Effects of Average Room Number, Low Income Family Proportion and Student-to-Teacher Ratio on Nearby Housing Price

Yancheng Zhou

Instructor: Jun Wang

Introduction

The objective of this paper is to understand the question of how do factors like the average number of rooms and low-income family proportion affect the housing price in a specific area. This paper explores the relationships between the primary variable and the secondary variables, and therefore come up with a best model to fit them. The result of this is important to customers and land agents for potential future references.

Literature Review

A research conducted by Limsombunchao, V compares the predicting powers of two prevalently-used prediction models hedonic model and artificial neural network model. The result shows that the artificial neural network model holds a better performance against the hedonic model in predicting the housing price in New Zealand. This article is related to our paper as we also compare the predicting powers of two models—Linear Regression Model and Regression Tree Model.

Another research conducted by Sibel Selim investigates the predicting power of Hedonic Regression Model. Although this study

uses different models than ours, we share similar variables such as the number of rooms.

Data Description

To answer our research question, we used a sample of housing price data named “housing.k” from the UCI machine learning repository. The dataset contains individual data related to housing price, having a total number of 489 observations. Our primary variable that we are interested in investigating is median value of owner-occupied homes of an area (MEDV). The secondary continuous variables are the average number of rooms (RM), lower status of the population (LSTAT), and the pupil-teacher ratio by town (PTRATIO).

Table 1. Summary statistics for key variables

Variable	Mean	Median	Maximum	Minimum	Range	IQR	Standard Deviation
RM	6.24	6.19	8.40	3.56	4.84	0.70	0.64
LSTAT	12.94	11.69	37.97	1.98	35.99	9.75	7.08
PTRATIO	18.52	19.10	22.00	12.60	9.40	2.80	2.11
MEDV	454343	438900	1024800	105000	919800	168000	165340

A comparison of means and medians of each continuous variables shows that there is no significant difference between them, and thus indicating there are no significant outliers exist in our variables.

