Exercise3

Zonghao Li

4/9/2021

This is my third exercise in Data Mining!

Problem1: What causes what?

Question 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? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)

Since every district has its own crime pattern, obviously high crime cities have an incentive to hire a lot of cops. So if just get data from a few different cities and run the regression of "Crime" on "Police", the data would be really messy. And this regression would not be able to achieve the goal of finding the casual effect of police on crimes.

Question2. How were the researchers from UPenn able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below, from the researchers' paper.

In their research, they used the district and day fixed effect so they controlled their analyses in Washington DC. They firstly regressed the daily total number of crimes in D.C. on the high alert level, and secondly regressed the daily total number of crimes in D.C. on both the high alert level and logged midday METRO ridership. In the first regression (results in 1st column), the coefficient -7.316 indicates that daily total number of crimes in D.C. would decrease by about 7.3 on the high alert days and it is statistically significant at 5% level. In the second column (results in 2nd column), after controlling the tourism, the coefficients indicate that the coefficient of the high alert level drops to about 6.05, and the number of daily crimes in D.C. would increase by about 1.7 if Metro ridership increases by 10%, and this estimated coefficient is statistically significant at 1% level. Since this increase is small, we can learn that the change in tourists is not strongly correlated with the change in number of daily crimes in D.C.

Question3. Why did they have to control for Metro ridership? What was that trying to capture?

Since after adding more cops, they made a hypothesis that the tourists were less likely to visit Washington during that particular time. And the number of tourists changed every day while the number of criminals was almost fixed. Then they checked that hypothesis by looking at ridership levels on the Metro system, and the number of tourists actually was not diminished on high terror days, so they suggested the number of victims was largely unchanged. They wanted to capture that the number of tourists was the same during the high horror days, and then the number of crimes was less likely to be related to the number of tourists.

Question4. 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?

The model here describes the effect of high alert in different districts on the number of crimes. The regression tells us that the expected number of daily crimes would decrease by about 2.62 in District 1 during the high alert days, and this estimated coefficient is statistically significant at 1% level. On the other hand, the expected number of daily crimes would decrease by about 0.57 in other districts but this estimated coefficient is not statistically significant at 5% level. And in this case, the expected number of daily crimes in D.C. would increase by about 2.48 if Metro ridership increases by 1 unit. To conclude, if cops decrease in other districts, it would not show that higher levels of crime in other districts during high alert days, and then they wanted to get the casual effect of police on crimes.

Problem2: Predictive model building: green certification

1. Overview

In recent years, real estate developers would consider constructing buildings with "green certifications" to realize environmental protection. Since the revenue from these buildings is also quite important to the decision made by these developers, our goal is to build the best predictive model possible for revenue per square foot per calendar year, and to use this model to quantify the average change in rental income per square foot associated with these green certifications, holding other features of the building constant.

2. Data and model

2.1 Data

2.1.1 Database Description

The data set we use contains data on 7,894 commercial rental properties from across the United States. LEED and Energystar are two specific kinds of green certifications. Among those properties, there are 685 have been awarded either LEED or EnergyStar certification as a green building.

The variables that can be refered are below:

```
(1) CS.PropertyID: the building's unique identifier in the database.

(2) cluster: an identifier for the building cluster, with each cluster
containing one green-certified building and at least one other non-
green-certified building within a quarter-mile radius of the cluster
center.

(3) size: the total square footage of available rental space in the
building.

(4) empl.gr: the year-on-year growth rate in employment in the
building's geographic region.

(5) Rent: the rent charged to tenants in the building, in dollars per
square foot per calendar year.

(6) leasing.rate: a measure of occupancy; the fraction of the
building's available space currently under lease.

(7) stories: the height of the building in stories.

(8) age: the age of the building in years.

(9) renovated: whether the building has undergone substantial
renovations during its lifetime.

(10) (11) class.a, class.b: indicators for two classes of building
quality (the third is Class C). These are relative classifications
within a specific market. Class A buildings are generally the highest-
quality properties in a given market. Class B buildings are a notch
down, but still of reasonable quality. Class C buildings are the least
```

desirable properties in a given market.

(12) green.rating: an indicator for whether the building is either LEED- or EnergyStar-certified.

(13) LEED, Energystar: indicators for the two specific kinds of green certifications.

(14) net: an indicator as to whether the rent is quoted on a "net contract" basis. Tenants with net-rental contracts pay their own utility costs, which are otherwise included in the quoted rental price.

(15) amenities: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.

(16) cd.total.07: number of cooling degree days in the building's region in 2007. A degree day is a measure of demand for energy; higher values mean greater demand. Cooling degree days are measured relative to a baseline outdoor temperature, below which a building needs no cooling.

(17) hd.total07: number of heating degree days in the building's region in 2007. Heating degree days are also measured relative to a baseline outdoor temperature, above which a building needs no heating.

(18) total.dd.07: the total number of degree days (either heating or cooling) in the building's region in 2007.

(19) Precipitation: annual precipitation in inches in the building's geographic region.

(20) Gas.Costs: a measure of how much natural gas costs in the building's geographic region.

(21)Electricity.Costs: a measure of how much electricity costs in the building's geographic region.

