House Rent Prediction Method Based on Decision Tree: Take India as an Example

James Jiayu Guo

Soong Ching Ling Domesticated High School Shanghai, China

* Corresponding Author Email: 202660132@stu.scls-sh.org

Abstract. In this essay, this paper use decision trees to predict house rent and compare it to other linear regressions to find out why decision trees are a good fit for house rent and use house rent data 2023. This paper used the data set from Kaggle which has many factors and is up-to-date for my project, employing some visualization methods on Python to show these factors and their graphs so we can understand some important circumstances about house rent in India. This paper employed three different methods of training models to train them to predict their rent. In the last section, this paper use R2_score, RMSE, MAE and MSE to compare their conclusions. Thus, this paper show why decision trees are the best model for predicting house rent. After all process, this paper proposes that data rent price prediction is very important, because many people don't have a good obedience on other things like other people's advice or other black heart house property agents.

Keywords: House Rent Prediction, Decision Tree, Time Series Prediction.

1. Introduction

The housing rental market is complex, and there are many factors: size of the house, locations in both locations of the houses in downtown compared to suburb area, first-tier cities versus third-tier cities that influence how the rental markets shakes out for renters and for landlords.[1] There are many key events in the history of the rental market, including the introducing of rent control in India in 1918 and the Rent Control Act of 1948 that covered the rights of landlords. These events caused policy changes like the introduction of control, which naturally favors tenants over landlords, and has led to India's reputation as being far more friendly to tenants than to landlords. Beyond these historical impacts, other factors that influence the rental market include a lack of affordable housing, imbalances between supply and demand, fluctuations in the market in price, and challenges related to tenant rights and protections. Thus, the prediction of house rent prices can be difficult, and there are many methods to predict future rent prices in the field of data analysis. This essay compares different methods and models and evaluates their accuracy and convenience.

Using predict learning to make a decision for rent property is very increasingly useful to buyers who want to rent a place or seller who want to correctly price rent probably. Most people make a decision on renting and these people generally rely on other people's opinions or realistic agents' recommendation. Employing this machine learning prediction model into these people, it can be more easily to rely on machine not other people. Some people use other ways like comparing house prices in similar locations or area to put a similar price or do a lot of research on internet to make a prices. However, these ways are not accurately to make a price for these houses' rent.[2]

Currently, there are a couple of strategies that are employed to predict prices for the housing and rental market. These strategies usually involve linear regression or random forest regression models. Researchers have typically used factors such as the area under study, methods of rent collection, infrastructure, transportation, and orientation (Ming 2020 when decision tree model) when designing their models. These studies typically release results where the models' fit 90 percent of the data. Additionally, the model's RMSE is under 15, MAE is under 7, and MSE is under 200. This is a pretty good model to predict house rent. Decision tree is a model that can combine many factors into one model, with ability to basically combine many complex nonlinear relationships to model.

This essay utilizes the Klaggle data set in Python to draw up different data visualization methods, which allows viewers to better easily access the data and promotes analysis. This paper is used for

predicting annual dataset. This work then compares the different outcomes of each data analysis method and determines that decision tree modeling better than linear regressions for predicting prices in the housing and rental markets. Decision tree modeling was more accurate than the linear regression, and it was also more convenient than linear regressions; both of these facts make decision tree modeling the best kind of analysis for predicting Indian rental market prices.

2. Experiment

2.1. Methods

The study used pandas to import a CSV from data scientist Sourav Banerjee and to see whether the data set had the correct information or not. This essay deleted columns of specific locations because it is too hard to work on decision tree model. Then using some method on data visualization were imported for data analysis use, and the data was divided into training and test sets. [3]The model was trained on the training sets and features were extracted to categorize, because training model need a lot of data to help them to learn and predict, and other test-set can help us to understand this model is good or bad. The method used was decision tree methodology.

For this experiment, first a data screening was done in order to check that the data did not have any missing information. The data did not have any missing information. The paper also did a data analysis in order to analyze the distribution of the data, the features of the data, and pinpoint columns that might be useful in predicting house rent and columns that would not be useful in predicting house rent. A decision tree model was trained to combine several factors of the data to see if the model had a high degree of accuracy when it came to fitting with the data. Finally, other methods were introduced to compare them with the new decision tree model to check whether the decision tree was more accurate than the other models. [4]

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \widehat{y_{R_j}} \right)^2 \tag{1}$$

Decision tree methodology is a “commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable” (Song 2015). Researchers typically use decision trees because they are easy to manage, as they break down large data into smaller parts that are easier to understand. Decision trees use different factors to reduce rents. ‘Visualization technique is used to display the first decision tree. At least one test example is selected from the test set and input into the first decision tree for testing, and the classification path of at least one test example is generated. [5][6][7]

