(F23) PSTAT 126: Project Step 4

Anthony Cu and William Mahnke

13 December 2023

Introduction

In our project, we explore the 1990 California Housing data set, providing information on a specified district in the state. The survey data describes homes in a district of California in 1990 in order to represent the larger population of this district during the 1990's overall. Each independent observation corresponds to a different block within the district. The data set comes from Kaggle's data repository.

We view the correlations between our explanatory variables.

```
                    housing_median_age total_rooms total_bedrooms population
housing_median_age                1.00       -0.39          -0.36      -0.32
total_rooms                      -0.39        1.00           0.95       0.88
total_bedrooms                   -0.36        0.95           1.00       0.92
population                       -0.32        0.88           0.92       1.00
households                       -0.34        0.94           0.99       0.94
median_income                    -0.09        0.22           0.03       0.01
median_house_value                0.13        0.19           0.11       0.02
                    households median_income median_house_value
housing_median_age       -0.34         -0.09                0.13
total_rooms               0.94          0.22                0.19
total_bedrooms            0.99          0.03                0.11
population                0.94          0.01                0.02
households                1.00          0.05                0.13
median_income             0.05          1.00                0.71
median_house_value        0.13          0.71                1.00
```

We begin by fitting a naive linear model to our dataset, by inputting all our variables of interest linearly.

```
fit <- lm(median_house_value ~., data = houseData[-c(1:3, 12)])
```

From this fitted model, we observe that $R^2 = 0.7026573$, which is moderately large.

We now check the eigen decomposition of $x^T x$

```
[1] 0.2652414 0.9405006 0.9810497 0.9003862 0.9829647 0.4722287 0.3408343
[8] 0.1837948 0.1124782
```

We obtain that $\sqrt{\frac{\lambda_1}{\lambda_p}}$ for each predictor is:

```
[1]     1.000000     7.819039    28.563437    90.229077   150.686255
[6]  2090.657990  7533.561862 11842.263618 14749.680428
```

Further, the R_j^2 for all predictors is:

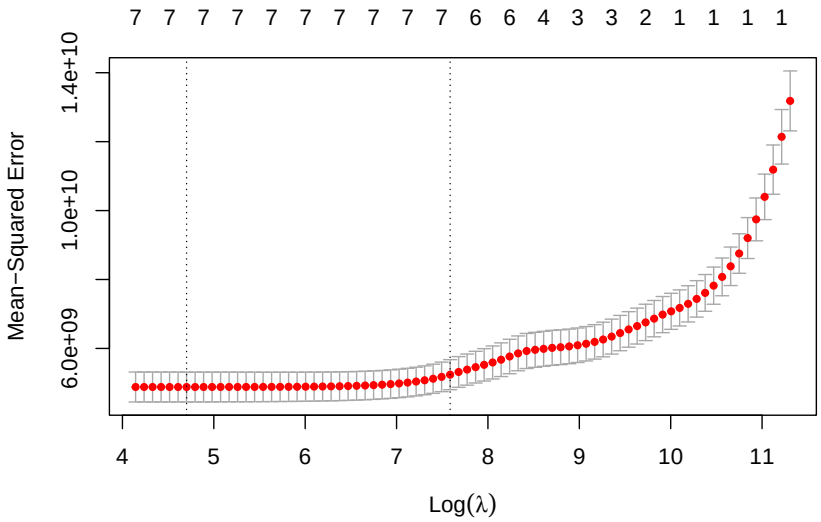
```
[1] 0.2652414 0.9405006 0.9810497 0.9003862 0.9829647 0.4722287 0.3408343
[8] 0.1837948 0.1124782
```

We now check the variance inflation factors:

housing_median_age	total_rooms	total_bedrooms
1.360991	16.806897	52.769514
population	households	median_income
10.038768	58.701809	1.894760
ocean_proximityINLAND	ocean_proximityNEAR BAY	ocean_proximityNEAR OCEAN
1.517069	1.225182	1.126733

We observe $\sqrt{\frac{\lambda_1}{\lambda_p}} \geq 30$ which indicates that there's collinearity in the variables. The variance inflation factor for some of the variables also indicates the presence of collinearity. Thus, ridge and lasso regression will be effective techniques to help us understand our data.