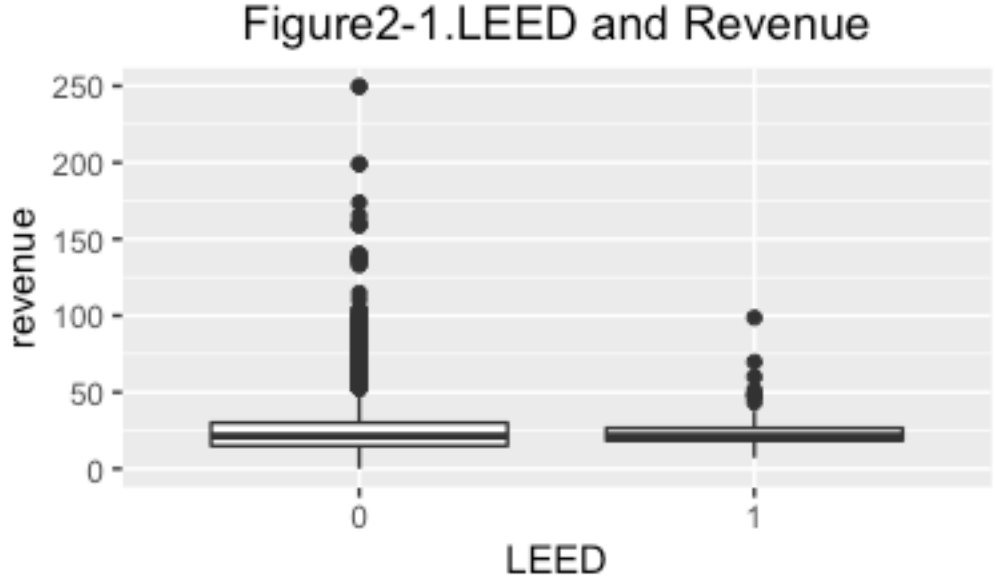
(22)City_Market_Rent: a measure of average rent per square-foot per calendar year in the building's local market.

This time, a “revenue” varibale should be created, which equals to Rent*leasing_rate, and it indicates revenue per square foot per year.

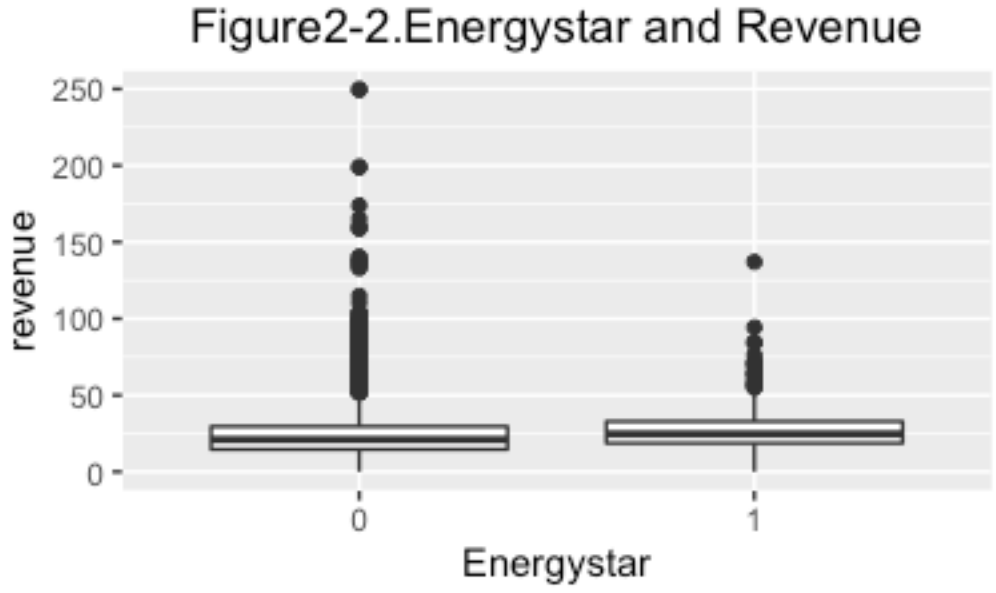
2.1.2 Several Data Descriptions of Actual Values

At first, we should modify LEED and Energystar by using ‘factor’, which tells R to treat a number as a category and make dummies.

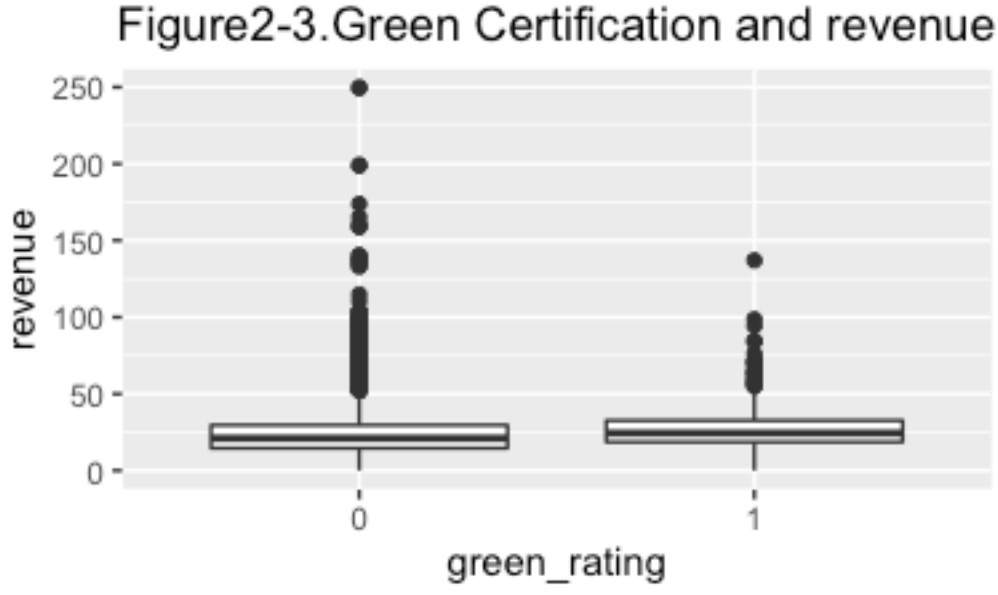
In our intuition, environmentally friendly building materials are more expensive than normal building materials, so these buildings with green materials cost more to build and therefore the rent charged may be higher. Here are some plots that shows the relationship between green certifications and revenue per square foot per calendar year.