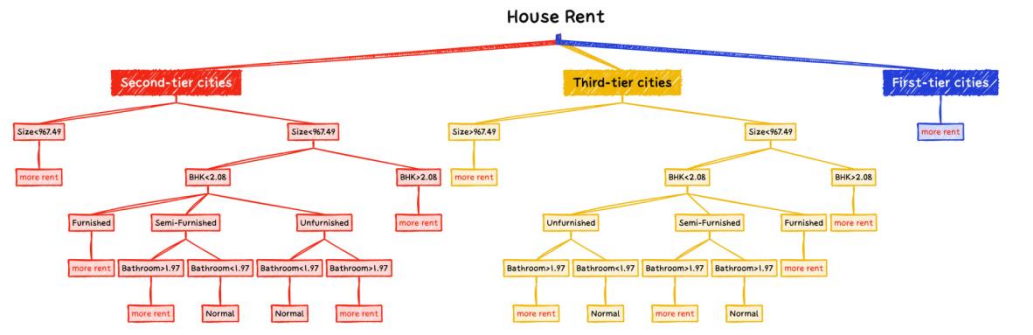


Fig. 1 Mind-mapping of my project

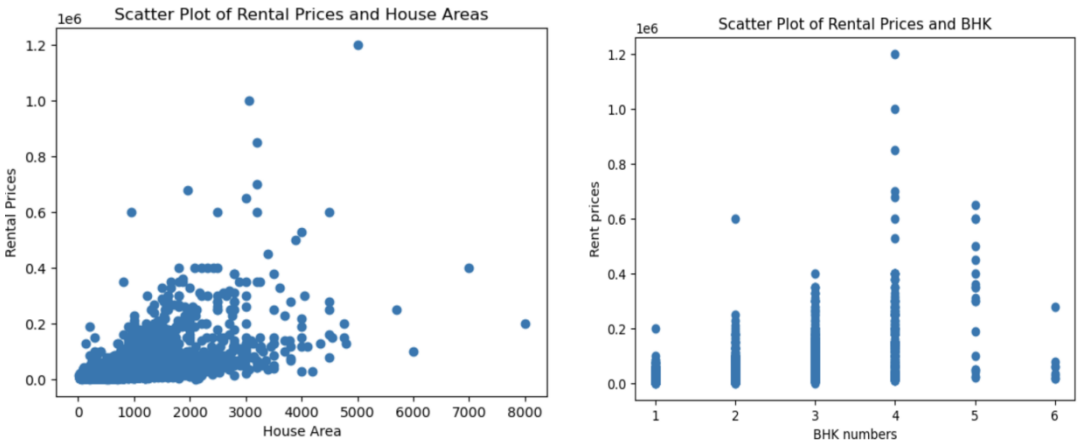
According to this mind map (Fig. 1), decision trees are very easy to apply. This paper wants to choose decision trees for my project. Decision trees can compress many factors into one analysis, whereas linear regressions cannot. This also enhances reproducibility because you do not need to run multiple tests; all information is collected in one tree. I compared decision trees with linear regressions and other methods to show the greater effectiveness of decision trees.

2.2. Data Preparation

This paper downloaded this dataset from kaggle.com. This essay chose this data because it is an updated dataset, and it has many columns that I can use (factors). Some of the factors in the data set include rent prices, the furnished or unfurnished state of the apartment the location, what kind of tenant was preferred, and how many bathrooms the rental had, among others. According to this paper's data-preparation, I used several techniques like histograms, boxplots and pie charts to visualize my data. I deleted some columns that may not help our prediction. Cleaning up the data is an essential part of preparing data for analysis, and there were some parts of the kaggle dataset that would not have been helpful towards my analysis, so they were deleted.

According to data preparation, my dataset has 4746 rows and 11 columns. While unfortunately, the sample size of the data set is low, it is still a good data set because of its recency. I find that this dataset has no missing values in it, so it is complete.

