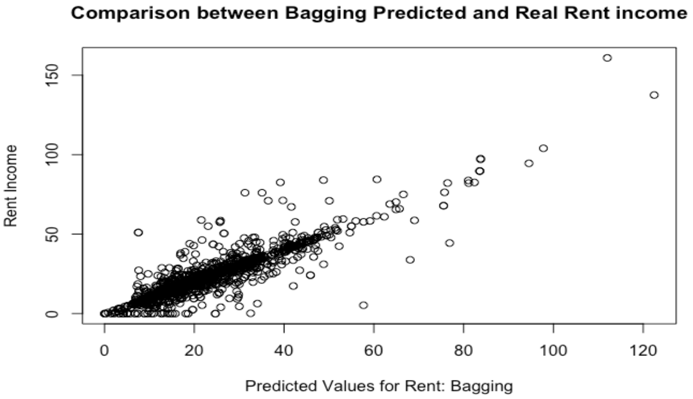
The right plot shows the accuracy of the random forest prediction model.



5. Bagging

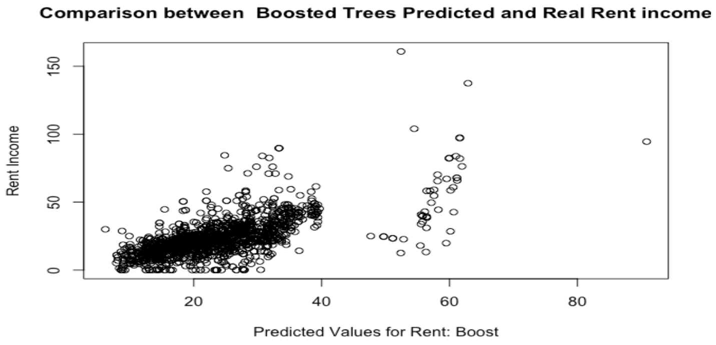
Now we using *bagging* to average predictions to reduce estimation variance without adding biases. The bagging procedure has created *150 trees* by bootstrapping, and all *15 variables* are considered at each split since the trees are not pruned.

The right plot shows the accuracy of the bagging procedure, and we can see it's more accurate than CART and random forest.



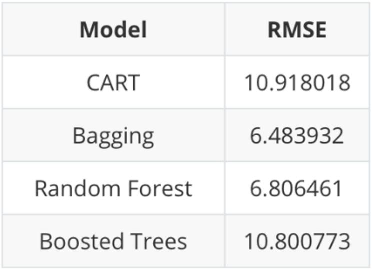
6. Boosted Trees

Like random forest, *boosting* is an ensemble method. The right plot below shows the accuracy of the boosting procedure.



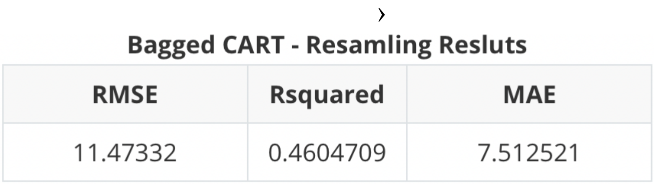
- **Compare 6 Predictive Models**

Now we compare the *RMSE* among these tree models, it shows that the Bagging model has the lowest *RMSE*, the second best model is the Random Forest, which much better than the forward selected linear model.



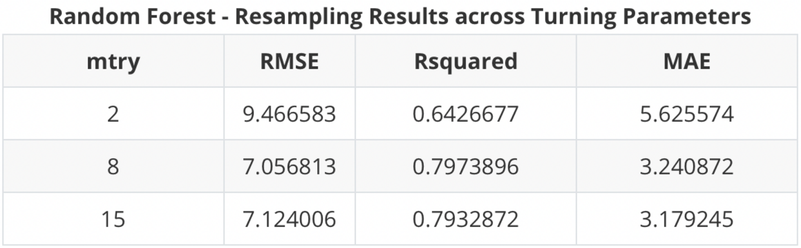
Model	RMSE
CART	10.918018
Bagging	6.483932
Random Forest	6.806461
Boosted Trees	10.800773

Now we calculate the *k-fold cross-validation standard error* for Bagging and Random Forest Tree models.


