A3_recovery

Yefu Chen

2021/4/8

QUESTION 1

First, listen to this podcast from Planet Money. Then use your knowledge of statistical learning to answer the following questions.

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

The primary reason is that the causality of increasing the number of cops in a city to the crime is unclear. Although the statistical model can indicate the one unit increase in the cops' number correlated with one unit decrease in the crime, it does not tell other factors that can affect this relationship. First, criminals can move to other cities if there are more cops in DC. Although researchers observe that the number of crimes in DC is decreasing, it is not confident to say that increasing cops can affect the crime through a generalized perspective. Second, criminals may apply crimes in hidden places to avoid the cops on the streets. Under this situation, although the number of victims on the streets is decreasing, it is not confident to say that increasing cops can reduce all crimes. Hence, to explore the real relationship between increasing the number of cops and crime, we should apply a dynamic and systematic perspective instead of focusing on a few empirical cases.

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 UPenn researchers introduced the metro ridership to control possible external effects. As present in the following table, they first chose the high-alert dummy variable as the independent variable to predict crime. They found that the crimes during high-alert days decrease by 7.316 than days that are not high-alert. However, they also notice possible external factors (e.g., crimes happening in hidden places and criminals moving out). Hence in the second model (column 2), they added log-transformed metro ridership as a control variable. Compared to normal days, the crimes in high-alert days decrease by 6.046 units. For each one-unit increase in log-transformed metro ridership, there can be 17.341 more units of crimes.

The results indicate fewer crimes during the high-alert days, while more metro ridership is positively associated with the number of crimes.

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

Choosing the control variable as metro ridership is a smart step since there are more cops on the streets during the high-alert days, while the number of cops in metro stations should be same. In other words, choosing the metro ridership can somehow control the effect of crimes in hidden places. Besides, the metro ridership can be a measurement for travel behaviors, which can relate to the population, or the number of criminals. Hence controlling the effect of metro ridership can control the effects of criminals moving out.

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?

TABLE 4 indicates the interaction between the high-alert dummy and District 1 and the interaction between the high-alert dummy and Other Districts, controlling metro ridership's effects. Compared to other districts, there are more decreases (2.621) in crimes in District 1 than others during the high-alert days, but the decreases in crimes in Other Districts are the same. For each one-unit increase in log-transformed metro ridership, there can be 2.477 more units of crimes.

From the above results, we can conclude that during the high-alert days, the number of crimes in District 1 decreases a lot during the high-alert days. Under this situation, this area may not be the priority for cops to prevent. Also, those areas with high metro ridership may be riskier than others, and they should be the areas that cops gather. Hence, I suggest policymakers wisely allocate the number of cops to maximum safety during the high-alert days.

QUESTION 2

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.

Answers:

Overview: This study explores the associations between green certifications and the revenue per square foot per year through the Gradient Boosting Machines. I hypothesize that green certifications can promote the revenue of commercial rental properties.

Data and Model: I used the dataset about green buildings. There are 7,894 commercial rental properties from across the United States. Of these, 625 properties have been awarded by EnergyStar certification, 47 by LEED certification, and 7 by both certifications as green buildings.

I applied the Gradient Boosting Machines (GBM) to explore the associations between green certifications and the revenue per square foot per year. Before analysis, I dropped all rows that included the NA value. At first, I calculated the product of two rent and leasing_rate as the dependent variable. I noticed a skewness in the distribution of the independent variable (**Figure 2-1**). I used the log-transformation to ease this effect.

