|  |  |  |
| --- | --- | --- |
|  | **Typed Name** | **Signature** |
| Contact Author | Rujun Dai |  |
|  | Yang Men |  |
|  | Xin Yu |  |
|  | Siqi Ren |  |
|  | Lan Yan |  |

house price prediction project

Data Mining for Business (BUDT – 758T)

Rujun Dai, Yang Men, Xin Yu, Siqi Ren, Lan Yan

We the undersigned certify that the actual composition of this report was done by us and is original work

Contents

[1. Executive Summary 2](#_Toc482655736)

[2. Data Description 3](#_Toc482655737)

[Data Source 3](#_Toc482655738)

[Data Preparation 3](#_Toc482655739)

[3. Research Question 5](#_Toc482655740)

[4. Methodology 6](#_Toc482655741)

[Data Partition 6](#_Toc482655742)

[Linear Regression 6](#_Toc482655743)

[Regression Tree 6](#_Toc482655744)

[Lasso 6](#_Toc482655745)

[Random Forest 6](#_Toc482655746)

[Neural Network 7](#_Toc482655747)

[5. Result and Finding 8](#_Toc482655748)

[Linear Regression 8](#_Toc482655749)

[Lasso Regression 8](#_Toc482655750)

[Regression Tree 9](#_Toc482655751)

[Random Forest 9](#_Toc482655752)

[Neural Network 9](#_Toc482655753)

[6. Conclusion 11](#_Toc482655754)

[Appendix 12](#_Toc482655755)

# 1. Executive Summary

Nowadays, many people need to purchase or sell houses. It’s important to do house price estimation so that you won’t pay or sell a house for an unreasonable price. This project studies the relationship between house price and house attributes such as number of bedrooms, level of construction and design. The team is devoted to dig out what house attributes have impact on house price and the degree of impact. The purpose of this project is to give people a house prediction model and key findings related to house price.

The main key findings:

* The 1-year period sale date does not impact on house price.
* Square footage of the land space does not impact on house price.
* Level of construction and design has strong impact on price, higher level leads to higher house price.

# 

# 2. Data Description

## Data Source

The dataset we used, named as “House Sales in King County”, is from Kaggle. The link is listed below:<https://www.kaggle.com/harlfoxem/housesalesprediction>**.**

## Data Preparation

The original dataset, consisting of 21 variables, contains 21,207 records of house sale information for King County, which locates in state of Washington, including the biggest city, Seattle, of the state.

For data preparation, we firstly removed outliers in dataset. Price that exceeds average price plus 3 standard deviations are identified as outliers. Then, ‘id’ and ‘zipcode’ are excluded because they do not have impact on model performance. Secondly, the ‘date’ column is eliminated because sale date does not have strong impact on price. (Figure 1) Thirdly, to evaluate the actual lives of the buildings, a new variable ‘Age’ is created by using current year minus the year built. While constructing prediction models, we used the newly created variable ‘Age’ to replace the ‘yr\_built’ column. Fourthly, excluded ‘long’ and ‘lat’ field, which indicate the longitude and latitude of the house. The reason is that all the houses in the dataset are in the same county. Given the geographical proximity, the difference in position can be ignored (Figure 2). Fifthly, we created a dummy variable for ‘yr\_renovated’. yr\_renovated=0 if the house has not been renovated, otherwise equals 1 regardless of the actual renovated year. In the last, ‘sqft\_above’ and ‘sqft\_basement’ fields, whose sum equals the value of ‘sqft\_living’, were removed to eliminate the strong linear dependency among the three variables. We kept ‘sqft\_living’ instead of ‘sqft\_above’ and ‘sqft\_basement’ because ‘sqft\_living’ has stronger correlation with price. (Figure 3) After data preparation, 13 variables are ready for the next step.

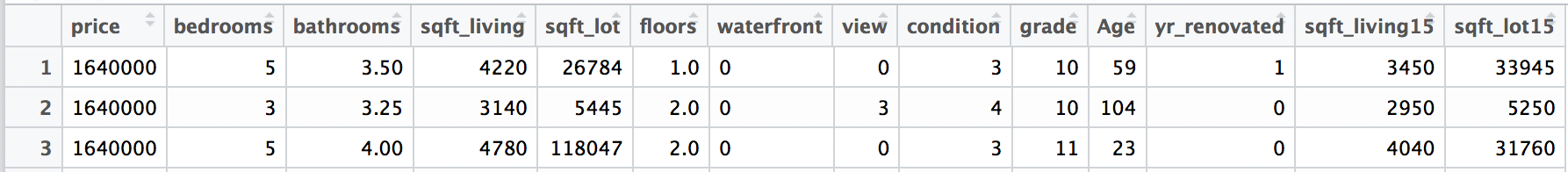
Dependent Variables

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Description |
| Price | Numerical | Price of each home sold, with a range from $75,000 to $7,700,000 |

Independent Variables

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Description |
| bedrooms | Numerical | The number of bedrooms |
| bathrooms | Numerical | The number of bathrooms |
| sqft\_lot | Numerical | Square footage of the land space |
| floors | Numerical | The number of floors |
| waterfront | Categorical | whether the apartment was overlooking the waterfront or not |
| view | Categorical | An index from 0 to 4 of how good the view of the property was |
| condition | Categorical | An index from 1 to 5 on the condition of the apartment |
| grade | Categorical | An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design |
| sqft\_living | Numerical | The square footage of the interior housing space |
| yr\_built | Numerical | The year the house was initially built |
| yr\_renovated | Categorical | Whether the house has been renovated or not |
| sqft\_living15 | Numerical | The square feet of interior living space for the nearest 15 neighbors |
| sqft\_lot15 | Numerical | The square feet of the land lots of the nearest 15 neighbors |

Sample data from R



# 3. Research Question

The purpose of this research project is to explore the hidden business value of the house sales data, obtaining profound insight in the housing market of King County, and hence provide the best model for house owners and local real estate agencies to estimate the values of the house.