The difference in revenue between owning and not owning LEED is quite small, and there is no difference in revenue between owning a LEED and not owning a LEED property, so this not corresponds to our intuition.



The difference in revenue between owning and not owning Energystar is small, but the properties with Energystar have a slightly higher revenue than those without Energystar. And this corresponds to our intuition too.

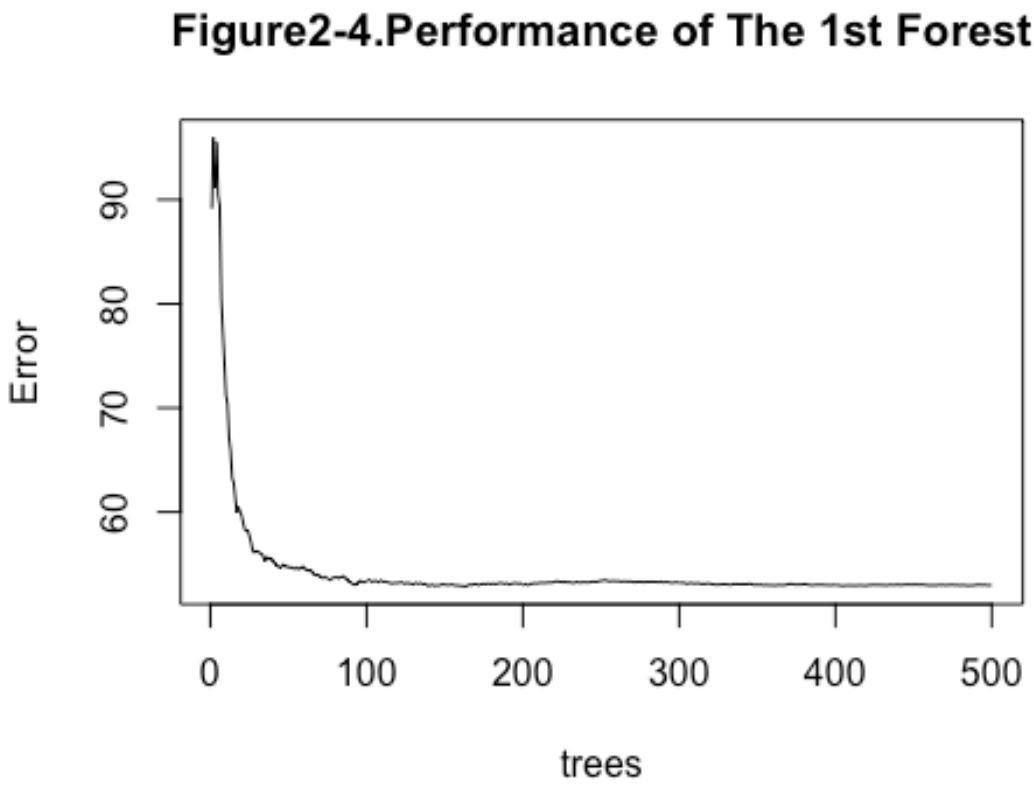


By the way, whether or not to own LEED or Energystar occupies a small revenue range, but owning LEED or Energystar has a little bit high revenues than not owning green certifications. And this also corresponds with intuition.