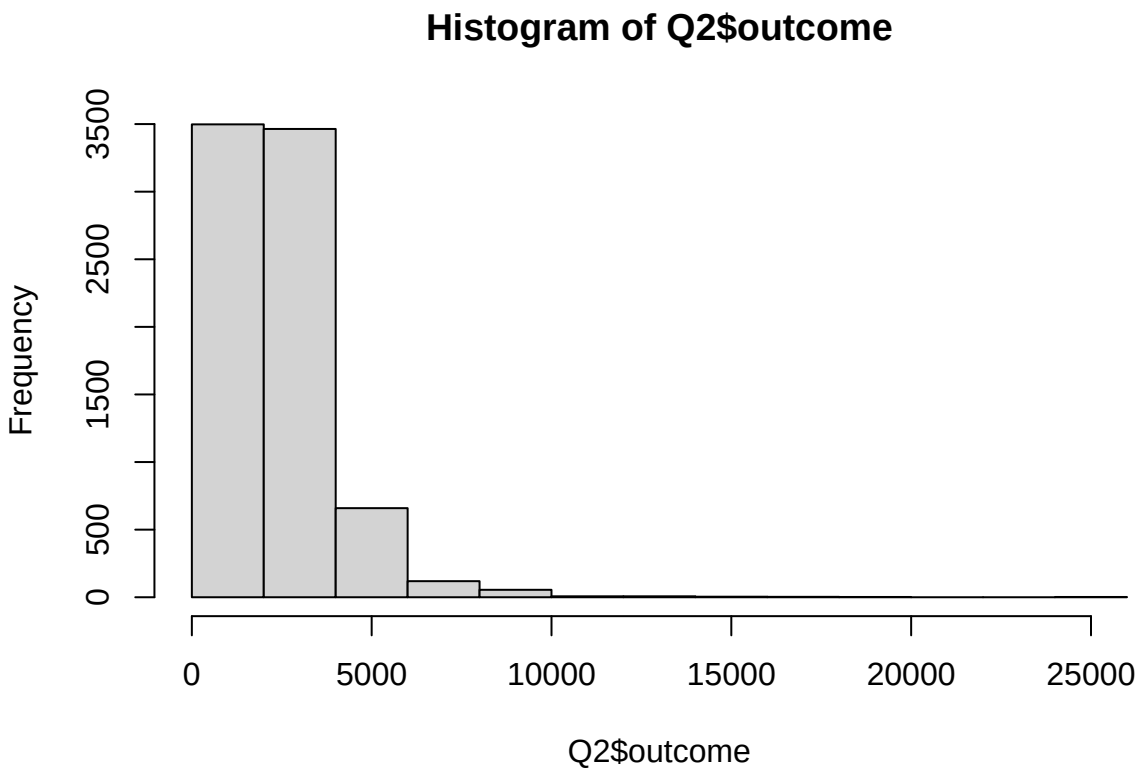


Figure 2-1: Histogram of the revenue per square foot per year

Then, I created two independent variables, GR and GR2. GR is a factor variable including four categories, "Not Green", "ES only", "LEED only", and "LEED and ES". These categories separately refer to no green certifications, certified by EnergyStar only, certified by LEED only, and certified by both. GR2 is a numeric variable, referring to the number of Green certifications the building has. Besides, other variables are cluster, size, empl_gr, stories, age, renovated, class_a, class_b, net, amenities, cd_total_07, hd_total07, total_dd_07, Precipitation, Gas_Costs, Electricity_Costs, and City_Market_Rent.

Before the GBM, I applied a grid searching to find the best hyperparameter combinations. I focused on the number of trees and interaction depth. For the number of trees, the grid searching includes 100, 1000, 2000, and 5000. For the interaction depth, the grid searching includes 1, 3, 7, 9, and 11. Another hyperparameter was set as consistent due to the laptop performance. Besides, I chose the Root Mean Square Error (RMSE) as the criteria of models' performance.

Results: The grid searching pointed out that when the number of trees is 1000 and the interaction depth is 11, the RMSE is the lowest (0.850). I used this combination for the final model. **Figure 3-2** presents the importance of the top ten variables. It indicates that the total available rental space, the building's local market, the age of the building, the height of the building, and the cluster occupied the top five significant roles in predicting the revenue per square foot per year of the building; however, neither GR nor GR2 is the top ten indicators in predicting the revenue per square foot per year of building. However, focus on the coefficient of GR2 (the number of Green certifications) indicates that one certification can be associated with an average 0.0572 increase in the revenue per square foot per year of buildings.

