Valuing Livability: Analyzing Attributes Influencing Housing Prices in Boston

Ruoxi Zhang, Zhisong Ye, Zuojun Zhang, Chenxiao Song

Summary: This report explores the relationship between housing attributes and market prices in Boston, using the 1970 Boston Housing dataset. The study addresses two key objectives. *For government policymakers*, the analysis aims to understand the factors that residents prioritize in their living conditions. It identifies key variables that significantly impact housing prices, helping policymakers prioritize public service improvements. The findings suggest that variables such as the proportion of lower-status population, accessibility, air quality, and crime rates are crucial considerations for residents in Boston. *For industry practitioners*, the report provides insights into accurate prediction models for housing prices. It compares the performance of linear regression models (full and reduced) with a random forest model. The strengths and weaknesses of linear regression models are stated under acknowledgement of the better predictive accuracy of the random forest model. *In conclusion*, this analysis offers valuable guidance for both government officials and industry professionals, helping them make informed decisions regarding housing policies, public service improvements, and accurate housing price predictions.

1 Case Background

The Boston Housing dataset contains information concerning owner-occupied housing for 506 census tracts in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. It was known to be developed by Harrison and Rubinfeld (1979) to illustrate the issues with using housing market data to measure consumer willingness to pay for clean air.

In the following analysis, we use the data with 506 observations on 13 variables (droping the categorical variable *chas*). The description of the variables used in the analysis could be found below.

Variables	**Description**
$crim$	per capita crime rate by town
zn	proportion of a town's residential land zoned for lots greater than 25,000 sq.ft
$indus$	proportion of non-retail business acres per town
tax	full-value property-tax rate per USD 10,000
$ptratio$	pupil-teacher ratio by town
b	$b = 1000(B - 0.63)^2$, where B is the proportion of blacks by town
$lstat$	percentage of lower status of the population
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
$medv$	median value of owner-occupied homes in the census tract (USD 1000's)

Table 1: Variables in the BostonHousing Dataset

medv, median value of owner-occupied homes in the census tract (USD 1000's), is treated as the response variable. For the 12 covariates, in reference to the original paper by Harrison and Rubinfeld (1979), we divided them into the following groups based on their indication in real life:

Covariates related to the **neighborhood**: $crim$, zn, $indus$, tax, $ptratio$, b, $lstat$

Covariates related to **pollution**: nox

Covariates related to the **housing structure**: rm, age

Covariates related to **accessibility**: dis, rad

2 Problem Description

2.1 Problem Formation

Housing price often reflects people's willingness to pay. Thus the relationship between housing price and its attributes could be a good indicator of people's willingness to pay for certain living conditions. For example, housing with better air, better accessibility, and better neighborhood are likely with higher price.

Based on this understanding, our analysis intends to resolve the following problems from the two parties: For the government, we want to show what people care about their surroundings, and furthermore, which living conditions are more important among all. While for the industry practitioners, we want to help them provide more reasonable and accurate housing price prediction.

2.2 Model Selection

As we can see, for the government, the problem is more about interpreting the housing data. In this way, we apply linear regression model on the full data, aiming at obtaining more interpretable results.

While for the practitioners, prediction accuracy is more valued. Therefore, in addition to the linear regression model, we also tried to use more advanced machine learning method like random forest, to aim for better prediction accuracy. We thus split our data into train and test to enable comparison and demonstration of the model performance.

3 Analysis Process

3.1 Descriptive Analysis

To gain a better understanding of the data, we first conducted descriptive analysis on the variables.

```
      crim                zn              indus            nox               rm             age       
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      dis              rad              tax           ptratio            b              lstat       
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73  
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95  
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65  
 3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95  
 Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97  
      medv      
 Min.   : 5.00  
 1st Qu.:17.02  
 Median :21.20  
 Mean   :22.53  
 3rd Qu.:25.00  
 Max.   :50.00
```

Figure 1: Descriptive Statistics of the Variables

From figure 1, we can clearly see that the variables are of quite different scales, suggesting that standardization should be applied before we fit them into the linear regression model. The detailed histograms of the variables could be found in the appendix.

