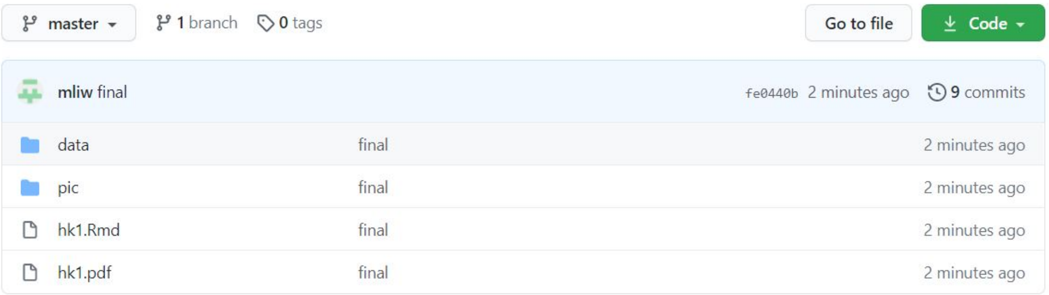
ECO395M STAT LEARNING Homework 3[*]

Mingwei Li, Xinyu Leng, Hongjin Long

Abstract

This document is the third homework of ECO395M STAT LEARNING.



mliw final fe0440b 2 minutes ago 9 commits

data	final	2 minutes ago
pic	final	2 minutes ago
hk1.Rmd	final	2 minutes ago
hk1.pdf	final	2 minutes ago

[*] Mingwei Li, Xinyu Leng and Hongjin Long are master students of economics, The University of Texas at Austin

Contents

1 What causes what?

(Q1_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.)

What we want to know is the causal effect of police on crime rate. This is of great policy significance.

It's possible that places with an inordinate amount of crime tend to employ a large police force. Which means crime rate is the causal effect of police.

The result of such regression can not credibly identify a causal effect of police on crime.

(Q1_2) 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.

How were the researchers from UPenn able to isolate this effect? Briefly describe their approach

They use the easily identifiable and clearly exogenous shock provided by changes in the terror alert level in Washington, D.C., to evaluate the causal effect of police on crime. A notable benefit of their research design is that their treatment, the terror alert level, turns on and off repeatedly during their sample.

The logi is: changes in the terror alert level ⇒ changes in the number of police ⇒ changes in different types of crimes. Therefore, they can estimate the causal effect of police on crime.

and discuss their result in the "Table 2" below, from the researchers' paper.

The table 2 is listed below. We would discuss the entry in the table one by one.

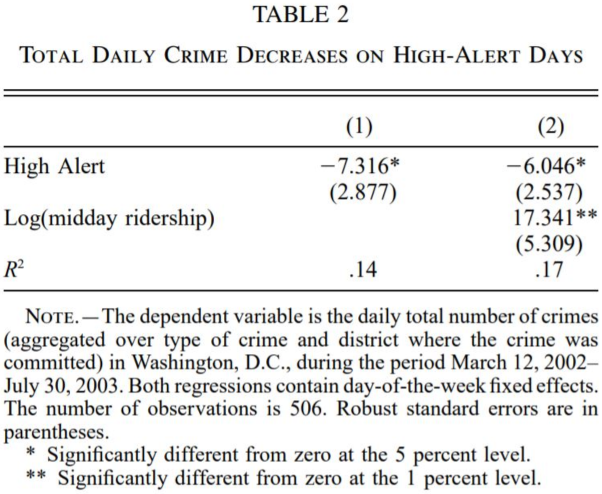


TABLE 2

Total Daily Crime Decreases on High-Alert Days

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

Note.—The dependent variable is the daily total number of crimes (aggregated over type of crime and district where the crime was committed) in Washington, D.C., during the period March 12, 2002–July 30, 2003. Both regressions contain day-of-the-week fixed effects. The number of observations is 506. Robust standard errors are in parentheses.

* Significantly different from zero at the 5 percent level.