›

Bagged CART - Resamling Resluts

RMSE	Rsquared	MAE
11.47332	0.4604709	7.512521

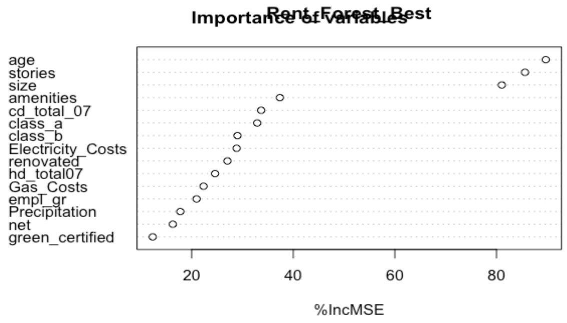

Random Forest - Resampling Results across Turning Parameters

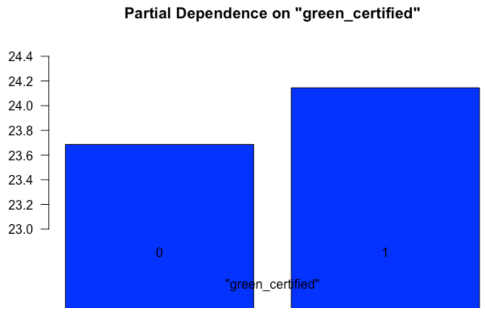
mtry	RMSE	Rsquared	MAE
2	9.466583	0.6426677	5.625574
8	7.056813	0.7973896	3.240872
15	7.124006	0.7932872	3.179245

It concludes that the *Random Forest Mode*l with **mtry=8** is the best prediction model.

Next we measure the *importance* of each variable in the Random Forest model.

It shows that the **'size','stories',and 'age'** are the top 3 important variables, and **green_certification** doesn't seem to have much influence on rent income, and then we plot the partial influence of *green_certification*.





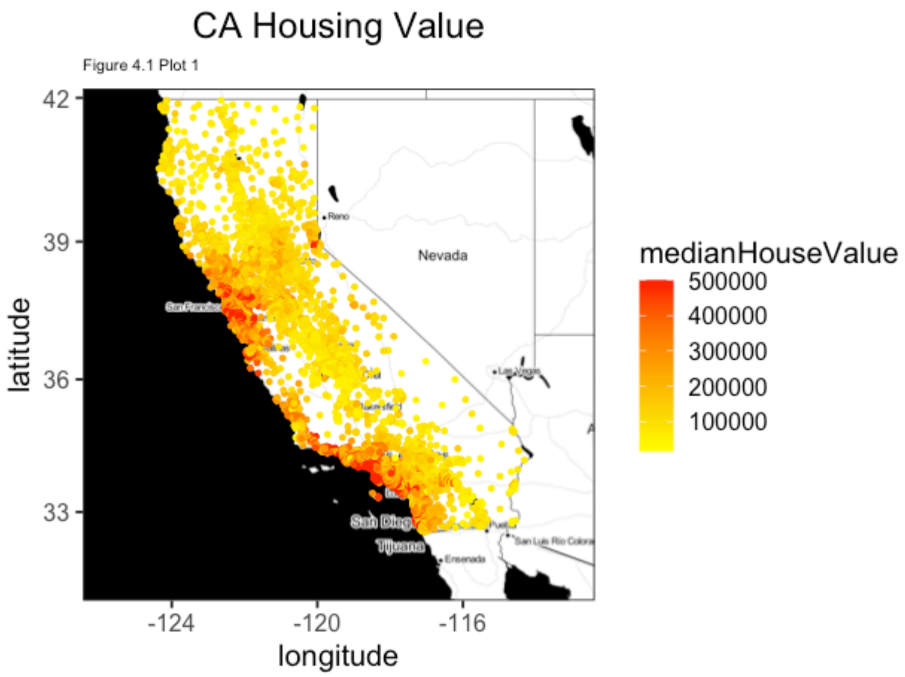
From the upper right plot, we can see that the average effect of green certification on the rent income is about 0.5.