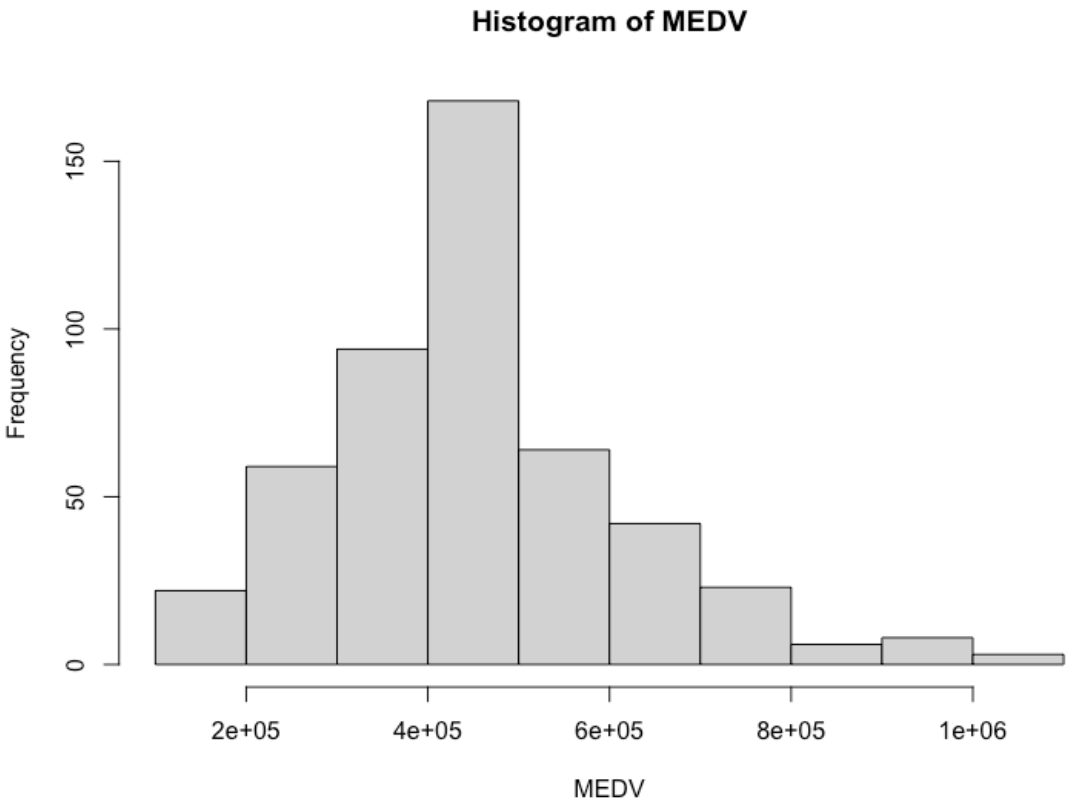

Figure 1. Histogram of the primary variable MEDV

(Note: this histogram looks normally distributed with skewed slightly to the right.)