Finding best lambda

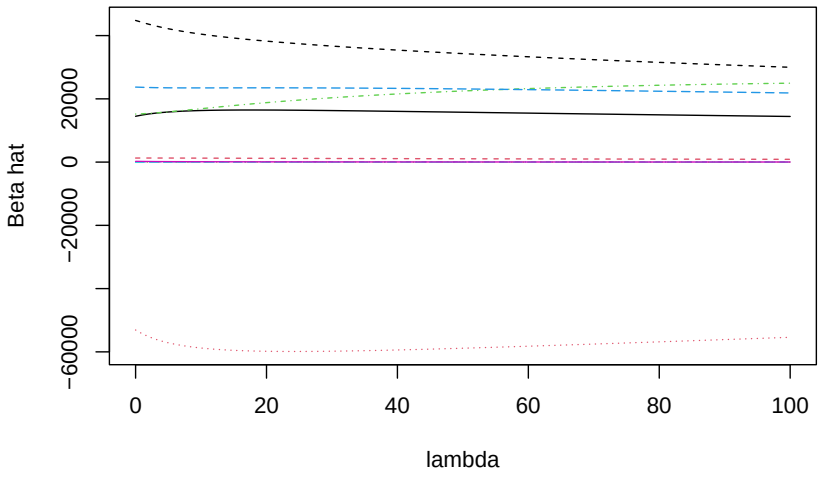


Our plot shows the test MSE by λ value. Using cross-validation, we obtain that our best λ value is 110.1849793.

Ridge Regression

We use the `MASS` package to perform ridge regression. Our regression coefficients (after normalization) should not be very large, so that we should bound/restrict the size of the coefficients (shrinkage).

We plot the Ridge Regression model coefficients for each value of λ.

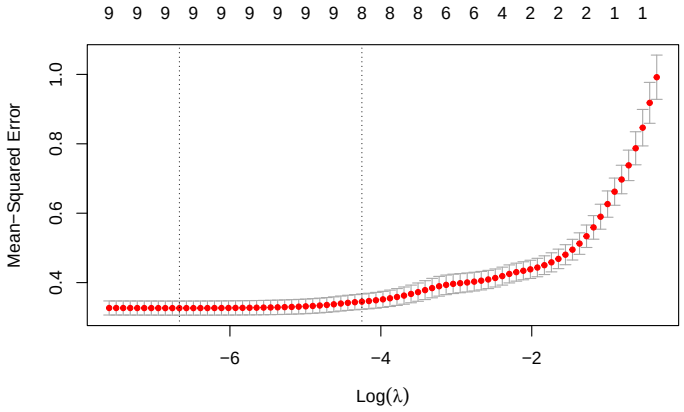


Using ridge regression, we obtain the coefficients for a fitted linear model:

	housing_median_age	total_rooms
14899.81360	1287.20343	-20.47217
total_bedrooms	population	households
121.06277	-66.08796	200.98956
median_income	ocean_proximityINLAND	ocean_proximityNEAR BAY
44171.97970	-54273.96676	15277.64940
ocean_proximityNEAR OCEAN		
23643.57523		

LASSO Regression

We proceed with performing a LASSO regression. Using the `glmnet` package, we find a best λ value to fit our model using Lasso Regression.



Using Lasso regression, we obtain the coefficients for a fitted linear model:

```
11 x 1 sparse Matrix of class "dgCMatrix"
                                    s0
(Intercept)                 0.1294886
housing_median_age          0.1438327
total_rooms                -0.3735547
total_bedrooms              0.4089639
population                 -0.6400794
households                  0.7117889
median_income               0.7032746
ocean_proximity<1H OCEAN    .
ocean_proximityINLAND      -0.4692958
ocean_proximityNEAR BAY     0.1284347
ocean_proximityNEAR OCEAN   0.2017338
```

We obtain a λ value of 0.0012685. Further, `ocean_proximity<1H OCEAN` is not shown as a coefficient because the lasso regression shrunk it all the way to zero. This means it was completely dropped from the model because it wasn't influential enough. Ridge regression shrinks all coefficients towards zero, but lasso regression has the potential to remove predictors from the model by shrinking the coefficients completely to zero.

MLR, RR, LASSO Visualization

We construct a single graph that superimposes the three different predictions. We reintroduce the MLR model that we deduced from *Project Step 3*, using backward selection