Then, we conducted correlation analysis on the covariates to examine possible multicollinearity. The results were printed out in the form of hotmap, where the darker blue represents for stronger positive correlation, and the darker red represents for stronger negative correlation.

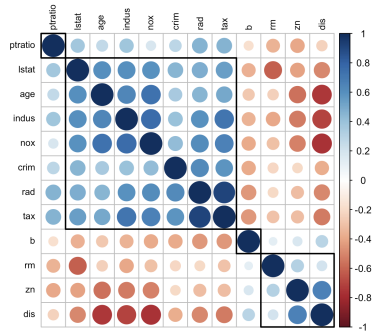


Figure 2: Correlation Coefficient Hotmap of the Covariates

From figure 2, it can be seen that some of the housing attributes are correlated to some degree. For example, the two variables $ptratio$, $lstats$ have positive correlation. The is consistent with the common sense, i.e. housing tends to be both around lower status population (higher $lstats$) and with less education resourses(higher $ptratio$). This analysis made us worried about the problem of multi-collinearity among the covariates, which, if left unsolved, would harm the identifiability of the model and lowering the estimation accuracy. We thus test for multicollinearity in the next step.

3.2 Diagnoses of Multicollinearity

To test for multicollinearity within the covariates, we first fit an intial model, and obtained the Variation Inflation Factors for each of the covariates. Then we also checked the condition number for the correlation matrix of the covariates.

```
VIF of full model:
    crim       zn    indus      nox       rm      age      dis      rad      tax  ptratio        b    lstat
1.787705 2.298257 3.949246 4.388775 1.931865 3.092832 3.954961 7.397844 8.876233 1.783302 1.344971 2.931101


Condition Number of full model:
[1] 94.77388
```

Figure 3: VIF and Condition Number of the Full Model

As the results of VIF(<10) and Condition Number(<100) not passing the empirical bar for multicollinearity, we could conclude that there is no strong multi-collinearity between the covariates in the full model. Therefore, we will not apply ridge regression or other specific methods designed solely for handling multi-collinearity, but rather focus on the diagnoses and refinement of the model from other aspects.

3.3 OLS and Variable Selection

We fit the initial full model by OLS, and then tried to conduct variable selection via stepwise regression under the criteria of AIC. Both backward and stepwise regression suggested the deletion of *indus* and *age*.

Initial full model:

```
Call:
lm(formula = medv ~ ., data = df_scale)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45663 -0.30556 -0.07019  0.20812  2.86780

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.986e-15  2.314e-02   0.000 1.000000
crim        -1.058e-01  3.097e-02  -3.417 0.000686 ***
zn           1.193e-01  3.511e-02   3.398 0.000734 ***
indus        3.007e-02  4.603e-02   0.653 0.513889
nox         -2.188e-01  4.852e-02  -4.509 8.13e-06 ***
rm           2.942e-01  3.219e-02   9.137  < 2e-16 ***
age          8.520e-03  4.073e-02   0.209 0.834407
dis         -3.401e-01  4.606e-02  -7.383 6.64e-13 ***
rad          3.108e-01  6.300e-02   4.934 1.10e-06 ***
tax         -2.521e-01  6.901e-02  -3.653 0.000287 ***
ptratio     -2.333e-01  3.093e-02  -7.542 2.25e-13 ***
b            9.670e-02  2.686e-02   3.600 0.000351 ***
lstat       -4.147e-01  3.965e-02 -10.459  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5205 on 493 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.7291
F-statistic: 114.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

Stepwise Regression ——————> By AIC

Initial reduced model:

```
Call:
lm(formula = medv ~ crim + zn + nox + rm + dis + rad + tax +
    ptratio + b + lstat, data = df_scale)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45389 -0.30382 -0.05989  0.20596  2.87027

