Problem Set 3

Jack Cunningham, Ken Noddings, Ali Fazl

2023-03-22

Question 1: What causes what?

A) 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 most obvious reason that we can't simnply regress crime on police to understand the causal effect of police on crime is that causation could very well go the other way. It's not a stretch to hypothesize that neighborhoods with higher crime rates see more police activity *because* of the higher crime rates - not the other way around. But if we ran the crime on police regression, we might find that there is a significant positive effect of police on crime, ie. having more police presence increases rather than decreases crime. This would be akin to regressing health outcomes on hospital visits, then saying that hospitals cause worse health outcomes - people go to the hospital because they are unhealthy! So there is obvious selection bias in this example.

There could also be confounder bias; something that causes both high crime and high police rates, whose causal effect we would be able to tease out with the simple regression above. For instance, consider a college neighborhood like West Campus in Austin. We might see higher crime rates there because younger people are more likely to commit crimes than older people and younger people are a higher percentage of the neighborhood's residents; we might also see more police there because the university has its own police system. So being near the college could have causal effects on both crime and police presence that make it much harder to estimate the effect of police presence on crime.

B) 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.

The researchers from UPenn took advantage of a Terrorist Alert warning system. Since Washington, D.C. is at a high risk of terrorist attacks relative to other cities, the District of Columbia puts more uniformed officers on the streets on days with a higher threat level (Orange) of terrorist attack. Therefore, there was almost a natural experiment such that the increased level of police presence was uncorrelated with local crime (terrorist attack risk being a different, uncorrelated conceptual category). They then seek to determine whether street crime goes down on Orange alert days, with the most obvious causal interpretation for such a result being the increased police presence.

The results table contains two columns; column 1) gives the effect of an Orange alert day on crime rates. The researchers find a reduction of, on average, 7.316 total crimes in DC on Orange alert days (HAEC), a result that is significant at the 5% level. In column 2), the researchers control for log midday Metro ridership to test whether the reduction in crime was a result of fewer people out in public. They found a significant negative relationship again, -6.046, meaning that the reduction in crime was still explained well by increased police presence instead of reduced ridership. (The coefficient on log midday Metro ridership gives the average increase in crime, 17.341 per day, for a 100% increase in Metro ridership.)

C) Why did they have to control for Metro ridership? What was that trying to capture?

The researchers control for Metro ridership because they wanted to test whether the reduction in crime noted on Orange alert days was explainable by a reduction in people out in public on those days. They find that there is in fact no reduction in Metro ridership on Orange alert days; thus, the hypothesis that fewer people in public was driving the reduction in crime rates could be ruled out, making the increased police presence explanation more likely.

D) 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 estimated in this table further disaggregates by location in DC - the first row represents average crime reduction in District 1 on Orange Alert days, the second row represents average crime reduction in all other districts on Orange Alert days, and the third and fourth rows are the control for Metro ridership and the intercept (all with robust standard errors). Focusing on the first two rows, we can see that the reduction in crime in District 1 is much greater in magnitude (-2.621) than the reduction in all other districts (-.571); moreover, the reduction in District 1 is significant at the 1% level, while the reduction in all other districts is not statistically significant.

District 1 is the district of DC containing the National Mall, the White House, and most of the likeliest terrorist targets; hence, on Orange alert days, the police presence increase in District 1 is much higher than the police presence increase in other DC districts. The conclusion from this result is that the most likely explanation for the drop in crime in DC on Orange alert days is due to increased police presence - and the larger the increase, the bigger the effect on crime.

Question 2: Tree modeling - dengue cases

Your task is to use CART, random forests, and gradient-boosted trees to predict dengue cases (or log dengue cases – your choice, just explain) based on the features available in the data set. As we usually do, hold out some of the data as a testing set to quantify the performance of these models. (That is, any cross validation should be done only on the training data, with the testing data held as a final check to compare your best CART model vs. your best random forest model vs. your best boosted tree model.)

Table 1: RMSE of each model, first pass

model	RMSE
lm1 (baseline)	33.46797
lm2	33.09153
lm3	33.04024
dengue_tree1	33.81737
dengue_tree2	33.74887
dengue_forest1	31.68799
dengue_boost1	33.54293
dengue_boost2	37.69619
dengue_boost3	37.74484

