Predicting Housing Rent Prices Using House Characteristics

Taron Kui, Iftikhar Ramnandan, Jenny Tran

Abstract

The real estate market's significant expansion in the past few decades has increased the demand for tools for property price prediction. The purpose of this study is to predict rental prices using the physical characteristics of houses (including the area, the number of rooms, the number of bathrooms, the number of parking spaces, the availability of furniture, allowance for pets, and the location of the house). We use data collected from the Brazilian house-renting market in 2020 to fit a pruned-tree regression model, a bagged model, and a random forest model. Our models did not support our hypothesis. Contrary to our prediction, not all the loaded-in factors were useful in predicting rental prices; however, some factors such as the area, the availability of furniture, and the location of the house do seem useful. Our project contributes to the existing literature on rental-price prediction and provides an additional model to predict real-time rental prices.

1. Introduction

The real estate industry has recently taken up a giant share of the financial market in almost every nation. Therefore, real estate price prediction has become a hot topic for both academic and market research. Case, Quigley, and Shiller (2005) show that “variations in real estate prices have had a significant effect on aggregate consumption in the US, in fact, more significant than the stock market, even before the recent volatility in the residential market” [1]. Housing prices, including property prices and rental fees, have been increasingly investigated using statistical and artificial intelligence methods. Experts of the field, as well as interested stakeholders, have invested a significant amount into developing price-predictive models. Balode and Kamols (2019) study the rental housing market in Riga (Latvia) and reveal that “distance from the city center, neighborhood safety, quality of housing and transport infrastructure, access to shopping malls, and employment opportunities" affect rent level [2]. A different study conducted by Yuan et. al（2017）based on data collected from Nanjing, China finds that “construction costs, household income, floor area and structure, transportation, market rents in the same district and public facilities" are statistically significant in predicting the rental price [3].

The existing literature suggests a wide range of attributes that can affect rental prices, resulting in increased difficulty in determining exactly which factors most significantly influence renting costs. It is indeed because of this reason, together with the interesting nature of the real estate market and its relations to other sectors of the economy as well as its realistic applications to society, that we decide to focus our project on rental prices. In this project, we will investigate the house-renting market in Brazil to predict rental prices using the house’s physical characteristics such as area, number of bedrooms, etc. via developing a regression tree model.

2. Data

This study uses a dataset collected by Rubens Junior and was published on Kaggle in 2020 [4]. The dataset included information of houses to rent in Brazil in 2020 posted online, which contains 13 attributes, including houses’ geographical locations, characteristics of these houses, renting prices, and taxes. The dataset contains 10962 observations and the unit of observation is an individual house to rent. We intend to generalize our analysis to study what factors might influence the prices of houses to rent with similar features.

Variables

The response variable in our analysis is the total rent prices of a house (measured in Brazillian Reals (R$)), excluding homeowner association tax, rent amount, property tax, and fire insurance. The price is a numerical variable ranging from R$450 to R$45000. To assess the relationship between housing rent prices and housing characteristics, we include area (a numerical variable indicating the property area of each house), rooms (a numerical variable showing the number of rooms), bathrooms (a numerical variable for the number of bathrooms), parking spaces (a numerical variable for the number of parking spaces), city (a categorical variable indicating whether the house is located at Belo Horizonte, Campinas, Porto Alegre, Rio de Janeiro, or São Paulo), animal (a binary variable reflecting whether the owner accepts pets),

and furniture (a binary variable representing whether the house is well-furnished or not) as the main explanatory variables.

3. Methodology

We begin our exploratory data analysis by calculating summary statistics and examining the multicollinearity between variables. We then randomly split our dataset into a training set with 6000 observations and a test set with 4692 observations. As our problem involves a continuous, non-discrete response, we create a pruned regression tree, a bagged model, and a random forest model using the training dataset. We then check the accuracy of these models using the testing dataset.

Pruned Regression Tree Model

We use a greedy approach via binary splitting to minimize the total sum of the squares with a penalty proportional to the size of the tree when creating the pruned regression tree. This methodology mimics that of LASSO regression in terms of its use of regularization. We firstly use the “rpart” package in R to create a full deep tree, consisting of all the attributes present in the data. Then, we conduct 10-folds cross validation to find the complexity parameter, a parameter that controls the size of the regression tree, that yields the lowest cross-validated error. We employ the “printcp” function in R to come up with the optimal prunings based on the complexity parameter of the tree (Figure 1). The best complexity parameter is 0.00137. We prune the tree with this complexity parameter to avoid any overfitting of the data, aiming to have an optimal regression tree with the least cross-validated error.

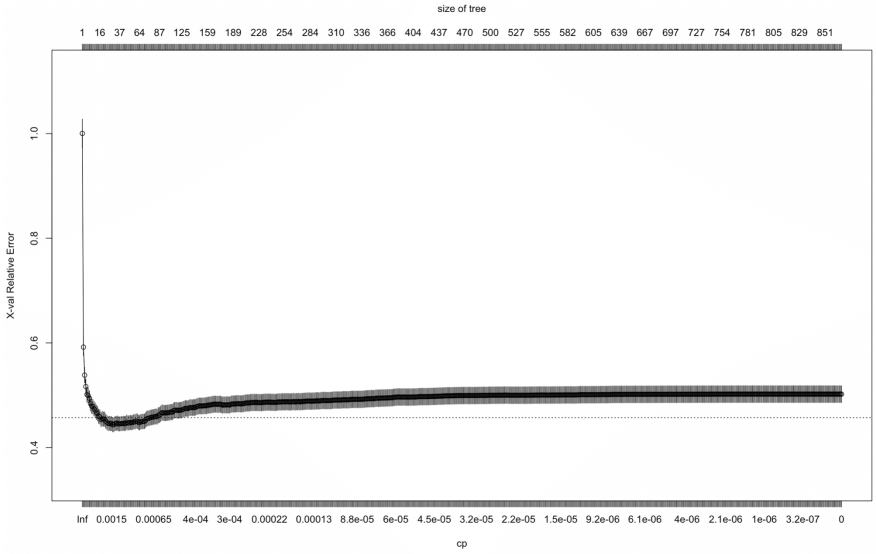


Figure 1: *The relationship between the complexity parameter and the cross-validated error.*

Bagged Model and Random Forest Model

We apply other methodologies such as bagging and random forests to the house rent data, using the “randomForest” package in R. Bagging is a special case of a random forest where the number of predictors randomly sampled, m is equal to the number of total predictors, p.

