ECO395M_Exercise3

Youngseok Yim (EID: yy9739)

2023-03-22

1. What causes what?

1.1 Cities with high crime rates may have deployed more police officers in an attempt to reduce crime. However, this correlation between crime rates and police presence can lead to false conclusions that more police leads to increased crime. Thus, it is not appropriate to simply analyze data from a few cities and run a regression between "Crime" and "Police."

1.2 The researchers aimed to determine the effect of increasing police presence on crime rates, controlling for factors unrelated to crime. During high alert days in Washington D.C., when the threat of terrorism is high, police increase their presence. The researchers evaluated the impact of this increased police presence on crime rates. Results from Table 2 showed that during high alert days, crime rates in Washington D.C. decreased by an average of 7 crimes per day, a statistically significant decline at the 5% level. The decline remained significant at 5% level even when controlling for the Log of midday metro ridership, with a decrease of 6 crimes per day.

1.3 It is possible that the decline in crime during high alert days is due to a decrease in the number of tourists, and therefore potential crime victims. To address this, the researchers controlled for midday metro ridership as a proxy for tourist numbers.

1.4 During high alert days, the average number of crimes in the first police district area decreased by 2.621 crimes per day, a statistically significant decline at the 1% level. In other districts, the average number of crimes decreased by .571 crimes per day, but this decline was not statistically significant.

2. Tree modeling: dengue cases

3. Predictive model building: green certification

I analyzed a data set on green buildings with the goal of constructing the best predictive model for forecasting revenue per square foot per calendar year. The process began with data cleaning, where I identified and removed all null values. Afterwards, I created the target variable, "revenue per square foot per calendar year," by multiplying the rent and leasing rate.

After creating a base model, I constructed five additional models to find the best predictive model. I compared their performance and chose the one with the highest accuracy. The base model excluded the interaction between rent and leasing rate as a feature variable, as it was deemed meaningless. Additionally, "CS_propertyID," the building's unique identifier, was removed as it did not contribute to the model. The variable "total_dd_07" was deleted due to its collinear relationship with "cd_total_07" and "hd_total07" (total_dd_07 = cd_total_07 + hd_total07). The impact of cluster on rent was considered to be reflected in "City_Market_Rent," the average rent per square-foot per calendar year in the building's local market, and so cluster was not included in the model. The model only considered "green_rating" and did not separate LEED and EnergyStar, resulting in their removal as well.

Forward Selection Model

Forward selection model starts with a model having no variables and add all possible one-variable additions to it, including every interaction. The model with the lowest AIC which we get from forward selection process

is:

revenue_persquarefoot ~ cluster + size + class_a + class_b + amenities + cd_total_07 + green_rating + age + hd_total07 + Electricity_Costs + net + cluster:size + amenities:green_rating + size:amenities + green_rating:age + size:Electricity_Costs + cluster:hd_total07 + cd_total_07:hd_total07 + hd_total07:Electricity_Costs + size:class_a + size:class_b + size:age + class_a:age + class_a:cd_total_07 + size:cd_total_07 + cluster:Electricity_Costs + cluster:age + age:Electricity_Costs + cd_total_07:Electricity_Costs + class_a:Electricity_Costs + amenities:Electricity_Costs + cd_total_07:net + class_b:amenities + size:green_rating

The AIC for this model is 108159.9 and the number of variables is 34.