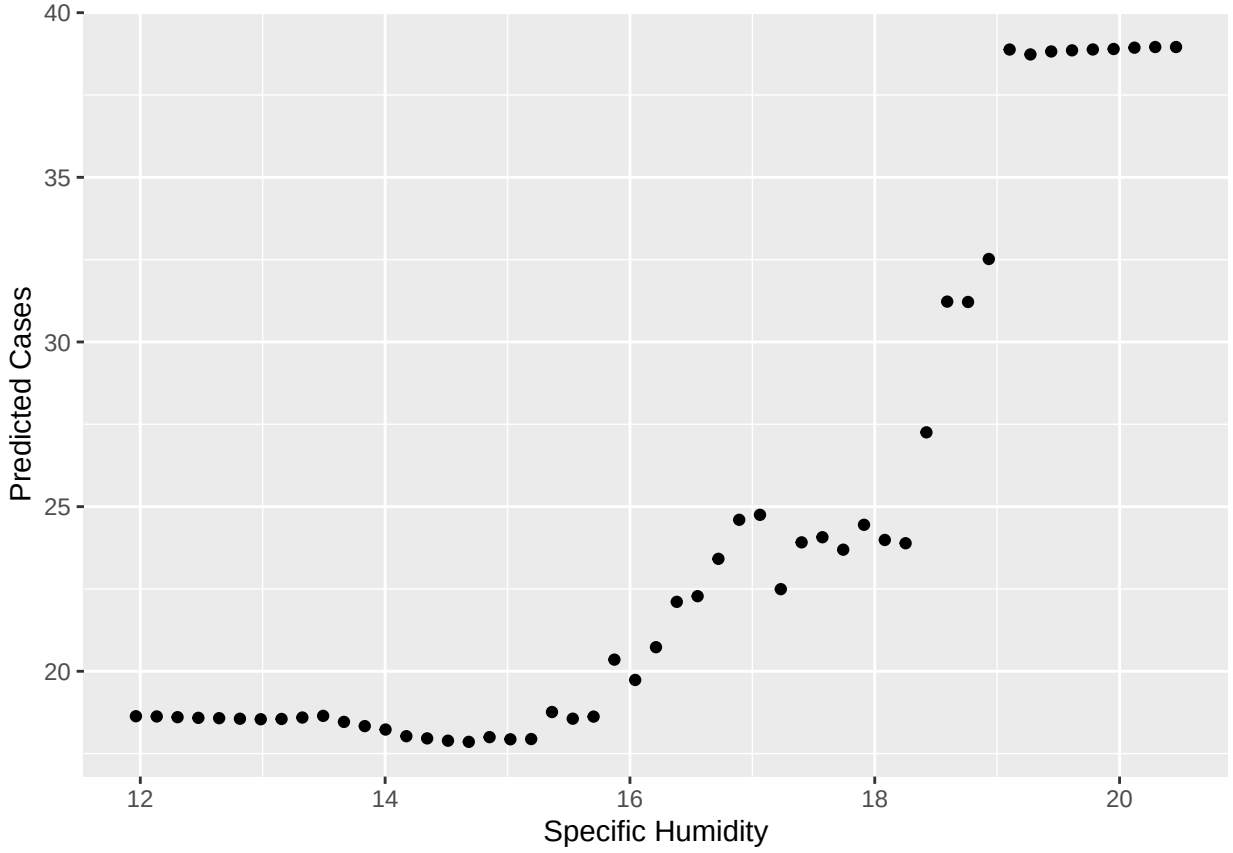
This table represent the out-of-sample root mean squared error for each model considered; three linear models for calibration purposes, two CART models, one random forest, and three gradient-boosted models.

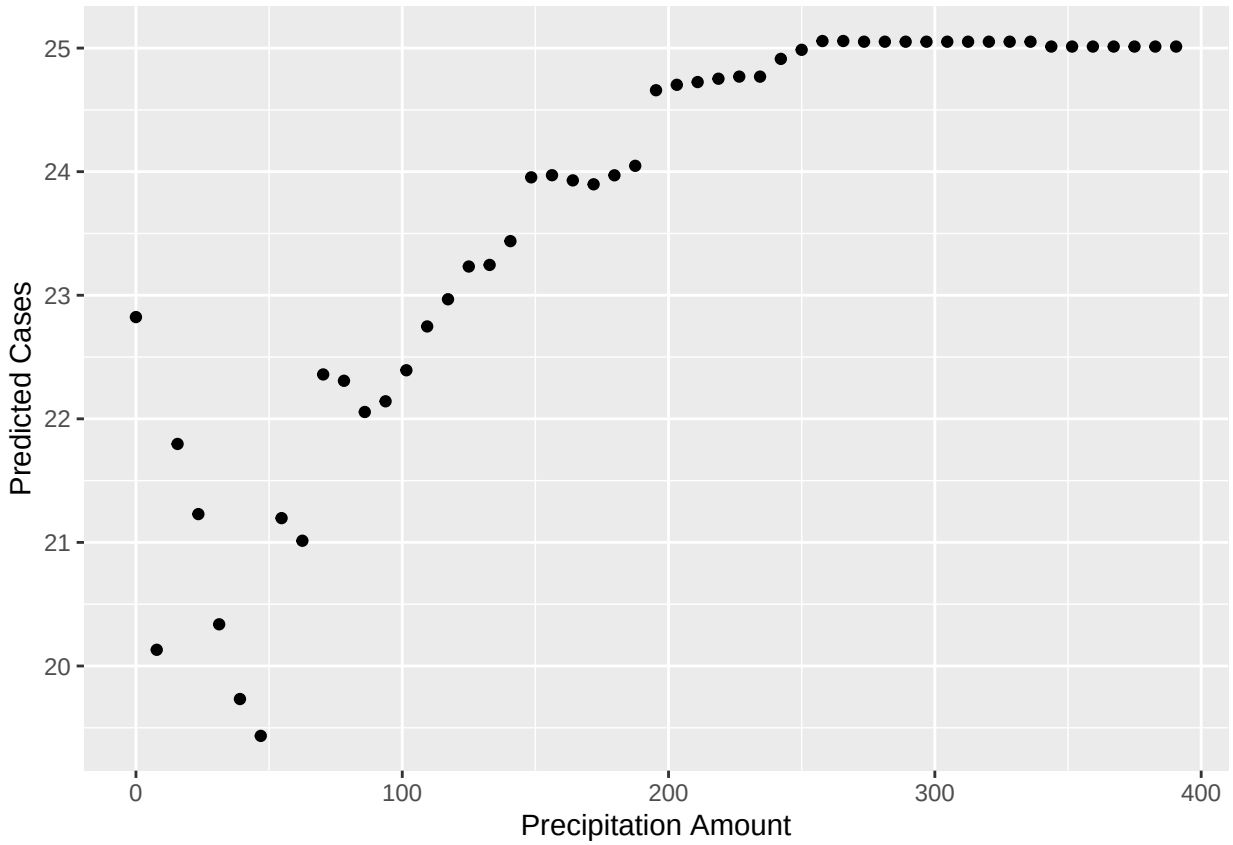
Table 2: RMSE of three best models, cross-validation