Therefore, we use the randomForest() function to perform both random forests and bagging. Given the continuous nature of our response variable, we create a random forest composed of regression trees. Rather than use a single pruned decision tree, we use 500 bagged unpruned trees, B = 500. By not pruning the trees we keep bias low and variance high, which is when bagging has the greatest effect. Growing a random forest is the same process as bagging, except that we use a smaller value of the number of predictors randomly sampled. When choosing the number of predictors randomly sampled, since some of our variables are strongly correlated, we employ 2 predictors.

Unfortunately, due to the bagging process, models that are normally perceived as interpretable are no longer so. However, we can still make inferences about how features are influencing our model using varImpPlot() function in R to capture the importance measure. For each tree, we compute the sum of the reduction of the loss function across all splits. We then aggregate this measure across all trees for each feature. The features with the largest average decrease in RSS are considered most important.

4. Results

The pruned regression tree model described above has a training MSE of 4489163 while the variance of the response variable, total rent, in the training dataset is 11051089. Thus, the (Training MSE)/ (Training response variable's variance) ratio is 0.406. Implementing this tree with the test dataset, the value of the test MSE is 5616494. The variance of total rent in the test dataset is 12337766. The (Test MSE)/ (Test response variance) ratio is 0.455. The $pseudo\ R^2$(1 - (Test MSE)/ (Test response variance)) of the pruned regression tree model is 0.545. As suggested by the Pruned Regression Tree, "area" is the most important variable (Figure 2). The pruned tree also shows that "furniture", "bedroom" are important variables splitting the tree (Appendix).

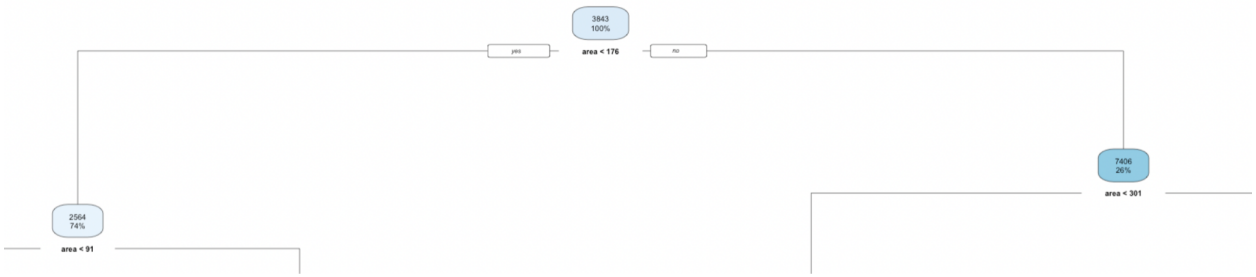


Figure 2: *The first split of the pruned regression tree (full pruned regression tree shown in the appendix)*

The bagged model has a test MSE of 5625727, which yields a (Test MSE)/ (Test response variance) ratio of 0.456. The $pseudo\ R^2$of the bagged model is 0.544. The variable importance chart of the bagged model shows that "area" is the most important variable in decreasing MSE, followed by "furniture" and "city" (Appendix).

The random forest model has a test MSE of 5139713, which yields a (Test MSE)/ (Test response variance) ratio of 0.417. The $pseudo\ R^2$ of the bagged model is 0.583. The variable importance chart of the bagged model shows that “city” is the most important variable in decreasing MSE, followed by “furniture” and “area” (Appendix).

The value of the $pseudo\ R^2$ of each model indicates that these models are moderately accurate in predicting the housing rent. Comparing the value of $pseudo\ R^2$ of each model, the random forest model yields the most accurate prediction of housing rents. We also create plots comparing the predicted response value of each model to the real value of the response variable (Appendix). These plots suggest that the predictions are moderately accurate with exception of predicting some outliers with high housing rent. These models also suggest that “area”, “furniture”, and “city” are the most important variables in decreasing the MSE of each model.

5. Discussion

While the model suggests that some physical characteristics of the house (which include the area, availability of furniture, and the location of the house) are important in determining the rental rate, it fails to lend support to our hypothesis because each house-characteristic variable has a different effect on the dependent variable, and the majority of them has a fairly small effect on rental prices. For example, we expect the number of rooms/bathrooms, parking spaces, and allowance for pets to be essential factors in determining rental rates, but our model suggests they are not.

The variable importance chart of these models indicates that “city”, “furniture”, and “area” are the most important variables in the prediction. The “city” variable represents which city the house is located in. The pruned regression tree model shows that houses that aren’t located in Belo Horizonte or Porto Alegre tend to have higher prices than houses with similar characteristics but that are located in these two cities. Comparing these cities, São Paulo, Campinas, and Rio de Janeiro are more urban than Belo Horizonte and Porto Alegre. This finding is consistent with our expectation that houses in metropolitan areas tend to be more expensive. Whether the house is well-furnished is also an important factor determining the housing rent. If the house is well-furnished, the housing rent tends to be higher than similar houses that are not well-furnished. Since well-furnished houses provide the tenants with more convenience and they do not necessarily need to spend much on buying furniture, the landlord can charge higher rents. The area of the house is also important in predicting the housing rent and houses with larger areas tend to have higher rents. When all other characteristics are similar, larger houses are more expensive to purchase than smaller houses. Therefore, it makes sense for landlords to charge higher prices as larger houses also give tenants more usable spaces. Surprisingly, our model shows that the number of parking spaces together with the allowance for pets are not important factors to change rental prices. Logically, the availability of parking spaces can significantly increase rental prices, especially in metropolitan areas where finding parking spaces can be very difficult. Also, allowing for animals would often increase

rental prices because such allowance tends to be accompanied by additional service fees for cleaning or compensating other tenants for the noises caused by pets.