Backward Selection Model

Backward selection model starts with the full model that has all the variables including all of interactions, then improves its performance by deleting each variable. The model with the lowest AIC we get from backward selection process is:

revenue_persquarefoot ~ size + empl_gr + stories + age + renovated + class_a + class_b + green_rating + net + amenities + cd_total_07 + hd_total07 + Precipitation + Gas_Costs + Electricity_Costs + cluster + size:empl_gr + size:stories + size:age + size:renovated + size:class_a + size:class_b + size:green_rating + size:cd_total_07 + size:hd_total07 + size:Electricity_Costs + size:cluster + empl_gr:stories + empl_gr:renovated + empl_gr:class_a + empl_gr:class_b + empl_gr:Gas_Costs + stories:age + stories:renovated + stories:class_b + stories:cd_total_07 + stories:Precipitation + age:class_a + age:green_rating + age:cd_total_07 + age:hd_total07 + age:Electricity_Costs + age:cluster + renovated:hd_total07 + renovated:Precipitation + renovated:cluster + class_a:amenities + class_a:hd_total07 + class_a:Precipitation + class_a:Gas_Costs + class_a:Electricity_Costs + class_b:hd_total07 + class_b:Precipitation + class_b:Gas_Costs + class_b:Electricity_Costs + green_rating:amenities + amenities:Precipitation + amenities:Gas_Costs + amenities:Electricity_Costs + cd_total_07:Precipitation + cd_total_07:Gas_Costs + cd_total_07:Electricity_Costs + hd_total07:Precipitation + hd_total07:Gas_Costs + hd_total07:Electricity_Costs + Precipitation:cluster + Gas_Costs:cluster + Electricity_Costs:cluster

The AIC for this model is 108044 and the number of variables is 68.

Stepwise selection Model Stepwise selection model starts with our base model lm(revenue_persquarefoot ~ . - Rent - leasing_rate - CS_PropertyID - cluster - LEED - Energystar - total_dd_07)' and we considered all possible one-variable additions or deletions including interactions. The model with the lowest AIC we get from stepwise selection model is:

revenue_persquarefoot ~ size + empl_gr + stories + age + renovated + class_a + class_b + green_rating + net + amenities + cd_total_07 + hd_total07 + Precipitation + Gas_Costs + Electricity_Costs + cluster + size:cluster + stories:class_a + size:Precipitation + empl_gr:Electricity_Costs + green_rating:amenities + Precipitation:cluster + hd_total07:Precipitation + amenities:Gas_Costs + amenities:Precipitation + stories:Gas_Costs + renovated:Precipitation + size:age + cd_total_07:Precipitation + stories:class_b + age:green_rating + class_a:Gas_Costs + class_a:Electricity_Costs + age:cluster + age:Electricity_Costs + renovated:cluster + Electricity_Costs:cluster + cd_total_07:hd_total07 + age:class_a + renovated:hd_total07 + class_a:Precipitation + stories:renovated + size:renovated + size:Electricity_Costs + size:stories + size:hd_total07 + class_a:hd_total07 + empl_gr:renovated + age:hd_total07 + amenities:Electricity_Costs + class_a:amenities + renovated:Gas_Costs + size:green_rating

The AIC for this models is 108070.3 and the number of variables is 53.

The backward selection model has the lowest AIC when compared to the other three models, making it the best performing model according to AIC. Additionally, I conducted k-fold cross-validation to compare the performance of the four models (base, forward selection, backward selection, and step-wise selection). The average root mean squared error (RMSE) calculated using 10-fold cross-validation was 1037.316 for the base model, 1006.796 for the forward selection model, 1006.068 for the backward selection model, and 1006.601

for the step-wise selection model. The lowest RMSE, belonging to the backward selection model, further confirms that it is the best among the four.