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.003e-15  2.310e-02   0.000 1.000000
crim        -1.067e-01  3.089e-02  -3.453 0.000602 ***
zn           1.160e-01  3.461e-02   3.352 0.000864 ***
nox         -2.075e-01  4.480e-02  -4.631 4.65e-06 ***
rm           2.937e-01  3.131e-02   9.381  < 2e-16 ***
dis         -3.494e-01  4.285e-02  -8.155 2.89e-15 ***
rad          2.987e-01  6.039e-02   4.947 1.04e-06 ***
tax         -2.323e-01  6.215e-02  -3.737 0.000208 ***
ptratio     -2.303e-01  3.057e-02  -7.535 2.34e-13 ***
b            9.658e-02  2.675e-02   3.611 0.000337 ***
lstat       -4.100e-01  3.714e-02 -11.042  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5197 on 495 degrees of freedom
Multiple R-squared:  0.7353,    Adjusted R-squared:  0.7299
F-statistic: 137.5 on 10 and 495 DF,  p-value: < 2.2e-16
```

Figure 4: Initial Full Model and Reduced Model Summary

Here, we provide some explanation for why these two variables are deleted at the very beginning. From the initial full model in figure 4, we can see that the coefficients of these two variables are not that large, while their standard errors could be quite large, resulting in the large t value and thus insignificance. This is probably due to large variance in the distribution of the two covariates themselves.

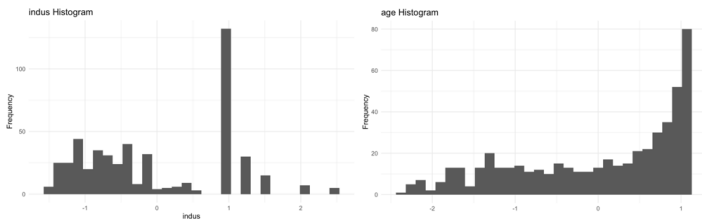


Figure 5: Histogram of *indus* and *age*

3.4 Regression Diagnostics

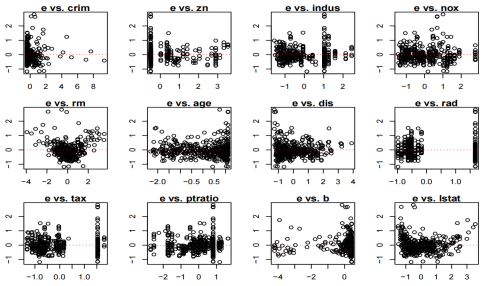
The diagnoses and correction process are the same for the reduced and the full model. We show the process for the full model as example. The results for the reduced model could be found in the appnedix.

3.4.1 Heteroscedasticity

For the heteroscedasticity test, we conducted the rank correlation test on each of the covariates with the absolute values of the residuals, and also show the scatter plots of the residuals with covariates.

Variable <chr>	Spearman_Correlation <dbl>	Spearman_P_Value <dbl>
zn	0.09147063	0.0397049085
rm	0.15833075	0.0003497853
age	0.12307461	0.0055679808
dis	-0.15510651	0.0004624398
rad	0.11346964	0.0106377560
ptratio	-0.11498021	0.0096364316
lstat	-0.09895071	0.0260274953

(a) Spearman Correlations under threshold 0.05



(b) Residual Plots

Figure 6: Heteroscedasticity Diagnoses Results

Figure (a) shows the covariates with a significant ($p < 0.05$) correlation with the absolute residuals. We could see from the test results and the plots that quite a few covariates show significant correlation with the residuals, indicating heteroscedasticity.

3.4.2 Autocorrelation

Based on the case background that the observations were arranged by their census indices and the census tracts have spatial correlation. We conducted autocorrelation diagnostics based on the order of the observations in the original data. We draw the residual plot between variables and conducted the quantitative Durbin-Watson test.

```
        Durbin-Watson test

data:  lm_full
DW = 1.0159, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 7: Results of DW test on Initial Full Model

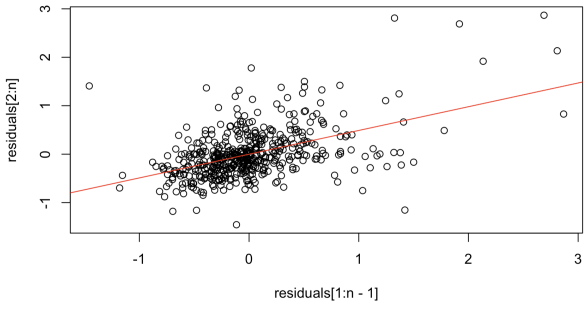


Figure 8: e_t and e_t-1 Residual Plot of the Full Model

Both the DW test (significant, rejecting the null hypothesis) and the plot supports that autocorrelation exists.