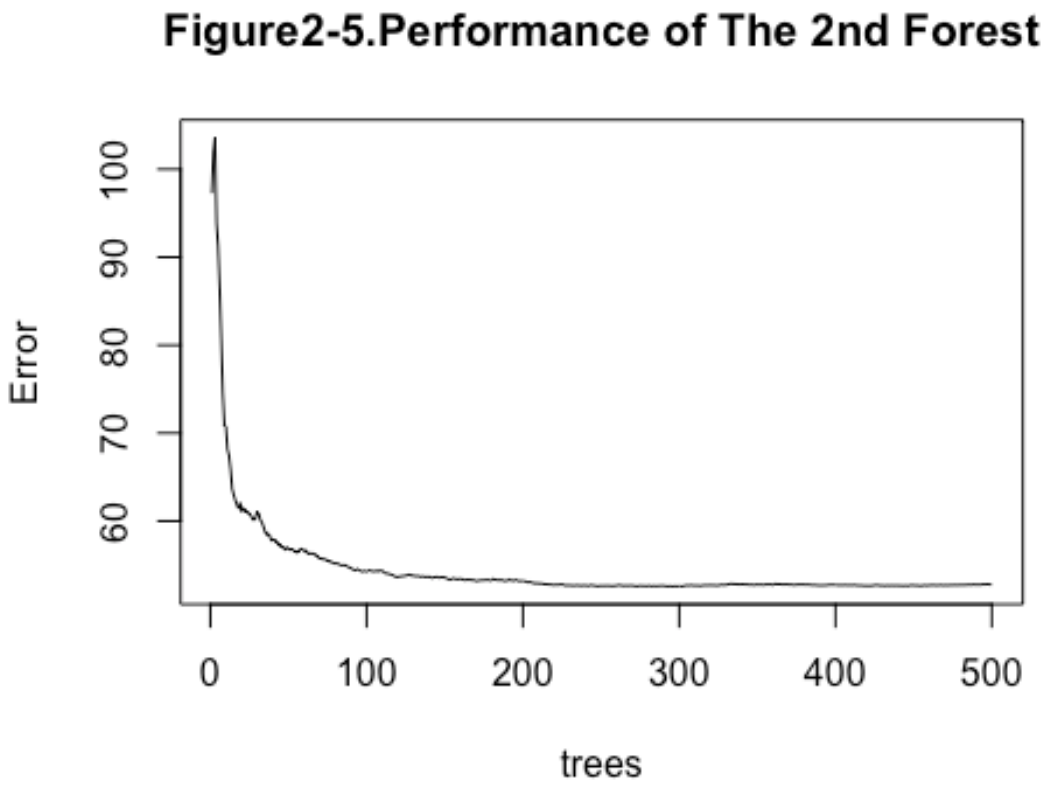
2.2 Model Selection

Our goal for this problem is to build the best predictive model possible for revenue per square foot per calendar year, and to use this model to quantify the average change in rental income per square foot (whether in absolute or percentage terms) associated with green certification, holding other features of the building constant.

By measuring out-of-sample performance, there is random variation due to the particular choice of data points that end up in our train/test splits. To make sure that our script addresses this, such as by averaging the estimate of out-of-sample RMSE over many different random train/test splits. The Root Mean Squared Error is used to measure the deviation of the observed value from the true value. We will compare the RMSEs of several models and choose a comparatively proper as our best predictive model. And in the end, the random forest should be selected as our predictive model.



It can be seen that the error curve turns to be lastly flat around 250-300, so the number of trees will be set by 300.

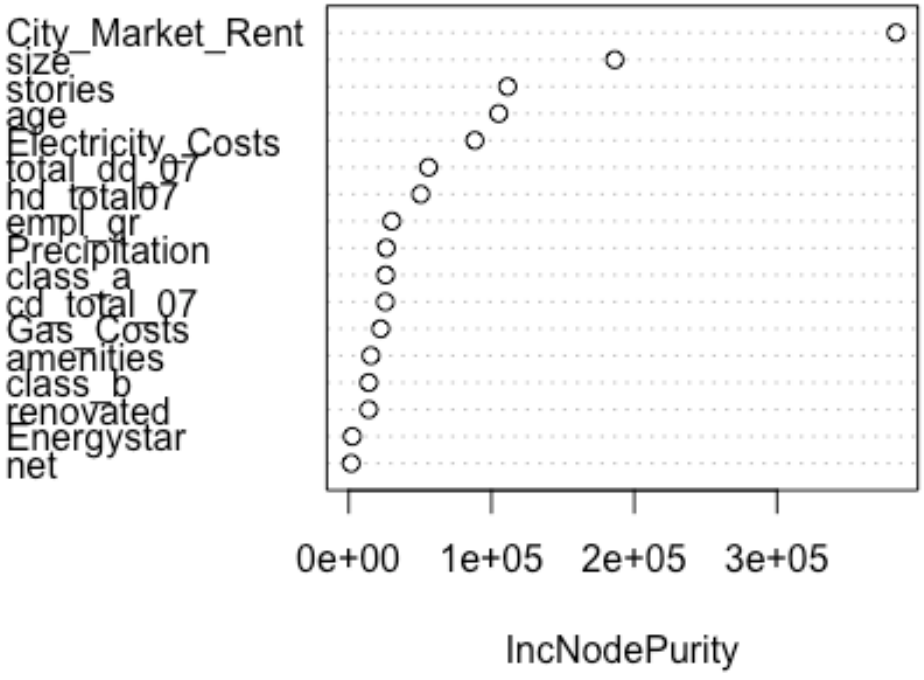


It can be seen that the error curve turns to be lastly flat around 300-400, so the number of trees will be set by 400.

After comparing these RMSEs, the value of errors of random forest model is about 6.5 and is a bit lower than the errors of boosted regression trees, so the random forest model would be better. Then we can fit a random forest model to predict revenue that include several variables as predictors, such as City_Market_Rent, size, stories, and Energystar or LEED. Energystar and LEED, as two green certifications, will be analyzed seperately.

3. Results

Figure3-1.The 1st Forest



So， in the first random forest, City_Market_Rent is the most important variable, and size and stories are two relatively important variables in our considering variables. However, 'Energystar' variable ranks very low, which indicates that it is not very important.

