Exercise 3

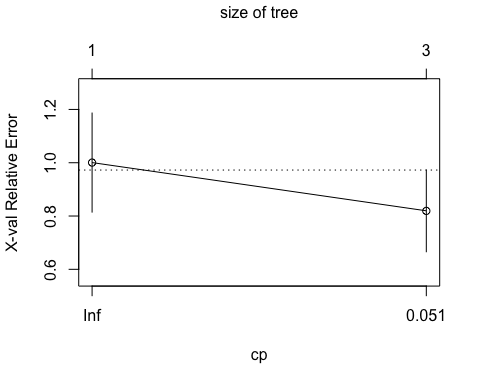
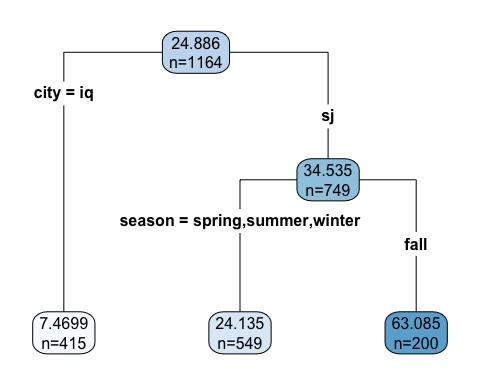
Takehiro Hasimoto (UT EID - TH33985), Avijit Mallik (UT EID - AM99484), Arindam Chatterjee (UT EID - AC83995)

2023-03-21

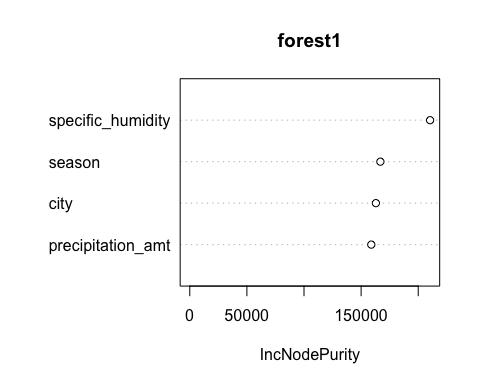
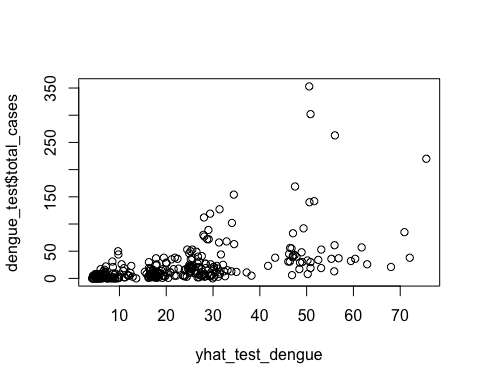
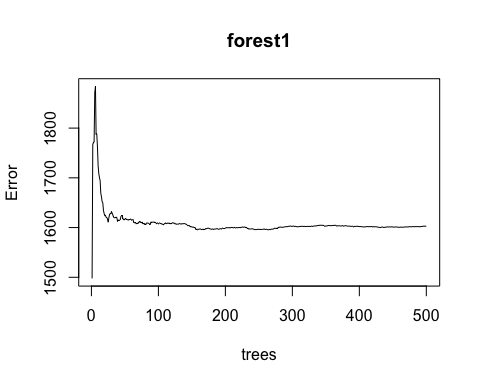
## Problem 2