3.4.3 Resolutions

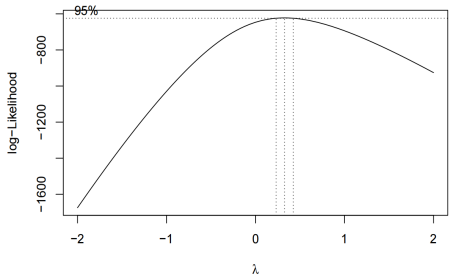
We first tried to solve the problem with Box-Cox transformation directly, as it was suggested to be a comprehensive correction for the assumptions for normal linear regression. However, the result shows that neither the issues of heteroscedasticity or autocorrelation was solved. Thereby we first performed Iterative Method to handle the first-order autocorrealtion in the data.

```
        Durbin-Watson test

data:  lm_it
DW = 2.0082, p-value = 0.4695
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 9: DW Test Results after Iterative Method

As we could see, autocorrelation problem is largely solved.




```
        Durbin-Watson test

data:  lm_it_bc
DW = 2.1791, p-value = 0.9681
alternative hypothesis: true autocorrelation is greater than 0
```

Spearman Correlations Under Threshold 0.05

Variable <chr>	Spearman_Correlation <dbl>	Spearman_P_Value <dbl>
dis	-0.1759536	6.91255e-05
rad	0.1059751	1.70943e-02
tax	0.1339857	2.52685e-03

(a) Log-likelihood Function and Choice of $lambda$ in Box-Cox Transformation

(b) DW test and Rank Correlation Test Results after Box-Cox Transformation

Figure 10: Box-Cox Transformation Results

Then, we perform Box-Cox transformation to handle the rest. We can see that after iterative method and Box-Cox transformation, the problem of autocorrelationa nd heteroscedasticity are largely improved, though some variables still show quite high correlation with the abse of the model.

3.5 Outliers and Influencial Points

Then we identified the outliers by cook distance and influential points F-test on SRE. And we deleted datapoints that are the intersection of outliers and influential points in the full model and reduced model respectively.

4 Final Models

Here we summarize the procedures to obtain the final full model final reduced model and their results.

4.1 Final Full Model

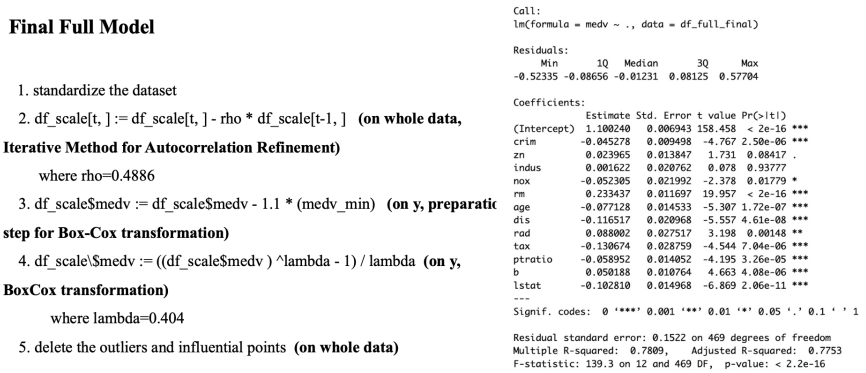


Figure 11: Steps and Results of Final Full Model

4.2 Final Reduced Model

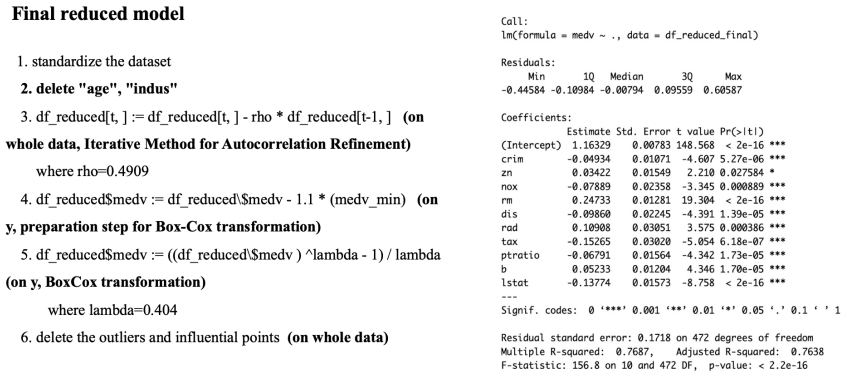


Figure 12: Steps and Results of Final Reduced Model

It is somewhat unexpected to find that the final full model shows better adjusted R^2 than the final reduced model. Upon reflection, this might be the case that after correcting for the GM(Gauss-Markov) conditions, the issues with the two variables *indus*, *age*, especially *age*, with initial large variance are mitigated in the final full model. Also, there are more point lying in the intersection of outliers and influential points for the full mode, as we can see in the by the resulting Degree of Freedom of full model less than that of the reduced model($469 < 472$), which may also help boosted the adjusted R-square of the full model.

5 Interpretation of Results

Since the difference in goodness-of-fit of the two final linear regression models is quite minor, while the variables in the final reduced model are all significant, we then base our interpretation on the results of the final reduced model, i.e. not considering the two deleted varaibles *indus*, *age*.

Category	Variables	Expected	Model
Neighborhood	*crim*	-	-0.05
	zn	+	+0.03
	indus	+	delete
	tax	-	**-0.15**
	ptratio	-	**-0.07**
	b	-	+0.05
	lstat	-	**-0.13**
Pollution	*nox*	-	**-0.08**
Structure	*rm*	+	**+0.25**
	age	-	delete
Accessibility	*dis*	-	**-0.10**
	rad	+	**+0.10**

Table 2: Coefficients from the Final Reduced Model

5.1 Coefficients of the final model

After standardization and transformations correcting for GM conditions, the resulting model has coefficient estimates and standard errors of the estimates of about the same magnitude, which provides convenience and reliability for the interpretation of values of the coefficient estimates.

The signs of these coefficients are all consistent with our common sense and expectation, i.e. roughly speaking, higher housing prices are related with variables indicating a more ”quality” neighborhood, less pollution, better housing structual attributes, and better accessibility. We could then focus our interpretations on the absolute values of the coefficients.