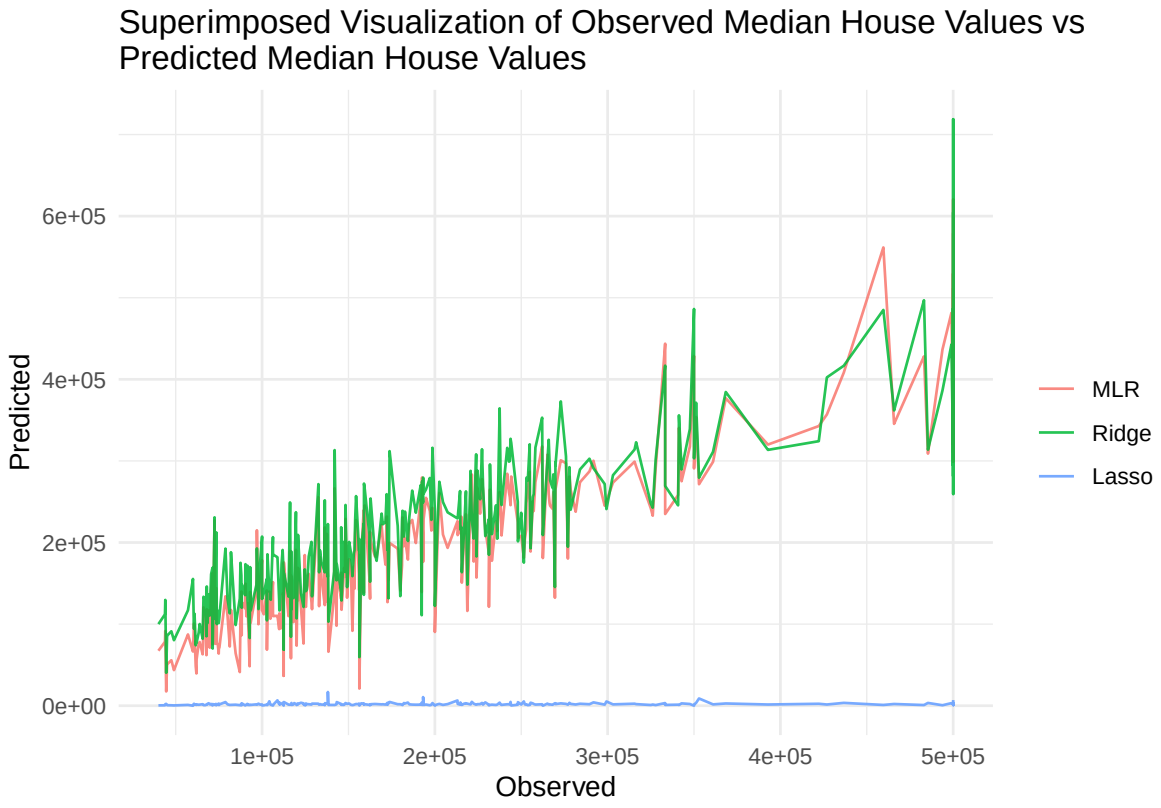
```
fitMLR <- lm(median_house_value ~ housing_median_age + total_rooms +
    total_bedrooms + population + households + median_income +
    ocean_proximity + housing_median_age:total_rooms + housing_median_age:population +
    housing_median_age:households + total_rooms:total_bedrooms +
    total_rooms:population + total_bedrooms:population + total_bedrooms:households +
    total_bedrooms:median_income + population:median_income +
    median_income:ocean_proximity, data = houseData[, -c(1:3, 12)])
```

The coefficients for the fitted MLR model are:

```
                          (Intercept)                   housing_median_age
                        -8.743954e+03                         1.723569e+03
                          total_rooms                      total_bedrooms
                         7.009904e-01                         6.792427e+01
                           population                         households
                         6.076805e+01                        -1.203920e+02
                        median_income                ocean_proximityINLAND
                         3.915215e+04                        -8.503690e+04
              ocean_proximityNEAR BAY            ocean_proximityNEAR OCEAN
                        -8.347252e+04                        -9.142805e+04
      housing_median_age:total_rooms          housing_median_age:population
                        -1.366026e+00                        -4.273274e+00
       housing_median_age:households             total_rooms:total_bedrooms
                         1.789022e+01                         4.540968e-03
              total_rooms:population              total_bedrooms:population
                         4.001061e-03                         4.200561e-02
          total_bedrooms:households           total_bedrooms:median_income
```

```
                          -2.303486e-01                                 5.022440e+01
      population:median_income     median_income:ocean_proximityINLAND
                          -1.945786e+01                                 6.348165e+03
median_income:ocean_proximityNEAR BAY median_income:ocean_proximityNEAR OCEAN
                           1.865680e+04                                 2.895120e+04
```

We create a visualization with the observed median house values on our x-axis, and predicted mean house values on the y-axis. We superimpose the three different predicted values that we yield: our MLR model, Ridge Regression model, and Lasso Regression model.