Conclusion: This study finds that the associations between whether the building has green certifications and the revenue per square foot per year are insignificant. Still, a green-certified building within a quarter-mile radius is significant. These findings deny the hypothesis that green certifications can directly promote the revenue of commercial buildings. Still, they prove that close to a green-certificated building, the commercial buildings can gain more revenue. I suggest that commercial rental property owners work together and ensure

a green-certificated building within a quarter-mile to maximize the revenue.

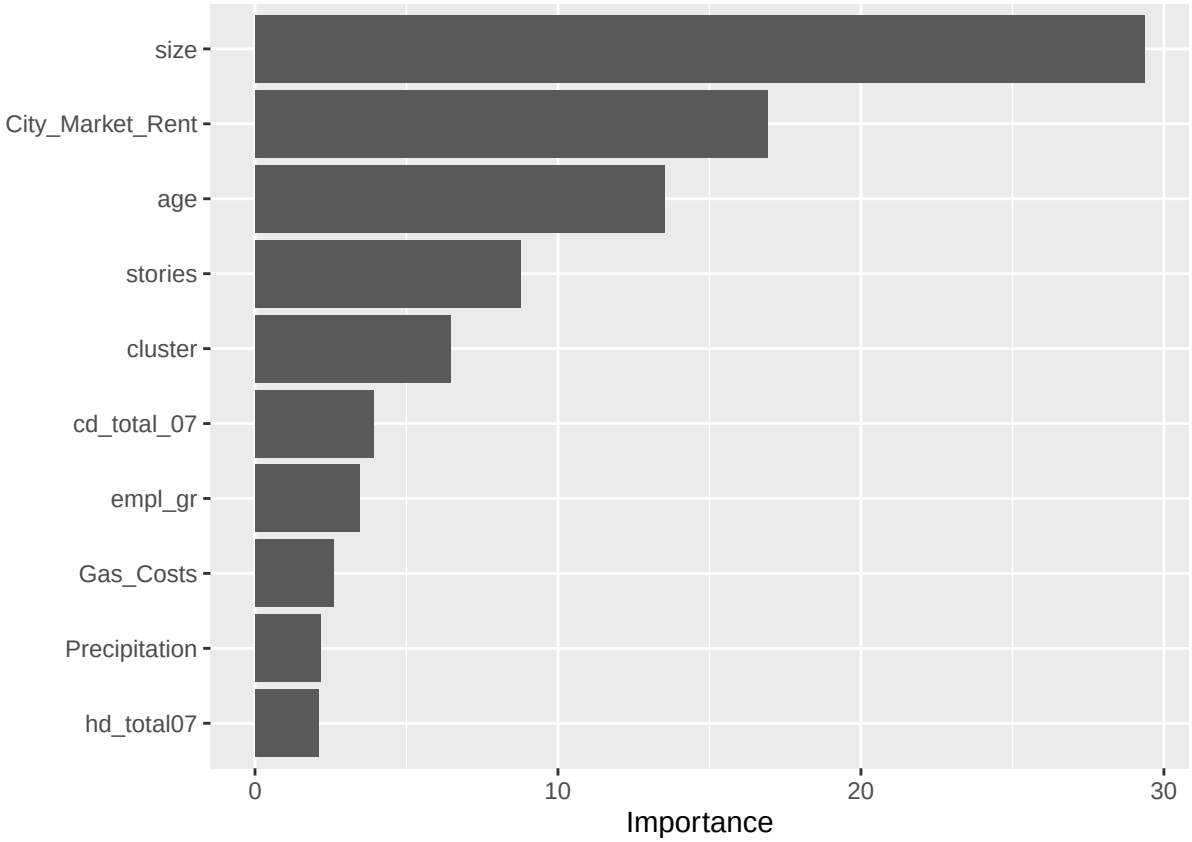


Figure 3-2: Top 10 indicators of the revenue per square foot per year

QUESTION 3

Build the best predictive model you can for median house value, using the other available features, and create three figures.