As the the variables have already been standardized, here each of the coefficients represents **the average effect on the standardized response variable, *medv*, for per standard deviation change in the covariate, at fixed levels of all other covariates in the model**.

The highest one being *rm*, and as it is left as the only variable reflecting the intrinsic structural attribute of the housing, it is no wonder that it has the most impact on the housing price. Then is *tax*, of which a higher value means property owners in that area are paying a larger percentage of their property’s value in taxes. This can be seen as a financial burden for the owner and could be directly reflected in market prices, which is why it is represented as a negative value.

5.2 Guidance for the government

For the government, the absolute values of the coefficients, mainly the rest of the variables in the ”neighborhood”, ”pollution” and ”accessibility” categories, could thus reflect the relative importance people place on different attributes for living conditions in Boston, 1970. We can see $lstat$, percentage of lower status of the population, has the largest absolute value, 0.13 under the neighborhood category, other than tax, which is also quite intuitive. The two variables reflecting accessibility of the housing the employment centers and radial highways sharing about the same relative importance, 0.10. Then comes the variables indicating public resources of clean environment and educational resources, namely nox and $ptratio$, with 0.08 and 0.07 importance, followed by the security variable, crime rates $crim$, with 0.05 importance.

Guidance for the government could be to prioritize the improvement of different aspects of public service based on the ranking of these variables that reflects the preferences of the residents. Though the status composition of the population might be hard to change directly, the government could consider devoting most on improving the accessibility of the community, i.e. attracting more factories and companies to the town, building or extending infrastructure like radial highways. The second tier of priorities could then be controlling pollution and increasing educational resources, followed by controlling crime rates.

6 Model Prediction

For industry practitioners, while the sensible interpretation of the underlying model offers reliability, it is sometimes more valued to give an accurate prediction of the housing price. Thereby we first utilize the two linear models we have fitted, and then experimented with other machine learning models to improve the prediction results.

6.1 Prediciton results

We randomly split the data into 80% training set and 20% test set, applied the two initial linear models (full and reduced) before correction for GM conditions (as it is closely related to the arrangement of the observations and the distribution of the training data, which we think would result in overfitting), and the random forest algorithm. Then we compared their performance in terms of the Mean Squared Error on the test set.

Model	**test MSE**
Full Linear Regression	23.45
Reduced Linear Regression	23.20
Random Forest	13.19

Table 3: Prediction Performance of Different Models

As we can see, while the initial reduced linear model shows slightly better generalizablity compared with the full model, the random forest model provides much better prediction for housing prices with regard to the metric of test MSE.

6.2 Guidance for practitioners

Practitioners might utilize the linear regression model for rough estimates while better interpretability, if explanations to its customers on the price decision are required. If the only goal is to obtain more accurate prediction of the housing price, other machine learning methods, like the random forest model we implemented here, could yield better performance. Though the variable importance in these models could be interpreted to some degree, as we include in the appendix, they might still be not as transparent as the linear models.

A Random Forest Model

1. Training Process

