Exercise 2

2023-03-19

1) What causes what?

1-1 Why we can't regress crime on police?

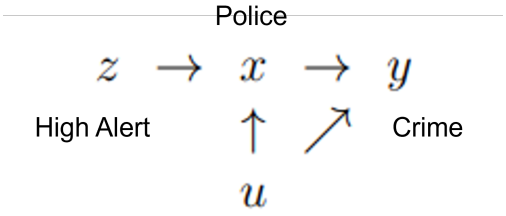
Because we can't misunderstand the causal effect between crime and police. As the podcast said, in Washington DC, a lot of extra police officers are hired even if the crime rate is low because of preparing terrorism. However, in general, high crime cities have an incentive to hire a lot of cops. Therefore, if you regress crime on police in a few different cities like DC, you probably misunderstand the effect. So you can't do this.

1-2 How can a professor identify the causal effect?

They found that when the terror alert level goes up, extra police are put on the mall in other parts of Washington to protect against terrorists has nothing to do with street crime. Also then, the streets were safer(the number of murder, robbery, assault goes down).

To show this causality, they regressed the crime on the High Alert, and got the result of the Table 2. This background of the regression is as follow:

At first, they thought this structure, where the outcome y is crime, a variable that want to shows the causality x, and a instrument variable z is the High Alert.



Because of the endogeneity, they cannot direct regress y on x, but they selected the High Alert as z that is positive correlated with x. And then they used regress y on z with reduced form like:

$$y_{crime} = \gamma_0 + \gamma_1 z_{\{High\ \ Alert\}} + \varepsilon$$

where $\beta_1\pi_1$ and $x_{police} = \pi_0 + \pi_1 z_{\{High\ \ Alert\}} + v$. Since the x and z are positive correlated, the coefficient of the High Alert on Table 2 shows $\beta_1 < 0$. This means that on the high-alert days, total crimes decrease by an average of seven crimes per day, or approximately 6.6 percent. Also, it means that the more police causes the less crime.

1-3 Why Metro ridership? What was captured?

According to their talks, they concerned about the possibility that tourist were less likely to visit Washington DC if the High Alert was announced, and what tourist were less likely to visit caused less crime. To check this hypothesis, they added a variable of the scaled Metro Ridership into the regression model.

And then, they found that the coefficient of the High Alert is still negative even if they added its variable. Therefore, they concluded that if the same number of tourists, the more police causes the less crime as well as the result 1-2.

1-4

This model is DID model, which control group is other district and treatment group is District 1.

And, the difference between the High Alert × District One and the High Alert × Other Districts coefficients represents the effect of District 1 on the crime under the setting that controls for all common factors between the districts. They found that even after controlling for all such factors and recognizing that our assumption is too strong, we still find that crime decreases in District 1 during high-alert periods by some two crimes per day, or more than 12 percent.

2) Tree modeling:dengue cases

1. Overview

- Our goal is to use CART, random forests, and gradient-boosted trees to predict dengue cases

2. Data and Model

2-1 Data

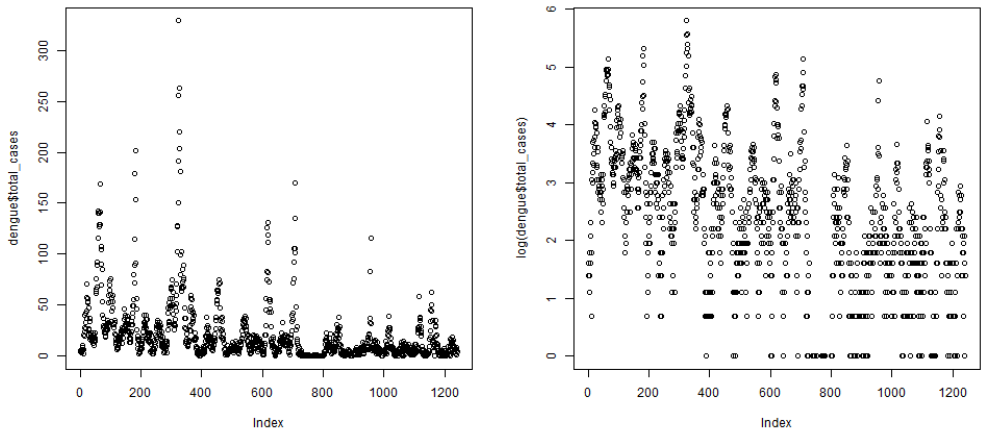
- dengue.csv
- The detailed explanations of each variables are in the prompt

2-2 Model

We took CART, random forests, and gradient-boosted trees to predict dengue cases as follow:

$$total\ cases = city + season + specific_humidity + tdtr_k + precipitation_amt$$

Note: we did not take log for total cases because we thought total cases did not look like it had any trend term as follow:



Note: the left side shows the total case and right side does the logarithm of the the total case.

