Predicting Boarding House's Price in Jakarta

Sofian Fadli

June 18, 2020

I. Introduction

I.1. Background

Jakarta, as the capital of the Indonesian state, has become the center of trade and technological progress in Indonesia. Because of this, many people from the regions migrated to Jakarta to look for jobs and established careers. However, sometimes because not everyone has relatives in Jakarta, they have to rent a boarding house for them to stay while working in Jakarta. And sometimes they are often wrong in choosing a boarding house, so choosing a boarding price that is too expensive compared to the facilities provided. And besides that, sometimes the owner of a boarding house in Jakarta is often confused in determining the right price for the boarding house to rent.

I.2. Problem

Sometimes there are rental fees that are not reasonable. This project will help in determining the right boarding price based on the facilities provided

I.3. Interest

This project is suitable for people who are looking for boarding prices in Jakarta based on the facilities they want. Then it can also be useful for boarders in determining the rental rates of the boarding house.

II. Data acquisition and cleaning

II.1. Data sources

I got these data from the website www.mamikos.com by scrapping using BeautifulSoup and Selenium. I'm looking for boarding in the Jakarta area and limit the max price range to 3 million rupiah. Due to the price above that much, most have entered the apartment segment which of course is different from the boarding house.

II.2. Data cleaning

After I got the data, I saw many ambiguous features, such as 'including electricity costs, hot water' and 'including electricity, hot water' which actually belong to the same facility. Then many features that are not common, or only owned by that boarding house, such as a 'gazebo' or 'view overlooking the morrissey hotel'. Then there are facilities that cannot be measured, such as 'comfortable boarding'. There are also features that are more complete than other features, such as 'gas and stoves available' and 'stoves'. Then the writing of the facilities is not standard in writing uppercase and lowercase letters, so all of them need to be lowercase. For the size, some use '3X3' using capital x. Then there are those who use decimal size using commas, for example '2,2' which should be '2.2'. Some add 'm' to the size, like '3x3m'.

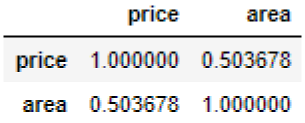
II.3. Feature Selection

From 155 features of the facility, I chose 6 features, namely "ac", "wifi", "indoor bath", "Shower", "TV", "Cleaning service", "Cable TV". These 6 features after I explore using Data Visualization make the most sense compared to other features. And these features are often used as parameters in determining the price of rent.

III. Exploratory Data Analysis

III.1. Relations between area size and price

I calculated the size area by multiplying the width and length (the size in meters). Then I visualize using regression plot. The correlation score obtained is quite high, which is **0.503678**. This is make sense, because the more size area of your room, the more price that you must pay for the rent.



	price	area
price	1.000000	0.503678
area	0.503678	1.000000

Table 1. The correlation score for the area size and price

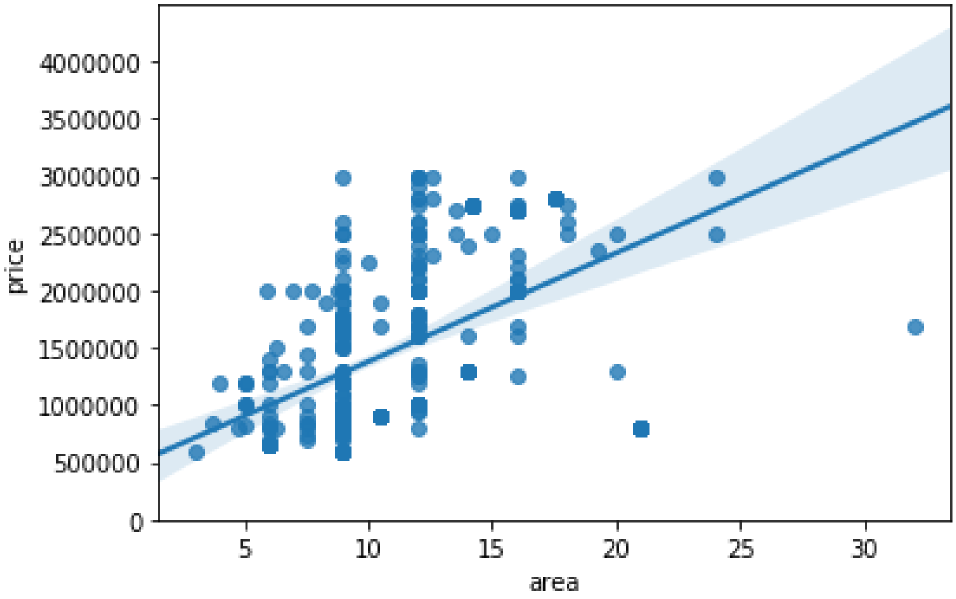


Figure 1. Regression Plot of the area size data and price

III.2. Relationship between 'AC' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it is good to be a predictor. This makes sense, considering that in Jakarta the temperature is quite hot and quickly makes people sweat, so finding a boarding house with AC facilities is a priority. And of course this is used by the boarding house owner to raise the rent price. If boarding houses that do not have air conditioning facilities, usually use a fan that we carry alone or provided by the boarding house.