```
Random Forest

404 samples
 12 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 404, 404, 404, 404, 404, 404, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    3.756040  0.8383117  2.451236
   7    3.486237  0.8519735  2.307367
  12    3.707115  0.8314256  2.411519

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 7.

Call:
 randomForest(formula = medv ~ ., data = df_train, mtry = 7, ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 7

          Mean of squared residuals: 9.936269
                    % Var explained: 88.23
```

Figure 13: Training final models of Random Forest Model

- **Ntree:** the nunber of decision trees
- **Mtry:** number of variables under each decision tree
- **Method of adjusting mtry:** the smaller the RMSE, the better the model fitting. So we just need to find the smallest RMSE to select the optimal model

After parameter tuning, it was found that the optimal parameter mtry was 7, and the error had stabilized when ntree exceeded 200. In order to be accurate, 500 was chosen here, and the final MSE was 9.94, with a variance interpretation rate of 88.23%.

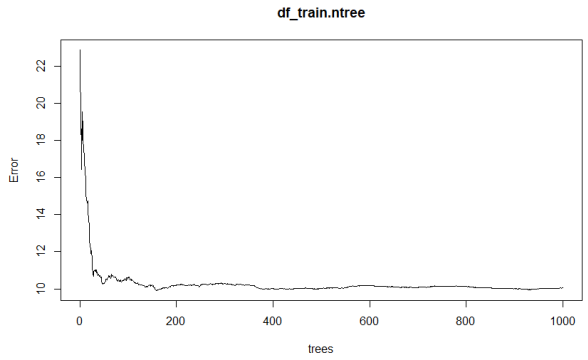


Figure 14: Error with the Choice of ntree

2. Explanation

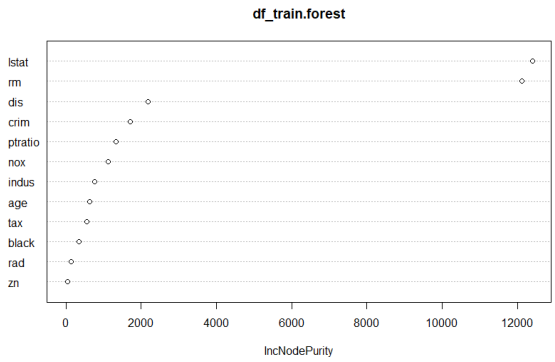


Figure 15: Variable Importance in Random Forest Model

According to the varImpPlot(Dotchart of variable importance as measured by a Random Forest) function, it can be concluded that the lstat and rm parameters are the most important, which are the number of rooms and the proportion of lower class population. This is consistent with cognition, as housing positioning and area are absolute factors that determine housing prices. Nox (air quality), dis (distance to work), ptratio (pupil-teacher ratio), and crim (crime rate) rank second, while age, indus, tax, b, zn, and rad are the least important factors, which is consistent with the insignificant factors age and indus in multiple linear regression. Due to the randomness of random forests, the results may vary each time, but overall, lstat and rm are the most important factors.

3. Prediction