Conclusion

The best predictive models possible for rent income is the *Random Forest Model*.The average change in rental income per square foot related to green certification, holding all else fixed, is **0.5** dollars per square foot.

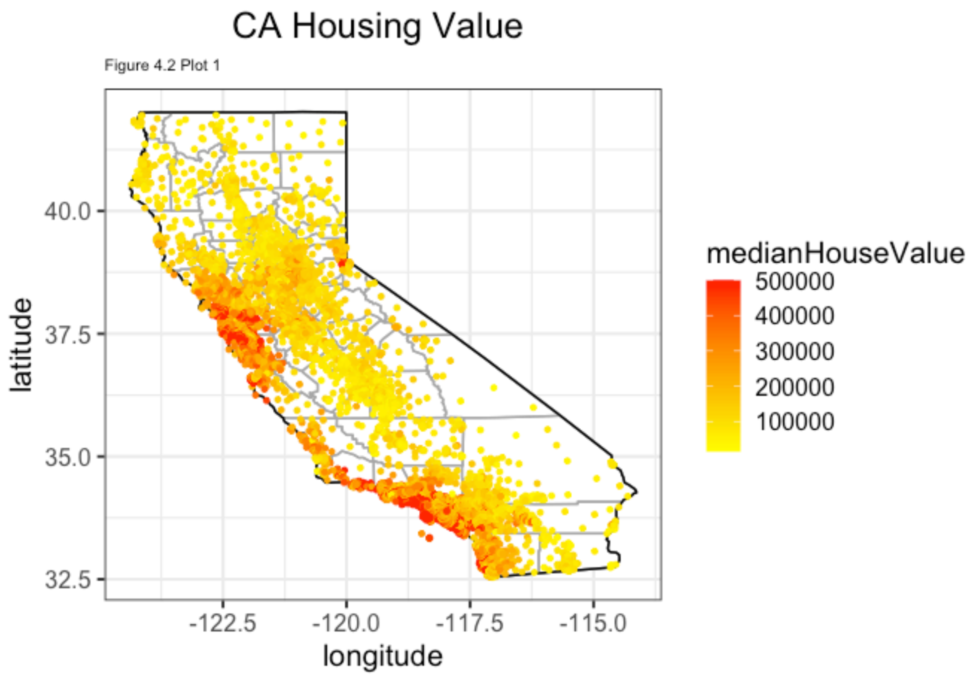
Problem 4

1-1. Plot the map between medianhousevalue in California



Note: The redder the color of the point is, the larger the median housing value is

1-2. Plot the map between medianhousevalue in California (or If API does not work)



Note: The redder the color of the point is, the larger the prediction value is

2. Prediction

（1）linear model

Based on the intuition and lasso regression, we set **"housingMedianAge medianIncome longitude latitude"** as prototype. Then, we use **step** to find a linear model with minimum AIC.

lm_step1 includes all variables in linear model

lm_step2 includes all variables in quadratic model

We add lasso regression to reduce the influence of overfitting and over-complexity. Combined all the models mentioned above, calculate the average rmse under cross validation with 10 folds. The result illustrates that *RMSE_lm_step2* is minimum, which means *quadratic model*
with step is better off.

Table 4.1 Linear Model RMSE

Model	RMSE
RMSE_lm1	73531.05
RMSE_lm_step1	69496.45
RMSE_lm_step2	65453.70
RMSE_lm_lasso	78740.24

（2）Decision Tree

We use *decision tree* to find the optimal model involving *tree model, bagging, random forest and boosting*.

For tree, cross-validation is required to find almost best model with less complexity. We use **xval** to set 10 folds in default.

For bagging,random forest model and boosting, we don’t necessarily do cross-validation again because the mechanism involves the principle of cross-validation when decides for each optimal tree.