Limitations

The first noticeable limitation of our model is there are outliers in the data. These outliers can result from errors in the data-collection process. Our data include 10692 observations and it was collected quite recently (in 2020); therefore, it can be expected that there are some mistakes in the collecting process. Furthermore, the outliers can be exceptional case studies for future research. In another model, analysts can separate these outliers from the rest of the data to study the predictive patterns of each group.

Another limitation is the high prediction error of our model. Out of the three models, our best one is the random forest model, which yields only roughly 58% accuracy in its explanation of the dependent variable. Rental prices may not be well-predicted using the house's physical characteristics but rather other factors that are in the literature (including access to transportation, neighborhood safety, etc.). It is also possible that the tree-based method is not suitable for predicting rental houses or house prices in general. In other words, the non-optimal nature of the tree-based method can be the cause of such a high error rate. In future work, researchers can attempt to apply other models (for example, a multivariate linear regression model or a polynomial regression) to predict rental prices. In addition to the high error rate, we also notice another restriction of our model – namely, high multicollinearity. Because most physical features of the house are logically related to another feature, it is expected that the variables would be highly collinear. In the future, this problem can be solved using the boosting model, which would increase randomness in the sampling process and, thus, help decrease some of the effects of the highly correlated data.

Last but not least, our data is collected specifically from the Brazillian house-renting market. Thus, we can only apply our prediction model to either the Brazilian rental market itself in the future or the rental market in countries in the same region or of the same size. If it is possible, we would be interested in building a similar model using another dataset collected from a different country (or region) to observe the model's ability to predict rental prices in a different context.

Conclusion

In this study, we explore the relationship between rental prices and rental houses' physical characteristics by analyzing data collected from the Brazilian real estate market. We use a pruned regression tree model, a bagged model, and a random forest model to predict rental prices using seven variables on house characteristics. Contrary to our expectations, many variables have small effects on rental rates; the three most important variables are the area, the availability of furniture, and the city in which the house is located. We did not expect the number of parking spaces and allowance for pets to have such little effect on rental prices. Our project both adds to the current literature on real-estate-rental price prediction and provides an alternative model to predict real-time rental prices in different markets worldwide.

References

[1] Case, K. E., Shiller, R. J., & Quigley, J. M. (2001, November 01). Comparing Wealth Effects: The Stock Market Versus the Housing Market. Retrieved from https://www.nber.org/papers/w8606

[2] Balode, S., & Kamols, U. (2019). Rental Housing Market in Riga: Price Determinants and Lesson Keys of Helsinki. Baltic Journal of Real Estate Economics and Construction Management, 7(1), 6-17. doi:10.2478/bjreecm-2019-0001

[3] Yuan, J., Zheng, X., You, J., & Skibniewski, M. (2017). Identifying Critical Factors Influencing the Rents of Public Rental Housing Delivery by PPPs: The Case of Nanjing. Sustainability, 9(3), 345. doi:10.3390/su9030345

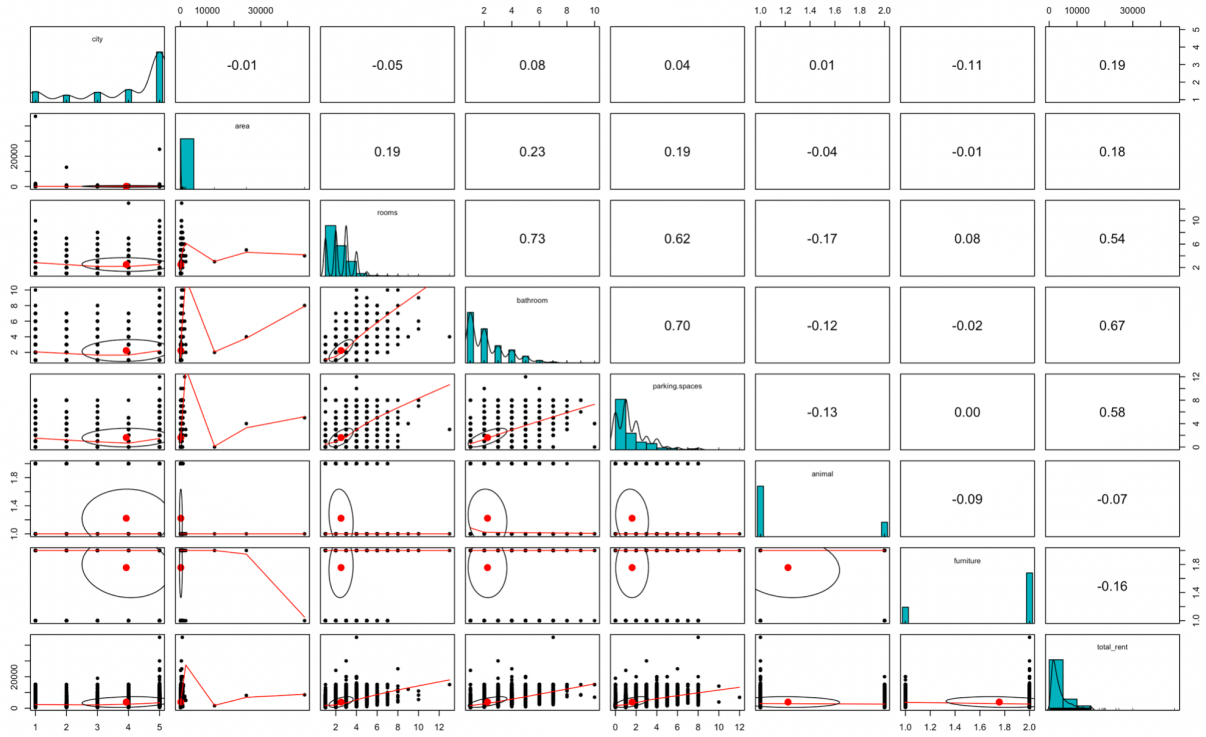
[4] https://www.kaggle.com/rubenssjr/brasilian-houses-to-rent?select=houses_to_rent_v2.csv