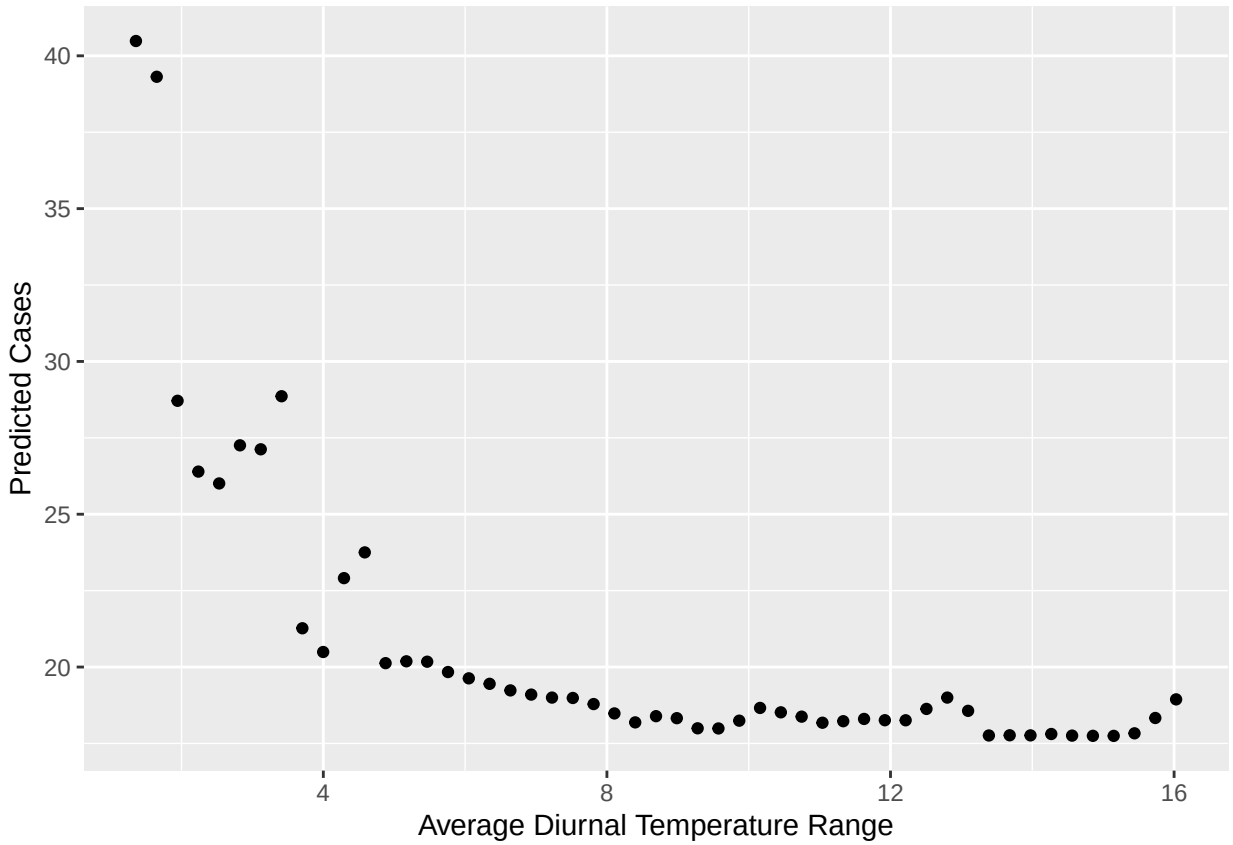
model	RMSE
dengue_tree1	30.04853
dengue_forest1	28.41947
dengue_boost1	29.41456

Here we've taken the best-performing CART, forest, and gradient-boosted models from above and cross-validated them against a testing set unused in the model development phase. It's clear that the random forest model has the best out-of-sample performance of all the models.