In Random forest, we only use 3 variable.

The result shows that decision tree has advantage over traditional linear model and **bagging** is slightly better off (similar like random forest), significantly improving the performance.

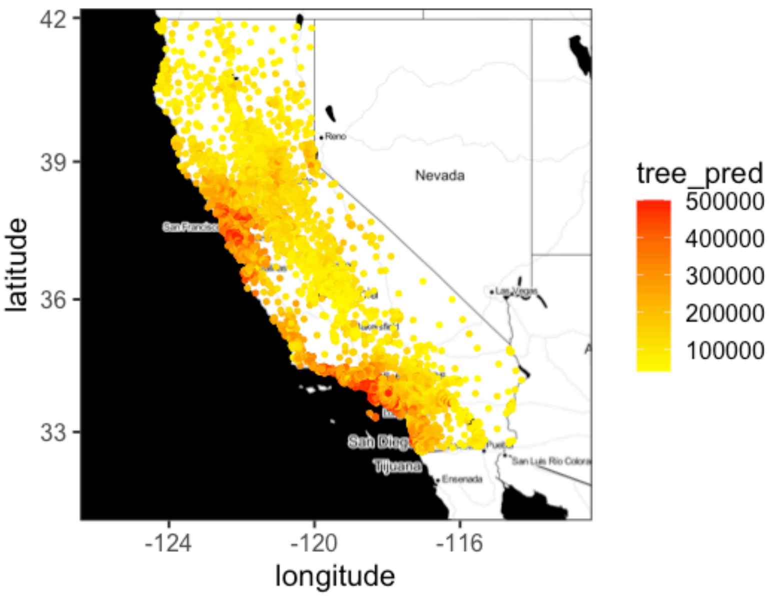
Table 4.2 Decision TREE RMSE

Model	RMSE
Decision Tree	57335.44
Bagging	48737.99
Random Forest	49270.44
Boosting	62014.40

Based on the best **rmse**, **bagging** is chosen to predict. The plot of error and predicting value are as follows.