RMSE from k-folds cross validation

Baseline model : Mean RMSE

```
## [1] 1038.912
```

Forward selection model: Mean RMSE

```
## [1] 1227.271
```

Backward selection model: Mean RMSE

```
## [1] 1186.186
```

Stepwise selection model: Mean RMSE

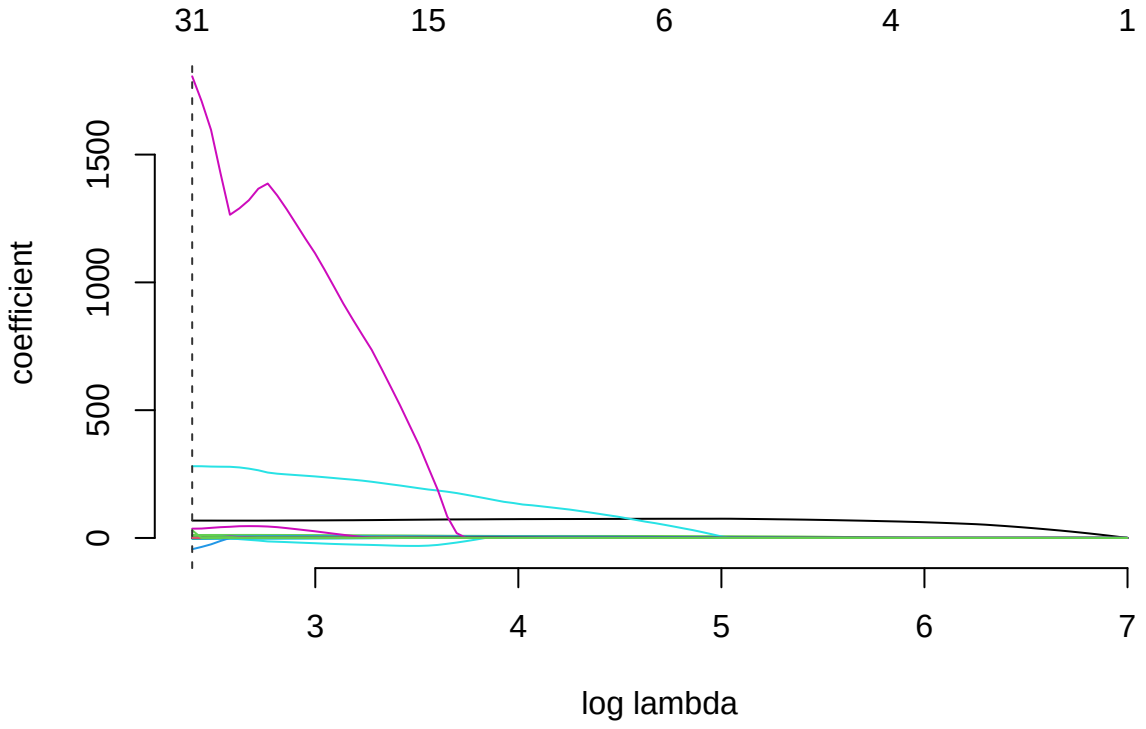
```
## [1] 1199.553
```

Lasso Regression

I then applied lasso regression to determine if it could outperform the best model obtained from backward selection. The full model, including all variables and two-way interactions, was used for lasso. The resulting path plot from running the lasso regression is shown below.

Figure 3.1 Path plot of lasso regression

```
## Warning: as(<dsyMatrix>, "dspMatrix") is deprecated since Matrix 1.5-0; do as(.,
## "packedMatrix") instead
```




```
##   seg100
## 2.393814
```

```
## [1] 31
```

The optimal value of lambda in a log scale is 2.39. The lowest AIC value is 108547.3 and the corresponding number of variables is 31 including the intercept.