Looking at the graph of the three different model predictions superimposed, we see that the MLR and Ridge regression models are very similar while the Lasso regression model's predictions are significantly smaller than both of the other models. This is further reinforced by the comparison of the coefficients for the models, where it's evident that the coefficients in the lasso model are significantly smaller than its ridge regression counterparts. Additionally, we can see the similarity between the MLR and Ridge Regression reflected in the comparison between their coefficients. The distinction of the Lasso regression (apart from the other two models) may be caused due to inadequate scaling of the dataset.

Innovation: Random Forests

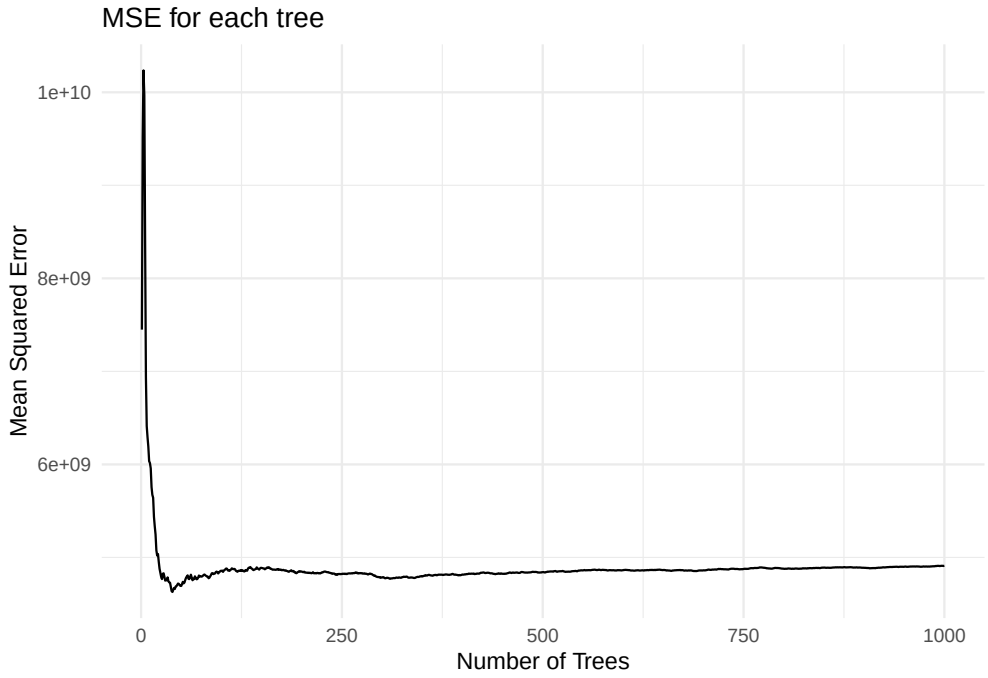
The technique we chose to learn for the project was using random forests to predict observation values. Random forests are appropriate when it comes to predicting observation values, especially when the observation variable is continuous. Additionally, random forests have high accuracy, often better than linear regression models.

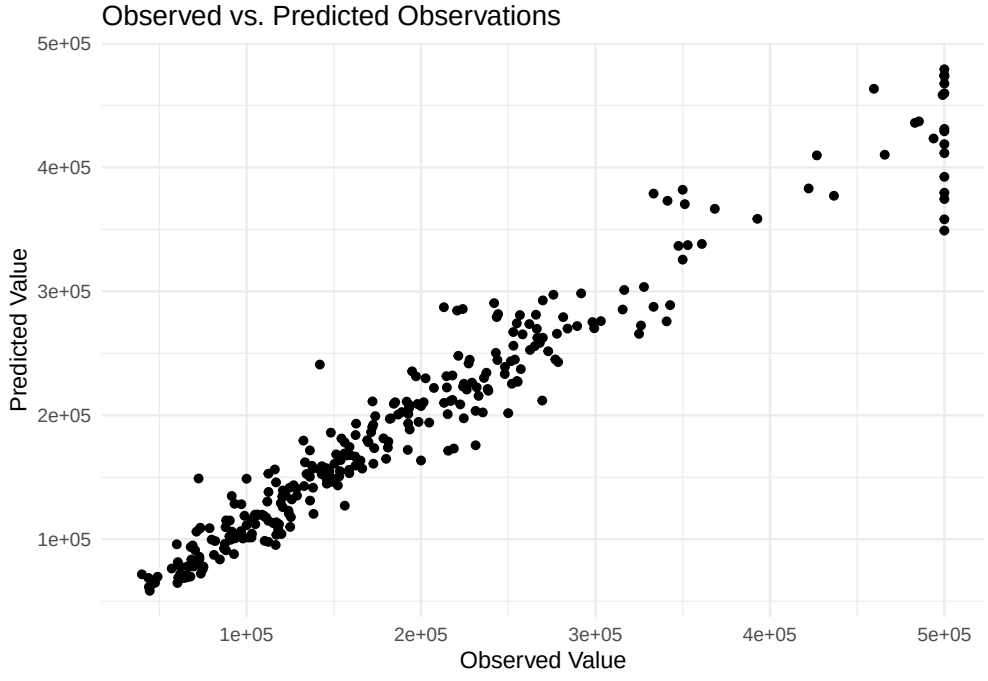
Random forests consist of a large quantity of decision trees that use a specified number of random vectors and a specified number of features from the data. Decision trees are a tree structure with levels of nodes