CA Housing Prediction

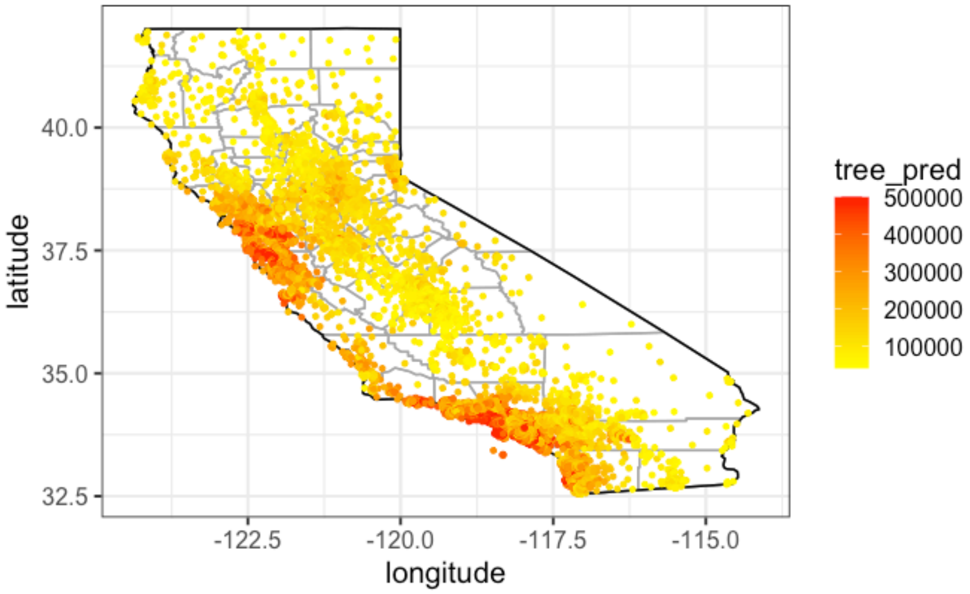
Figure 4.3 Plot 2



Note: The redder the color of the point is, the larger the prediction value is

CA Housing Prediction

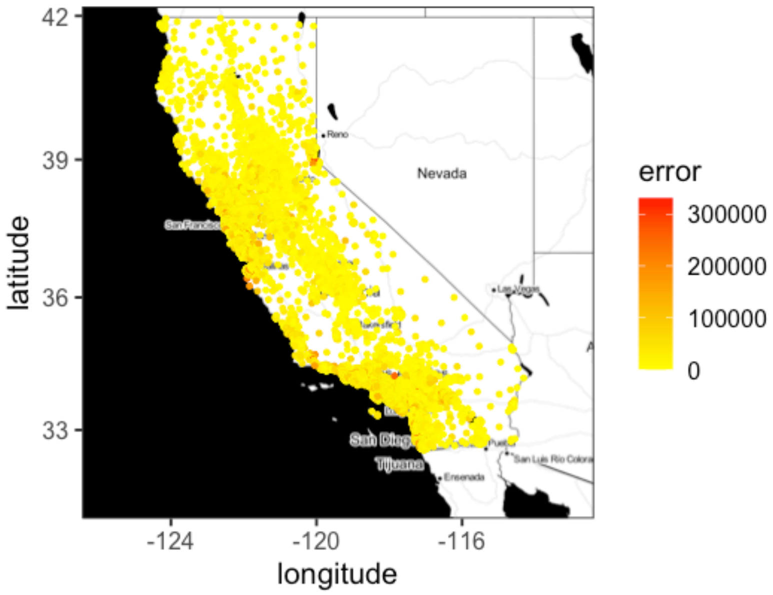
Figure 4.4 Plot 2



Note: The redder the color of the point is, the larger the prediction value is

CA Housing Prediction Error

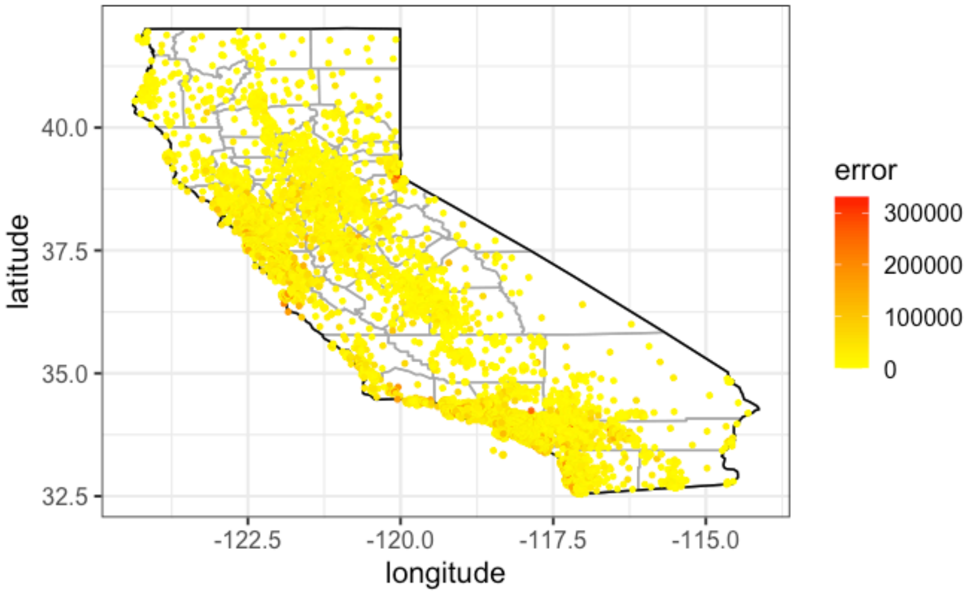
Figure 4.5 Plot 3



Note:The redder the color of the point is, the larger the prediction error is

CA Housing Prediction Error

Figure 4.6 Plot 3



Note:The redder the color of the point is, the larger the prediction error is