Answers:

Overview: This study explores variables affecting the median house values. I hypothesized that median household income plays a significant role in predicting the median household values.

Before introducing the data process, as a preannouncement, I planned to apply the ggmap package for data visualization but failed due to the google API. Hence, I used the ggplots for the three figures' visualizations.

Data and Model: I applied the Gradient Boosting Machines (GBM) for this analysis based on a dataset including median house values, location information (longitude and latitude), the median age of all residential households, total population, the total number of households, the total number of rooms, the total number of bedrooms, and median household income. Before analysis, I dropped all rows that included the NA value. I noticed a skewness in the distribution of the independent variable (**Figure 3-1**). I used the log-transformation to ease this effect.

The independent variables include location information (longitude and latitude), the median age of all residential households, total population, the total number of households, the total number of rooms, the

total number of bedrooms, and median household income. There was no preprocessing in independent variables.

I chose grid searching to determine the best hyperparameter combinations. I focused on the number of trees, interaction depth, and the ratios of test data and train data. For the number of trees, the grid searching includes 1, 10, 50, and 100. For the interaction depth, the grid searching includes 1, 3, 7, 9, and 11. For the ratios of train data and test data, it includes 0.6, 0.7, and 0.8. Besides, I chose the Root Mean Square Error (RMSE) as the criteria of models' performance.