To unearth the underlying relationship between house price and other affecting factors, several complicated methodologies were adopted to construct prediction models for house price. The report specifies the following three critical questions:

The first problem will be addressed in this report is that what kind of correlation is hidden between the house price and listed attributes. It is important for house owners to recognize the positive factors and the negative ones.

Secondly, to which degree is the house price affected by these factors. The dataset covers a wide range of information, including the building life, floor area, renovation condition, surrounding view and neighborhood. Some of these elements have stronger influence, while the change in the others can hardly be reflected in the final price of the house. For house owners and real estate agencies, the answer of this question allows them to set a better standard for house transaction.

Thirdly, the report intends to reveal the potential interactions among the factors. The combination of several attributes may add more value to the house. Realizing the underlying relationship gives house owners a deeper understanding of housing market and hence make better decision in house transaction.

# 4. Methodology

## Data Partition

To judge predictive performance, the dataset is divided to training and test dataset (7:3). The training dataset is used to estimate model parameters, and the test dataset is used to evaluate model performance in prediction by calculating AE, MAE and RMSE.

## Linear Regression

Multiple linear regression is used to predict the value of a variable based on the value of two or more independent variables and explain the relationship between them. To achieve the project goal, multiple linear regression model is a necessary option to find the basic relationship between price and different house attributes, such as the number of bedrooms, bathrooms, floors, square footage of land space and living space, the age of house, etc.

## Regression Tree

Decision tree is one of the most common model to explore predictive analytics. As our dependent variable, “price”, is numeric, choosing regression tree other than classification tree is more meaningful. Also, regression tree in nonparametric and therefore does not depend on data belonging to a data type. By pruning the tree to the optimal size, this approach reduces the complexity of the final classifier and improves predictive accuracy by the reduction of overfitting.

## Lasso

Lasso is a shrinkage approach for feature selection and regularization to improve accuracy of model. It can be easily find which variables contribute to the model. Different from subset selection approach, Lasso imposes a shrinkage penalty for choosing value of coefficient different from zero. The advantage of the penalty is that some coefficients can be end up being set to exactly zero.

## Random Forest

Random forest, one of the Ensemble Method, combines the results from regression tree models with the goal of improving prediction accuracy. By building a number of decision trees on bootstrapped training sample, it is constructed using a random subset of the variables for each split in the tree. Thus, random forest model could have good predictive performance, even when the data is very noisy and generally does not overfit.

## Neural Network

Neural network is a widely-used tool for both classification and prediction. The advantage of the model is that it can capture complex relationships but it does not define a specific relationship between response and dependent variables. Neural network can be used for making predictions, comparing predictors with actual values in training dataset, calculating cost or penalty and finding the “best” values for parameters.

# 

# 5. Result and Finding

## Linear Regression

The linear regression model was built based on all variables. Basically, the model has a relatively high adjusted R-squared, which is 0.6325. (Figure 4) Most variables are positively correlated with the dependent variable while only ‘bedrooms’ and ‘sqft\_lot15’ have negative impacts on the house price. According to the p-value of each variable, most of them are statistically critical except ‘sqft\_lot’ and ‘yr\_renovated’.

The Actual-Predicted graph (Figure 5) gives a clearer insight in the performance of the linear regression model. The x-axis of the graph represents the actual house price from the test dataset, while the y-axis indicates the predicted house price given by the linear regression model. As shown in the graph, most records are distributed around the regression line when the actual house price is under $1,000,000, which means that the linear model has a better performance when the actual house price is lower. However, when the actual house price increases, the model gives a lower predicted house price when compared with the actual value.

Overall, simple as and less flexible it is, the linear regression model gives a relatively precise prediction for the house whose actual value is lower than $1,000,000. However, it tends to underestimate the price of the house when its value grows higher.

## Lasso Regression

To further explore the correlation between independent variables and the house price, lasso regression was the second methodology adopted to build a model.

The best lambda, 413.8925, was selected based on the minimum cross-validation error. The coefficients of all the variables was then generated to give a overall look. (Figure 6) The coefficient of ‘sqft\_lot’ was driven to 0, which means that this variable has little contribution to the prediction of house price and is excluded in the lasso regression model.

As shown in the trace plot (Figure 7), it is obvious that most variables have positive influence on the house price, which is consistent with the result of linear regression. Two variables appear to have stronger impacts on the house price. One is ‘waterfront’, which is represented by the pink line, and the other one is ‘grade’, which is represented by the green line. In the other words, the relative position of the house towards water and the inner design and structure of the house should be taken into consideration when estimate the value of a house.