Then, for whichever model has the better performance on the testing data, make three partial dependence plots: specific_humidity, precipitation_amt, and wild card/writer's choice: you choose a feature that looks interesting and make a partial dependence plot for that.




Predicted Cases
25
24
23
22
21
20
0
100
200
300
400
Precipitation Amount



These partial dependence plots depict a clear relationship between dengue cases and the three meteorological variables considered. As specific humidity rises (particularly beyond 18 grams of water per kilogram of air), predicted dengue cases rise. The same is true for weekly precipitation in millimeters, especially beyond 200 mL/week. And for average diurnal temperature range, predicted dengue cases fall steadily between 1 and 6 degrees, plateauing afterwards. The best-performing random forest model predicts that high rainfall, high humidity, and low temperature range are associated with higher levels of dengue cases.

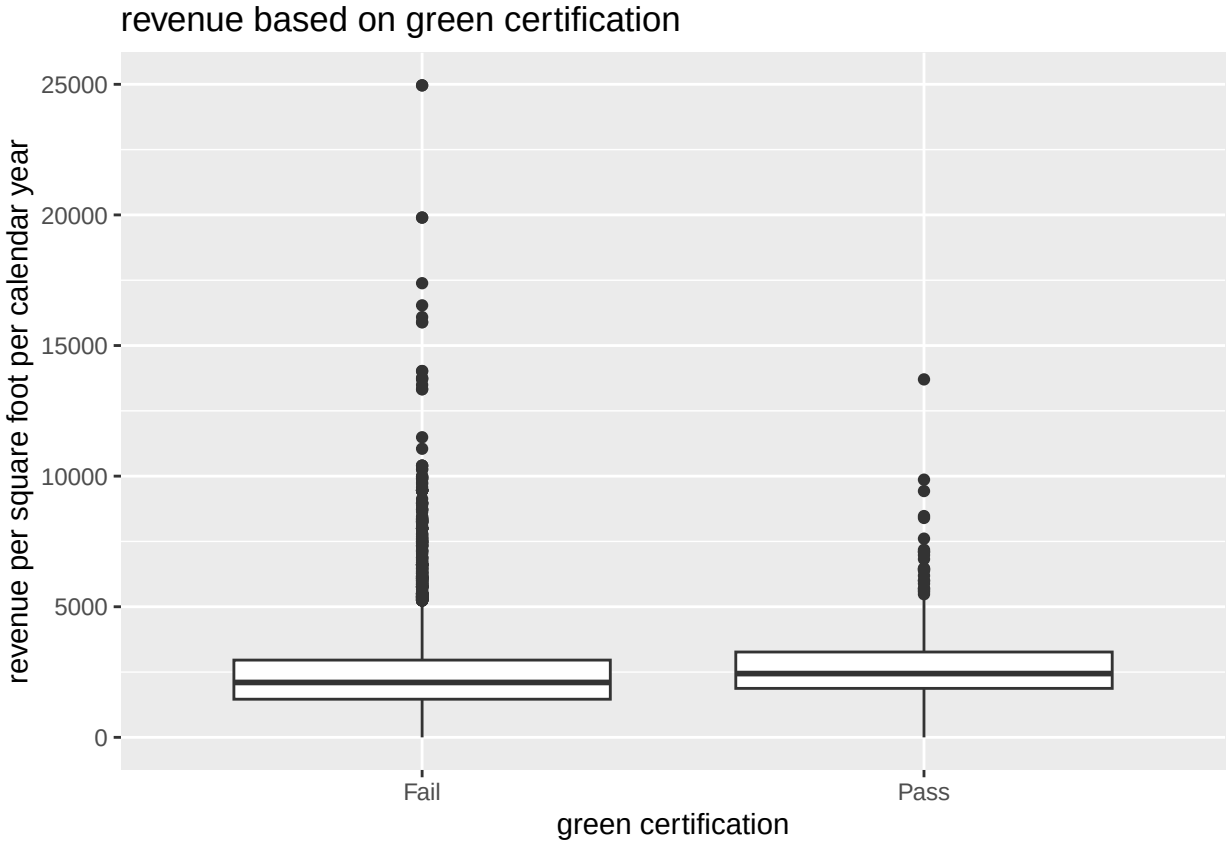
This prediction has implications for policymakers. In particular, when these indicators align, the population should be warned that dengue cases are likely to be elevated and admonished to take the proper precautions. Hospitals should also be aware of the meteorological indicators in order to predict periods of elevated dengue cases at their facilities.

Question 3: Predictive model building - green certification

Your 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 (whether in absolute or percentage terms) associated with green certification, holding other features of the building constant. (This might entail, for example, a partial dependence plot, depending on what model you work with here.) Note that revenue per square foot per year is the product of two terms: rent and leasing_rate! This reflects the fact that, for example, high-rent buildings with low occupancy may not actually bring in as much revenue as lower-rent buildings with higher occupancy.

You can choose whether to consider LEED and EnergyStar separately or to collapse them into a single "green certified" category. You can use any modeling approaches in your toolkit (regression, variable selection, trees, etc), and you should also feel free to do any feature engineering you think helps improve the model. Just make sure to explain what you've done.