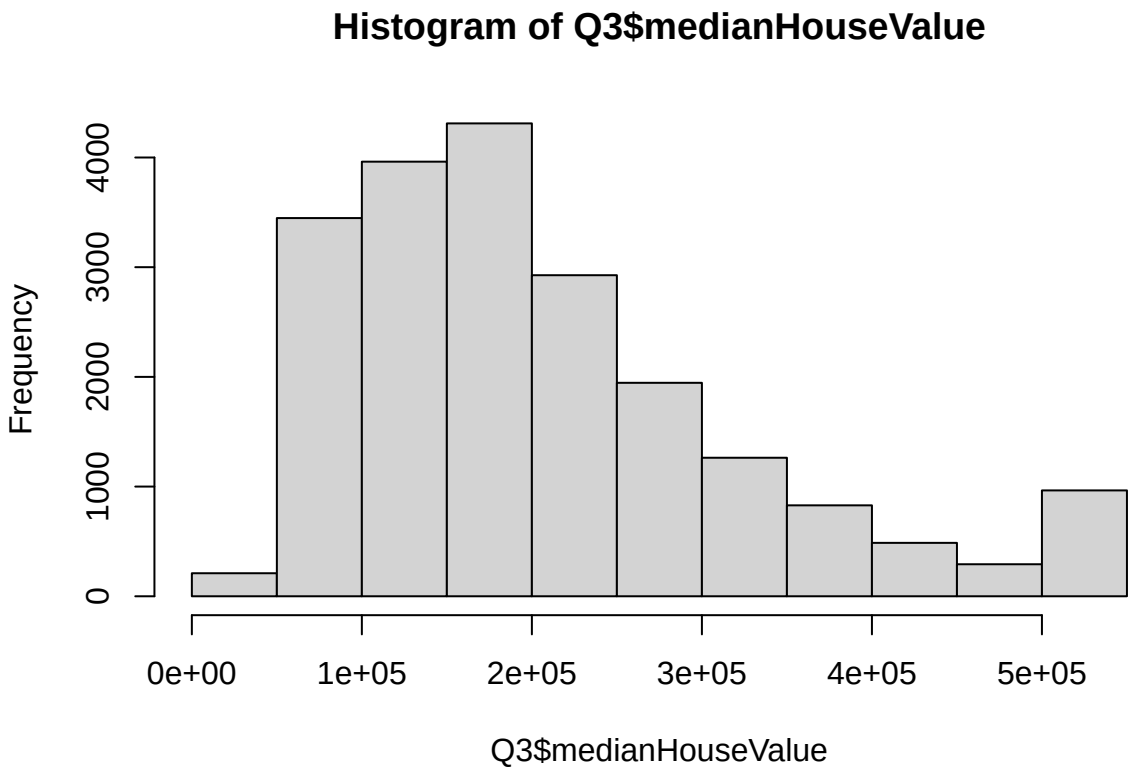


Figure 3-1: Histogram of median house value

Results: The grid searching pointed out that when the number of trees is 100, the interaction depth is 7, and the training ratio is 0.7, the RMSE is the lowest (0.332). I used this combination for the final model. **Figure 3-2** presents the importance of the top ten variables. The top five variables are median household income, location information (longitude and latitude), the median age of housing, and population. Focus on the effects of median household income, the coefficient is 0.658. It indicates that for one unit increase in median household income, the median house value in this region can gain 98.3% increases.