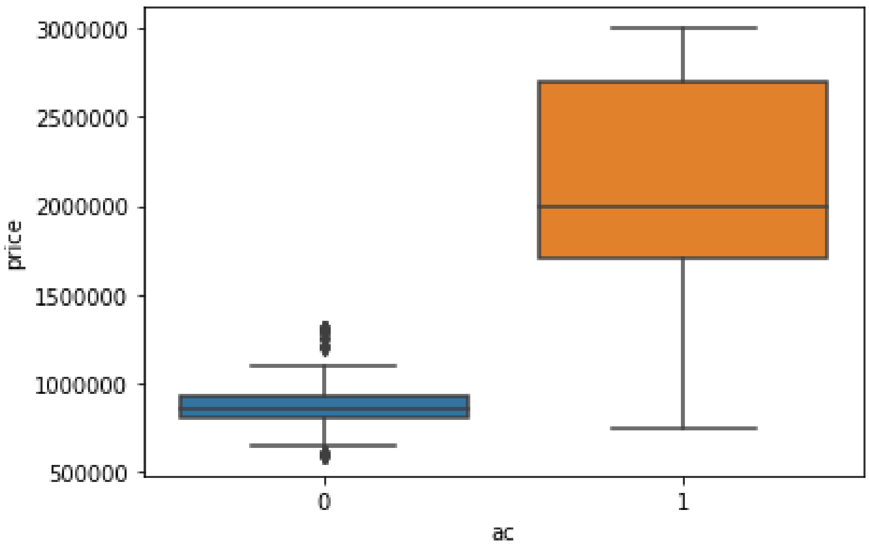


Figure 2. Box Plot from the “AC” features

III.3. Relationship between 'WiFi' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it is good to be a predictor. Wifi is undeniable for everyone's needs nowadays. Even though people can use the quota, but still if there is a Wifi facility in the boarding house, this is quite beneficial to save money in buying data packages. However, this is actually used by the owner of the boarding house to raise the price of rent. Sometimes, the sad thing is that there are boarding houses that have poor Wifi facilities and cannot be used.

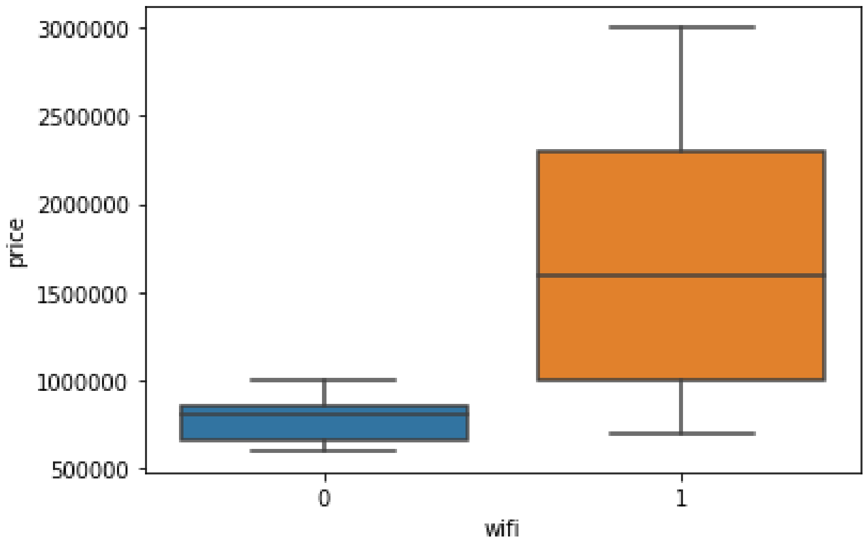


Figure 3. Box Plot from the “WiFi” features

III.4. Relationship between 'Indoor Bathroom' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it is good to be a predictor. The bathroom outside is sometimes poorly maintained by the owner of the boarding house, let alone used by many people. If people who use it can not maintain cleanliness, then we will feel disgusted when using the public toilet. So, having bathroom facilities in our rented room is certainly beneficial. Because we alone are responsible for maintaining the cleanliness of our bathrooms. In addition, it makes it easy to use the bathroom for office people, so there is no need to queue to take a shower. But, of course having your own bathroom in your room will increase the rent.