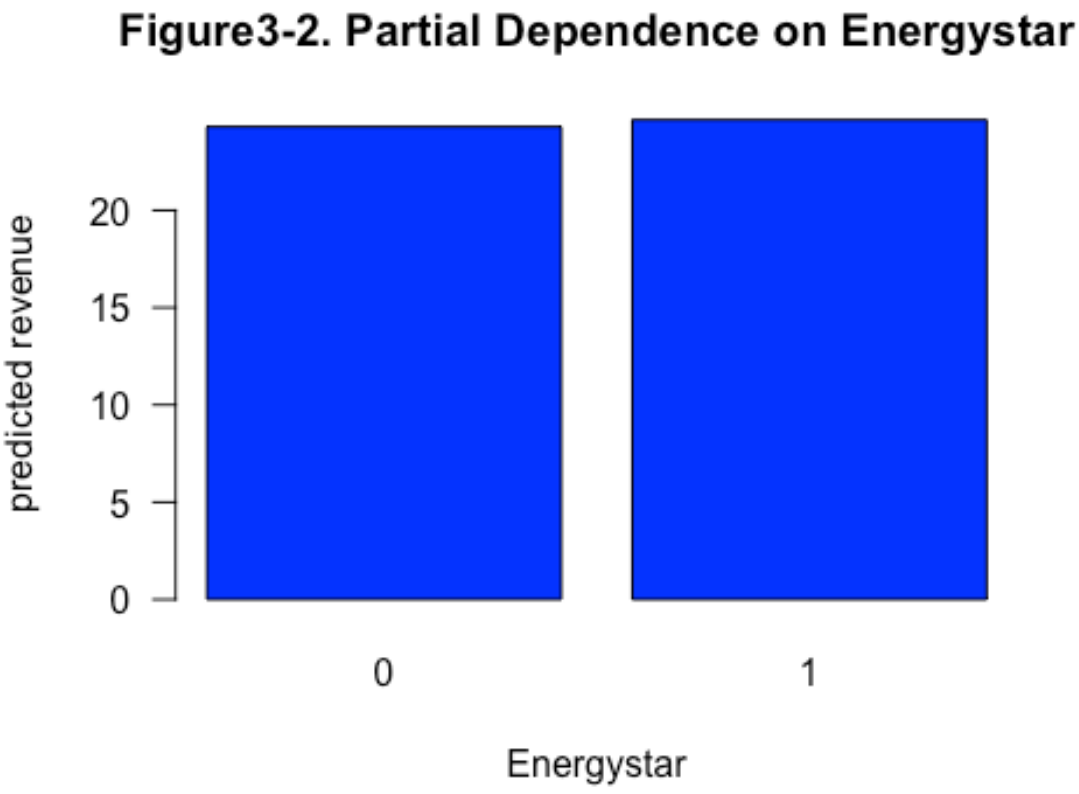
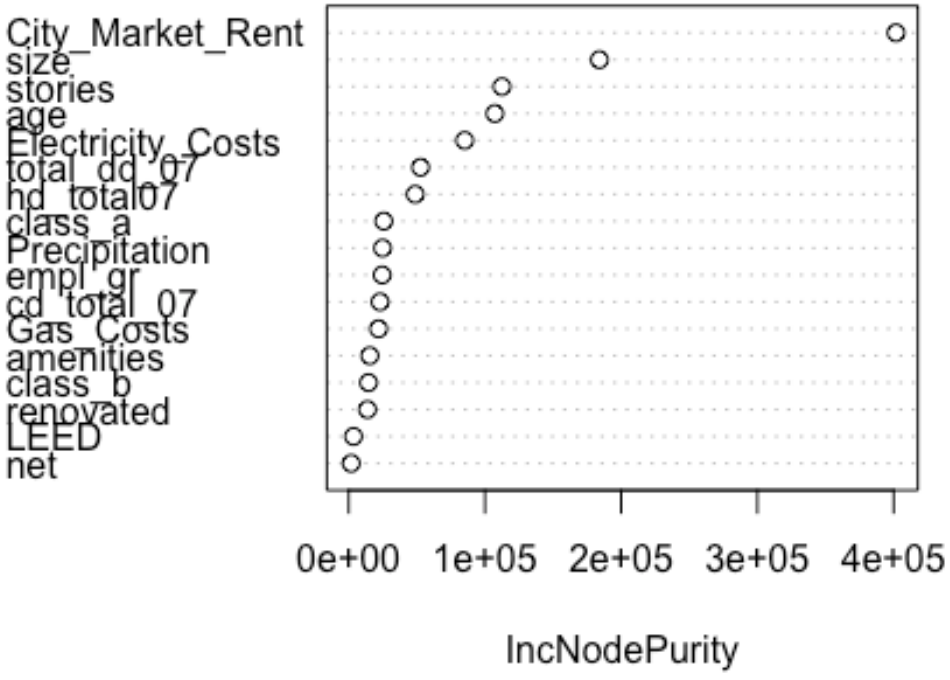


Figure.3-2 shows that having Energystar feature will have positive effect on the revenue based on prediction.

Figure3-3. The 2nd Forest



So, in the second random forest, City_Market_Rent is the most important variable, size is the second relatively important variable in our considering variables. Like 'Energystar' in the first random forest, the 'LEED' variable is not very important among these variables.