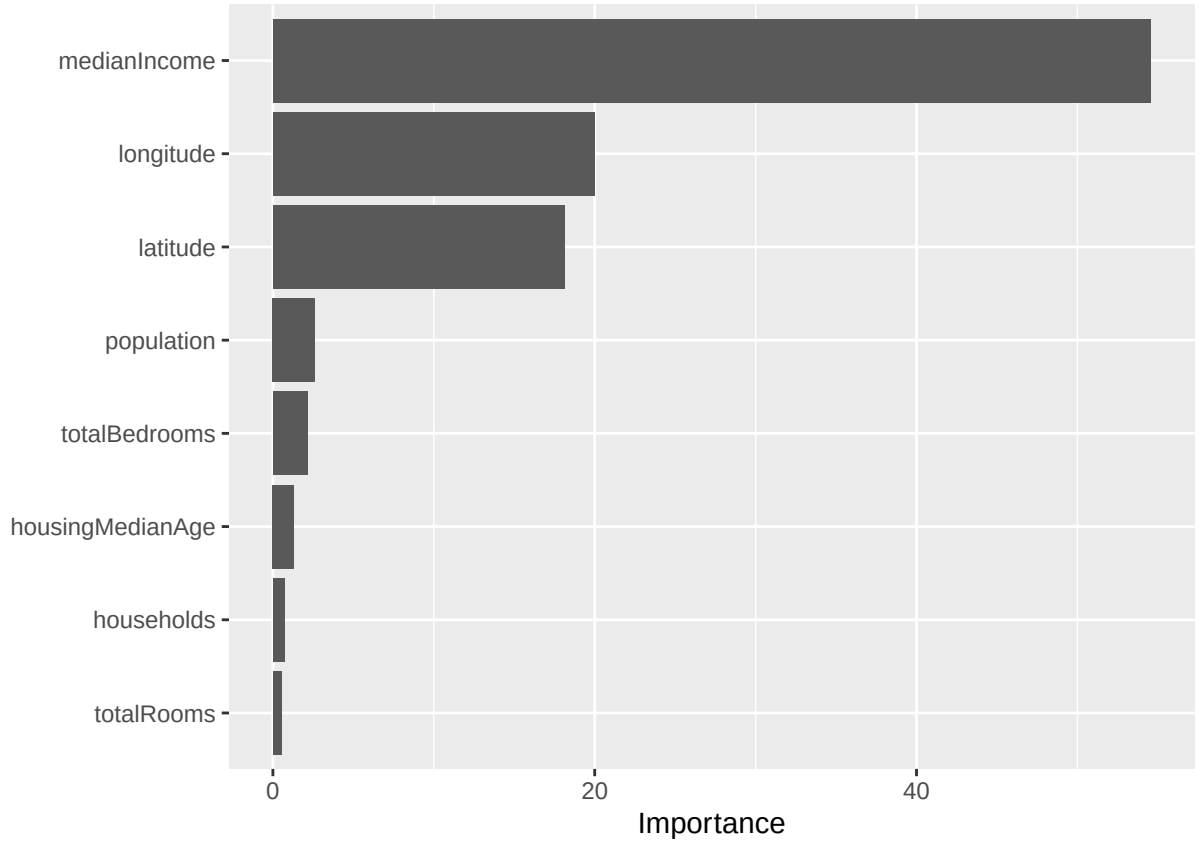
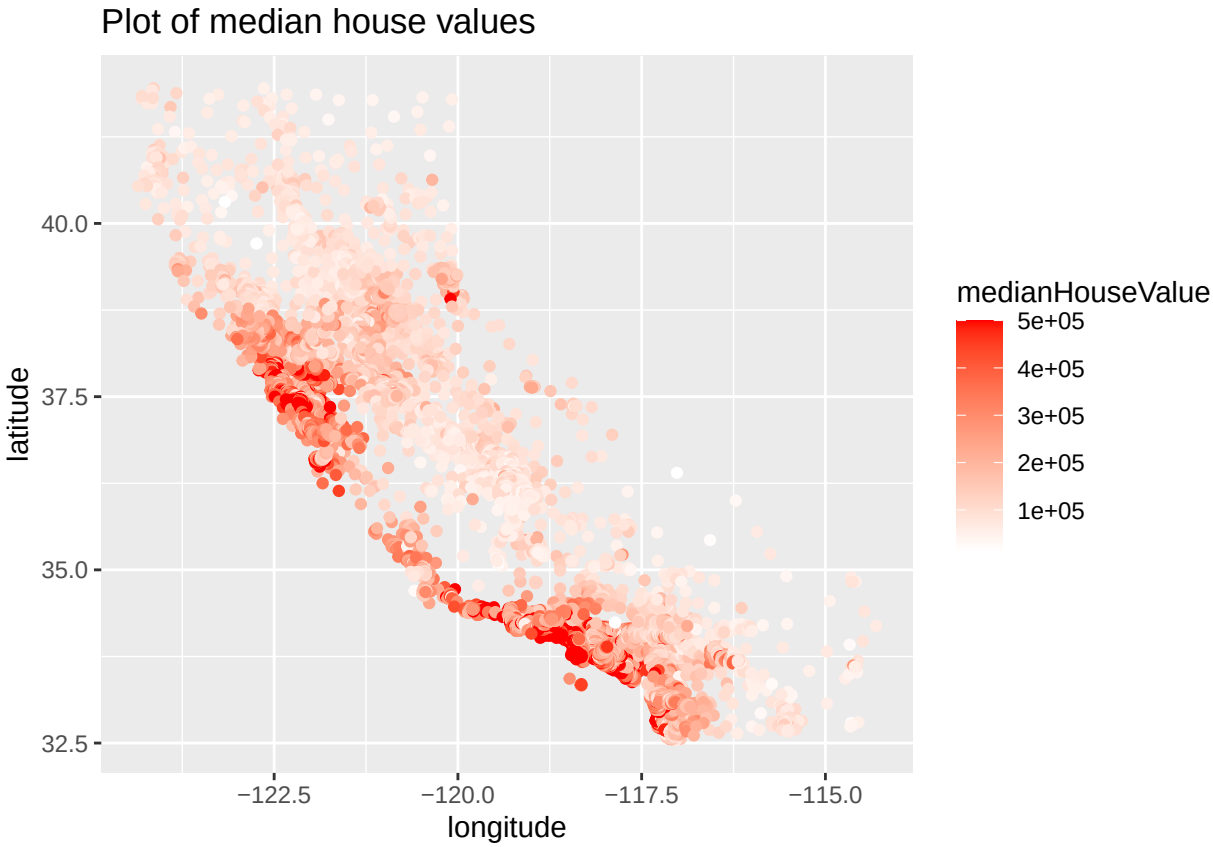
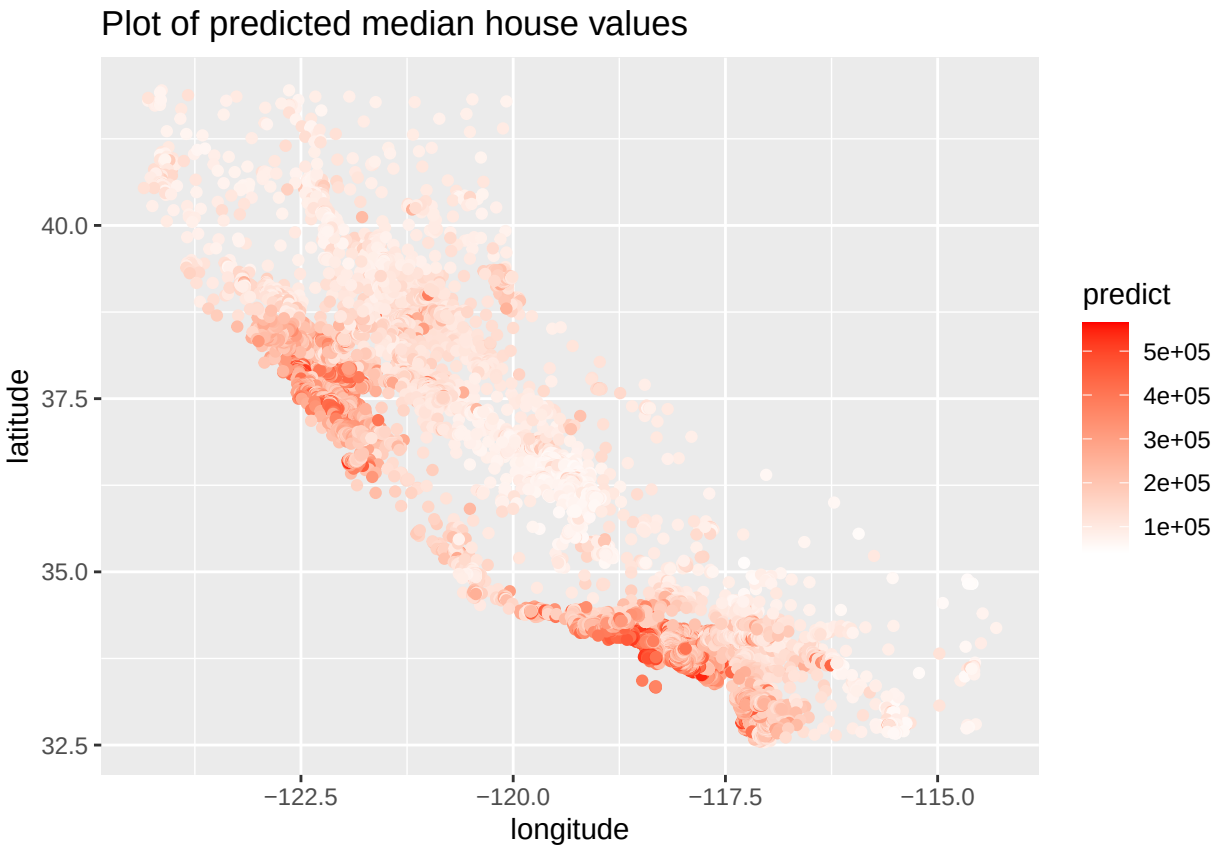