Lasso : Mean RMSE

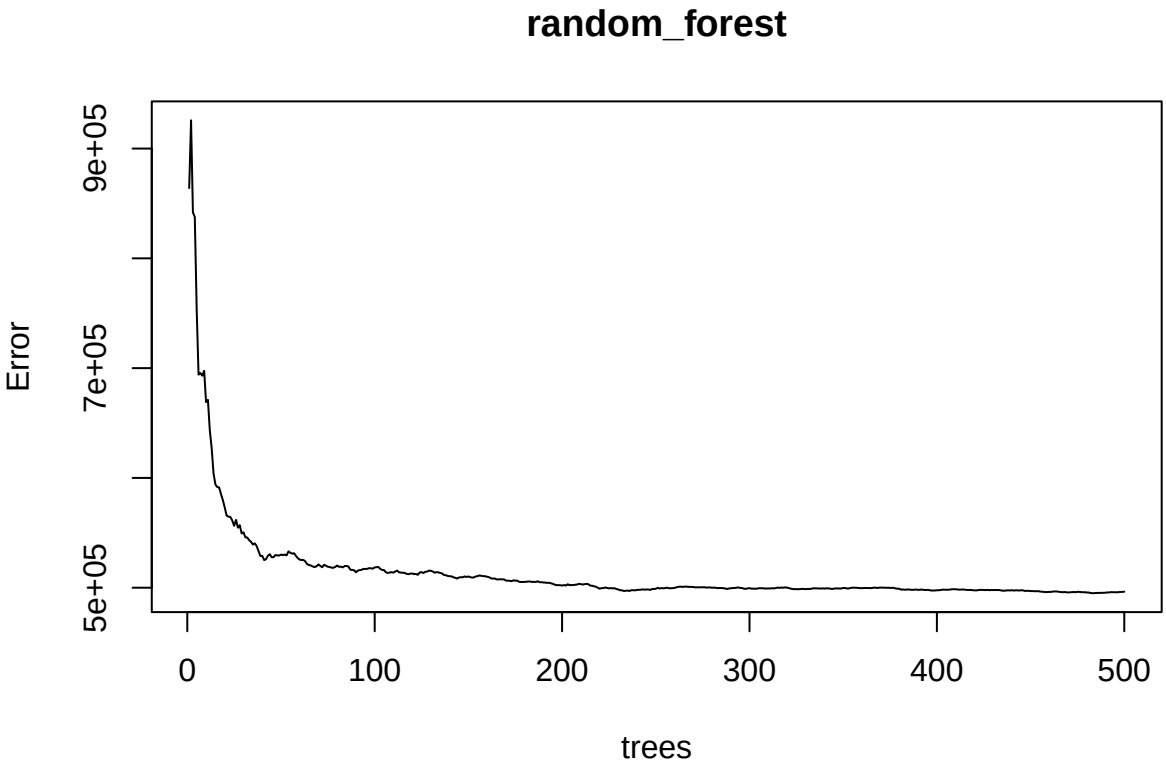
```
## [1] 1300.427
```

I performed k-fold cross validation on the lasso regression model. The results showed that the RMSE for lasso regression was higher compared to any of the models derived from step-wise selection, indicating that the step-wise selection model performs better than lasso regression in this case.

Random Forest

Lastly, I applied a random forest model using the base model with 500 trees. Figure 3.3 shows that 500 trees were sufficient to reduce errors.

Figure 3.3 Out of bag MSE as a function of number of trees



Random Forest : Mean RMSE

```
## [1] 697.7968
```

The RMSE from the k-fold cross-validation for the Random Forest model was lower than the RMSE of any of the models we used above. Therefore, we can conclude that the model derived from the Random Forest performs the best. To determine the average change in rental income per square foot per calendar year associated with green certification while holding other building features constant, we used the 'partial' function in the 'pdp package'.

```
##   green_rating     yhat
## 1            0 2403.894
```

```
## 2            1 2471.181
```

The change in rental income per square foot per calendar year resulting from green certification, while keeping all other building features constant, can be determined by finding the difference between the predicted value (yhat) when green_rating is 1 and the predicted value (yhat) when green_rating is 0.

4. Predictive model building: California housing

For this exercise, my goal was to create the most accurate model for predicting the median market value of houses in a specific census tract. I began with a baseline linear regression model that included all relevant variables without any interactions. Next, I developed two additional models: a Random Forest regression model and a boosting model. These models were designed to improve upon the baseline model and provide more accurate predictions.