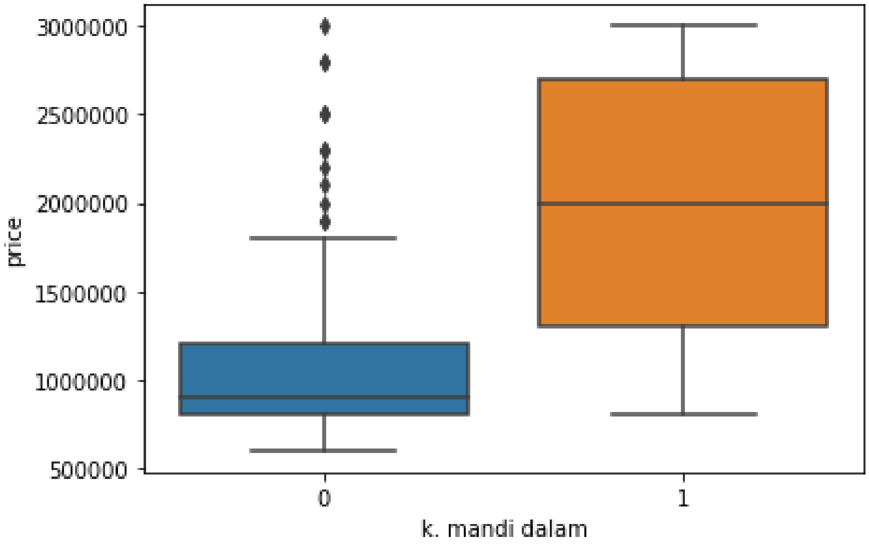


Figure 4. Box Plot from the "Indoor Bathroom" features

III.5. Relationship between 'Shower' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it is good to be a predictor. This is a unique feature that I get after I explore the existing facilities. Maybe the bathroom using a shower makes it easier to take a shower, so the rent is raised by the owner. And quite a number of boarding houses have shower facilities in Jakarta.

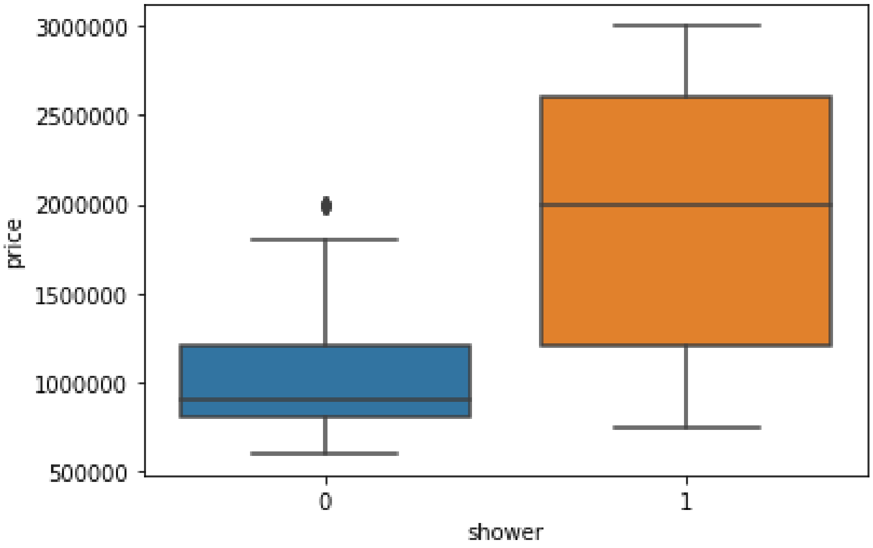
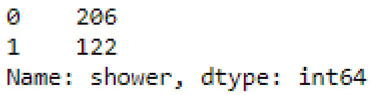


Figure 5. Box Plot from the "Shower" features



```
0    206
1    122
Name: shower, dtype: int64
```

Figure 6. Comparison of the number of boarding houses that have showers and not (from the data collected)

III.6. Relationship between 'TV' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it is good to be a predictor. This is unique. In this day, people are more using gadgets to look for entertainment, there are owners who still include a TV in the boarding facilities for rent. But this affects the cost of boarding.