** Significantly different from zero at the 1 percent level.

(row1 column1 -7.316): The results from their most basic regression are presented in Table 2, where they regress daily D.C. crime totals against the terror alert level (1 high, 0 elevated) and a day-of-the-week indicator. The coefficient on the alert level is statistically significant at the 5 percent level and indicates that on high-alert days, total crimes decrease by an average of seven crimes per day, or approximately 6.6 percent

(column2 -6.046 17.341): To investigate the effect of tourism more systematically, in column 2 of Table 2 they verify that high-alert levels are not being confounded with tourism levels by including logged midday Metro ridership directly in the regression. The coefficient on the alert level is slightly smaller, at -6.2 crimes per day. Interestingly, they find that increased Metro ridership is correlated with an increase in crime. The increase, however, is very small—a 10 percent increase in Metro ridership increases the number of crimes by only 1.7 per day on average. Thus, given that midday Metro ridership is a good proxy for tourism, changes in the number of tourists cannot explain the systematic change in crime that they estimate.

(Q1_3) Why did they have to control for Metro ridership? What was that trying to capture?

Why did they have to control for Metro ridership?

What has been confirmed from basic regression is high alert level ⇒ less crime.

There are 2 hypotheses:(1) high alert level ⇒ high police level ⇒ less crime. (2) high alert level ⇒ less tourism ⇒ less crime. They want to rule out the second hypothesis, therefore they have to control for Metro ridership.

What was that trying to capture?

They are trying to capture the causal effect of tourism on crime.

(Q1_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?

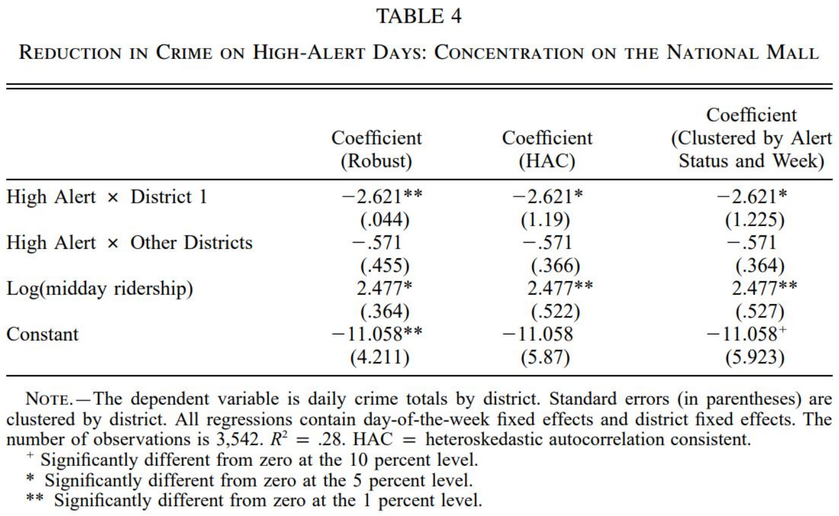


TABLE 4

Reduction in Crime on High-Alert Days: Concentration on the National Mall

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	−2.621**	−2.621*	−2.621*
	(.044)	(1.19)	(1.225)
High Alert × Other Districts	−.571	−.571	−.571
	(.455)	(.366)	(.364)
Log(midday ridership)	2.477*	2.477**	2.477**
	(.364)	(.522)	(.527)
Constant	−11.058**	−11.058	-11.058^{+}
	(4.211)	(5.87)	(5.923)

Note.—The dependent variable is daily crime totals by district. Standard errors (in parentheses) are clustered by district. All regressions contain day-of-the-week fixed effects and district fixed effects. The number of observations is 3,542. $R^2 = .28$. HAC = heteroskedastic autocorrelation consistent.

$^{+}$ Significantly different from zero at the 10 percent level.

* Significantly different from zero at the 5 percent level.

** Significantly different from zero at the 1 percent level.

Can you describe the model being estimated here?

D.C has many districts. District 1 is the most important one, as White House is there. Therefore, the police would place a great amount of force in district 1 during high-alert period.

The regression with district fixed effects is in Table 4. During periods of high alert, crime in the National Mall area(district 1) decreases by 2.62 crimes per day. Crime also decreases in the other districts, by .571 crimes per day, but this effect is not statistically significant.

What is the conclusion?

Police has a negative causal effect on crime, after controling other factors similar across the districts.

We assume the police level in district 1 is much higher than other districts. In this case, the difference between the High Alert×District One and the High Alert×Other Districts coefficients is a differencein-difference estimator that controls for all common factors between the districts. If bad weather, for example, causes decreases in crime, a coincidental correlation with the timing of a high alert could confound their results. The difference-in-difference estimator controls for any factors such as weather, tourism, or other events that affect the districts similarly. Even after controlling for all such factors and recognizing that their assumption is too strong, they still find that crime decreases in District 1 during high-alert periods by some two crimes per day, or more than 12 percent.