Baseline linear regression model

medianHouseValue ~ longitude + latitude + housingMedianAge + population + households + totalRooms + totalBedrooms + medianIncome

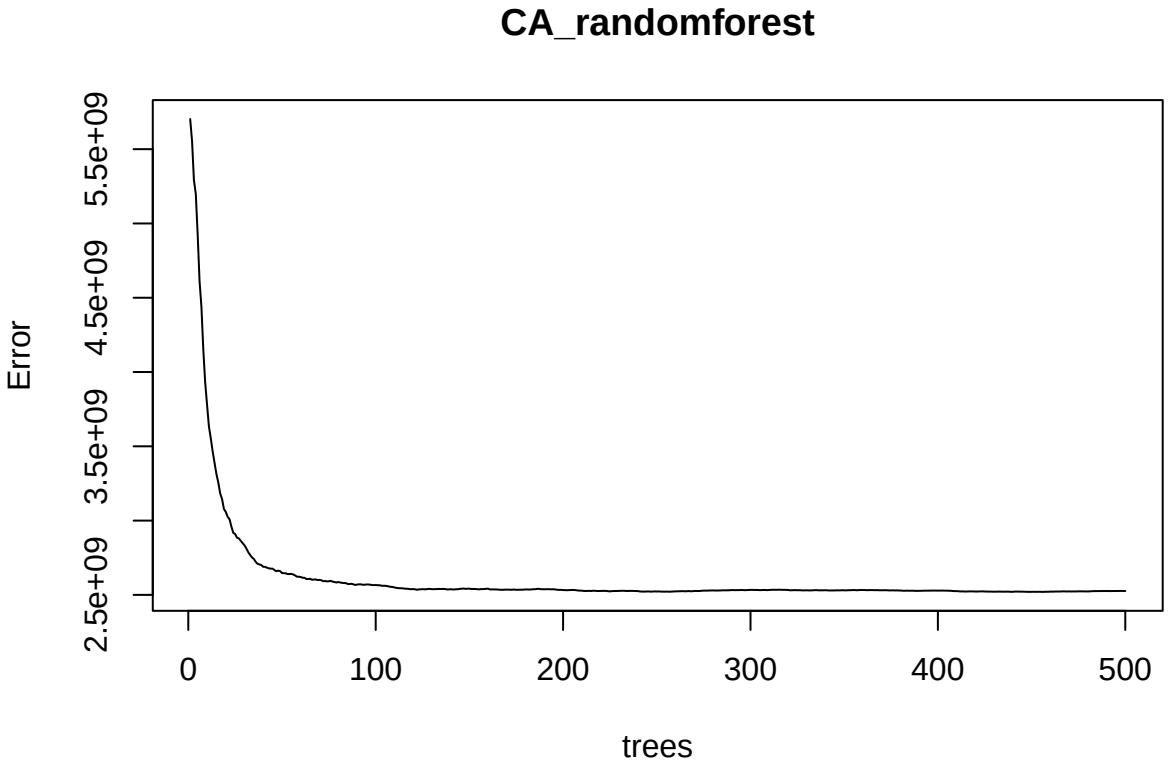
Baseline model : Mean RMSE

```
## [1] 69638.95
```

Random Forest model

I fitted a random forest model based on the base model. By examining the plot below, which displays the out-of-bag mean squared error (MSE) as a function of the number of trees, I determined that using 500 trees is sufficient to reduce errors. Therefore, I selected to use 500 trees for the model.

Figure 4.1 Out of bag MSE as a function of number of trees



Random Forest model : Mean RMSE

```
## [1] 50574.86
```

Boosting model

Finally, I fitted a boosting model, also starting with the base model, as was done with the Random Forest model. The root mean squared error (RMSE) from the k-fold cross-validation was slightly higher than that of the Random Forest model. Specifically, it was recorded as below:

Boosting model : Mean RMSE

```
## [1] 51823.46
```

Since the Random Forest model achieved the lowest RMSE value during k-fold cross-validation, it was selected for prediction purposes.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

Figure 4.2 Mean house value across various latitudes and longitudes in California