Write a short report, no more than the equivalent of about 4 pages, detailing your methods, modeling choice, and conclusions.

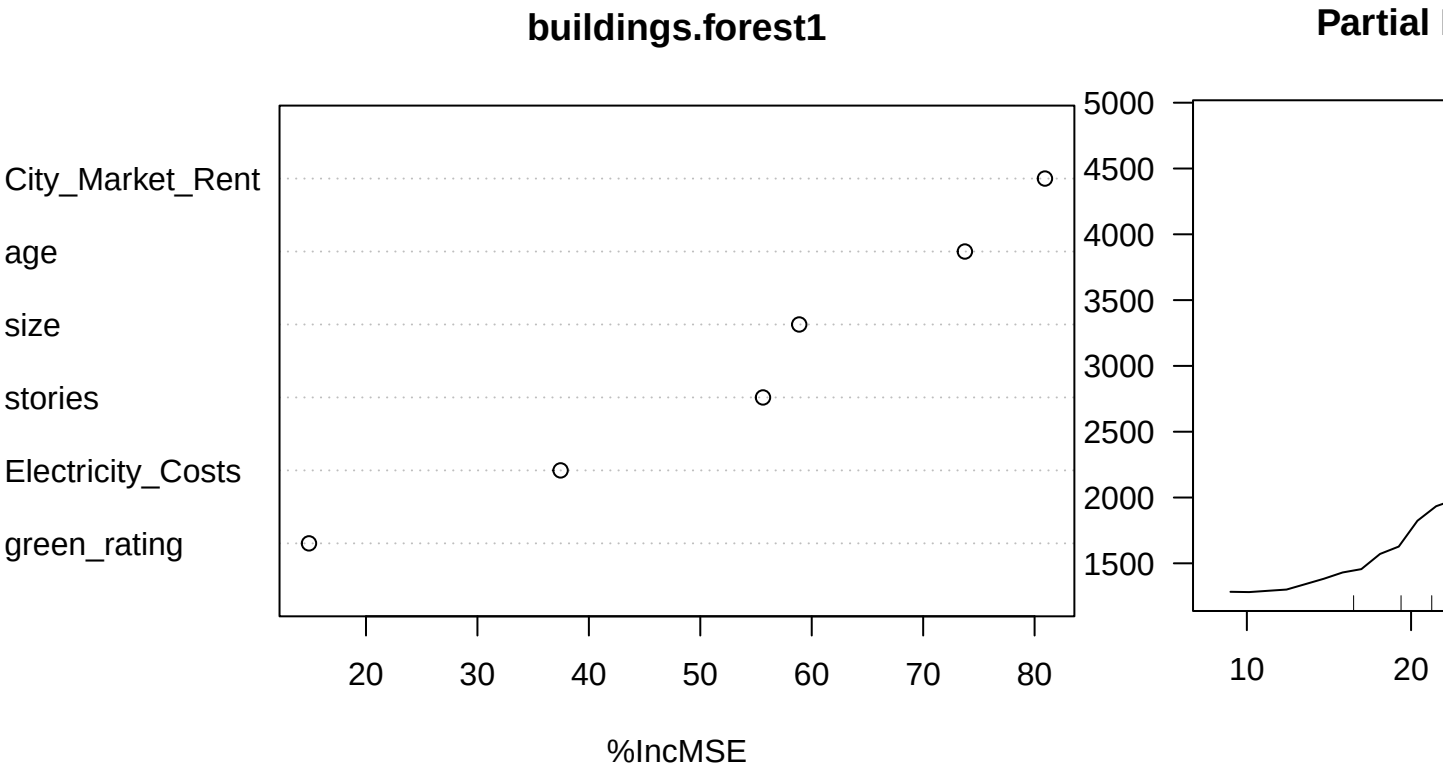


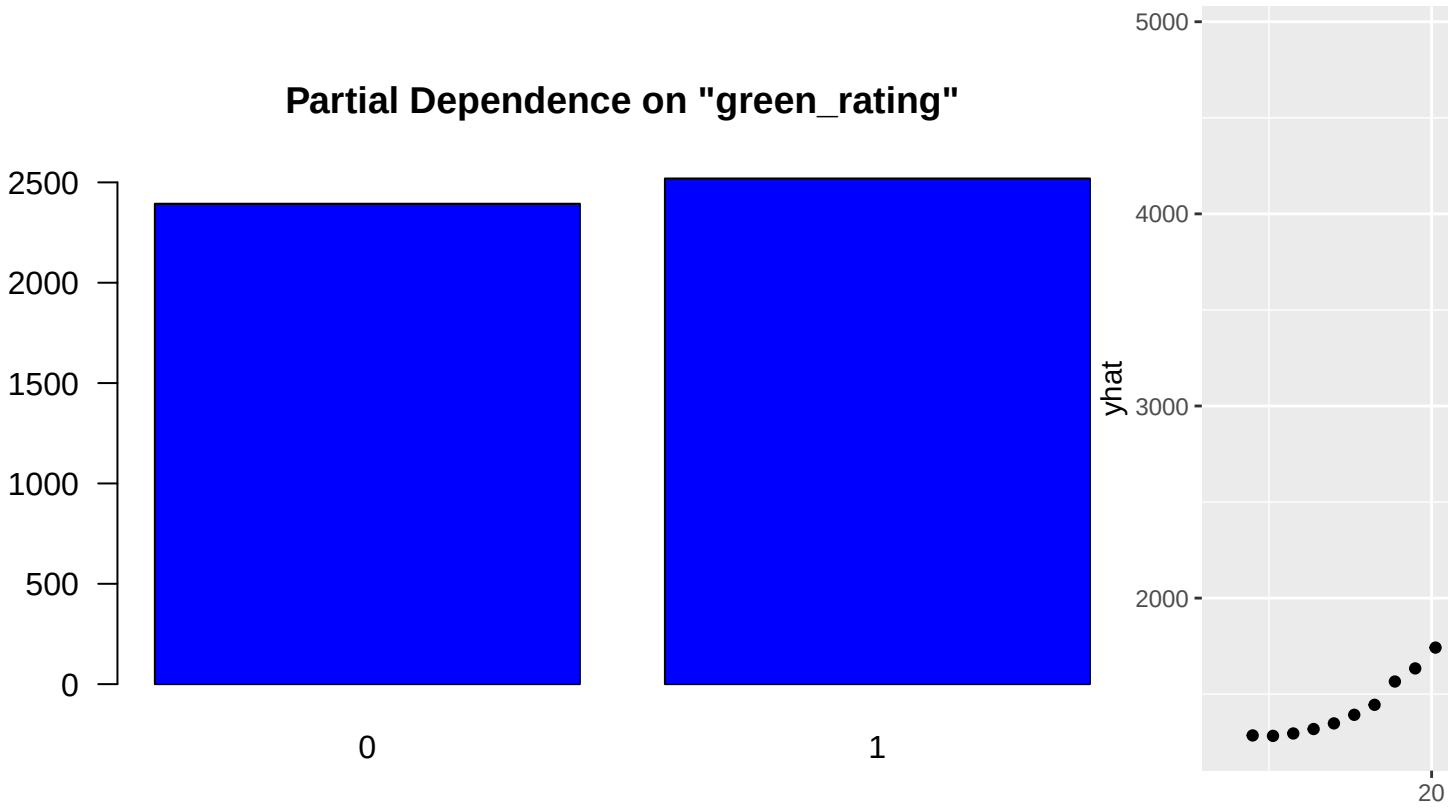