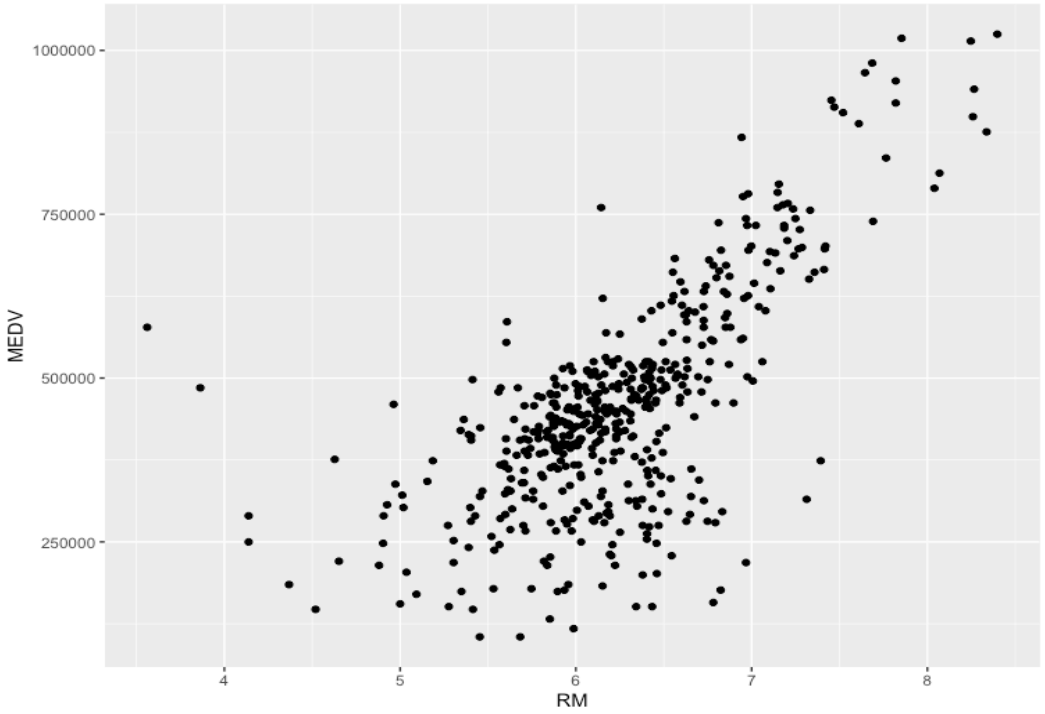

Figure. 2 Scatter plot with MEDV on the y-axis and RM on the x-axis

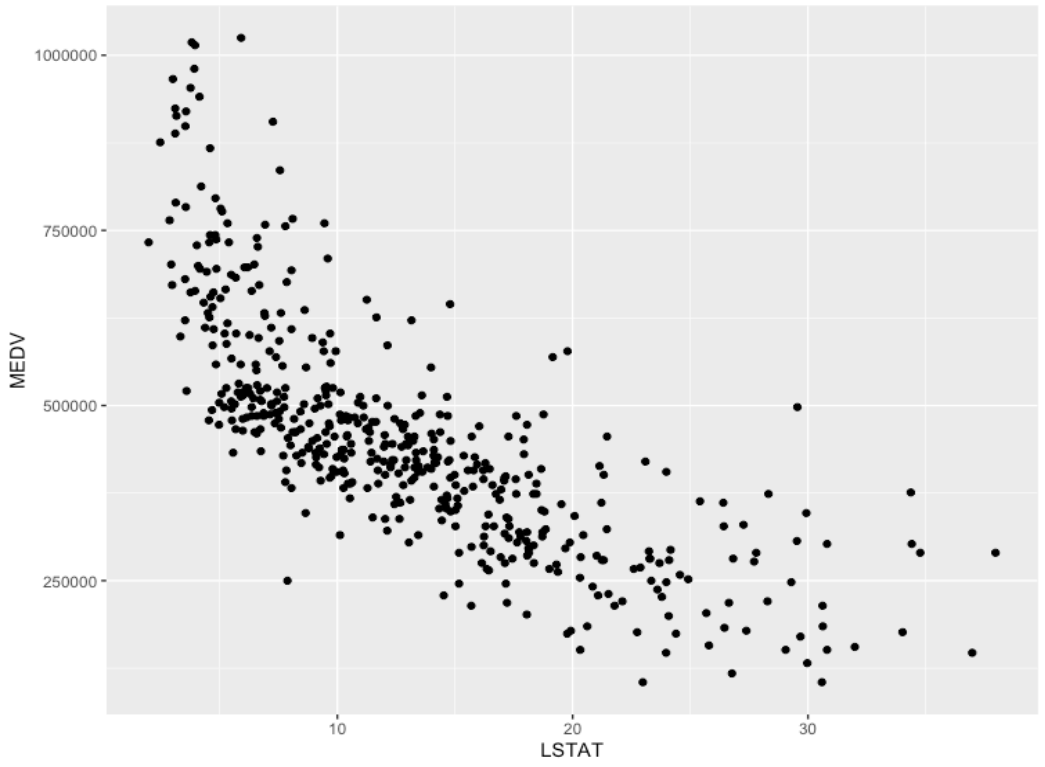

Figure. 3 Scatter plot with MEDV on the y-axis ad LSTAT on the x-axis

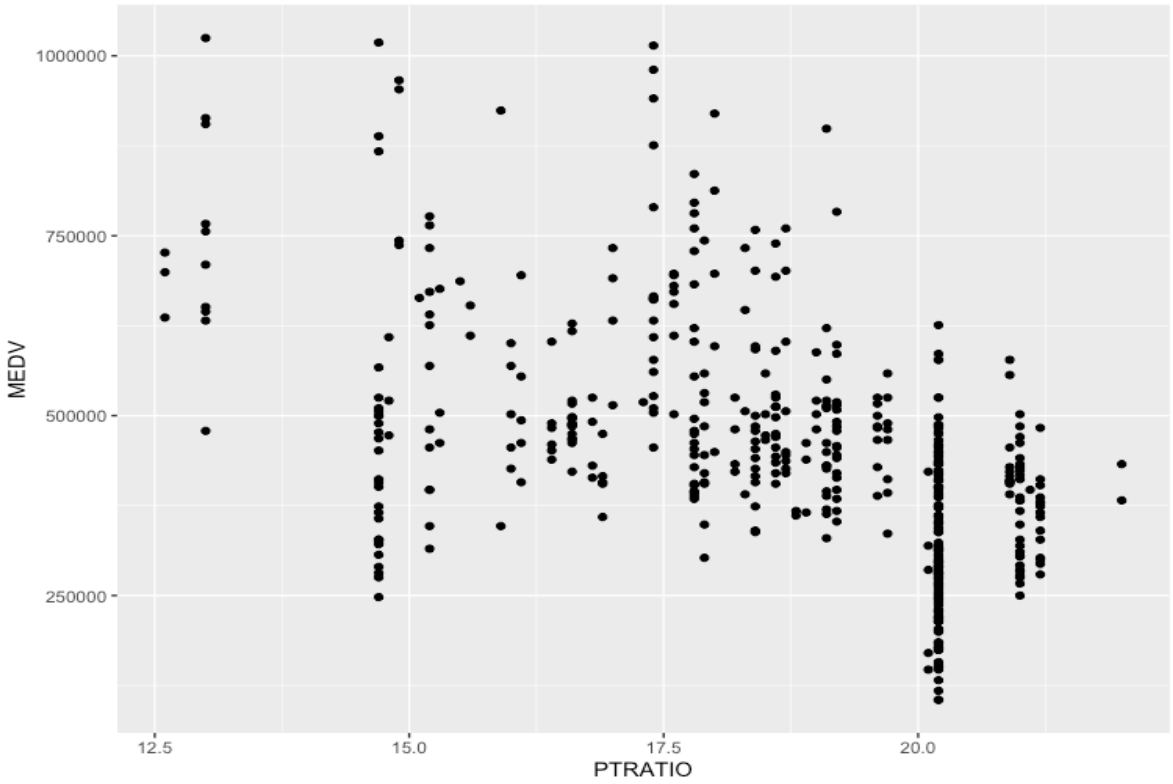

Figure. 4 Scatter plot with MEDV on the y-axis ad PTRATIO on the x-axis

Empirical Strategy

In the linear regression model, the dependent variable of interest is MEDV, and the independent variables are RM, LSTAT, and PTRATIO. To find the best model, we compute the mean absolute percentage error (MAPE) of the following regression models:

$MEDV=\beta0+\beta1*LSTAT$

$MEDV=\beta0+\beta1*PTRATIO$

$$MEDV=\beta0+\beta1*RM$$

$$MEDV=\beta0+\beta1*RM+\beta2*LSTAT$$

$$MEDV=\beta0+\beta1*RM+\beta2*LSTAT+\beta3*PTRATIO$$

In the process of selecting the best regression model, we use the method of stepwise variable selection. After figuring out the best regression model, we are going to do cross validation and out-of-sample testing to test how good the model actually is. To do that, we assembly 5 groups from the data and calculate the MAPE of each group.

In building the tree model, we are going to compare the prediction accuracy of the model before pruning and after pruning. The one with the higher accuracy will be compared with the best linear regression model we came out earlier.

Results and Analysis

Stepwise forward selection

By comparing the F-statistics and some other parameters, we found that $MEDV=\beta0+\beta1*LSTAT$ is the best single variable model. Then, we tested the addition of each variable to this model and

added the one whose inclusion gives the most statistically significant improvement of the fit. We repeated this process until none improves the model to a statistically significant extent. The result of this shows that MEDV=β0+β1*LSTAT is still the best model out of all other combinations. We used this model to test the data and got 19.25% average percentage difference between the predicted value and the actual value.

MEDV= 684138-17759*LSTAT

Intercept ~ Pr(>|t|) : <2e-16 ***

LSTAT ~ Pr(>|t|) : <2e-16 ***

P-value: < 2.2e-16 (<0.05)

R-squared: 0.5786

Adj R-squared: 0.5778

Residual Standard Error: 107400

F-statistics: 668.7

Figure. 5 linear regression model between MEDV and LSTAT

Cross Validation

The next step we did is to do cross validation with this model to see its out-of-sample performance. Each time, we took 20% of the data from the dataset to form the training data, and the rest 80% data formed the testing data, which was tested by the model. Figure 6 shows the mechanism of conducting cross validation.