## Regression Tree

The regression tree was built on all variables except for sqft\_lot, which was excluded in lasso model. According to the regression tree graph (Figure 8), there are total 7 nodes on it. the first split is that the house grade whether larger than 8.5 or not. Grade is an index measured the level of house construction and design, it has range from 1 to 13. 1 is the worst, 13 is the best. Based on the graph, the house with grade over than 8.5 usually has higher price. Therefore, if you were real estate developer, the house design and construction would be the highest priority attribute to invest so that the house could be sold in higher price. After that, the tree has been pruned. By looking at the figure, it can be concluded that tree with 7 notes has the best outcome. (Figure 9) Hence, unpruned tree is as good as the pruned tree.

## Random Forest

When running the random forest model, the number of variables have been chosen for each tree was defined as 8. There are total 12 variables, the reason it has been defined as 8 is because we thought at least half of the variables should be used. We tried from 6 to 11 to run random forest, 8 generated the best outcome in terms of lowest RMSE. According to the variables importance graph, for the ‘%incMSE’, house age is the most important one. It means if the value of age would be changed randomly but realistically, and keep everything else constant. Running the model again, a worse prediction in terms of higher MSE would be generated. For the increase in node purity, more useful variable achieves higher increase in node purity. According to this, grade is the most useful one.

## Neural Network

Before running neural network, each variable has been scaled. The hidden layer was set to 0 in the model because the dataset is linearly separable. The test RMSE was about 0.1 from the model. Because of using scaled data, the result of neural network is not comparable with other models. Linear regression model was implemented with the scaled data, the test RMSE was almost the same as neural network. However, changing the number of nodes and layers in hidden layer would affect the test RMSE. Therefore, we tried different number of nodes in hidden layer to calculate test RMSE. When hidden layer equals 3, the test RMSE equals 0.0959. When hidden layer equals 5, the test RMSE equals 0.0940. It seems like more nodes in hidden layer would decrease the RMSE a little bit. Overall speaking, the performance of neural network model is as good as the linear regression model.

Comparing each model’s accuracy (Table 1), random forest model has the best performance because of the smallest AE, MAE and RMSE. The result is not surprised because random forest picks subset of variables to run trees, this would avoid overfitting and strong correlation between variables.

# 6. Conclusion

To summaries, among all the methodologies discussed above, random forest shows the best result. Through the process modeling, there is no doubt that house price is affected by many variables and no single one of them can be used alone in predicting the price. Although only the limited variables provided in the dataset, it is obvious that there are also other factors may affect the price as well. For example, real estate agent may want to improve their sales performance by taking some promotions at holidays, which would affect the house price in a great way. Apart from that including more features in the models is a considerable way to improve performance, if possible, in the future it is meaningful to apply the methodologies into predicting of house prices in locations other than King county.

The final model is useful for both real estate agency and house owner in King county, since they can use it to predict the house price and therefore sale their house for a better and more appropriate price. Also, the model is suitable for person who want to use it to find an idea house of better attributes within budget.

Overall, our team collaborated very well, each of us contributed and learned a lot from this project. We find the importance of feature selection, and our project have proven that data cleaning and data preparation took most of the time of modeling.