2 Predictive model building: green certification

2.1 Overview of the Problem

With the increasing concentration on environmental protection, green certification for buildings become more and more important. The buildings with green certification are able to charge more rent as they can save energy cost and bring reputation to the tenant. However, the benefit of green certification needs to be quantified to provide more guidance for house owners. It's very difficult to extract the effect of green certification on revenue per square foot per calendar year, as there are many factors which can affect leasing rate and rent.

In this part, we (1) build a predictive model for revenue per square foot per calendar year based on 7894 commercial rental properties from across the United States, (2) and use this model to quantify the average change in rental income per square foot associated with green certification, holding other features of the building constant.

$$rental\_income\_per\_square\_foot = leasing\_rate \times Rent$$

2.2 Data and Modeling Approach

2.2.1 Data

Our data has 7894 observations, and each one of them is a commercial rental property. After removing useless columns, we have 18 features and one prediction target($rental\_income\_per\_square\_foot$).

In conclusion, X is a matrix of 7894×18, Y is a vector of 7894×1. Our goal is to build a model on $X \sim Y$ with the lowest cross-validation rmse.

The definitions of 18 features are listed below:

1 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.

2 size: the total square footage of available rental space in the building.

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

4 stories: the height of the building in stories.

5 age: the age of the building in years.

6 renovated: whether the building has undergone substantial renovations during its lifetime.

7,8 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.

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

10 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.

11 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.

12 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.

13 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.

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

15 Precipitation: annual precipitation in inches in the building's geographic region.

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

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

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

The definition of prediction target ($rental\_income\_per\_square\_foot$):

$$rental\_income\_per\_square\_foot = leasing\_rate \times Rent$$

2.2.2 Model

Xgboost is adopted to fit the data, as tree model can capture nonlinear relationship. The process of modeling can be summarized as follows:

(1) Fit on the whole data set to get feature importance ranks.

(2) Use top k features($k \in [1, 18]$) unioned with $green\_rating$ as input features. Then, we use grid-search to find the best parameters with the lowest cross-validation rmse. For instance, the top 3 features unioned with $green\_rating$ is

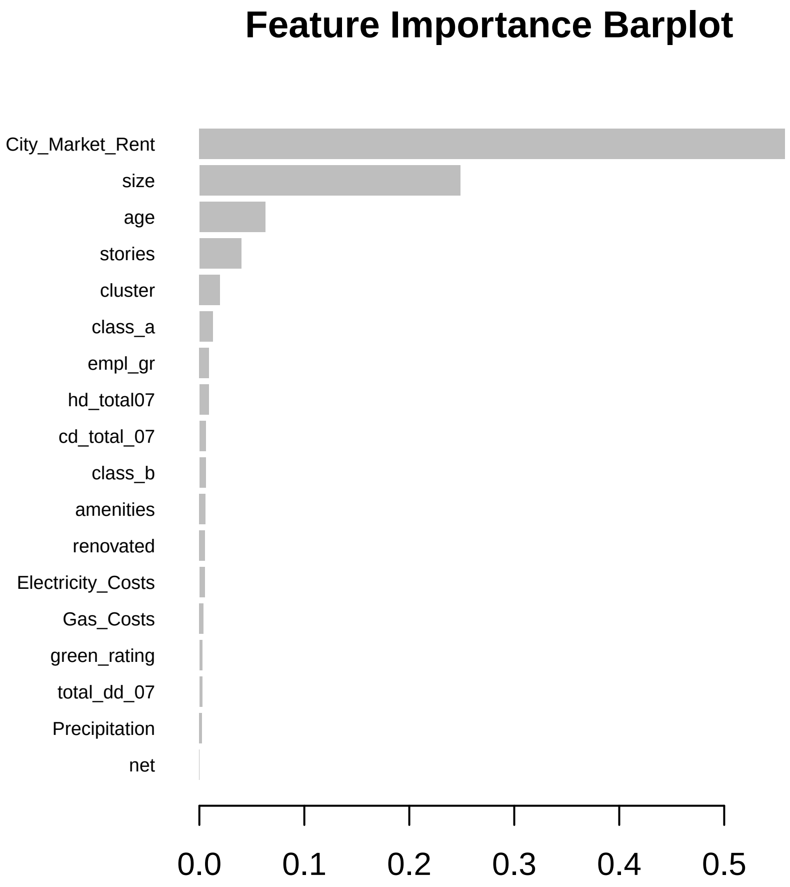
$$City\_Market\_Rent, size, age + green\_rating$$

(3) Finally, we find k_{best} features and best model parameters. We can 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.