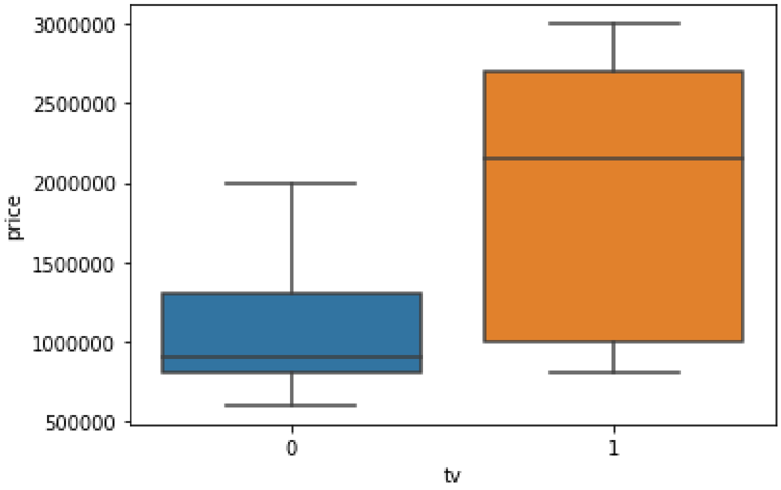


Figure 7. Box Plot from the "TV" features

III.7. Relationship between 'Cleaning Service' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it is good to be a predictor. This may be because the owner of the boarding house must pay the person in charge as a cleaning service, so it automatically affects the price of boarding rent.

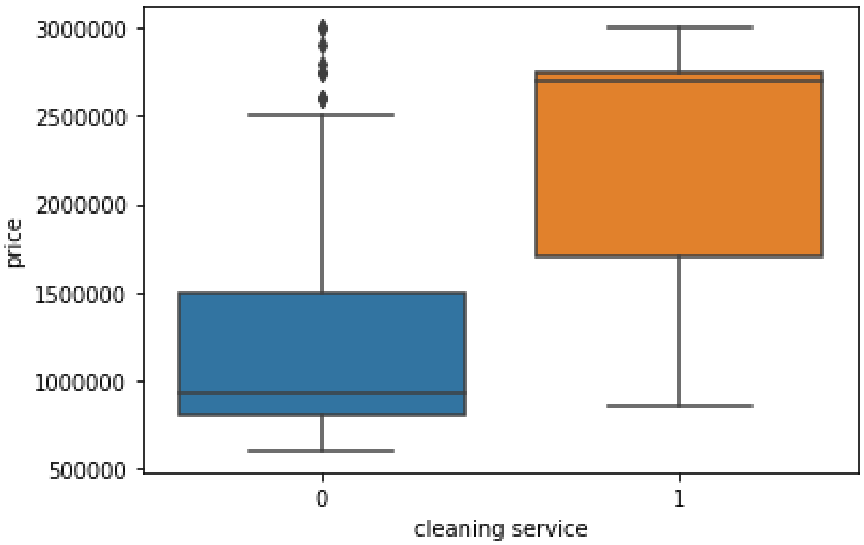


Figure 8. Box Plot from the "Cleaning Service" features

III.8. Relationship between 'Cable TV' facilities and prices

I visualize it using a boxplot. From the results of visualization, we can see there is no overlap between the values, so it might be able to be a predictor. But after checking, only a few boarding houses have TV cable facilities, so it's not good to be a predictor.

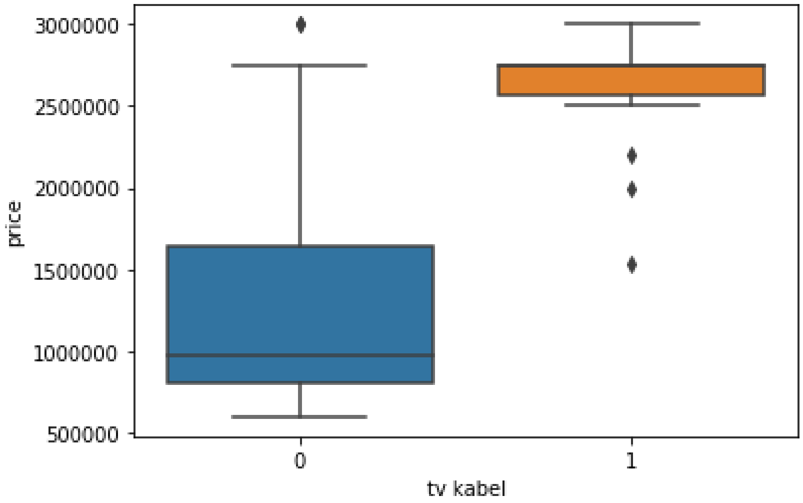
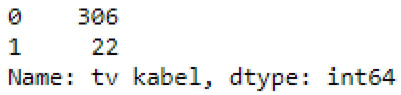