to determine an estimate of the observation using the nodes and random vectors (decisions in the tree are made as a result of evaluating the criterion at each node). Each tree generates an output, and the average of the outputs is returned as the final output or predicted value for the observation.

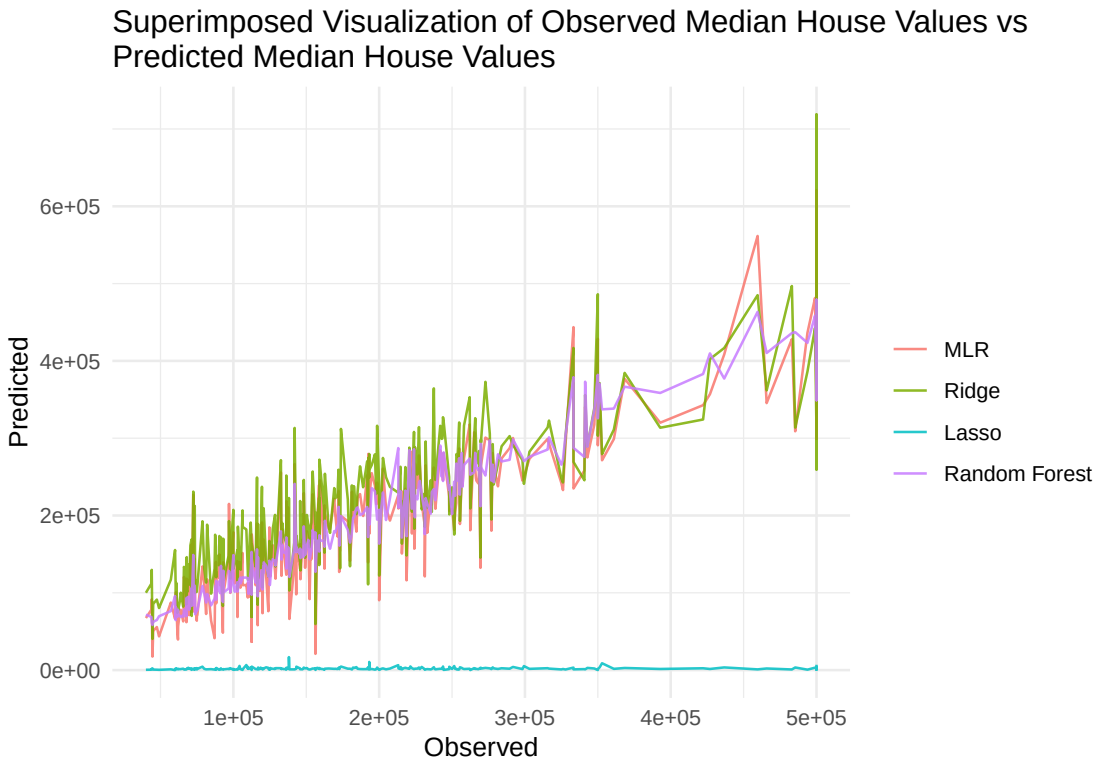
Random forests are beneficial in analysis because of their general ease-of-use and lack of technical conditions required for the model. We found very little pushback when making the model, so we didn't have to react to conditions being violated.

Using a random forest on our variables of interest, we graphed the error of each tree and a comparison of the observed value from the original data and the predicted value for the forest.





We visualize our random forest onto our superimposed visualization from before:



We observe that the Random Model produces similar predicted values as do the MLR backwards selection and ridge regression.

Conclusion

We conclude that lasso regression was less effective than our backwards MLR selection, ridge regression, and random forest models, for our dataset of California housing. If we were to repeat this analysis, we would try to find data where the number of predictors was larger than the number of observations so that lasso regression could be properly utilized and appreciated. Additionally, we would further explore how to change the random forest model to improve predictive accuracy and analyze other statistics about the random forest model.