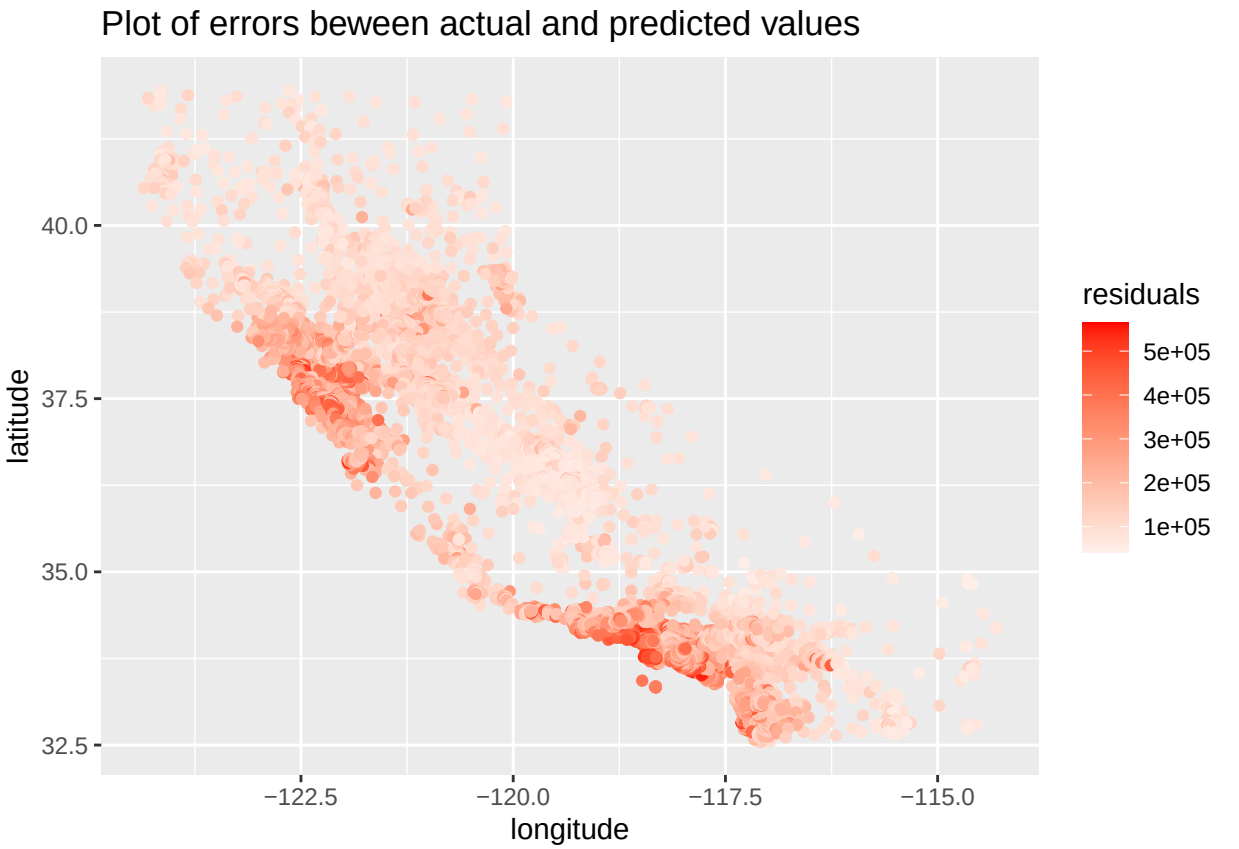
Figure 3-2: Top 10 indicators of median house values