Appendix

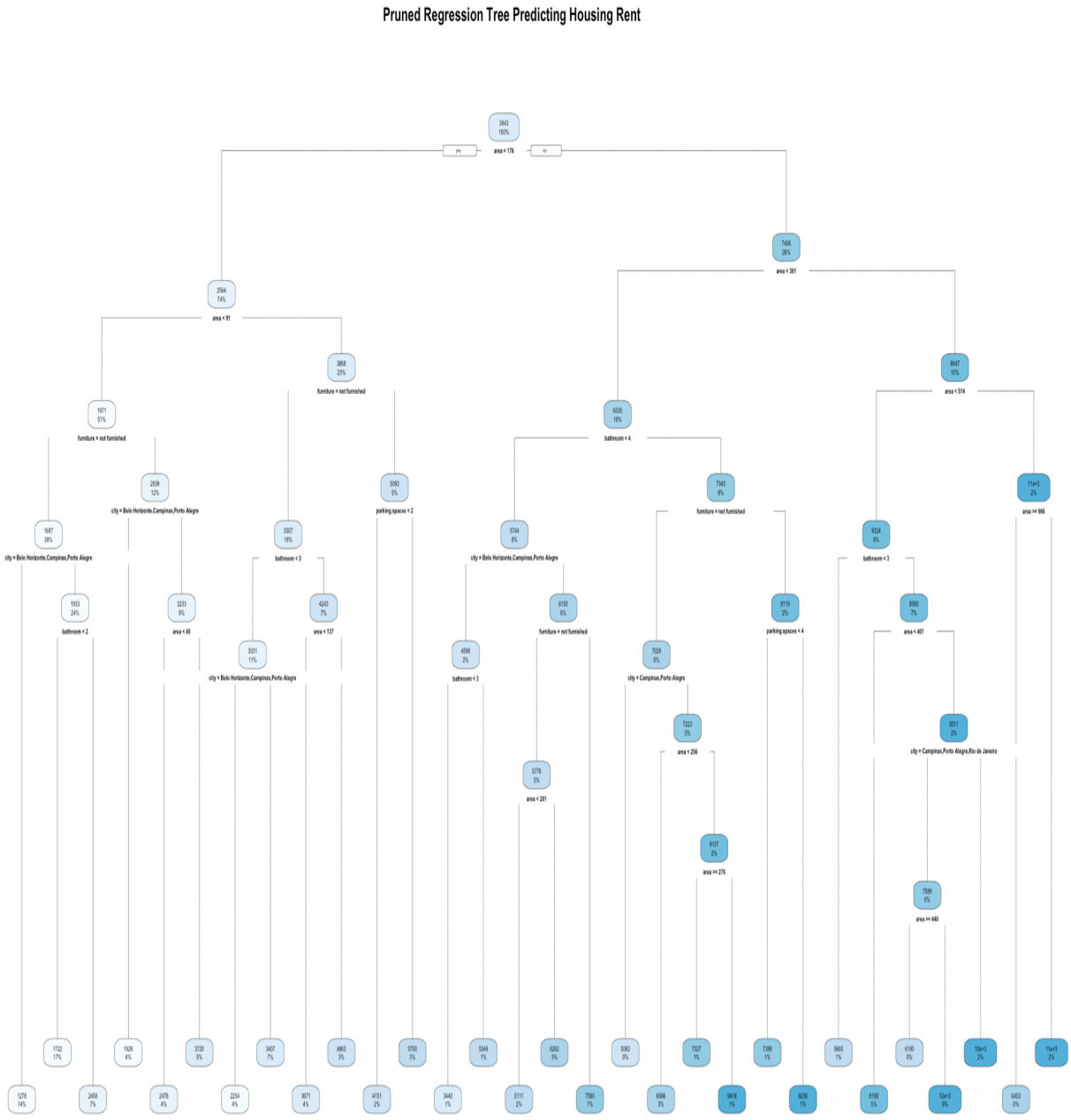
Summary Statistics

```
> summary(houses_rent)
            city           area            rooms             bathroom      parking.spaces
 Belo Horizonte:1258   Min.   :   11.0   Min.   : 1.000   Min.   : 1.000   Min.   : 0.000
 Campinas      : 853   1st Qu.:   56.0   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 0.000
 Porto Alegre  :1193   Median :   90.0   Median : 2.000   Median : 2.000   Median : 1.000
 Rio de Janeiro:1501   Mean   :  149.2   Mean   : 2.506   Mean   : 2.237   Mean   : 1.609
 São Paulo     :5887   3rd Qu.:  182.0   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 2.000
                       Max.   :46335.0   Max.   :13.000   Max.   :10.000   Max.   :12.000
      animal             furniture      total_rent
 acept    :8316   furnished    :2606   Min.   :  450
 not acept:2376   not furnished:8086   1st Qu.: 1530
                                       Median : 2661
                                       Mean   : 3896
                                       3rd Qu.: 5000
                                       Max.   :45000
```