# 

# Appendix

Figure 1: plot of price and sale date

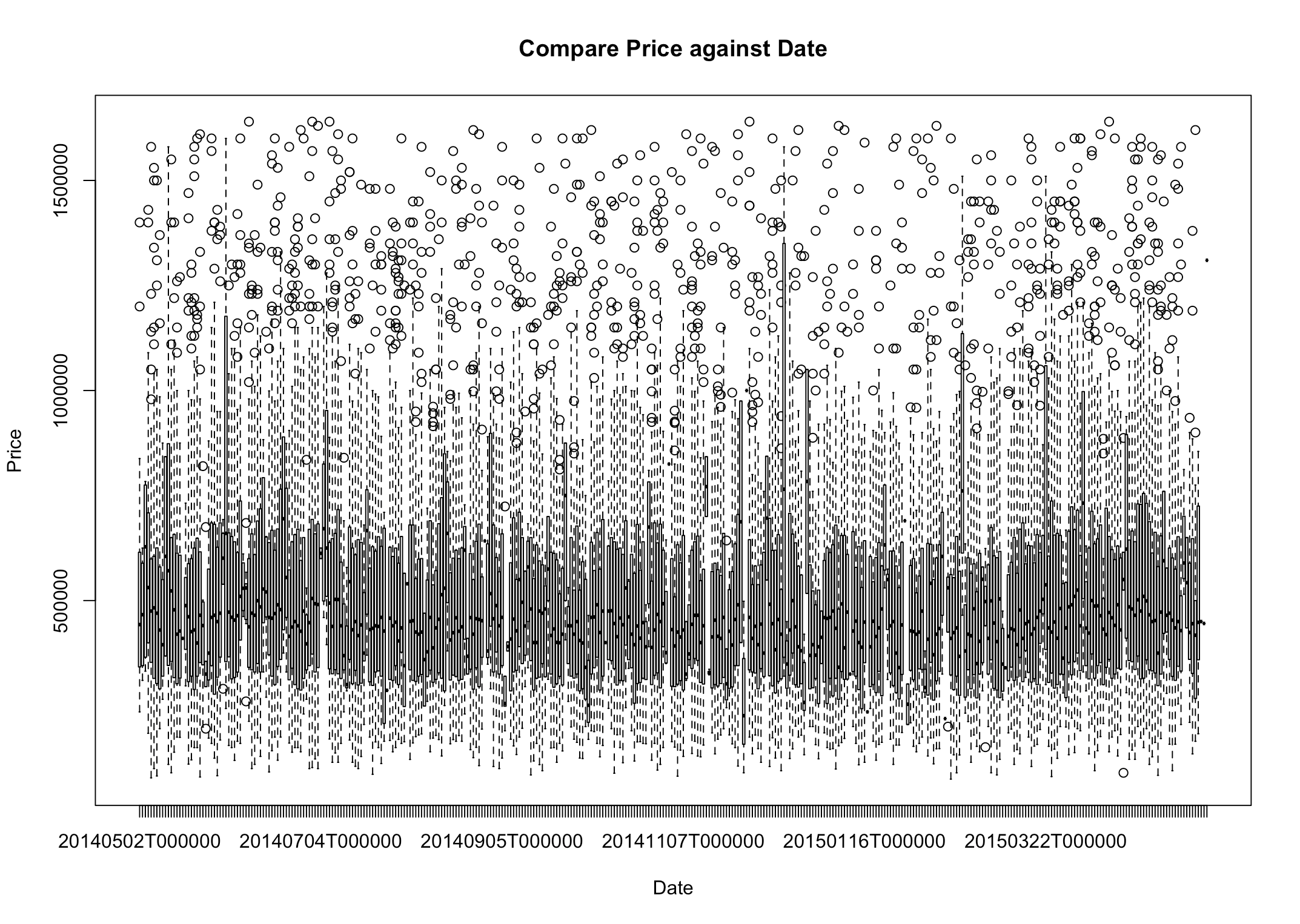


Figure 2: longitude and latitude in original dataset

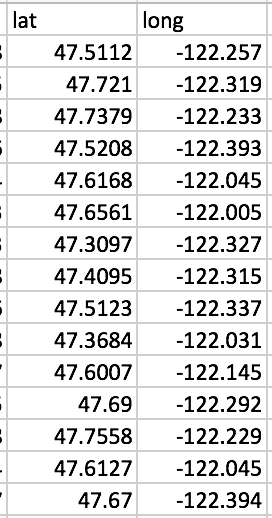


Figure 3: pairs correlation graph