## Tree modeling: dengue cases

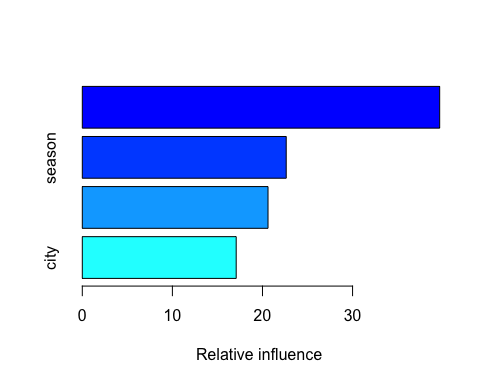
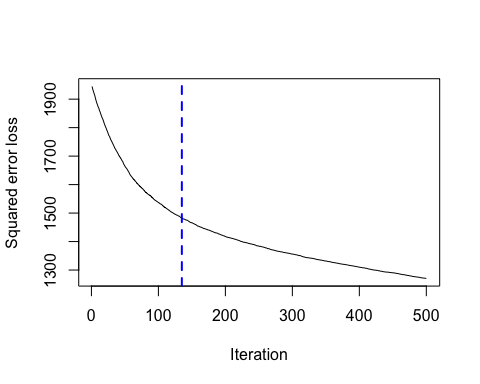
First, we use CART model.



**Now we use random forest model.**



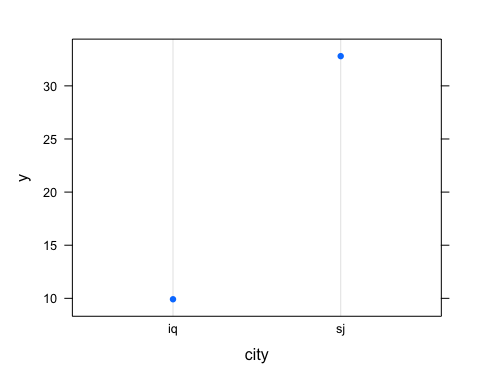
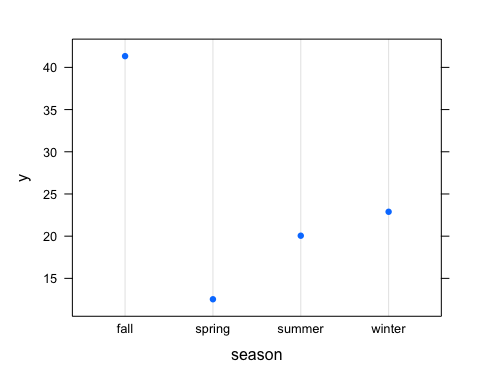
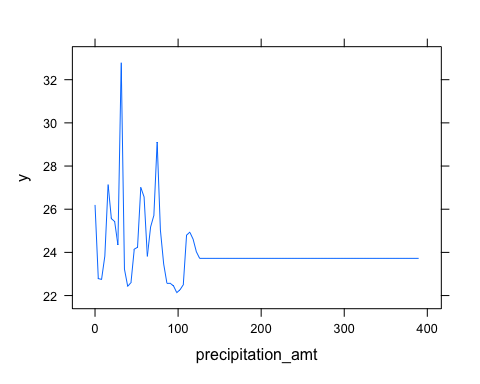
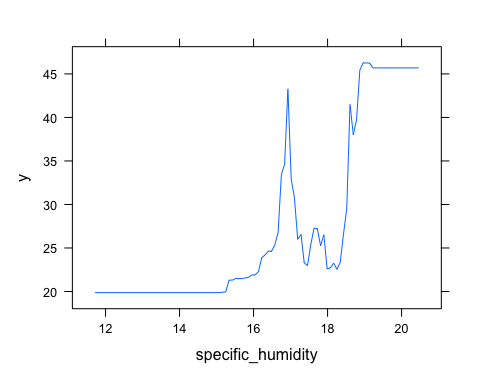
**Finally we model by using gradient Boosting model with Gaussian and Poisson distributions.**



## CART\_RMSE RForest\_RMSE Normal\_Boost\_RMSE Poisson\_Boost\_RMSE  
## 1 37.30277 35.88153 35.2054 35.53762

Based on the out of sample RMSE, the Gaussian Booster model seems to have the best prediction power.