2.3 Results of Data Analysis

2.3.1 Feature Importance

The 18 features are ranked in the following picture. We can see that *City_Market_Rent* is the most important feature in predicting revenue per square. *green_rating* ranks 15 among the 18 features, not very significant.



2.3.2 Best Features and Parameters

We use top k features($k \in [1, 18]$) unioned with $green\_rating$ as input features. Then, we use grid-search to find the best parameters with the lowest cross-validation rmse. For instance, the top 3 features unioned with $green\_rating$ is

$$City\_Market\_Rent, size, age + green\_rating$$

The cv-rmses of top k features($k \in [1, 18]$) unioned with $green\_rating$ are listed in table 1. Top 11 features unioned with $green\_rating$ is the best combination, with cv-rmse of 7.3272304. Corresponding $max\_depth = 9, eta = 0.25$

Table 1: Different Features and Parameters

features	max_depth	eta(learning rate)	cv-rmse
top 1 features	3	0.1	10.9966248
top 2 features	9	0.5	7.8924662
top 3 features	9	0.25	7.5659126
top 4 features	11	0.25	7.4700874
top 5 features	11	0.25	7.5681676
top 6 features	9	0.25	7.5822794
top 7 features	9	0.25	7.4596536
top 8 features	11	0.1	7.5252744
top 9 features	9	0.25	7.509116
top 10 features	9	0.25	7.4896898
top 11 features	9	0.25	7.3272304
top 12 features	9	0.25	7.3872686
top 13 features	11	0.1	7.4139402
top 14 features	9	0.25	7.3809
top 15 features	9	0.25	7.3809
top 16 features	11	0.25	7.4054108
top 17 features	11	0.25	7.3746844
top 18 features	11	0.1	7.3900858

The best feature combination is:

$$City\_Market\_Rent + size + age + stories + cluster + class\_a + empl\_gr + hd\_total07 + cd\_total\_07$$

$$+class\_b + amenities + green\_rating$$

2.3.3 Effect of Green Certification

After getting the best features and model-parameters, we fit on the **WHOLE** data set. The fitting rmse is 2.484052.

The we use the fitted model to predict an artifical example, this example is listed in table 2.

Table 2: An Artifical Example

Variable name	sample_data
City_Market_Rent	27.49728465
size	234637.7435
age	47
stories	13
cluster	403
class_a	1
empl_gr	3.206719949
hd_total07	3432.042311
cd_total_07	1229.354193
class_b	1
amenities	1
green_rating	1/0

The prediction is as follows. In this artificial example, the green certification house s' revenue per square foot per year is $27.75311 - 27.16981 = 0.5833$ higher than the house without green certification.

```
## [1] "prediction for green_rating==1"

## [1] 27.75311

## [1] "prediction for green_rating==0"

## [1] 27.16981
```

We do see Green Certification has a positive effect on the house revenue.

2.4 Conclusions

Conclusion 1: *green_rating* is NOT an important feature in determining revenue per square foot per year.

Conclusion 2: *green_rating* has a positive effect on revenue per square foot per year based on our artificial example. However, after we set *City_Market_Rent* = 34, the effect reverses.

```
## [1] "prediction for green_rating==1"

## [1] 29.31454

## [1] "prediction for green_rating==0"

## [1] 29.38739
```