Figure 9. Box Plot from the "Cleaning Service" features



```
0    306
1     22
Name: tv kabel, dtype: int64
```

Figure 10. Comparison of the number of boarding houses that have showers and not (from the data collected)

IV. Predictive Modelling (Regression)

IV.1. Applying standard algorithms and their problems

I applied multiple linear regression, polynomial regression, support vector regression (SVR), decision tree, and random forest to the dataset, using the root mean squared error (RMSE) and R^2 score as the tuning and evaluation metric.

First, for the multiple linear regression algorithm, the results are quite good. The resulting model is not overfitting. Although price predictions are quite off the mark.

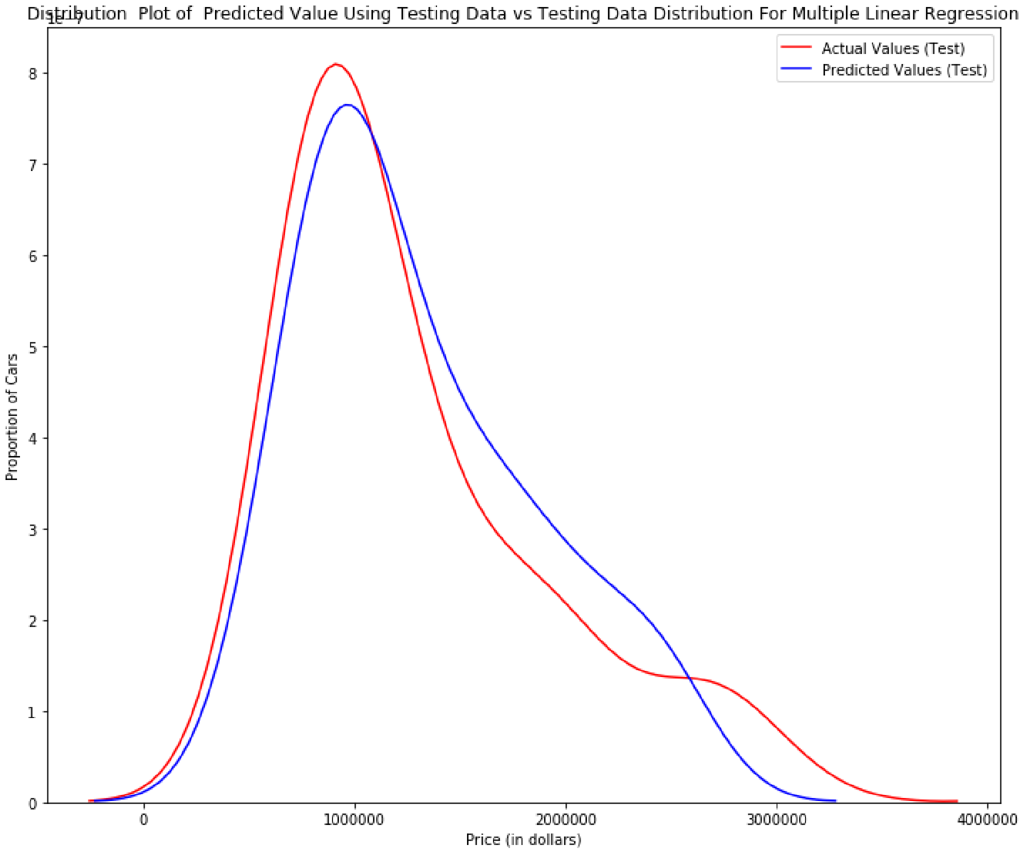


Figure 11. The distribution plot from the price test VS price prediction for linear regression

Second, we trying to using the polynomial. We trying to plot the R^2 score using variance of degree. And, we choose 2 degree. But, the result is not good enough, (the multiple linear regression more better). The model is underfitting, not learn anything from the data. So, we can know the data is not polynomial function.