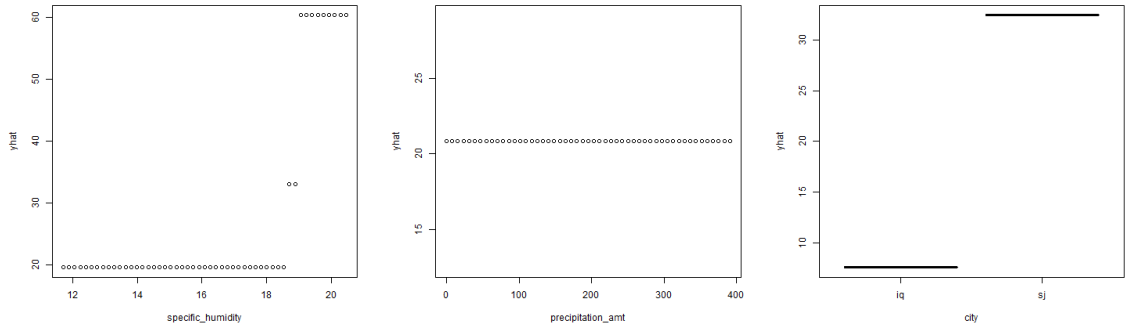
3. Results and Conclusion

From the result, these rmse of there models are

cart	forest	boost
24.5693	22.40354	18.50443

Therefore, the best model to predict dengue cases in this analysis is the gradient-boosted trees.

Also, three partial dependence plots on specific_humidity, precipitation_amt and city in the boost model are



From these graphs, we can get some interpretations in the following: - If average specific humidity(specific_humidity) is over about 19, total cases will increase by 40 - Rainfall for the week in millimeters does not relate to the total cases - the avereage cases in the San Juan will be larger than in Iquitos, Peru by about 30

3) Predictive model building: green certification

1. Overview

- Our goal is to build the best predictive model possible for revenue per square foot per calendar year
- "revenue per square foot per calendar year" is the product of rent and leasing_rate in the data

2. Data and Model

2-1 Data

- Green buildings in greenbuildings.csv (7894 commercial rental properties from across United States)
- Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building
- The detailed explanations of each variables are in the prompt
- "revenue per square foot per claendar year"(RPS), which will be dependent variable, is the product of rent and leasing_rate in the data.
- Excluded CS.PropertyID from the data because it has no meaning for this analysis

2-2 Model

We took 2 steps to get the best predictive model as follow:

2-2-1 Model Selection

At first, we did the stepwise selection and the Lasso regression to find independent variables that we should include the model. After we got both results, we compared two rmses and decided to include variables that has the lower rmse.

Note: Because of the limitation of the time(step wise selection are required to take a lot of time), we did not use cross validation to improve the quality of our analysis. And we considered LEED and EnergyStar separately thorough this analysis.

2-2-2 Compared Regressions and Trees

After we decided to use dependent variables in this model, we compared models of the "linear regression" and "Knn regression" and the models of the "CART", "Random Forest" and "Boost" from the perspective of the RMSE with K-CV(10 folds).

Note: in Tree models, we did not specify dependent variables like linear and knn regression, because they automatically consider interaction terms.

2-2-3 Partial Dependence(Additional)

We also got some partial dependences of the model that had the lowest rmse in our models from 2-2-2 to interpret our model.