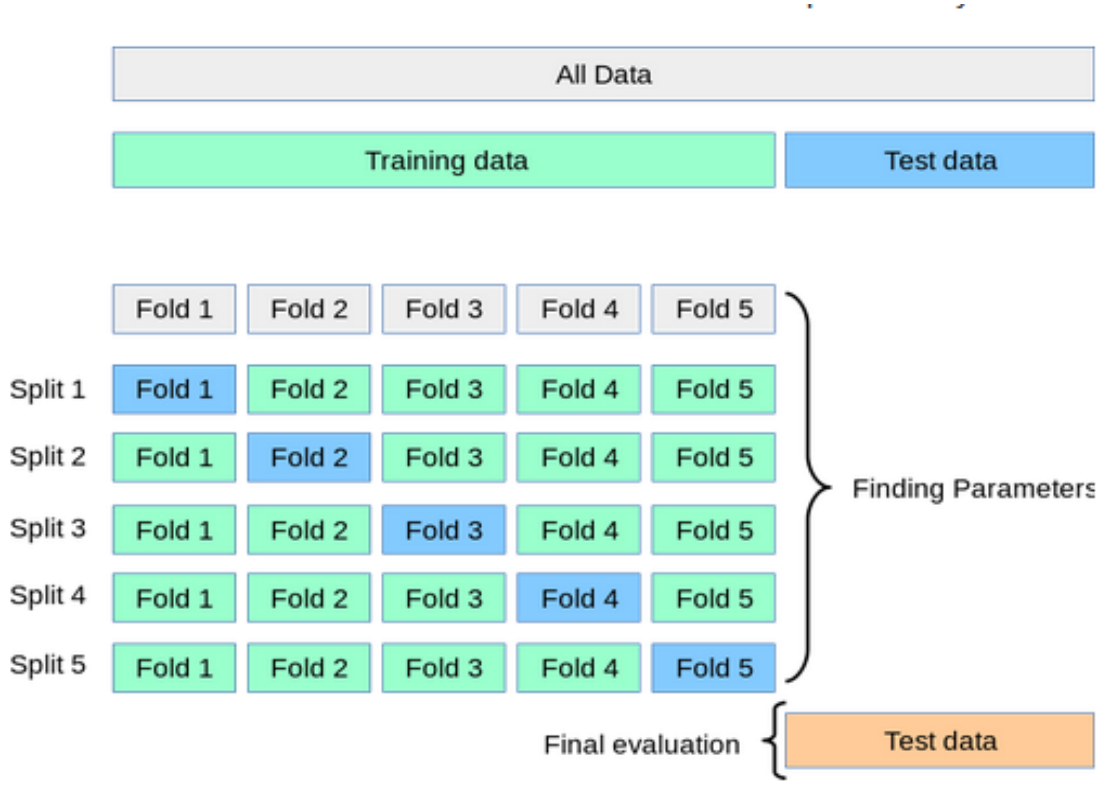

Figure. 6 the schematic diagram of cross validation

The respective average percentage differences between the predicted value and the actual value are shown below.

Table 2. Cross validation performance of linear regression model

	MAPE
All data	19.25%
Group 1	16.68%
Group 2	15.32%
Group 3	17.45%
Group 4	27.68%
Group 5	24.89%

Moving on, we built a regression tree model with all the secondary variables, and then we repeated the process of cross validation to get the below result.