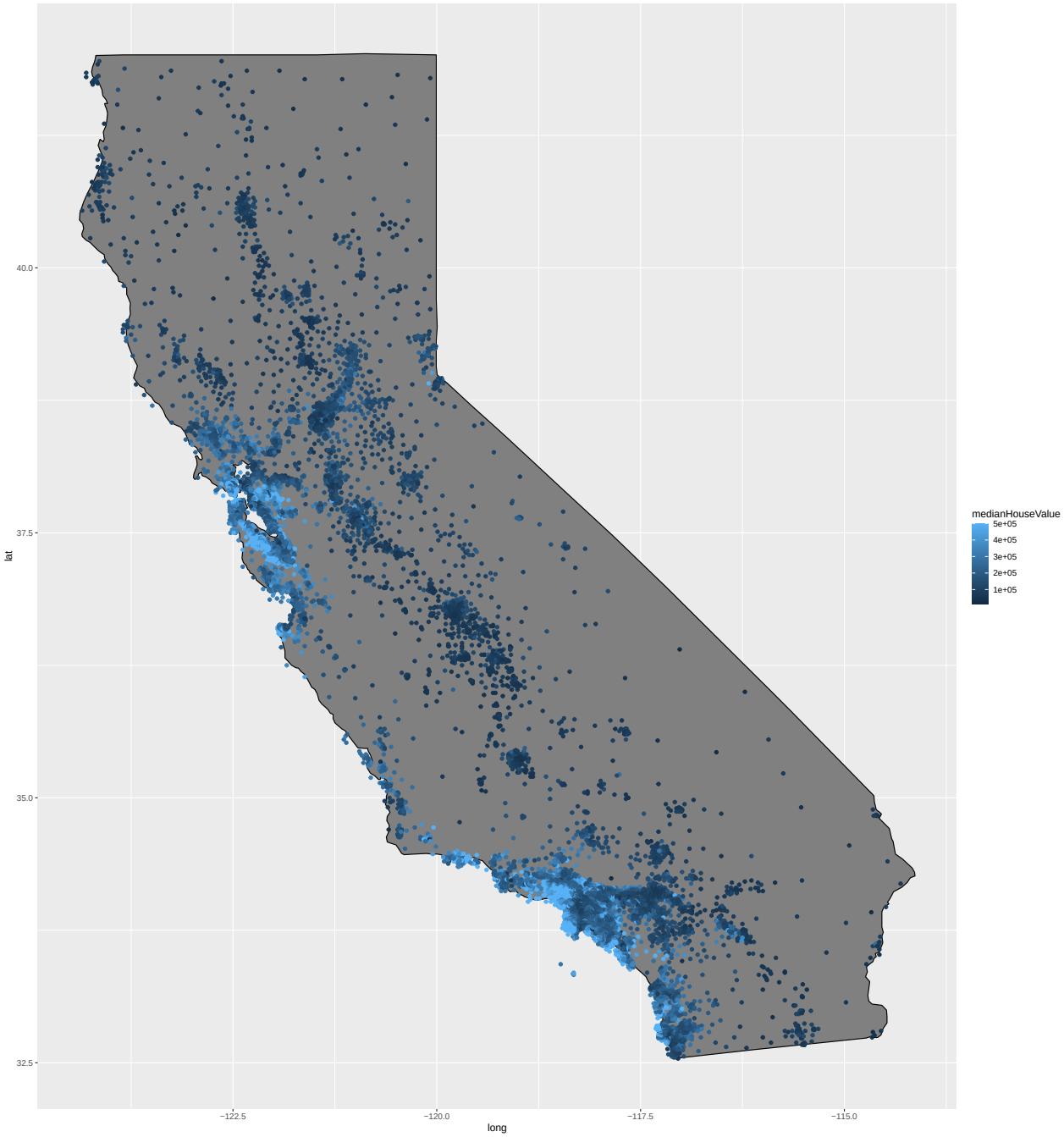


Figure 4.3 Prediction of Mean house value across various latitudes and longitudes in California using Random Forest