Figure 3-3 demonstrate the spatial distribution of log(median house value), predicted log(median house value), and errors between values and predicted values. Not surprisingly, both actual value and predicted value maps point out that the median house values are relatively high in the Bay Area and Los Angeles metropolitan areas. The plot of errors indicates that the model predicts well in the Bay Area and Los Angeles metropolitan areas. Still, there are underestimations in the north of Los Angeles metropolitan areas and overestimaions in the hinterland areas.

Conclusion: This study identifies possible indicators of median house values in California. Findings indicate that the impact of median household income is the most significant and confirm that the median house values are relatively high in the Bay Area and Los Angeles metropolitan areas. Also, the performance of models is different in coastal areas and hinterland areas. I suggest researchers need to be aware of these differences and separately investigate house values indicators in coastal areas and hinterland areas.


Plot of median house values
40.0
37.5
35.0
32.5
latitude
−122.5
−120.0
−117.5
−115.0
longitude
medianHouseValue
5e+05
4e+05
3e+05
2e+05
1e+05


Plot of predicted median house values
latitude
40.0
37.5
35.0
32.5
−122.5
−120.0
−117.5
−115.0
longitude
predict
5e+05
4e+05
3e+05
2e+05
1e+05



Figures 3-3: Spatial distributions of actual median house values, predicted median house values, and error between actual and predicted.