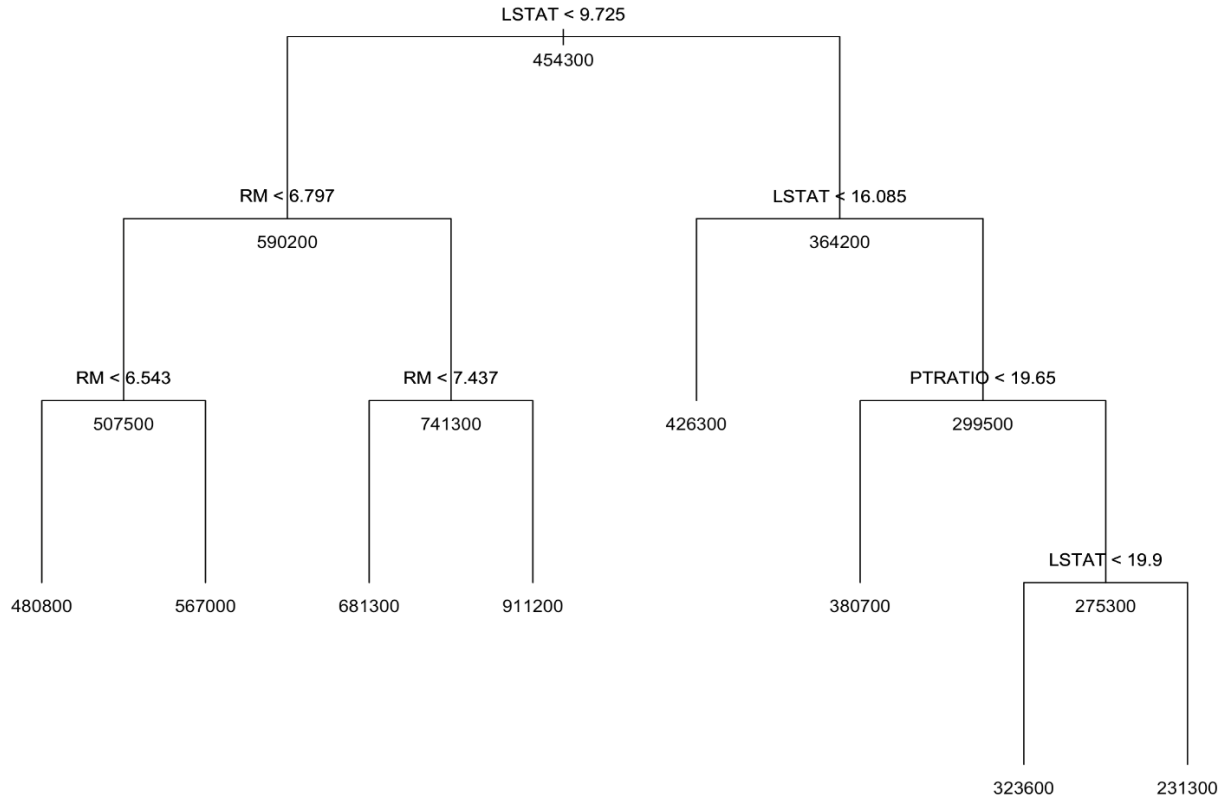

Figure. 7 Regression Tree Model without pruning

Table 3. Cross validation performance of regression tree model

	MAPE
All data	13.66%
Group 1	10.04%
Group 2	14.35%
Group 3	13.97%
Group 4	22.98%
Group 5	23.95%

By comparing the results, the regression tree model has better performance in its prediction accuracy against the linear regression model. Then, we removed the variable PTRATIO from our regression tree model to further improve the prediction power because PTRATIO has the most loose linear connection with the dependent variable MEDV out of all other secondary variables. The result of this is displayed below in table 4.

Table 4. Cross validation performance of regression tree model after removing the variable PTRATIO

	MAPE
All data	14.91%
Group 1	10.15%
Group 2	12.44%
Group 3	14.85%
Group 4	26.84%
Group 5	28.76%

The result was not what we expected—the prediction power decreased after the removal of the variable PTRATIO. However, this makes sense since PTRATIO has the worst linear connection with MEDV but tree model is not a linear model. This suggests that PTRATIO still has its contribution in the regression tree model.

The last step we took to improve our model was to do tree pruning. The logic behind this is to reduce overfitting between the model and our data by lowering the number of nodes. We manually set the number of nodes to 5 as opposed to the 8 nodes from the previous model.