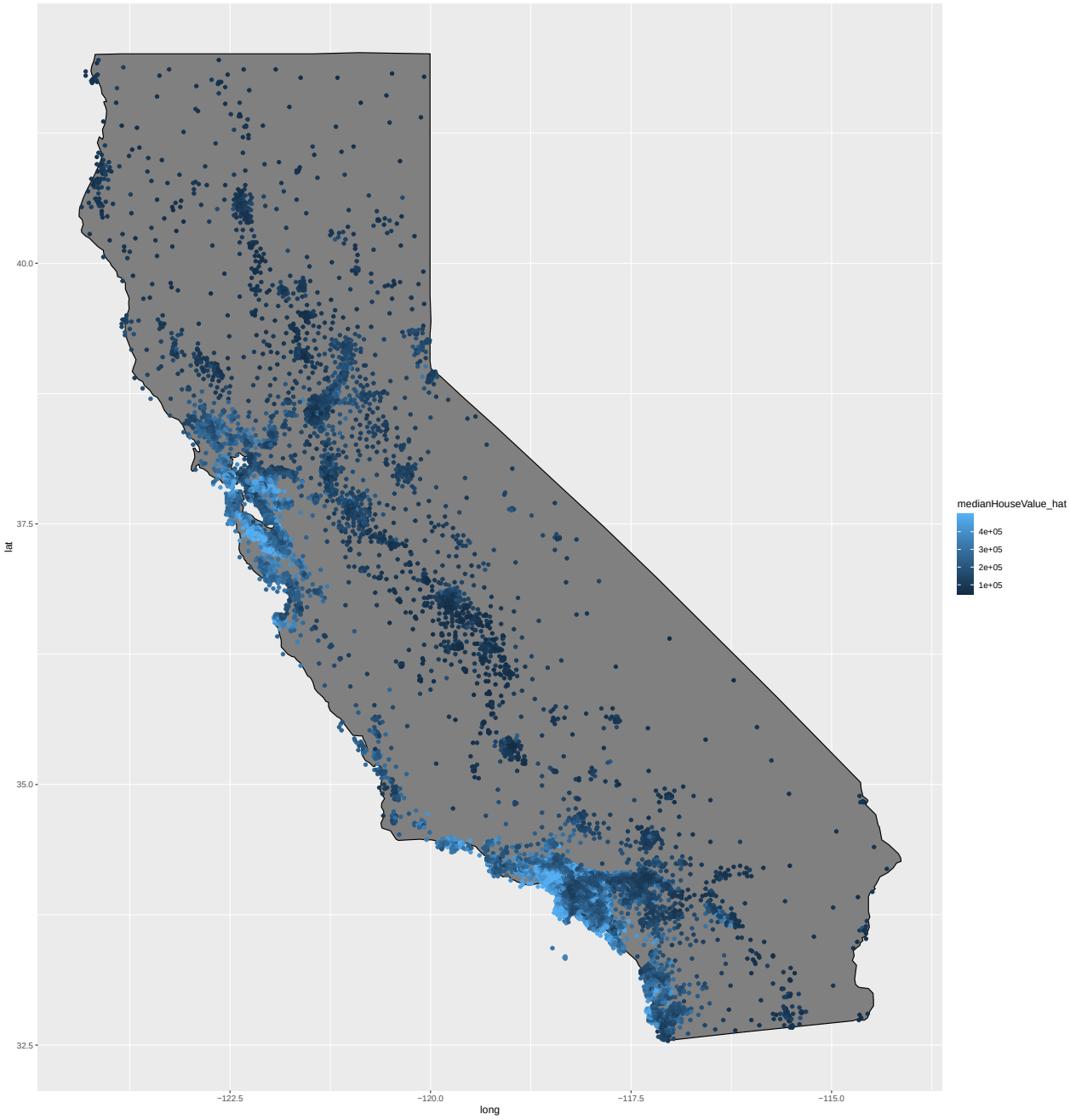


Figure 4.4 Residuals from prediction of mean house value using Random Forest