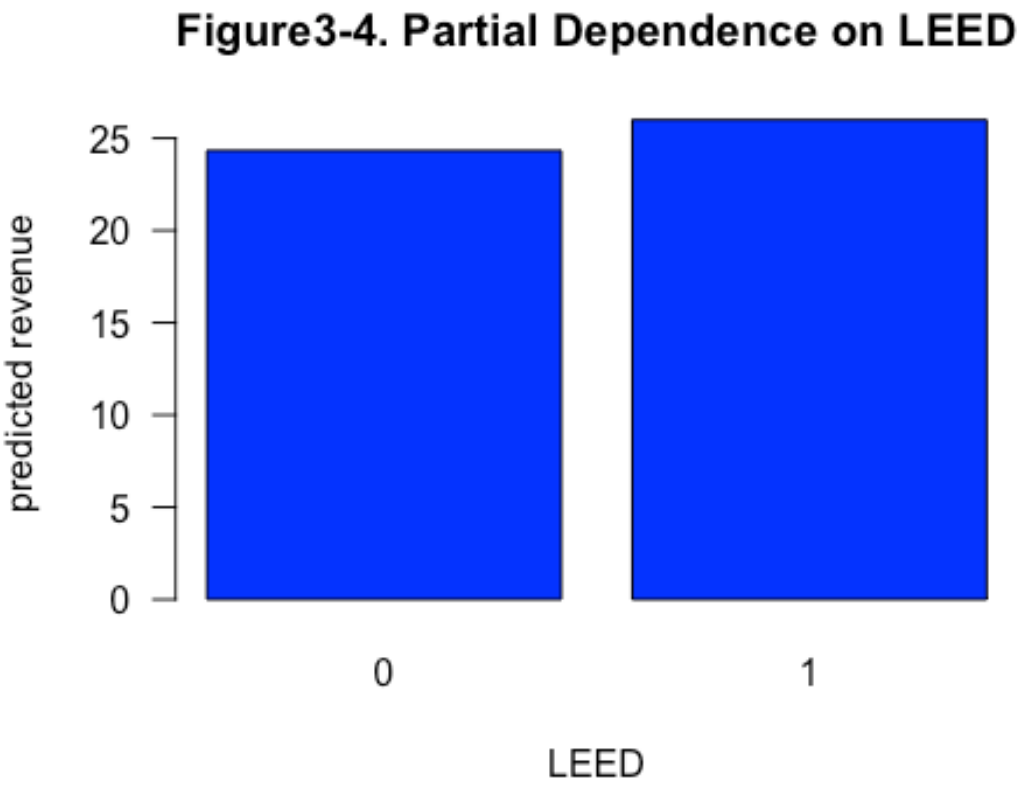


Figure.3-4 shows that owning LEED feature will have positive effect on the revenue based on prediction, and this effect seems stronger than the effect of Energystar feature.

4. Conclusion

To conclude, on the one hand, average rent per square-foot per calendar year in building's local market, building's total available rental space and building's height are the relatively very important variables. On the other hand, when considering the effect of green certifications (Energystar and LEED), both of them have positive effect on the average change in rental income per square foot holding other building's features constant, although these two are not very important among other features.

Problem3: Predictive Model Building: California Housing

FigureC1. Actual Median House Value in California

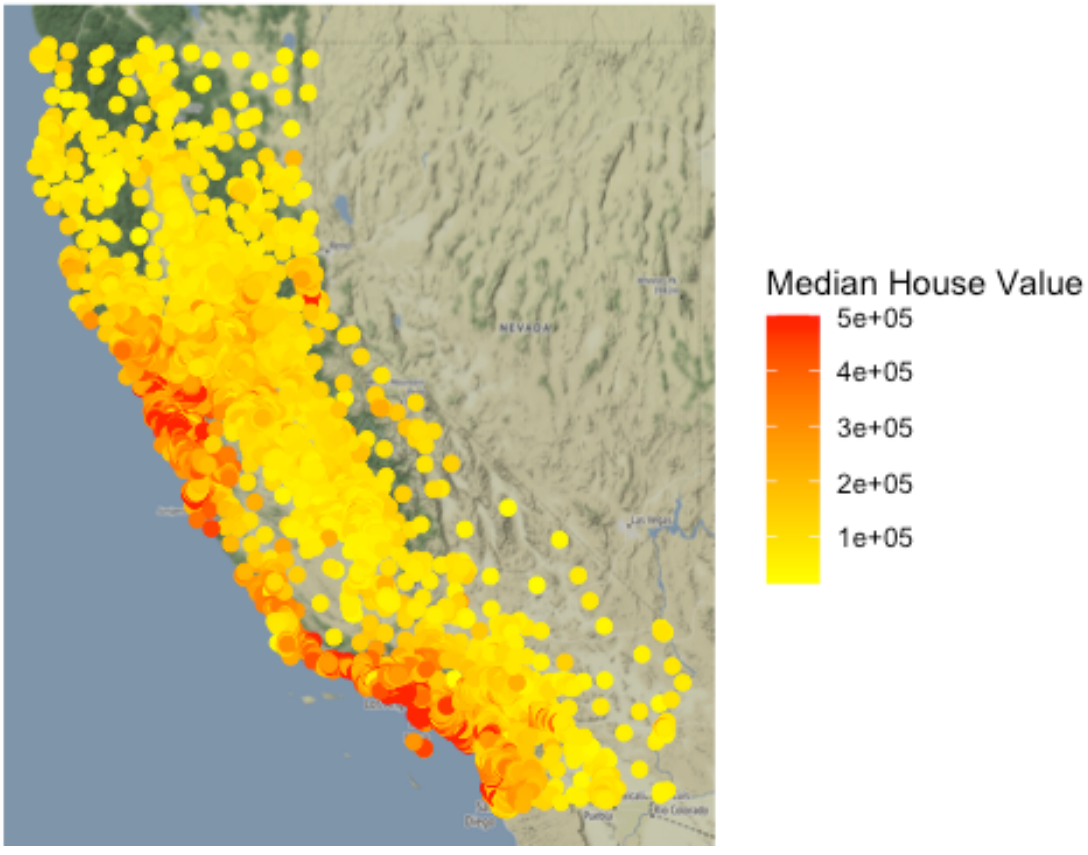
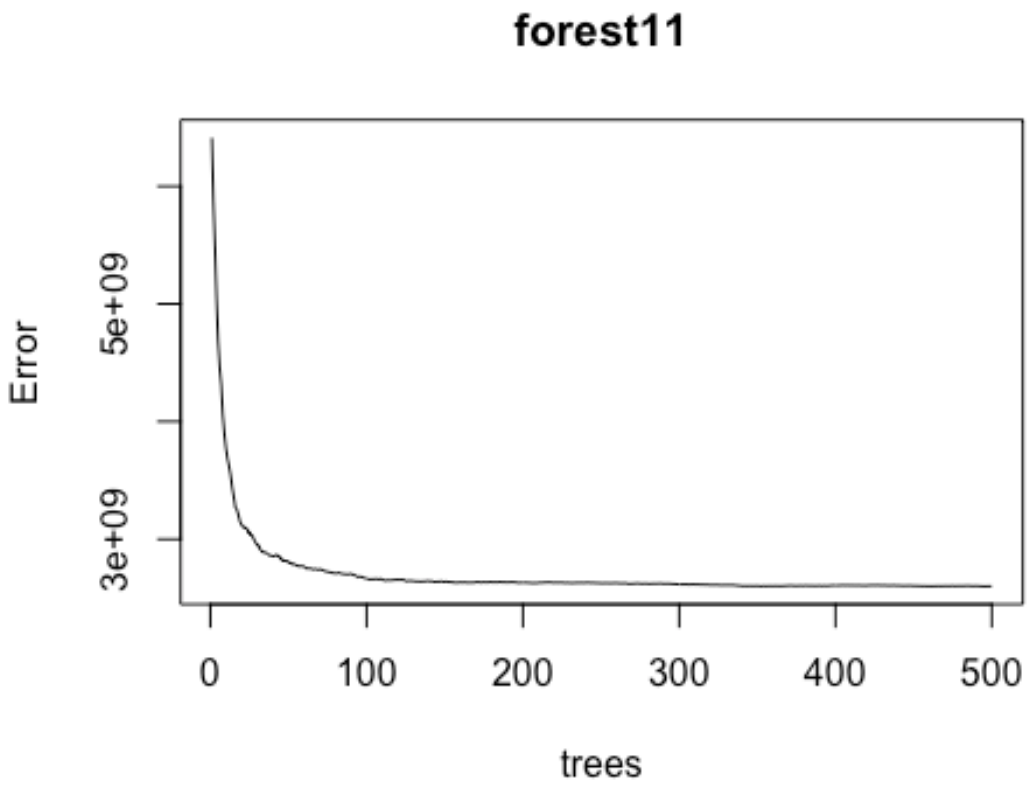