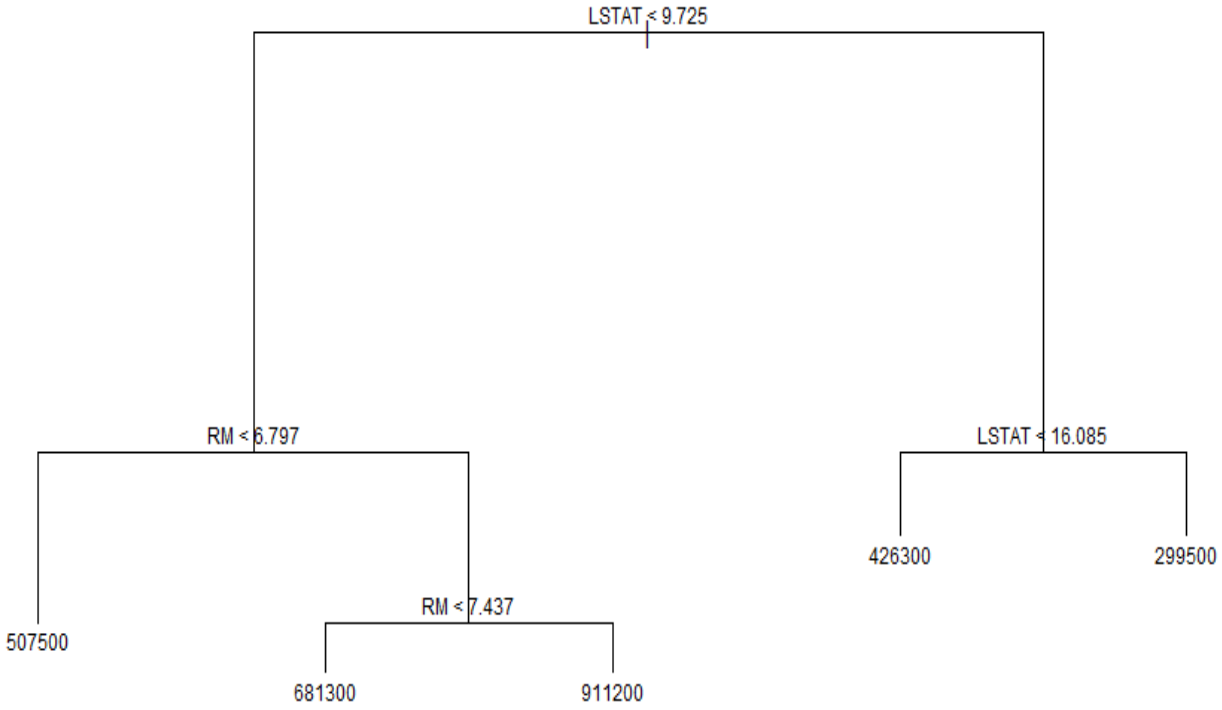

Figure. 8 Regression Tree Model after pruning

Table 5. Cross validation performance of regression tree model after pruning

	MAPE
All data	16.39%
Group 1	11.52%
Group 2	16.53%
Group 3	15.83%
Group 4	31.94%
Group 5	28.91%

The result shows that the prediction accuracy actually decreased after pruning, suggesting the previous model does not have overfitting problem. Therefore, we conclude that out of all the models we tested, the best model for predicting the housing price is the regression tree model with three secondary variables.

Future research could attempt to find the better model other than the two model we tested in this paper. Another direction for future investigation would be the importance of each individual variable on the housing price in different areas. As a result, real estate price would be better monitored by consumers and producers.

References

Limsombunchao, V, 2004-06. House price prediction: Hedonic price model vs. artificial neural network. *Research@Lincoln.* http://researcharchive.lincoln.ac.nz/handle/10182/5198

Sibel Selim, 2008. Determinants of House Prices In Turkey: A Hedonic regression Model. *Dogus University Research.* http://journal.dogus.edu.tr/index.php/duj/article/view/80

UC Irvine Machine Learning Repository. “housing.k” datasets. https://archive.ics.uci.edu/ml/index.php