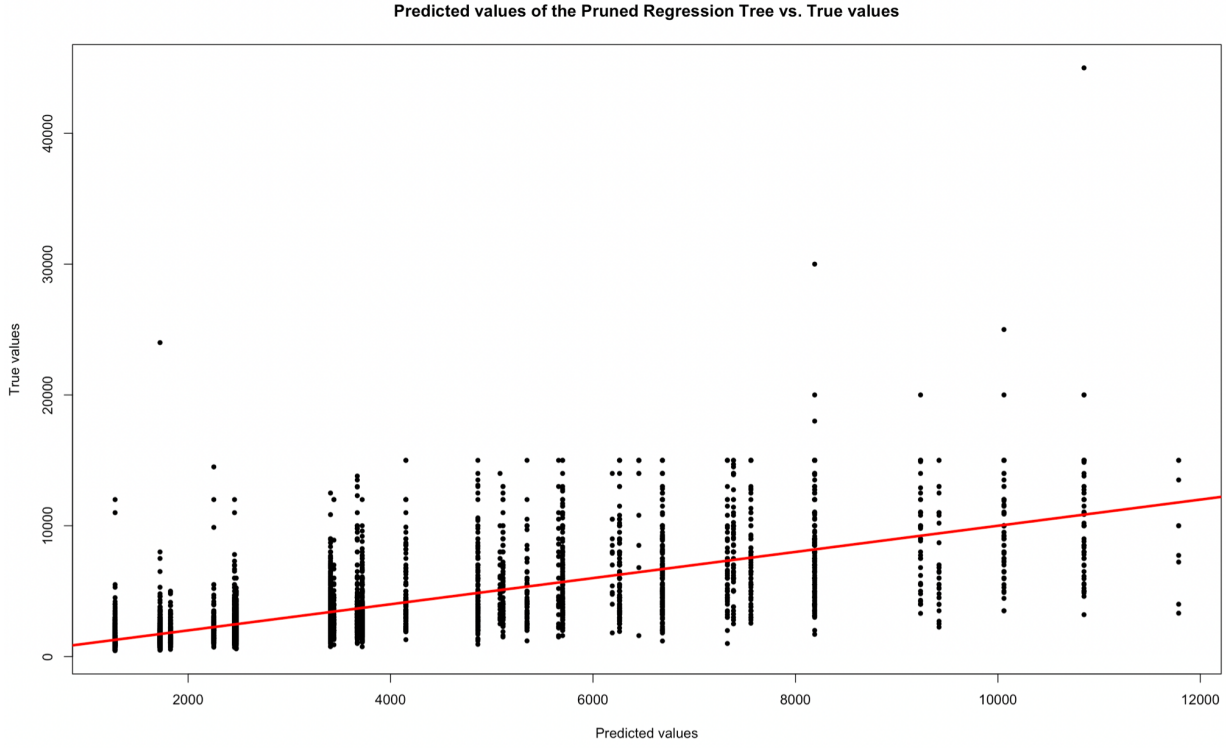
Check for Multicollinearity in the Dataset