3. Results

3-1 Model Selection

From the result, rmses of the stepwise and lasso are

step	lasso
991.152	2861.317

Therefore, we decided to use dependent variables according to the stepwise as follow:

$$
\begin{aligned}
RPS = \beta_0 + \beta(&cluster + size + empl_gr + stories + age + renovated + class_a + class_b \\
&+ LEED + Energystar + net + amenities + cd_total_07 + hd_total07 + Precipitation + Gas_Costs \\
&+ Electricity_Costs + City_Market_Rent + size \times City_Market_Rent + cluster \times size \\
&+ cluster \times City_Market\ Rent + stories \times class\ a + size \times Precipitation \\
&+ amenities \times Gas\ Costs + empl\ gr \times Electricity\ Costs + stories \times Gas\ Costs \\
&+ age \times City\ Market\ Rent + age \times Electricity\ Costs + renovated \times Precipitation \\
&+ renovated \times City\ Market\ Rent + renovated \times Gas\ Costs + size \times class\ b \\
&+ size \times class\ a + size \times age + age \times class\ a \\
&+ Electricity\ Costs \times City\ Market\ Rent + renovated \times hd\ total07 \\
&+ cluster \times Precipitation + class\ a \times Gas\ Costs + class\ b \times LEED \\
&+ size \times hd\ total07 + hd\ total07 \times Precipitation + cd\ total\ 07 \times Precipitation \\
&+ Precipitation \times City\ Market\ Rent + Gas\ Costs \times City\ Market\ Rent \\
&+ class\ a \times hd\ total07 + class\ a \times Electricity\ Costs + size \times Electricity\ Costs \\
&+ empl\ gr \times renovated + stories \times renovated + size \times renovated \\
&+ class\ a \times Precipitation + cluster \times Electricity\ Costs + cluster \times hd\ total07 \\
&+ cluster \times stories + size \times stories + stories \times age)
\end{aligned}
$$

3-2 Comparison

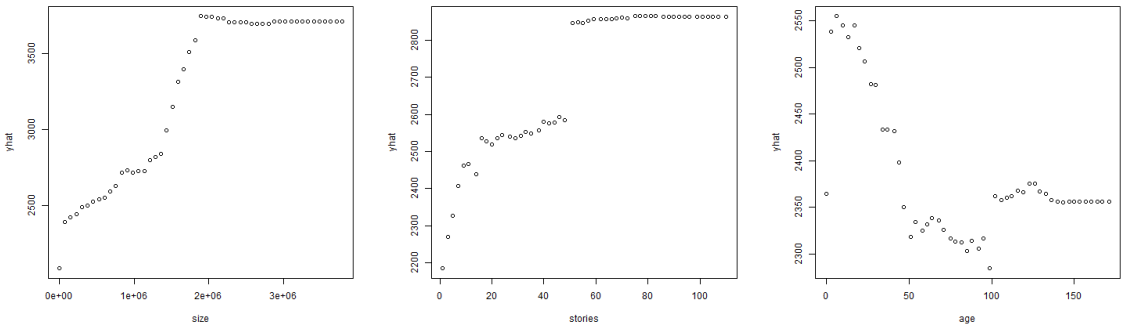
From the results, we got the rmse of the linear, knn, CART, Forest, Boost, which are

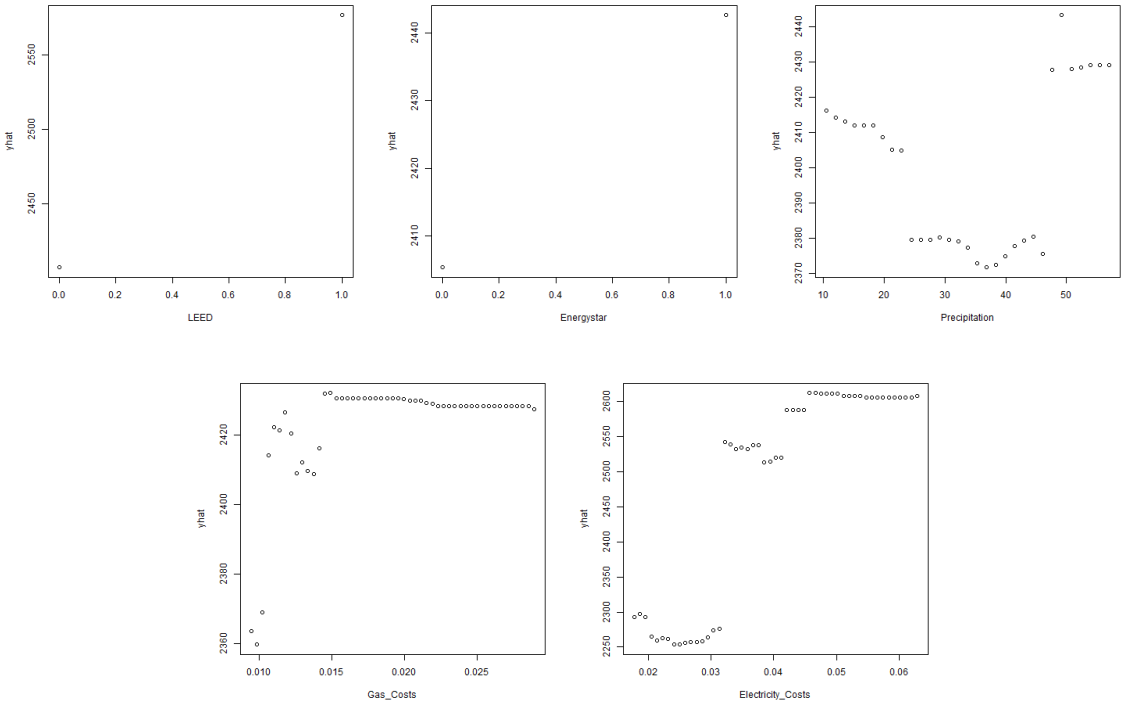
Lm	Knn	CART	Forest	Boost
1006.365	1375.176	940.1157	414.1079	817.5708