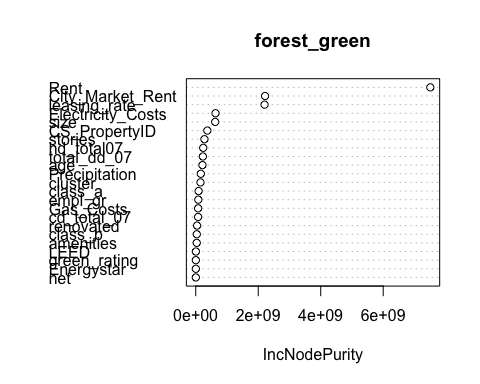
**Now we plot the partial dependence of 4 variables.**



The graphs above show the partial dependence (marginal effects) of the chosen variables on total cases of dengue based on the Gaussian boosting model. I have included all 4 variables since all of them seems interesting, especially with the high difference between the two cities, and the Fall season with the other seasons.

# Problem 3

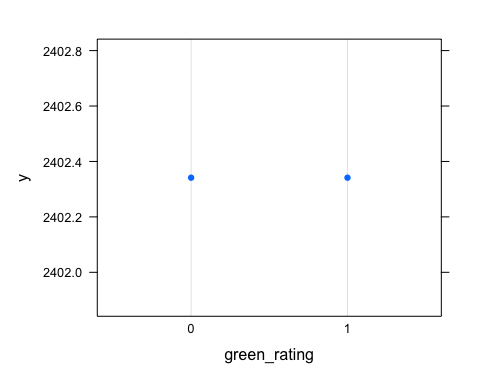
## So we used three random forest models, and one gradient boosting model to measure the efficiency of the predictions.



## Distribution not specified, assuming gaussian ...

## RFM1\_rmse RFM2\_rmse RFM3\_rmse Boost\_rmse  
## 1 169.3709 157.0977 188.3672 136.0743

**Now we check for the partial dependence of green rating based on the boosting model (the optimal model).**



## green\_rating yhat  
## 1 0 2402.341  
## 2 1 2402.341

The goal of this exercise is predict the revenue per square foot per calender year of about 8,000 commercial rental properties across the US. In addition, some of those properties are green certified which means they got green certification from either LEED or Energystar. Another question we want to answer is whether being green certified will raise your revenue or not. Now let’s move on the methodloly used to predict the revenue.

First of all, we have mutated a new column to calculate the revenue per square foot per calender year based on the original data. In order to do that, we took the product of rent and leasing\_rate. We need to do that to get unbiased prediction results since the occupancy or the rent\_rate alone won’t reflect the revenue.

Next, we needed to make sure that some of the variables are dummy variables, so we used the factor command on the 0/1 variables. Then, we start working on the model by splitting the data to training set (80%) and testing set (20%). We trained the data to predict revenue using random forest model. First model used is the base model, basically by regressing revenue on all variables, then check for the importance of each variable in order to try other models and compare them based on the results of their root mean squared errors.

Now we move to try other possible models based on the results of their importance. We can notice how green\_rating is not an important variable in the model which indicates there will not be significant partial effect of the green certification on the revenue. However, we have to include it in order to observe the real partial effect using the partial dependence algorithm.

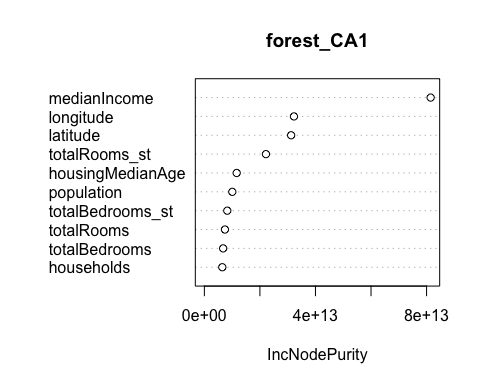
Now, after my base model, the 2nd model included 9 variables with different importance level for each one of them. The 3rd model had 12 variables with many more less important variables. We worried that it is going to overfit the model, so now we got to check the rmse for each one of them and compare it with what we got in the base model. So, the 2nd model with the 9 variables got slightly lower rmse than the first model which regressing revenue on all variables. However, should we stop now? since we are looking for the best predictive model, it is going to be worth it to try to model using gradient boosting model with the same variables of the best performing random forest model.