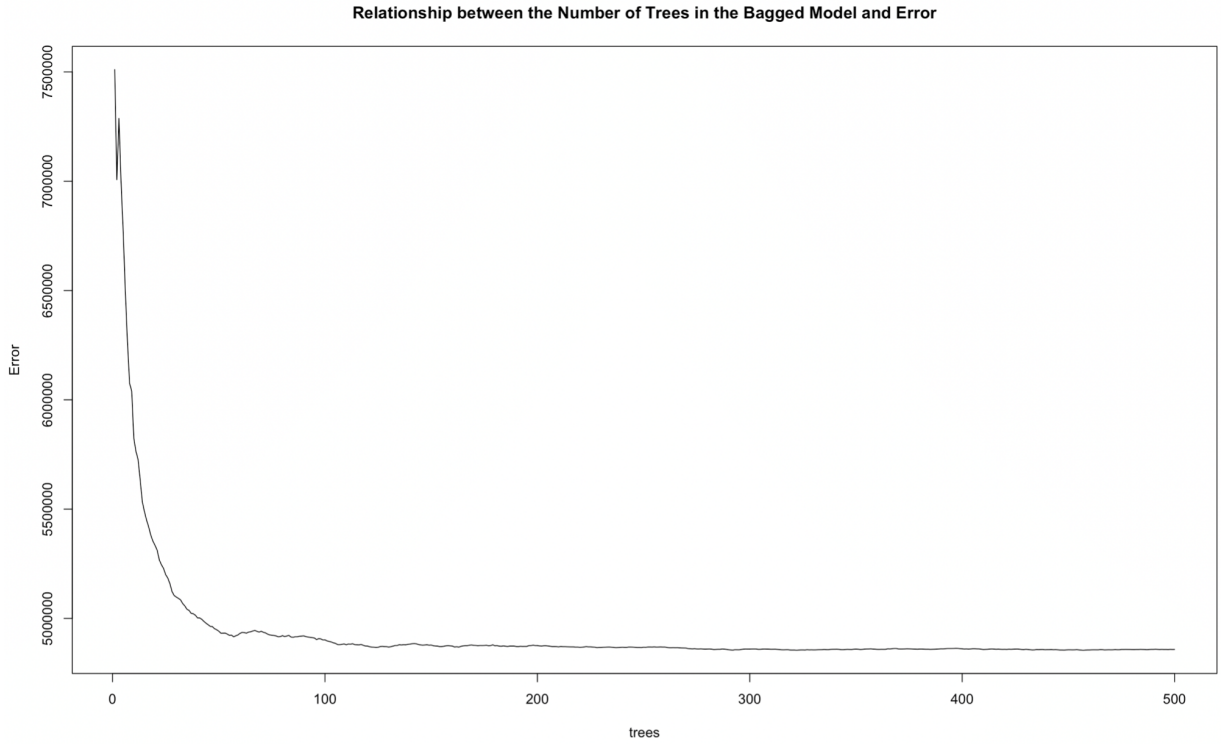


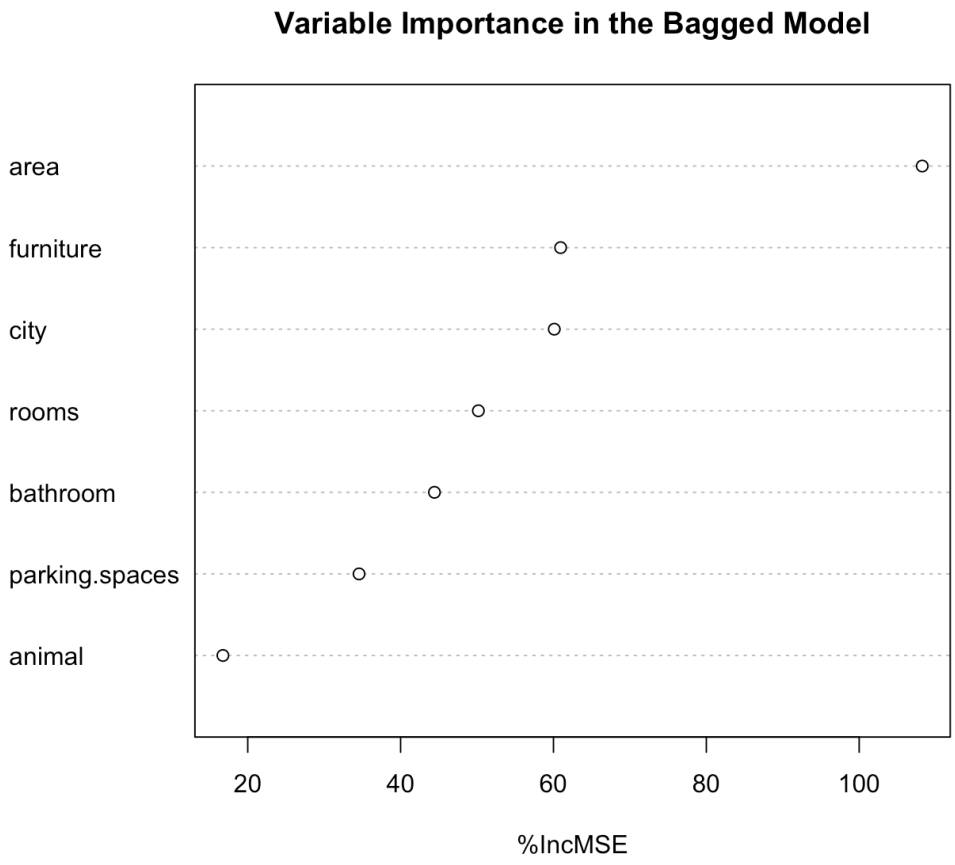
Pruned Regression Tree Predicting Housing Rent