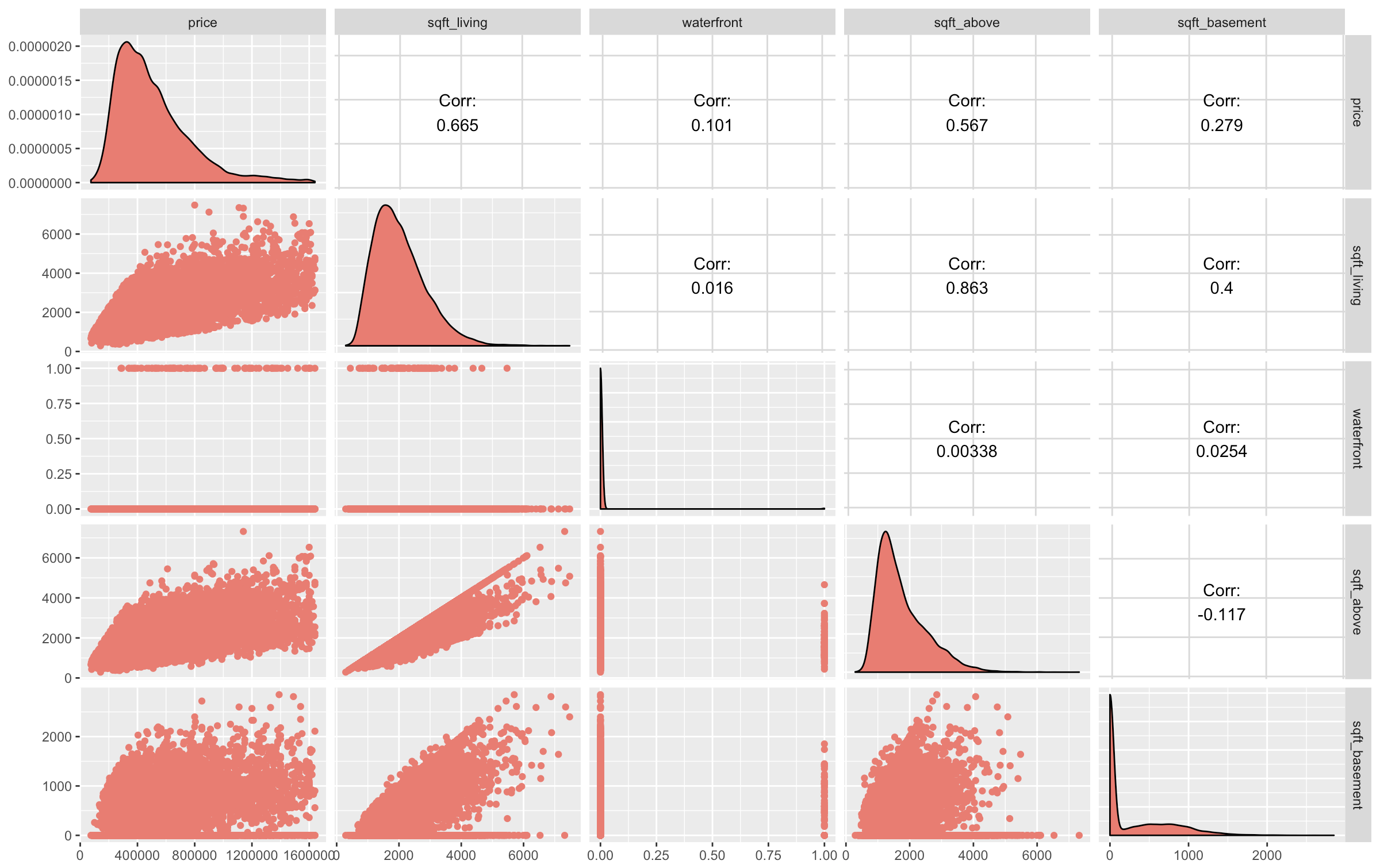


Figure 4: Linear Regression model result

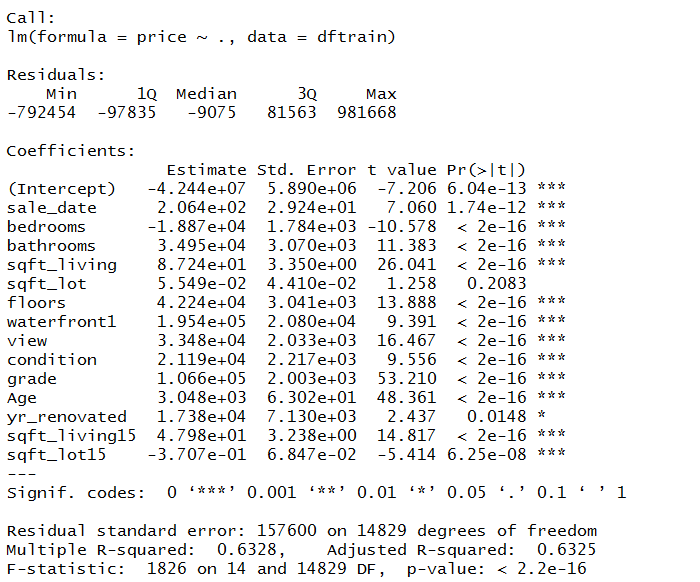


Figure 5: Plot of actual vs predicted

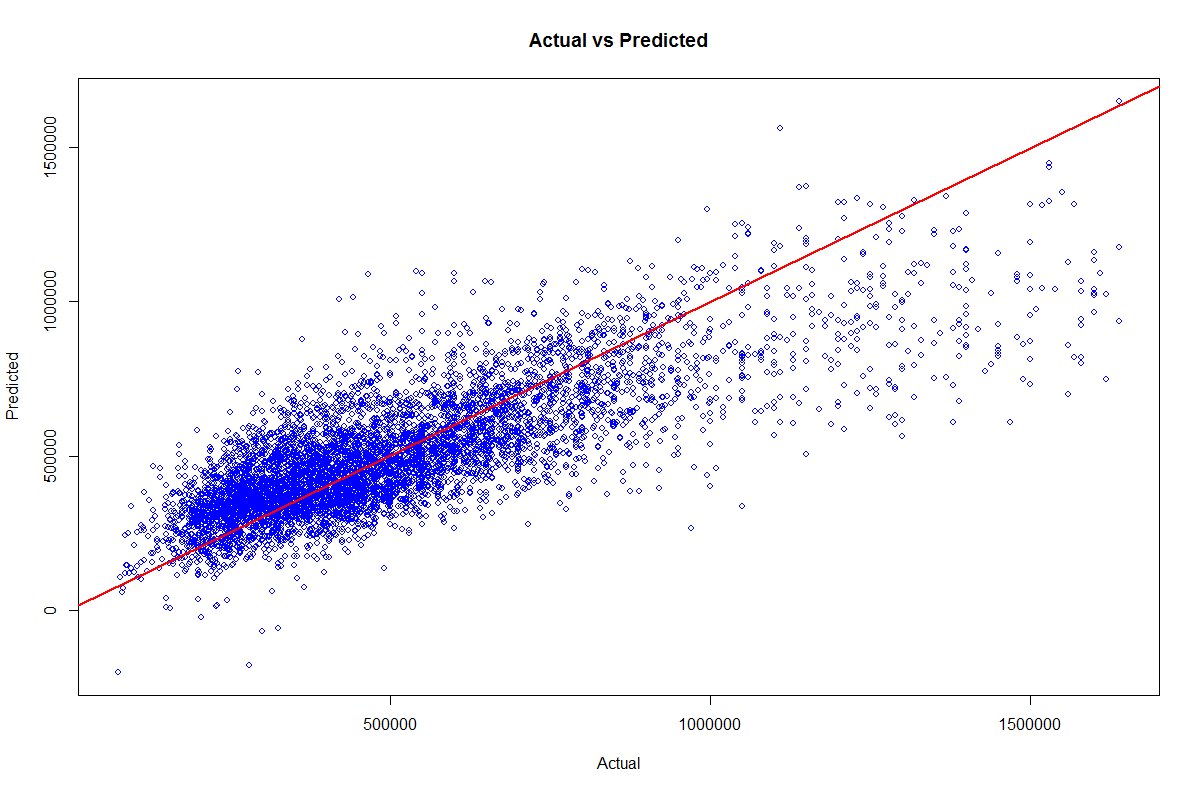