```
## [1] 1028.207
```

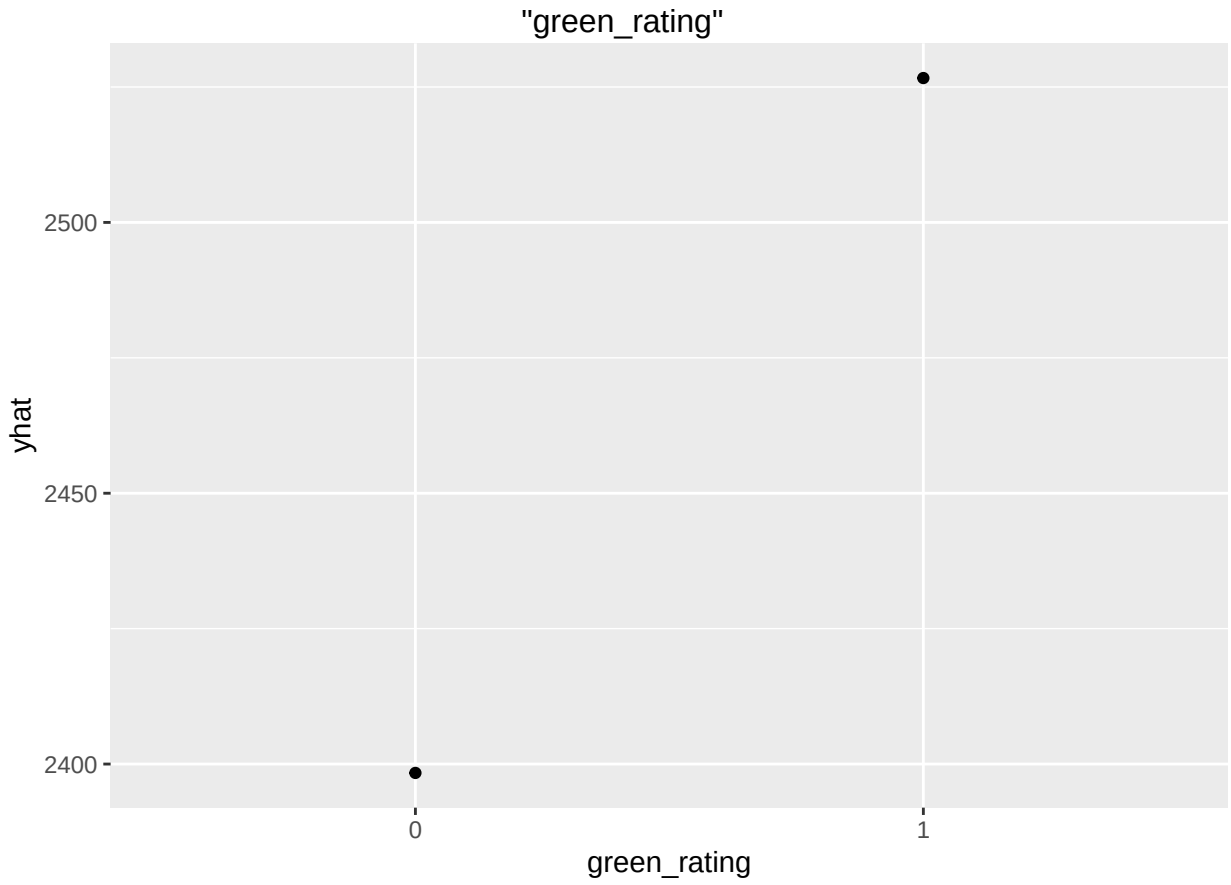
Table 3: RMSE of each model, first pass

model	RMSE
lm (baseline)	1013.1347
buildings.tree0	1028.2067
buildings.forest0	775.8514