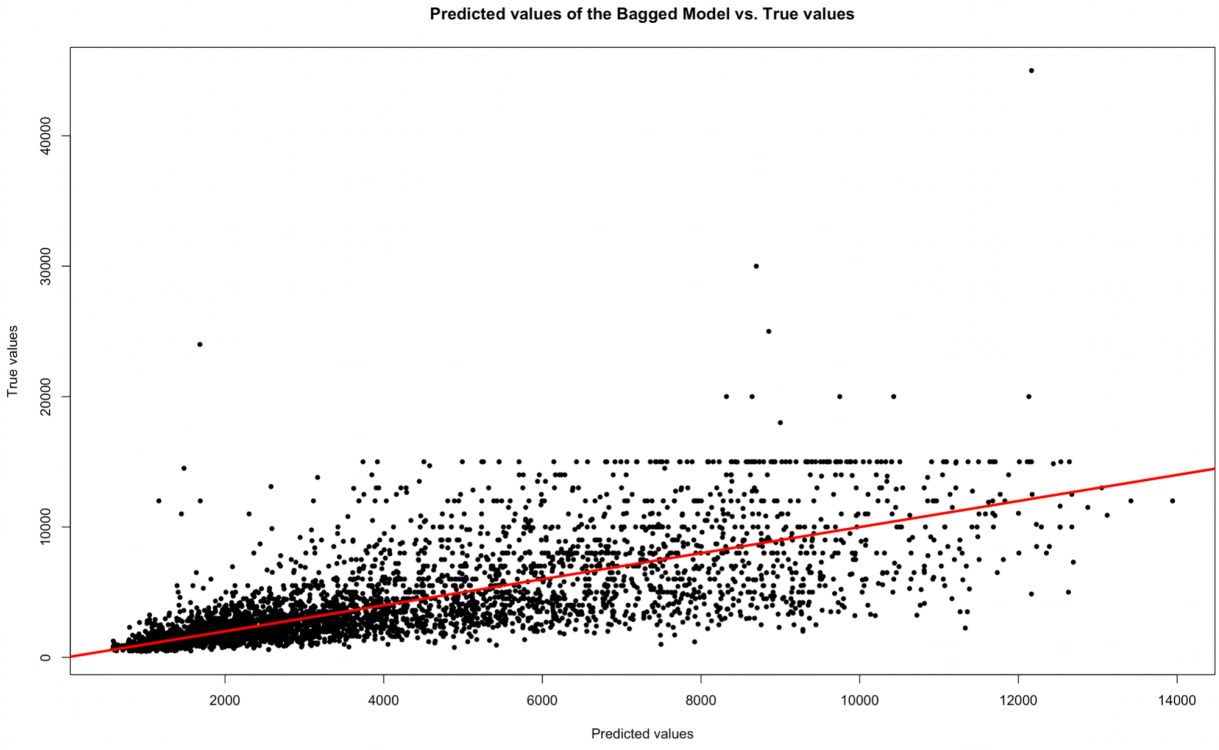


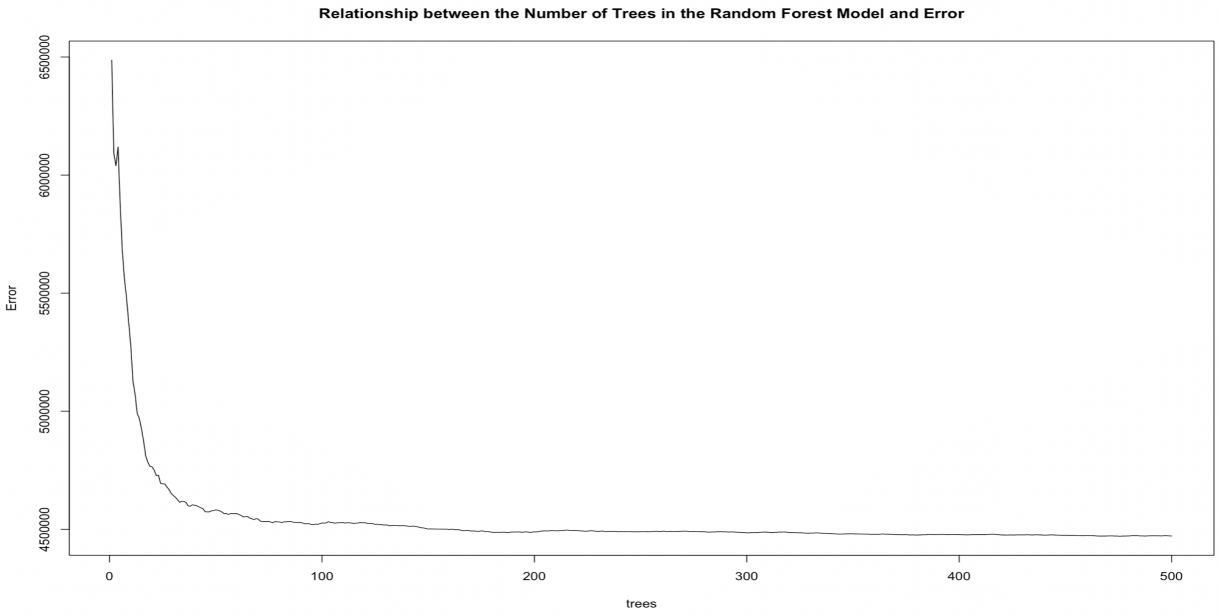
```
> print(prune_rent)
n= 6000

node), split, n, deviance, yval
      * denotes terminal node

  1) root 6000 66295300000  3843.372
    2) area< 175.5 4415 15557390000  2564.241
      4) area< 90.5 3034  4773703000  1970.977
        8) furniture=not furnished 2285  2013021000  1686.679
         16) city=Belo Horizonte,Campinas,Porto Alegre 860   278550600  1278.392 *
         17) city=Rio de Janeiro,São Paulo 1425  1504591000  1933.084
           34) bathroom< 1.5 1017   772689300  1722.389 *
           35) bathroom>=1.5 408   574220400  2458.270 *
        9) furniture=furnished 749  2012567000  2838.295
         18) city=Belo Horizonte,Campinas,Porto Alegre 210   137793000  1825.719 *
         19) city=Rio de Janeiro,São Paulo 539  1575570000  3232.805
           38) area< 44.5 211   170127700  2475.692 *
           39) area>=44.5 328  1206687000  3719.851 *
      5) area>=90.5 1381  7369804000  3867.618
       10) furniture=not furnished 1067  4137898000  3506.977
         20) bathroom< 2.5 648  1812288000  3031.289
           40) city=Belo Horizonte,Campinas,Porto Alegre 211   353365000  2253.863 *
           41) city=Rio de Janeiro,São Paulo 437  1269822000  3406.659 *
         21) bathroom>=2.5 419  1952214000  4242.647
           42) area< 136.5 218   631985800  3671.064 *
           43) area>=136.5 201  1171760000  4862.572 *
       11) furniture=furnished 314  2621556000  5093.108
         22) parking< 1.5 123   703704800  4151.228 *
         23) parking>=1.5 191  1738464000  5699.660
           46) area< 145.5 124   843422300  5192.839 *
           47) area>=145.5 67   804240600  6637.657 *
    3) area>=175.5 1585 23392610000  7406.377
      6) area< 301 975 11919180000  6504.978
       12) bathroom< 3.5 511  5523674000  5743.853
         24) city=Belo Horizonte,Campinas,Porto Alegre 143  1443103000  4587.629
           48) bathroom< 2.5 57   164208000  3440.368 *
           49) bathroom>=2.5 86  1154146000  5348.023
             98) area< 192.5 20    49409940  3556.250 *
             99) area>=192.5 66  1021070000  5890.985
              198) city=Campinas,Porto Alegre 28   247145400  4532.679 *
              199) city=Belo Horizonte 38   684199600  6891.842 *
         25) city=Rio de Janeiro,São Paulo 368  3815115000  6193.147
           50) furniture=not furnished 282  2720752000  5776.309
            100) area< 201 119   667244300  5110.739 *
            101) area>=201 163  1962307000  6262.215 *
           51) furniture=furnished 86   884694200  7559.988 *
       13) bathroom>=3.5 464  5773461000  7343.200
         26) furniture=not furnished 330  3991012000  7028.267
           52) city=Campinas,Porto Alegre 30   134817600  5081.533 *
           53) city=Belo Horizonte,Rio de Janeiro,São Paulo 300  3731132000  7222.940
            106) area< 256 189  1988483000  6685.852 *
            107) area>=256 111  1595298000  8137.441
              214) area>=275.5 68   862962600  7327.382 *
              215) area< 275.5 43   617150500  9418.465 *
         27) furniture=furnished 134  1669114000  8118.784
           54) parking< 3.5 81   855729200  7387.877 *
           55) parking>=3.5 53   703979200  9235.830 *
      7) area>=301 610  9414993000  8847.138
       14) area< 514 468  6763486000  8323.795
         28) bathroom< 2.5 41   359903800  5659.756 *
         29) bathroom>=2.5 427  6084662000  8579.593
           58) area< 401 301  4025827000  8189.748 *
           59) area>=401 126  1903807000  9510.889
            118) city=Campinas,Porto Alegre,Rio de Janeiro 28   424413400  7589.286
              236) area>=439.5 21   186624700  6190.476 *
              237) area< 439.5 7    73428570 11785.710 *
            119) city=Belo Horizonte,São Paulo 98  1346462000 10059.920 *
       15) area>=514 142  2100877000 10571.960
         30) area>=965.5 9    11294360  6452.889 *
         31) area< 965.5 133  1926549000 10850.690 *
```