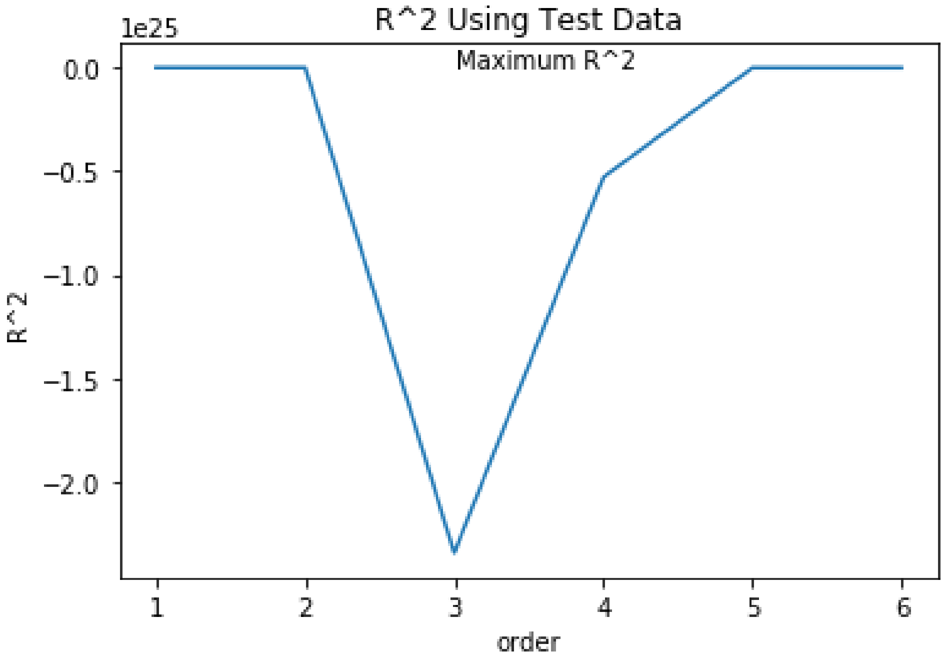


Figure 12. The plot of the R^2 scores using different polynomial degree

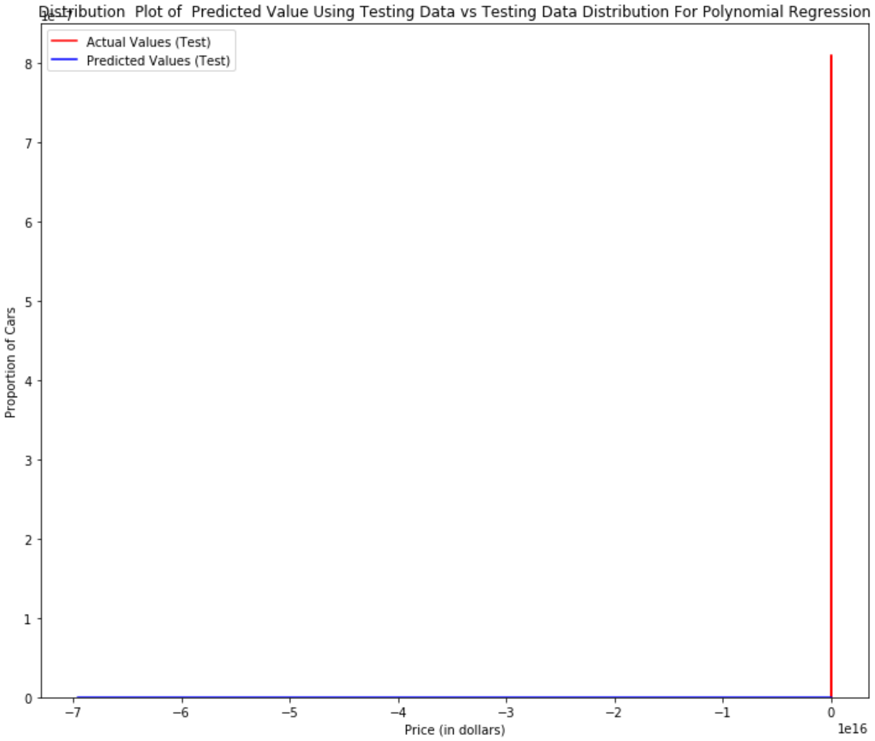


Figure 13. The distribution plot from the price test VS price prediction for polynomial

It’s the same for the SVR. The model is not learn anything from the data. (underfitting).