buildings.forest1	780.7535

model	RMSE
buildings.forest2	917.4243

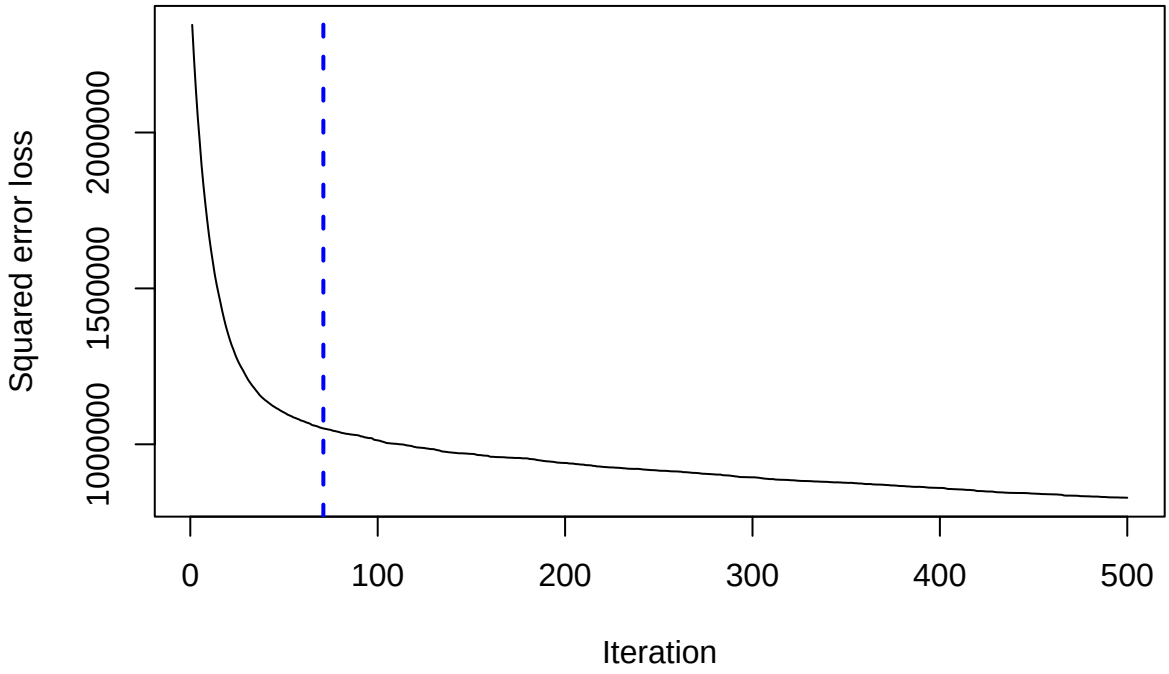



Partial Dependence on "green_rating"
2500
2000
1500
1000
500
0
0
1
"green_rating"
yhat
5000
4000
3000
2000
20