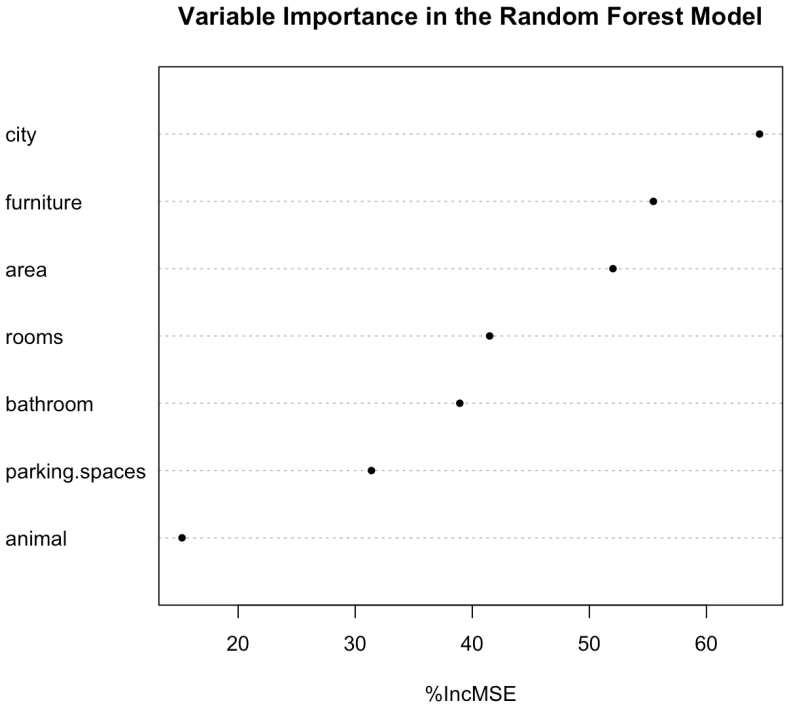

Predicted values of the Pruned Regression Tree vs. True values
True values
0
10000
20000
30000
40000
2000
4000
6000
8000
10000
12000
Predicted values

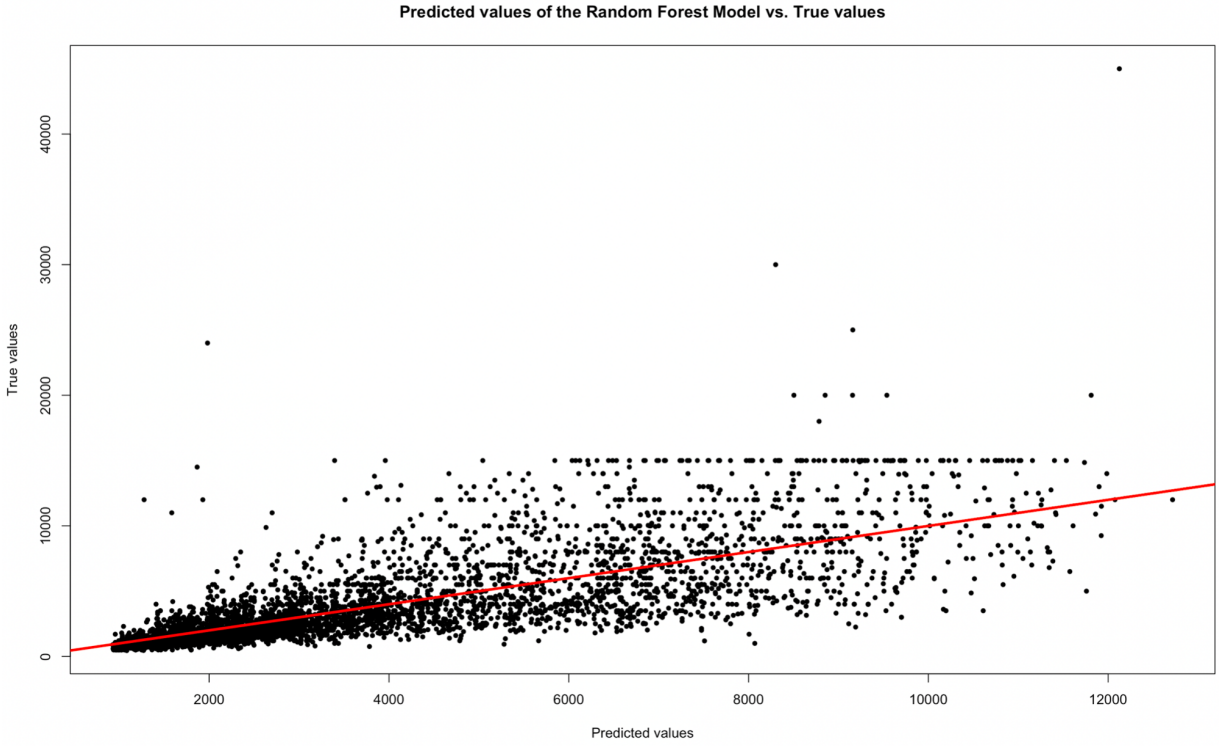

Relationship between the Number of Trees in the Bagged Model and Error
Error
5000000
5500000
6000000
6500000
7000000
7500000
0
100
200
300
400
500
trees


Variable Importance in the Bagged Model
area
furniture
city
rooms
bathroom
parking.spaces
animal
20
40
60
80
100
%IncMSE


Predicted values of the Bagged Model vs. True values
True values
0
10000
20000
30000
40000
2000
4000
6000
8000
10000
12000
14000
Predicted values


Relationship between the Number of Trees in the Random Forest Model and Error
Error
6500000
6000000
5500000
5000000
4500000
0
100
200
300
400
500
trees


Variable Importance in the Random Forest Model
city
furniture
area
rooms
bathroom
parking.spaces
animal
20
30
40
50
60
%IncMSE


Predicted values of the Random Forest Model vs. True values
True values
0
10000
20000
30000
40000
2000
4000
6000
8000
10000
12000
Predicted values