Figure 6: Lasso model coefficients

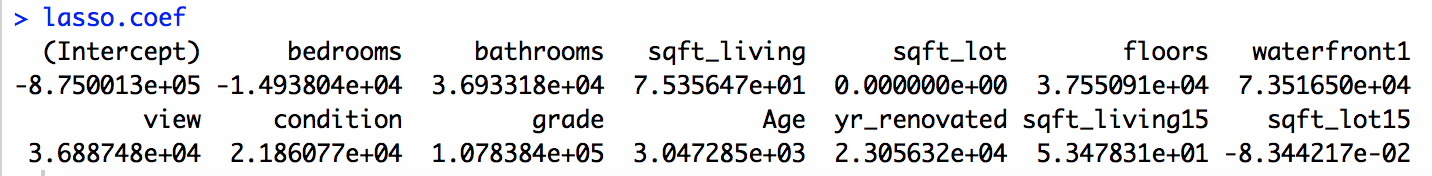


Figure 7: Lasso Trace plot

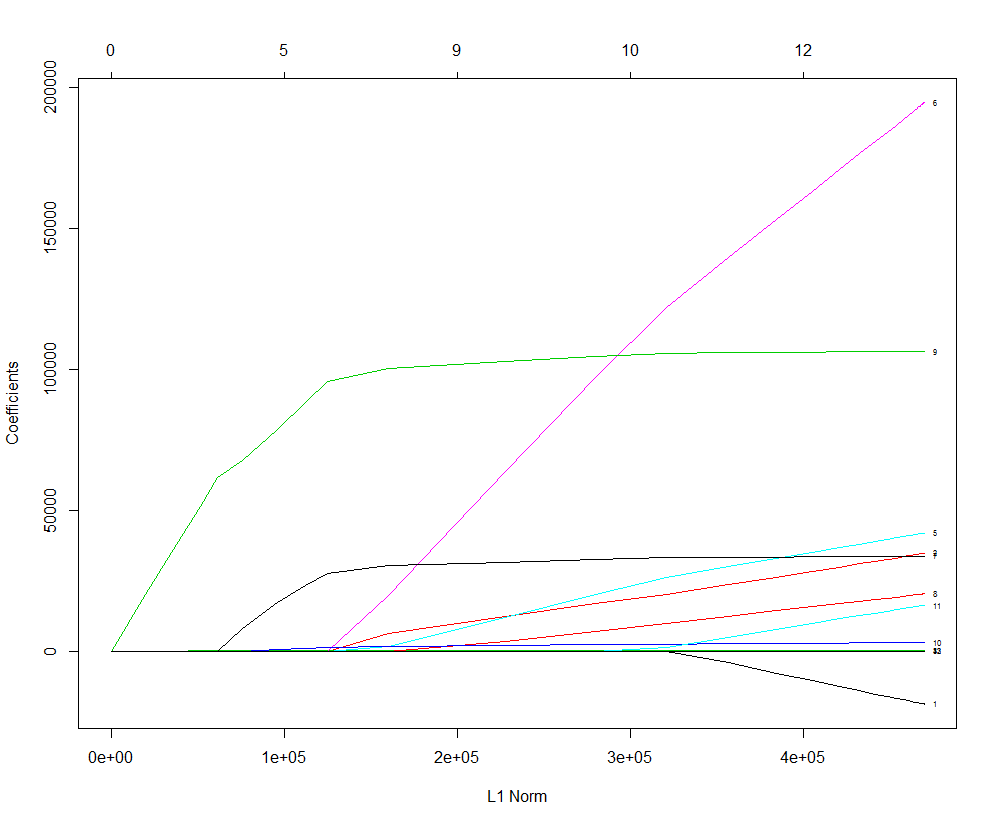


Figure 8: Regression tree model

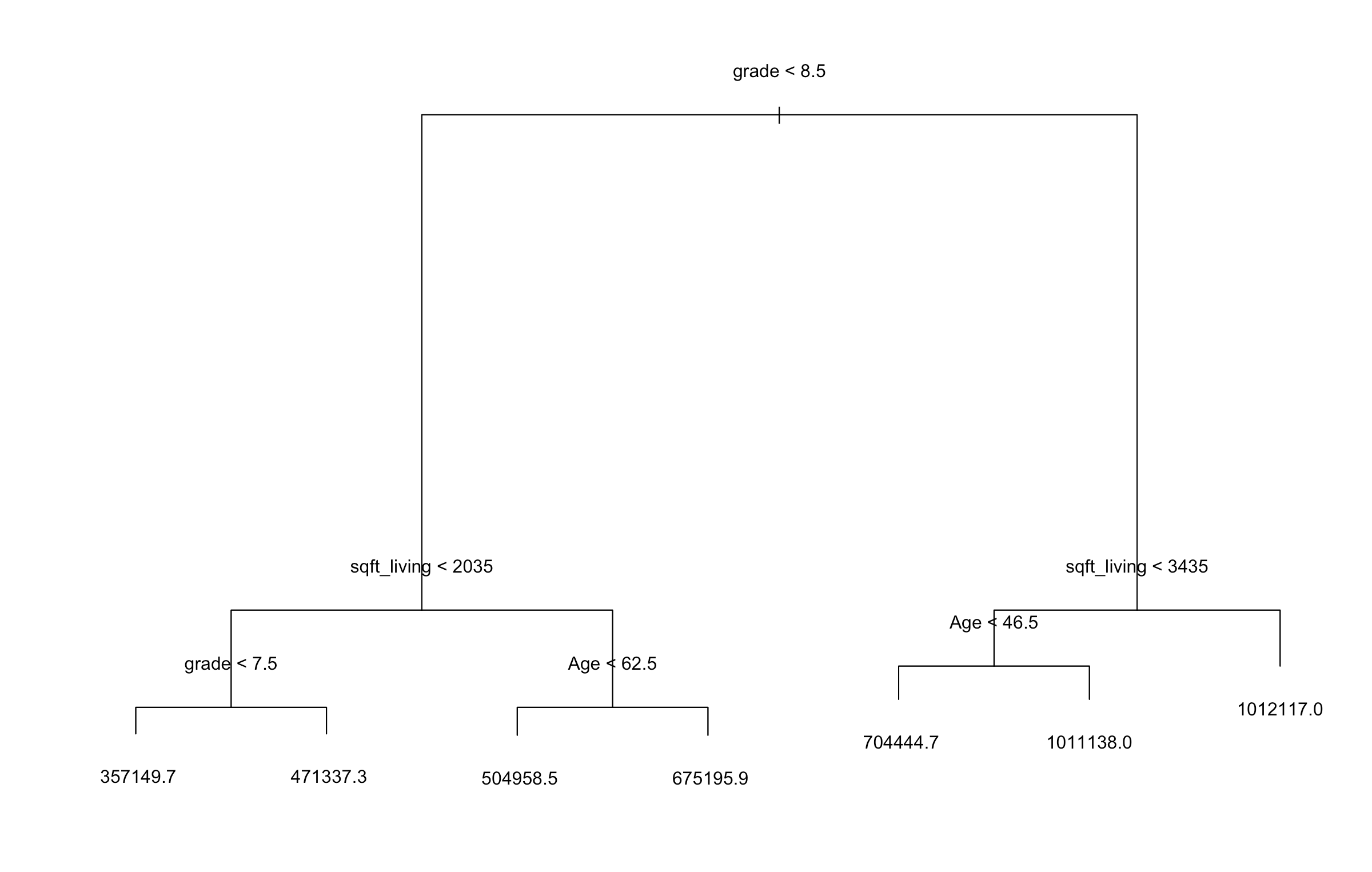


Figure 9: Prune tree