yhat
2500
2450
2400
0
1
green_rating

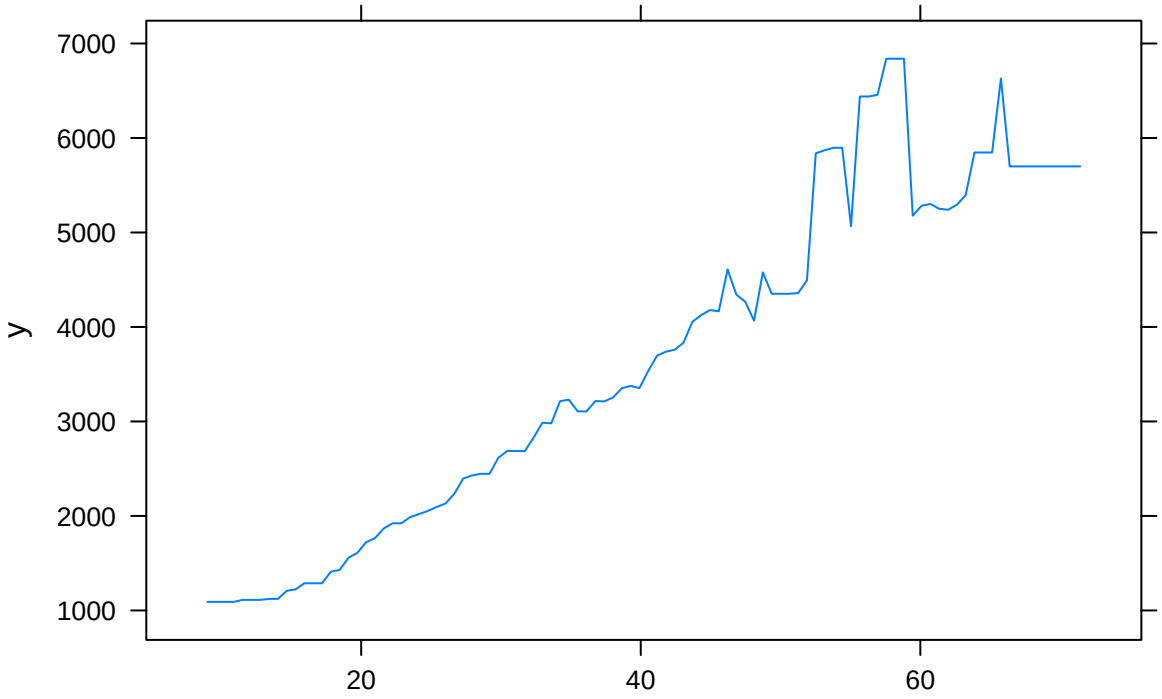
Our best model above is a random forest.

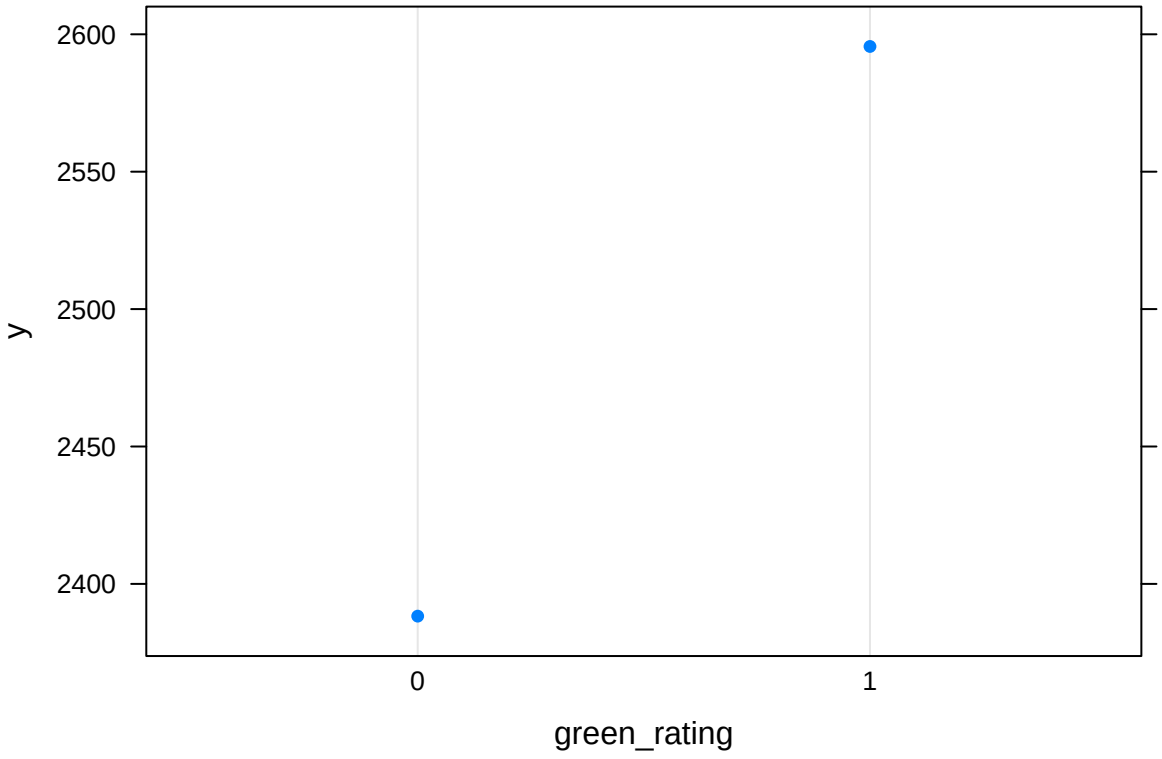
Comparing with a boosted regression tree:




```
## [1] 71
## attr(,"smoother")
## Call:
## loess(formula = object$oobag.improve ~ x, enp.target = min(max(4,
##     length(x)/10), 50))
##
## Number of Observations: 500
## Equivalent Number of Parameters: 39.85
## Residual Standard Error: 1073
```

```
## [1] 1036.371
```


7000
6000
5000
4000
3000
2000
1000
y
20
40
60
City_Market_Rent


2600
2550
2500
2450
2400
y
0
1
green_rating