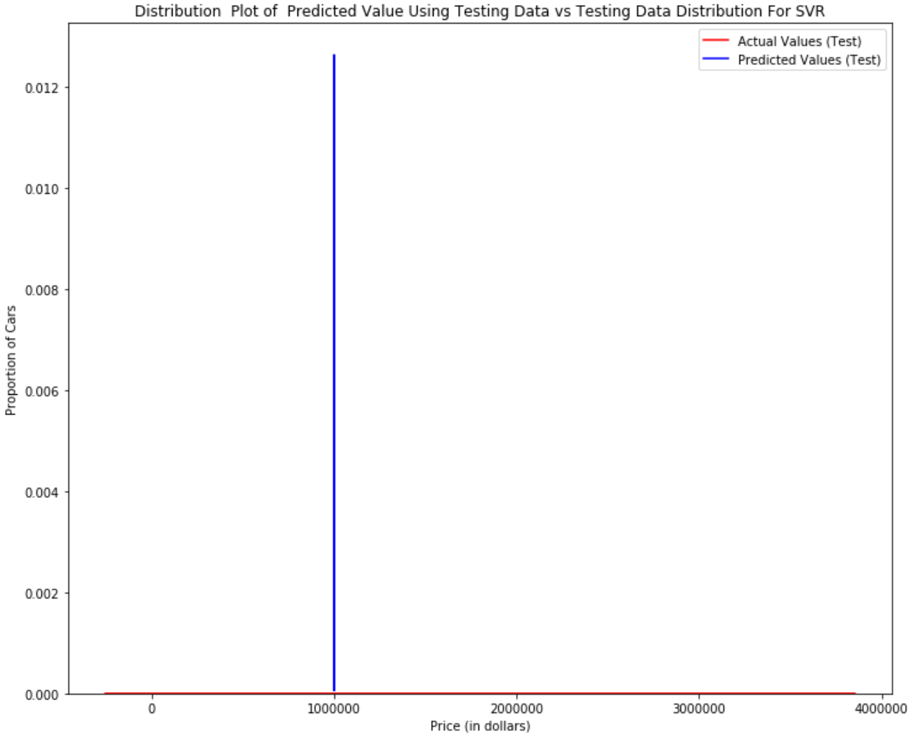


Figure 14. The distribution plot from the price test VS price prediction for SVR

Only, after we use the decision tree regression algorithm, we can see the results. This may be because its features are mostly categorical variables, so tree modeling is very suitable for this data.

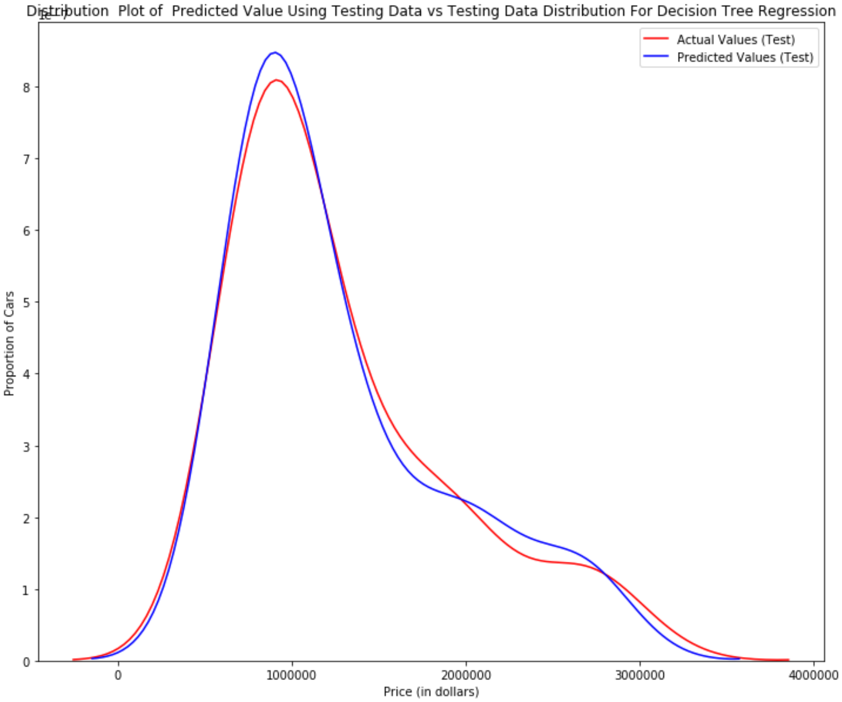


Figure 15. The distribution plot from the price test VS price prediction for Decision Tree

Finally, after we test using the Random Forest algorithm, the results are even better. (Parameter tree = 10).