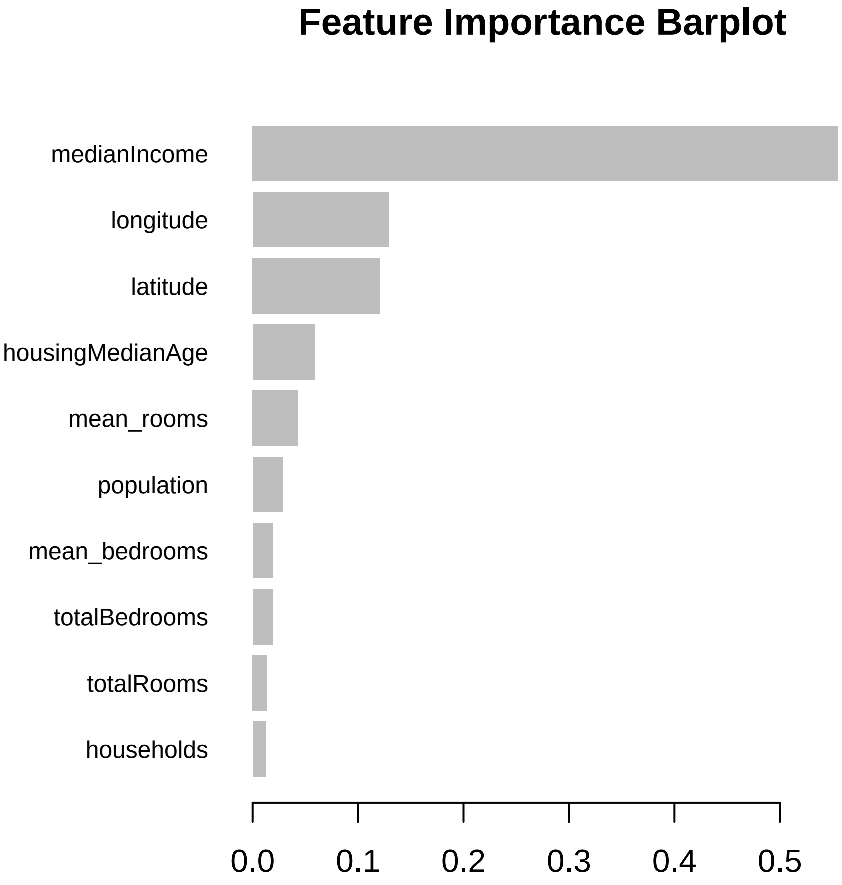
Note the fact that *City_Market_Rent* is the most important feature, the effect of *green_rating* relies heavily on other features.

3 Predictive model building: California housing

3.1 Build the Best Predictive Model for MedianHouseValue

Your task is to build the best predictive model you can for medianHouseValue, using the other available features. Write a short report detailing your methods. Make sure your report includes an estimate for the overall out-of-sample accuracy of your proposed model.

As Tree model can capture nonlinear relationship. We adopt xgboost to model the build the predictive model for medianHouseValue. The feature-importance figure is as follows:



The performance of different feature combinations is listed in the table 3.

Table 3: An Artifical Example

features	max_depth	eta(learning rate)	cv-rmse
top 1 features	3	0.25	82655.34844
top 2 features	5	0.25	69552.60156
top 3 features	11	0.25	47500.57578
top 4 features	9	0.25	48271.325
top 5 features	9	0.25	47906.44141
top 6 features	9	0.25	48124.78359
top 7 features	7	0.25	49340.41016
top 8 features	7	0.25	48800.51328
top 9 features	7	0.25	48629.82578
top 10 features	7	0.25	48478.21328

Top 3 features $medianHouseValue, longitude, latitude$, with $max\_depth = 11$, have the best out-of-sample performance($cv\_rmse = 47500.57578$).

Then, we use the best feature combination and best parameters to fit on the whole data set. The fitting rmse is 23846.53.