Therefore, Forest model had the lowest rmse in this analysis.

3-3 Partial Dependence(Additional)

We got some partial dependences of the random forest model that had the lowest rmse to interpret our model.





Conclusion

In our analysis, the forest model was the best predictive model, which means it had the lowest rmse in models(linear, knn, CART, random forest, boost).

From the partial dependence of the forest model, revenue per square foot per year(RPS) will goes up if its size, stories, gas costs and electronic costs. Also, if a building had LEED and Energy star, RPS will increase but its magnitude of LEED (over 2550)is larger that that of Energystar (under 2450) from graphs. The age and precipitation make RPS lower, but it looks like not going to do that if it goes over thresholds.

Appendix

The optimal coefficients in the stepwise function

Coefficients of the stepwise model		
(Intercept)	cluster	size
-3.13E+02	-7.85E-01	5.46E-04
empl_gr	stories	age
8.52E+01	-5.82E+00	-2.77E+00
renovated	class_a	class_b
-3.03E+02	9.14E+02	3.26E+02
LEED	Energystar	net
7.66E+01	3.82E+02	-1.43E+02
amenities	cd_total_07	hd_total07
-2.57E+02	2.12E-01	1.67E-01
Precipitation	Gas_Costs	Electricity_Costs
5.14E+01	-1.29E+05	-1.08E+04
City_Market_Rent	size:City_Market_Rent	cluster:size
5.46E+01	5.02E-05	7.54E-07
cluster:City_Market_Rent	stories:class_a	size:Precipitation
7.72E-03	-9.17E+00	-1.99E-05
amenities:Gas_Costs	empl_gr:Electricity_Costs	stories:Gas_Costs
6.39E+04	-3.78E+03	1.53E+03
amenities:Precipitation	Energystar:amenities	cd_total_07:hd_total07
-9.78E+00	-3.14E+02	-3.95E-05
age:City_Market_Rent	age:Electricity_Costs	renovated:Precipitation
-3.24E-01	2.97E+02	8.94E+00
renovated:City_Market_Rent	renovated:Gas_Costs	size:class_b
1.03E+01	-2.62E+04	-1.84E-03
size:class_a	size:age	age:class_a
-2.12E-03	-1.50E-05	5.04E+00
Electricity_Costs:City_Market_Rent	renovated:hd_total07	cluster:Precipitation
6.56E+02	3.43E-02	-4.41E-03
class_a:Gas_Costs	class_b:LEED	size:hd_total07
4.70E+04	5.79E+02	1.39E-07
hd_total07:Precipitation	cd_total_07:Precipitation	Precipitation:City_Market_Rent
-4.49E-03	-6.30E-03	-5.08E-01
Gas_Costs:City_Market_Rent	class_a:hd_total07	class_a:Electricity_Costs
2.46E+03	-6.13E-02	-2.05E+04
size:Electricity_Costs	empl_gr:renovated	stories:renovated
4.18E-02	7.91E+00	-1.64E+01
size:renovated	class_a:Precipitation	cluster:Electricity_Costs
6.25E-04	-5.47E+00	1.65E+01
cluster:hd_total07	cluster:stories	size:stories
5.93E-05	-7.45E-03	-3.72E-06
stories:age		
1.14E-01		

“‘