Figure.C1 shows that the houses with high actual median house value are distributed by the middle and south of the western coast of California.



It shows that the error curve turns to be lastly flat around 200-300, so the number of trees will be set by 300.

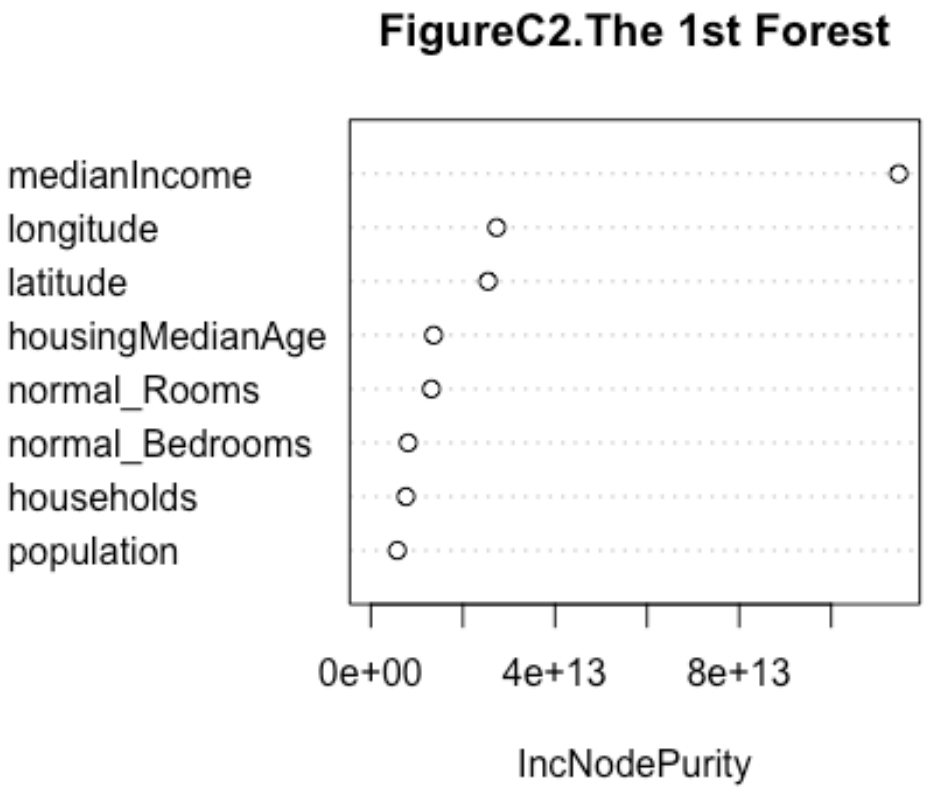
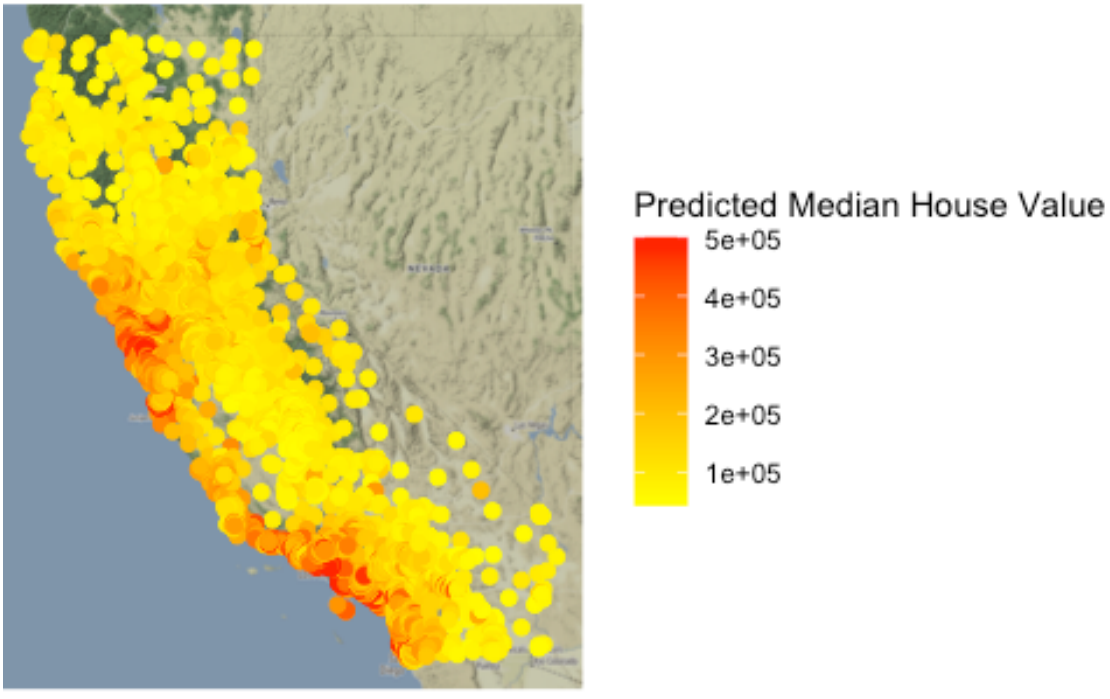