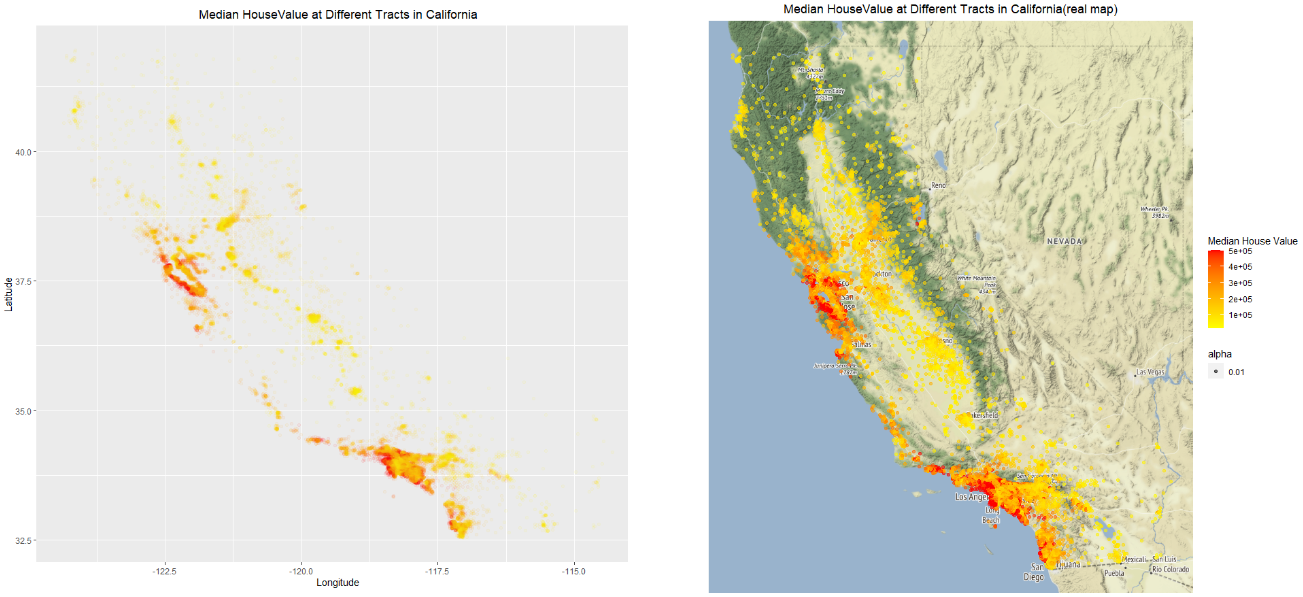
```
## [1] 23846.53
```

3.2 The Three Figures

Also include three figures:

(Q3__2__1):a plot of the original data, using a color scale to show medianHouseValue (or log medianHouseValue) versus longitude (x) and latitude (y).

From figure 1, we can see a lot of red points on the coastline, which means house price is higher in areas near sea.



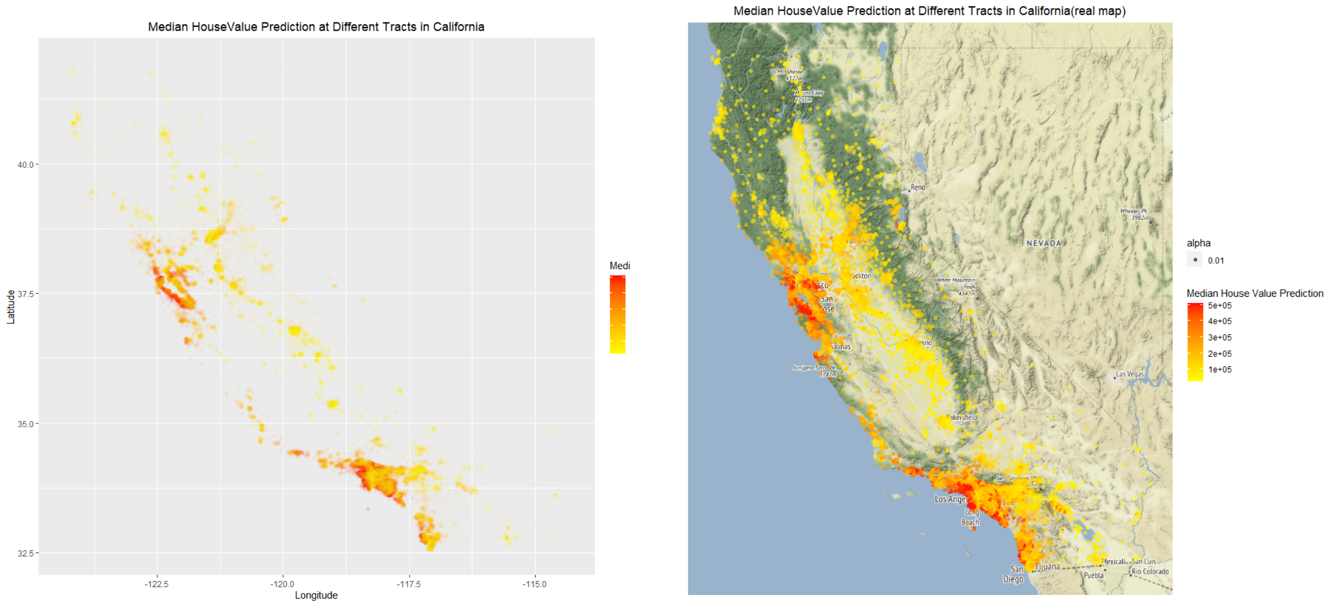
(a) Median HouseValue at Different Tracts in California

(b) Median HouseValue at Different Tracts in California(real map)

Figure 1: A Plot of the Original Data

(Q3__2__2):a plot of your model's predictions of medianHouseValue (or log medianHouseValue) versus longitude (x) and latitude (y).