2.3. Data Analysis



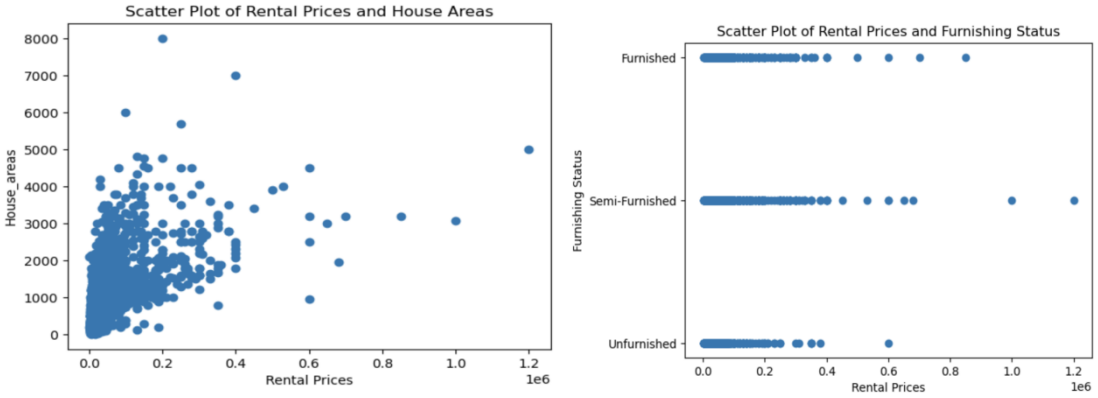
(a) (b)

Fig. 2 (a)Scatter Plot of Rental Prices (millions of dollars) and House Areas (Sq ft). (b)Scatter Plot of Number of BHK in Houses Available for Rent

According to Figure 2, people can see that most of the houses are around 1,000-3,000 square feet in area. Most of the prices are under 0.5 million dollars. We can conclude that the cost is very cheap and only some of the houses are very expensive. Thus, size (area) is a very important factor to influence the prices, because with a greater size (area), rent will be more expensive.

BHK is a convenient way to talk about Numbers of Bedrooms, Hall and Kitchen, because if you don't use BHK, you need to talk about "number of bedrooms", "number of halls" and "number of kitchens" separately. It will increase the work of models which waste time for model to predict.

According to Figure 3, we can see that the most common BHK number is two, and just a few are one and three. This shows that there are fewer people in one house; that is, no more than three or four people per house. Therefore, BHK is a very important factor that affects rent prices. It tells us how many rooms are in the house. BHK numbers see obvious in 1-4 BHK numbers of houses.

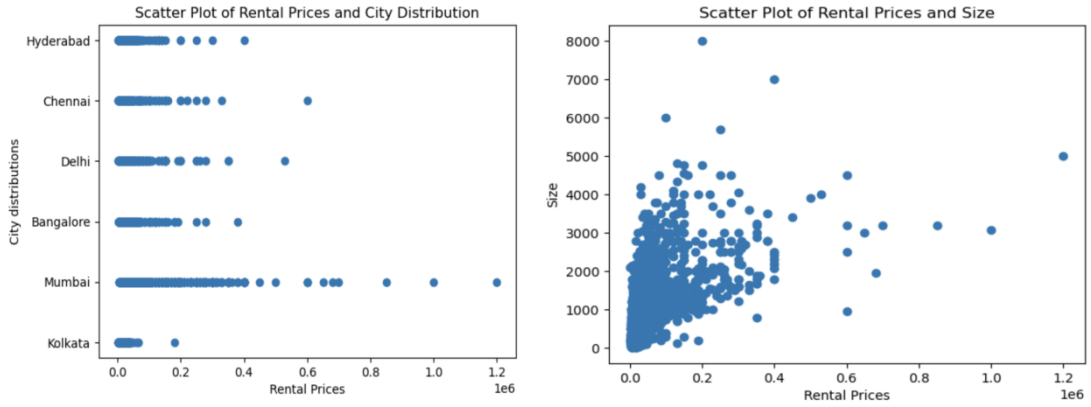


(a) (b)

Fig. 3 (a) Scatter Plot of Rental Prices (millions of dollars) and Area of Houses (Sq ft) (b) Scatter Plot of Rental Prices and Furnishing Status

According to Figure 3(a), The mean of the size of the houses is 967.49 square feet people can see that most houses'sizes are between 600 and 1,300 square feet. Therefore, if you have more size, your house rent will be higher. Thus, it is a very important factor to predict the house rent.

According to Figure 3(b), this data shows that semi-furnished status (2251) is the most frequent type of house rental. A semi-furnished home only has some items such as cupboards and chairs. Unfinished apartments do not add any money. Furnished homes add more money. Most people prefer semi-furnished because some people may not like the furnished style.



(a) (b)

Fig. 4 (a) Scatter Plot of Rental Prices and city distribution. (b) Scatter Plot of Bathroom Rooms and House Areas

According to Figure 4(a), not all the cities have the most people to rent. All of the first-tier, second tier and third-tier cities have an average number of people to rent. Thus, in an average city in India, the house rent is very popular to rent.