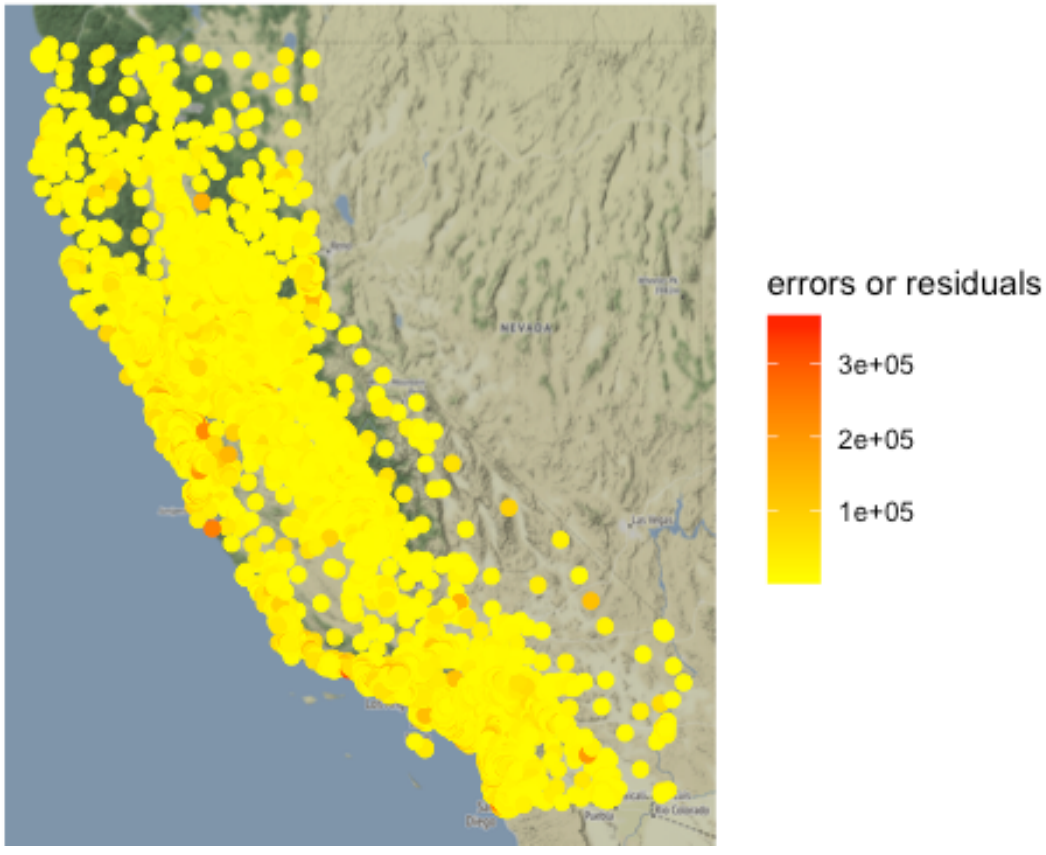
Figure.C2 tells us that the most important variable is medianIncome. Apart from longitude and latitude, housingMedianAge and normal_Tooms are the two relatively important variables.

FigureC3. Prediction of Median House Value in California



From Figure and Figure above, we can see that the distribution of predicted values looks similar to that of actual values. To be particular, from the colors of these predicted values, some red points turn to be orange, which means that predicted median house values are little lower than the actual values.

FigureC4. Model's erors



From the results shown in Figure, the absolute values of errors are mostly low, so the predictive model can be suitable and convictive to some extent.

Conclusion From the analyses above, the predictive model is effective, and the high median house values are distributed by the middle and south of western coast of California.