Figure 2 is similar to figure 1, which means our model fits very well. Cross-validation rmse is 47500.57578, not too far from fitting error of 23846.53 ⇒ our model is robust.


Median HouseValue Prediction at Different Tracts in California
Latitude
40.0
37.5
35.0
32.5
-122.5
-120.0
-117.5
-115.0
Longitude
Medi
Median HouseValue Prediction at Different Tracts in California(real map)
Reno
NEVADA
Los Angeles
Long Beach
San Diego
Tijuana
Las Vegas
alpha
0.01
Median House Value Prediction
5e+05
4e+05
3e+05
2e+05
1e+05

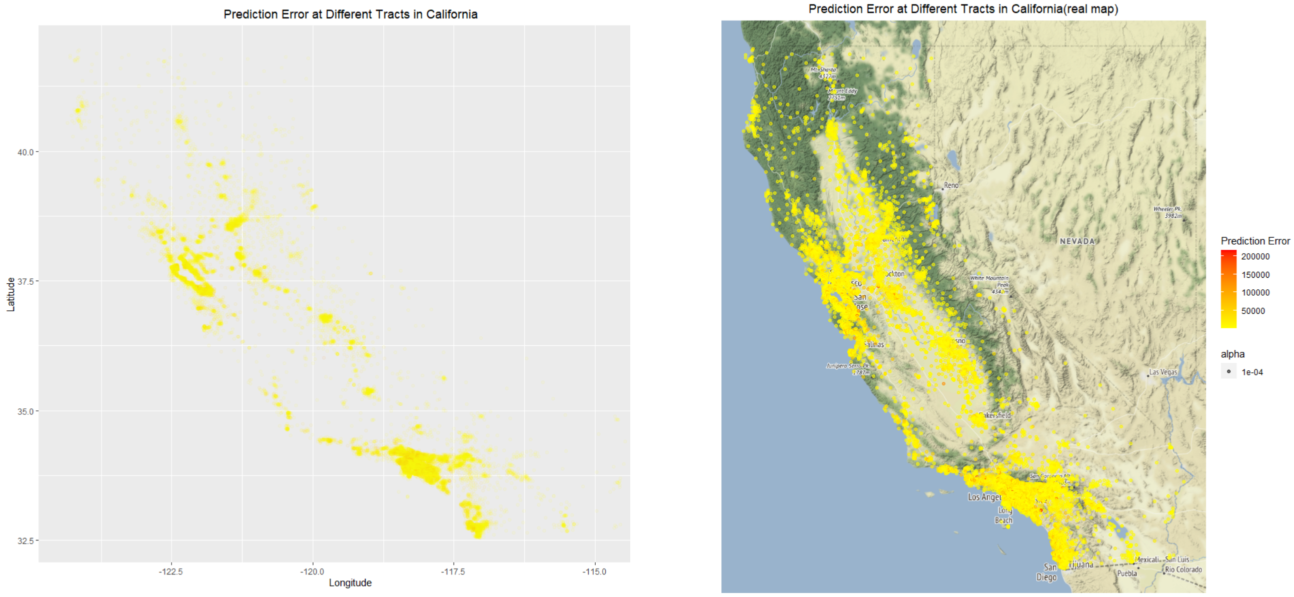
(a) Median HouseValue Prediction at Different Tracts in California

(b) Median HouseValue Prediction at Different Tracts in California(real map)

Figure 2: A Plot of my Model's Predictions

(Q3__2__3):a plot of your model's errors/residuals (or log residuals) versus longitude (x) and latitude (y).

Figure 3 is mainly consisted of yellow points, which means our model fits well across different tracts(absolute residuals are low). Note we've taken absolute value of those residuals to represent prediction errors.



(a) Prediction Errors at Different Tracts in California

(b) Prediction Errors at Different Tracts in California(real map)

Figure 3: A Plot of my Model's Errors/Residuals