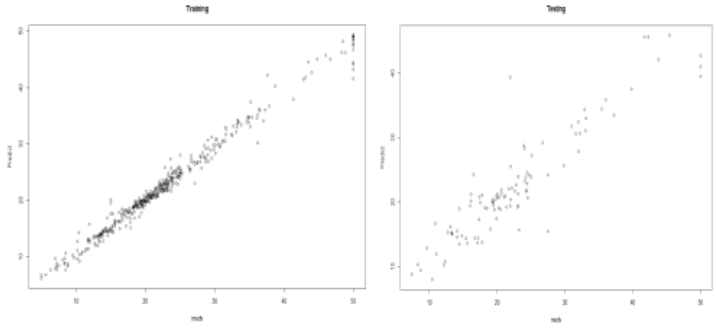


Figure 16: Training and Prediction Results

Machine learning models are difficult to explain and are mostly used for prediction. The above figures show the models used to predict housing prices using the random forest algorithm, calculated on the same training and testing sets. The test set predicted an MSE of 13.19, which is better than the two regression models. Generally speaking, when fitting a random forest, it randomly selects features and samples without repetition. Through multiple sampling and adjustments, overfitting can be avoided to a certain extent, and it has strong anti-interference ability.

B Graphics

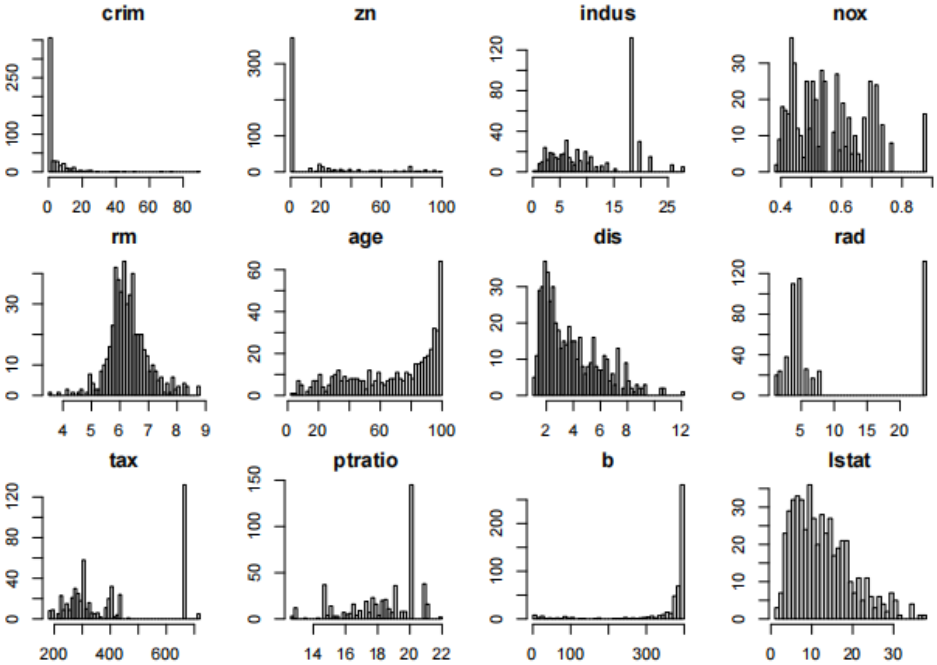


Figure 17: histogram

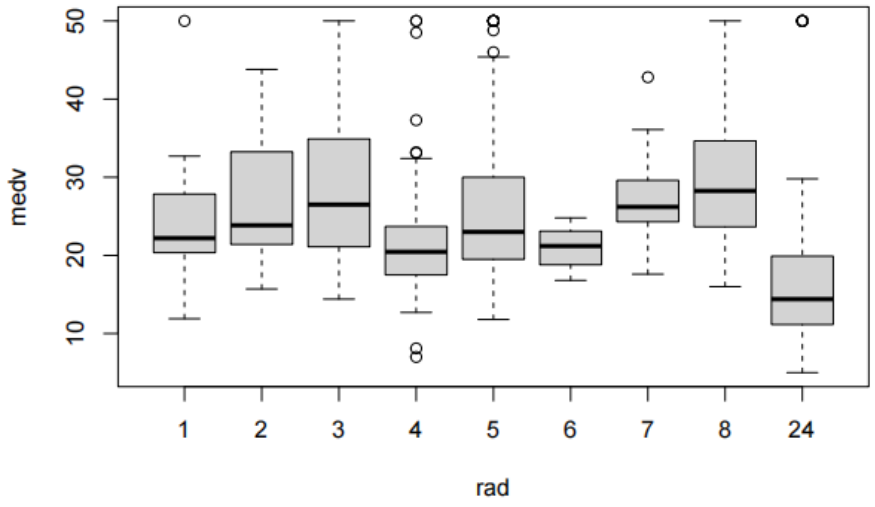


Figure 18: boxplot: *medv* vs *rad*

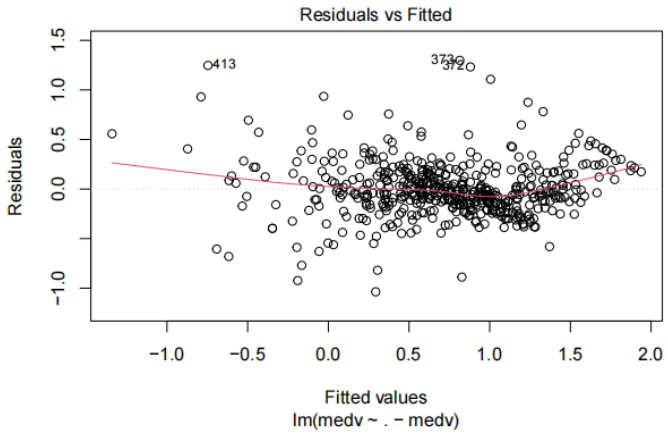


Figure 19: plot for Residuals vs Fitted (a)

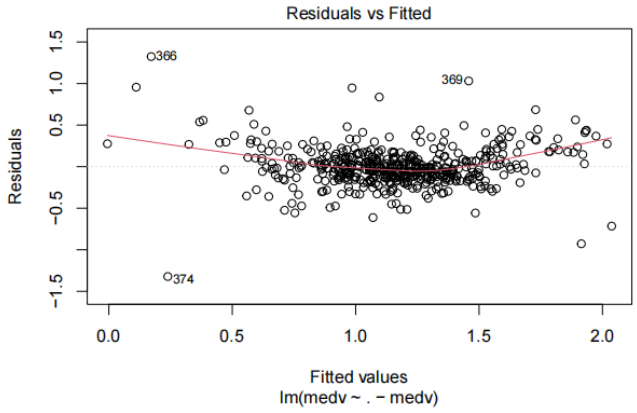


Figure 20: plot for Residuals vs Fitted (b)

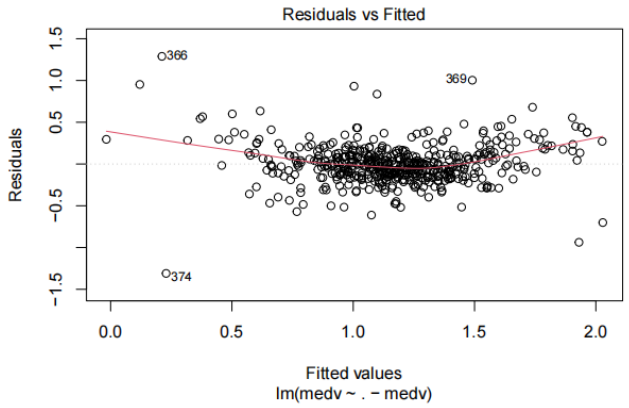


Figure 21: plot for Residuals vs Fitted (c)