According to Figure 4(b), most of houses which have larger sizes than others have more money than other prices of houses that have lower size. As a result, size of the houses is an important factor of higher house rent.

2.4. Data Prediction

We cannot use some data in the project, so we needed to convert this data to numbers.

This paper used the "groupby" to change these categories into numbers so we can use this data in our prediction process. In my opinion, this is a very important process. If I do not use this, I will miss a lot of columns.

I separated x and y to ensure these two variables are different. X refers to independent variables. Y refers to dependent variables.

80% of the data were used for training set. 20% were used for a test set. Random state is 4. In this paper, the model was set according to the train set and was trained by itself using the DecisionTree Regression from Sklearn. And the common evaluation methods were utilized in this paper, such as mean absolute error (MAE) [7][8] [9], root mean square error (RMSE) [10]

3. Results

As shown in Table1, the model's r2_score has a 92.95% data fit to the test set. By using formulas to calculate RMSE, MSDE, and MAE, the result is that its MAE is 8.02, MSE is 204.68, and RMSE is 14.30 comparing to SVR and Linear regression. In order to conclude that the decision tree model is the best model, this paper compares the decision tree model with another model: linear regression It shows that linear regression model using r2_score has 59.13% data fit to the test set. By using formulas to calculate RMSE, MSE and MAE, we find that its MAE is 27.49, MSE is 1186 and RMSE is 33.44. According to the data, the decision tree model has a lower RMSE, MSE and MAE compared with the linear regression model and its 92.95% data fit is higher than the linear regression model's data fit of 59.13%. Additionally, the decision tree model is an easier way to make a model than a linear regression model. Thus, the decision tree model is better than linear regression model.

Table 1. The Performance of Models

	Decision Tree	SVR	Linear Regression
MAE	8.02	33.65	27.49
RMSE	14.30	46.24	33.44
MSE	204.68	2138.77	1186
r2_score	92.95%	26.34%	59.13%

4. Conclusion

In conclusion, the decision tree model is the better model compared to SVR and the linear regression model. This is because its RMSE, MAE, and MSE is lower than the RMSE, MAE, and MSE of SVR and linear regression models. Additionally, the decision tree model's r2_score data fit is higher than the data fit of the SVR and linear regression models. In future studies, ordinary tenants can easily use software or websites using decision trees or random forests to predict rent prices, so they can't be deceived by other landowners in other parts of the world. The decision tree model could also be applied to other areas of the world and their rental markets; for example, researchers could examine whether the decision tree model or the linear regression model better represents the rental market in Europe or other countries.

References

[1] Redi Wang. Housing rental prices in guangzhou and its influencing factors analysis [D]. Fujian normal university, 2022. The DOI: 10.27019 /, dc nki. Gfjsu. 2021.000697. '

[2] Zhao Beigeng. Application of LDA algorithm based on R language in rent prediction [J]. Computer programming skills and maintenance, 2015 (04) : 67-68.doi:10.16184/j.cnki.com.prg. 2015.04.021.

[3] Zhiqiang Wang, Yong Fan. Analysis of Influencing factors of Real Estate Price in Beijing -- Real estate policy from empirical evidence [J]. Shopping Mall Modernization,2010(01):88-90.

[4] Chen Huanhua, Cao Guoxiang. A prediction method and device based on decision tree: CN201310131606.4[P].CN104111920B[2023-09-10]

[5] Dandan Wang, Mei Hua Li , JiaZhou CUI . Research on prediction model based on Decision tree induction [J]. China Strategic Emerging Industries, 2017, 000(38, 2017):82.

[6] YE X R. Data prediction method, device, equipment and storage medium [P]. Guangdong Province: CN113159175B,2023-06-06.

[7] GUO Lingling, Fan Simeng, Wang Mei et al. Behavior Analysis of Online Learning based on linear regression Algorithm [J]. Journal of Computer Technology and Development,202,32(07):191-195.

[8] Liu S Y. Analysis of influencing factors of urban rental prices based on machine learning method [D]. Nankai university, 2022. DOI: 10.27254 /, dc nki. Gnkau. 2021.000080.

[9] Zhaoxian Xie , Xingmin Zou , Wenjing Zhang. Efficient parameters of large data sets and pruning the decision tree algorithm research [J/OL]. Computer engineering: 1-11 [2023-09-10]. https://doi.org/10.19678/j.issn.1000-3428.0066519.

[10] Shen. Computer vision based on machine learning applications [J]. Computer programming skills and maintenance, 2023 (8) : 109-111.doi:10.16184/j.cnki.com.prg. 2023.08.038.