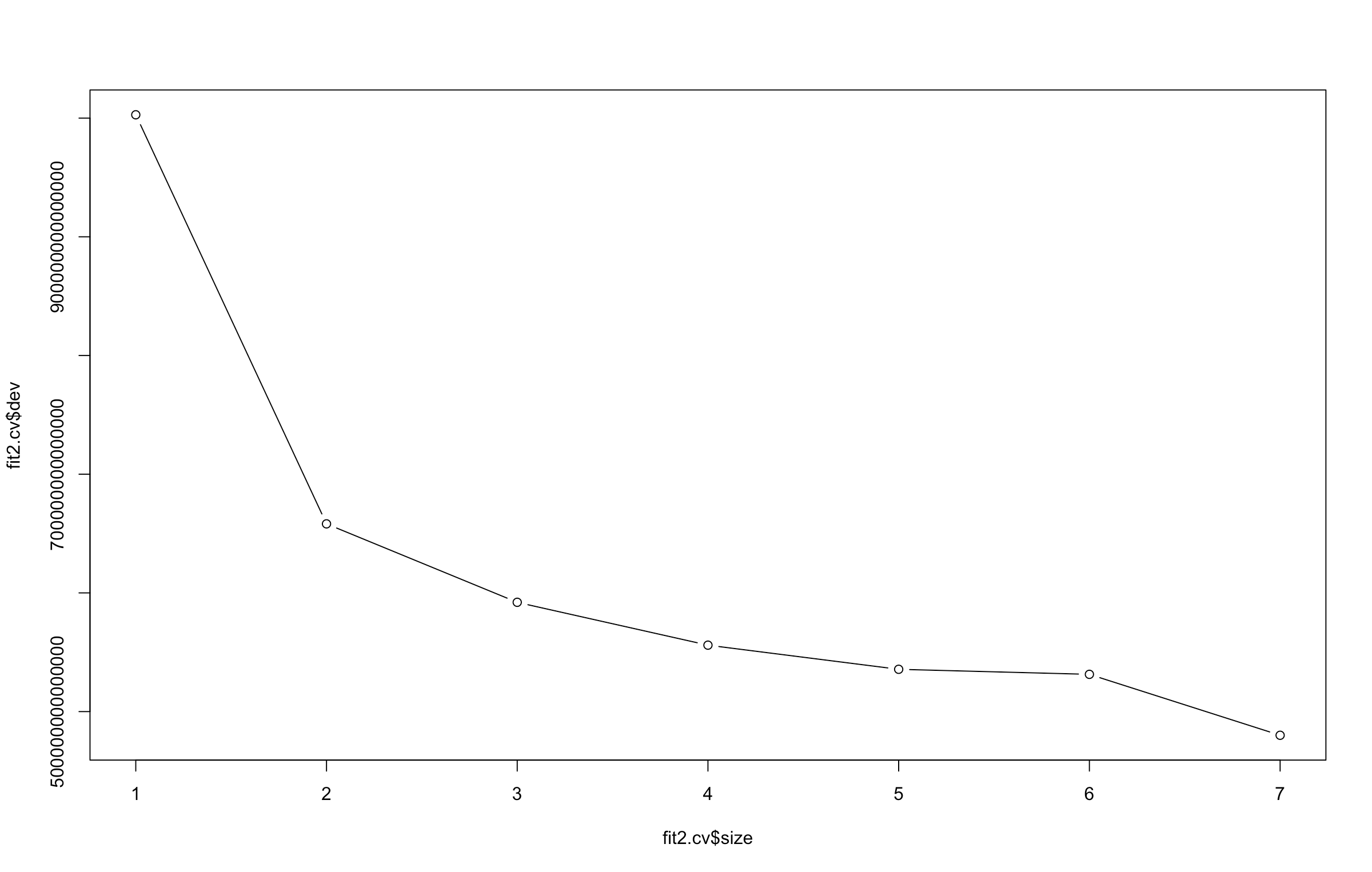


Figure 10: neural network

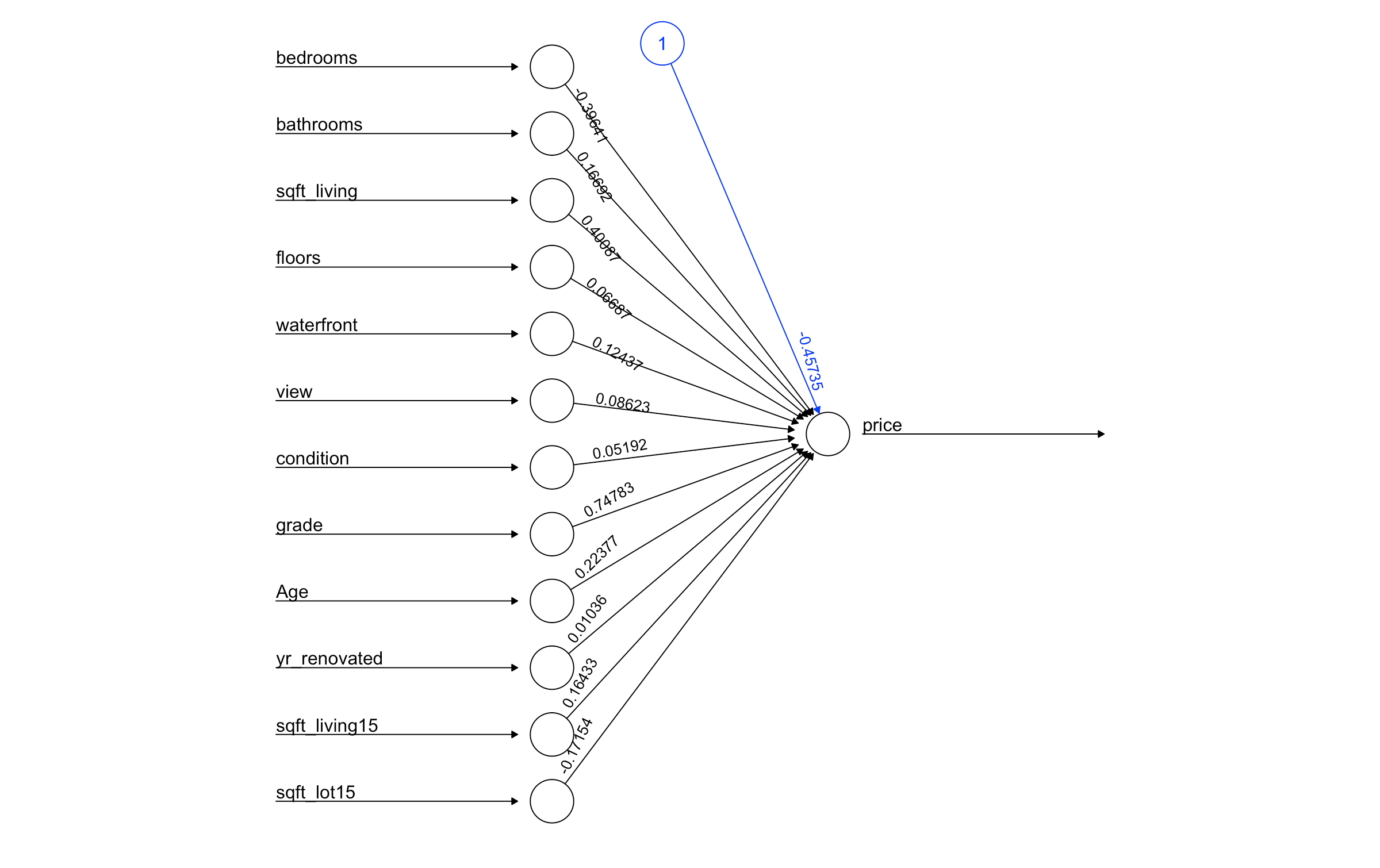


Table1: accuracy comparison between models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression | Lasso | Regression Tree | Random Forest |
| AE | 2,221.2178 | 2,239.1098 | 5,523.8211 | 1,132.4288 |
| MAE | 117,086.3290 | 117,037.2686 | 137,400.5400 | 101,818.6088 |
| RMSE | 158,093.5978 | 158,067.7688 | 182,565.3769 | 141,278.3726 |