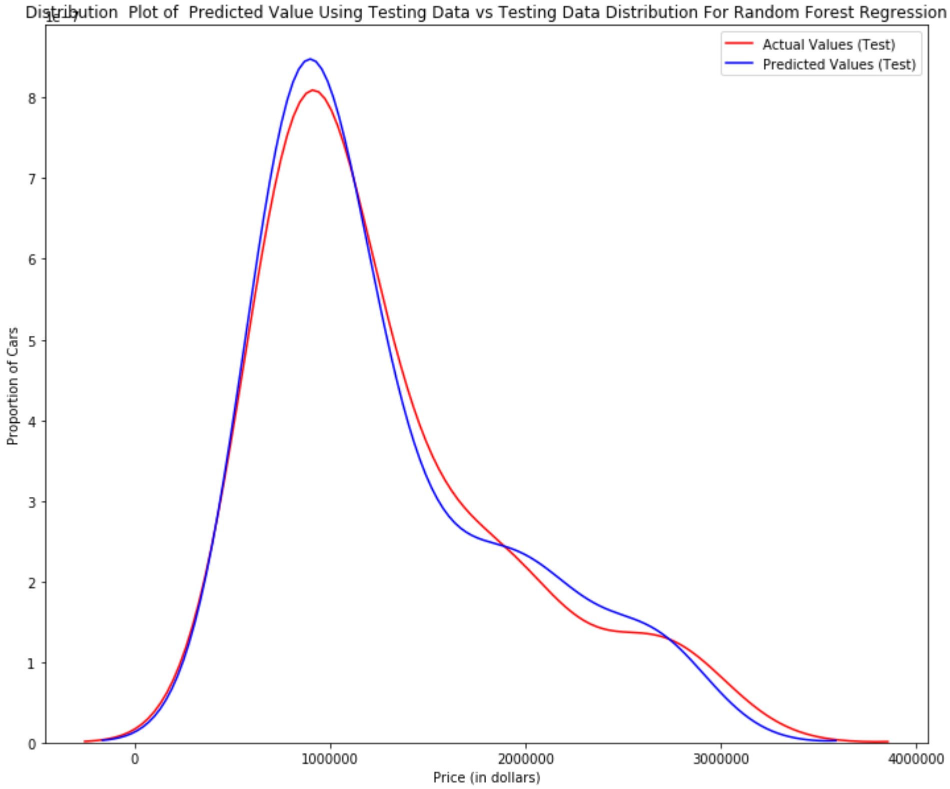
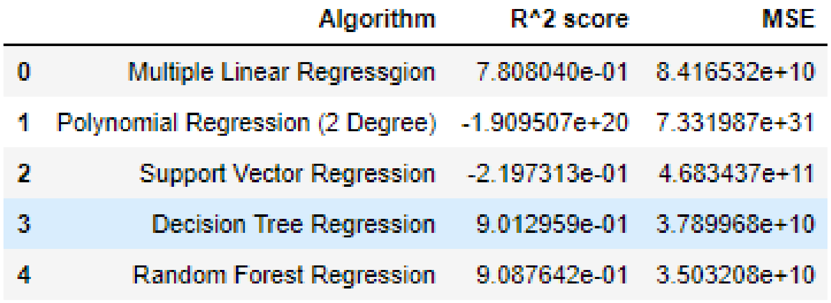


Figure 16. The distribution plot from the price test VS price prediction for Random Forest

Finally, after we test using the Random Forest algorithm, the results are even better. (Parameter tree = 10)

IV. 2. Performance of the different model

So, I tested each model using Mean Squared Error (MSE) and R ^ 2 scores. And as we can see in the graph, Decision Tree and Random Forest are quite good at predicting prices. The MSE value is also the lowest and has an R ^ 2 score close to 1, namely **0.9012959** for the Decision Tree algorithm, while **0.9087642** for the Random Forest algorithm.



	Algorithm	R^2 score	MSE
0	Multiple Linear Regressgion	7.808040e-01	8.416532e+10
1	Polynomial Regression (2 Degree)	-1.909507e+20	7.331987e+31
2	Support Vector Regression	-2.197313e-01	4.683437e+11
3	Decision Tree Regression	9.012959e-01	3.789968e+10
4	Random Forest Regression	9.087642e-01	3.503208e+10

Table 2. The Evaluation Score for Every Algorithm

V. Conclusions

In this study, I analyzed the relationship of prices and facilities owned by the boarding house to the monthly rental price. And it turns out these two things are very influential and can be used as a parameter in determining the rental price of a boarding house. I use regression modeling and this model can be useful for people who want to find a boarding price based on the facilities they want and help the owner of the boarding house in determining the rental price.

VI. Future Directions

Here, I only retrieve data based on boarding facilities and the rental price. While in Jakarta, the location can also greatly affect the price of boarding rental. For example, boarding houses that are close to the office usually have a higher rental price. Usually, location information is very helpful in predicting more accurate boarding prices.