After trying different shrinkage rates, we have succeeded in over-performing the 2nd model by having rmse = 134 compared to the best random forest model which was 167. So, we decided to select the boosting model to answer the question of the how much green certification is going to affect my revenue assuming all other variables are constant. So, we predicted the average value for both certified and certified, and as we can see, it has no partial effect at all. The values basically are the same and the plot gives us the same answer too.

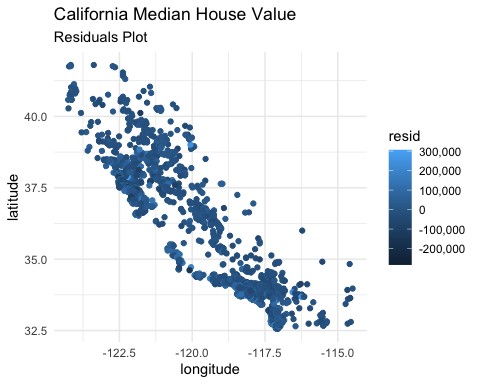
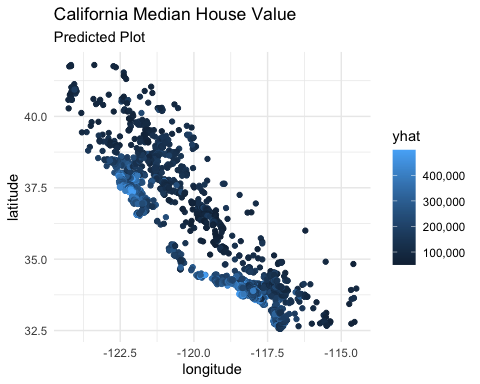
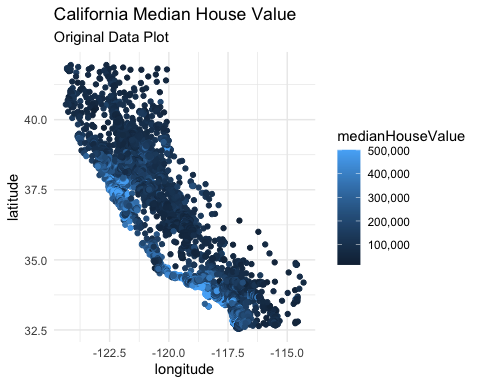
## Problem 4

## Predictive model building: California housing

We compare 4 different models to check which model is the optimal.

 **Now we check which model out of the 4 has the lowest root mean squared errors.**

## CA\_RFM1\_rmse CA\_RFM2\_rmse CA\_RFM3\_rmse CA\_Boost\_rmse  
## 1 48630.96 48572.76 47998.16 51663.24



For this model, the goal was to predict the median house value in California State. In order to do that, we have used machine learning tools to provide with reliable predictions. So, we have used the random forest model, which utilizes the interaction effects of the variables. First, I mutated to new columns to standardized the total rooms and total bedrooms by dividing each variable by households variable. Then, I split the data into 80% training set and 20% testing set and regress medianHousevalue on all the variables to test for the importance of each variables afterward. Next, we did two other specification models with different variables based on the results of the variables importance. The third model has the lowest root mean squared error which equals to 47,989. In order to check for room of improvements, we ran a gradient boosting model with many different shrinkage rates, but we could not have a lower rmse value than that found using the selected random forest model.

So, we decided to continue with the results of the optimal random forest model and predict the median housing values based on the testing set. Then we plottod the original observation which has the shape of California State, the predicted values based on the testing set, and the estimated residuals which is the difference between the two.