Housing Price Prediction using Linear Regression with One Feature

Jobb Q. Rodriguez
Ateneo de Naga University
Maryville Subdiv., Brgy. San Felipe
Naga City, Camarines Sur,
+639619116147
jobb.rodriguez22@gmail.com

ABSTRACT

In this paper, the author used Linear Regression with One Feature to identify the best feature to predict the house price. From the available features (Transaction Date, House Age, Distance to the nearest MRT station, Number of Convenience Stores, Latitude, and Longitude), Number of Convenience Stores is the best feature to use in predicting house price. The optimized cost was 62.236, (0, 0) as values of the initial thetas and 0 as value of the learning rate. The scatterplot is close to the feature's optimized regression line.

CCS Concepts

• Computing methodologies→Machine learning and Modeling and simulation

Keywords

Housing Price Prediction; Linear Regression; One Feature; Machine Learning; Supervised

1. INTRODUCTION

Houses have features that describe the house. Some features influence the amount of the price. When a feature influences the price, it can be a reliable variable for predicting the price. Linear Regression with One Feature helped identify the best feature to predict the price on the given dataset, *Real estate price prediction* [1].

2. RESULTS

2.1 Transaction Date vs Price

For the pair, the initial thetas are (0, 0), and the learning rate is 0.0000001.

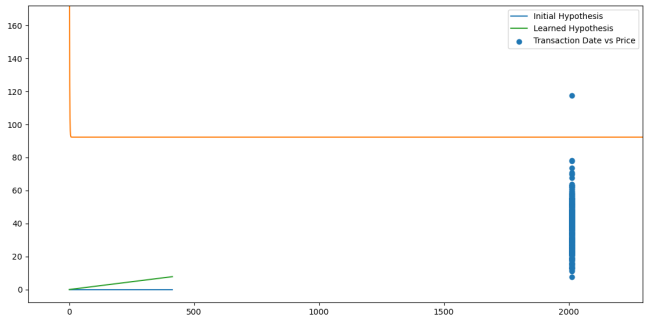

Figure 1. Graph for Transaction Date vs Price

The scatterplot is far from the optimized regression line. The optimized cost is 92.338.

2.2 House Age vs Price

For the pair, the initial thetas are (0, 0), and the learning rate is 0.02.

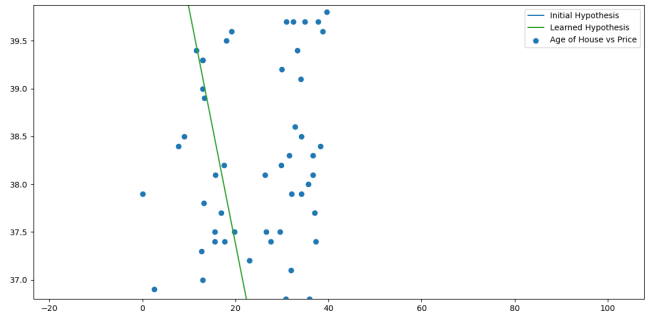

Figure 2. Graph for House Age vs Price (Zoomed In)

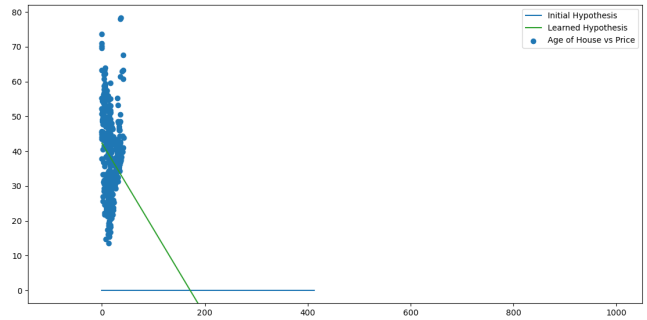

Figure 3. Graph for House Age vs Price (Zoomed Out)

A small segment of the optimized regression line is close to the scatterplot. The optimized cost is 88.252.

2.3 Distance to the nearest MRT station vs Price

For the pair, the initial thetas are (0, 0), and the learning rate is 0.0000001.

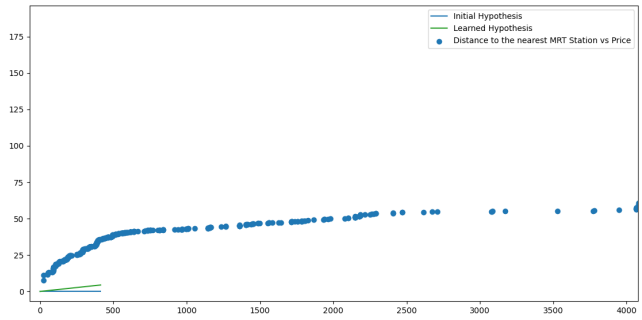

Figure 4. Graph for Distance to the nearest MRT station vs Price

The optimized regression line is far from the scatterplot. The optimized cost is 654.114.

2.4 Number of Convenience Stores vs Price

For the pair, the initial thetas are (0, 0), and the learning rate is 0.002.

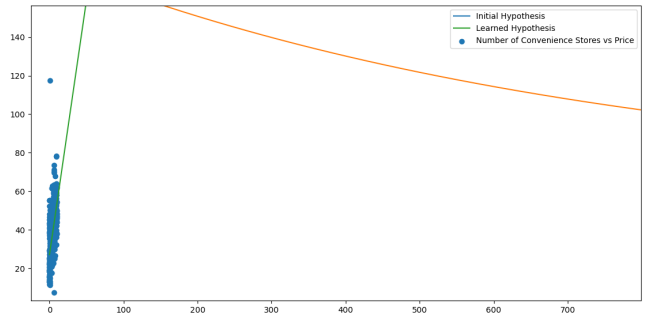

Figure 5. Graph for Number of Convenience Stores vs Price

The scatterplot is near the optimized regression line. The optimized cost is 62.236.

2.5 Latitude vs Price

For the pair, the initial thetas are (0, 0), and the learning rate is 0.000001.

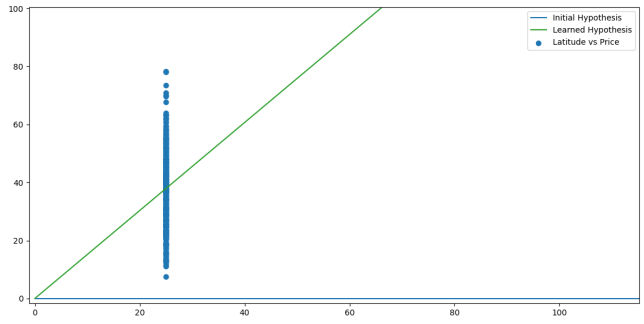

Figure 6. Graph for Latitude vs Price

The scatterplot does not match the optimized regression line. The optimized cost is 92.208.

2.6 Longitude vs Price

For the pair, the initial thetas are (0, 0), and the learning rate is 0.0000001.

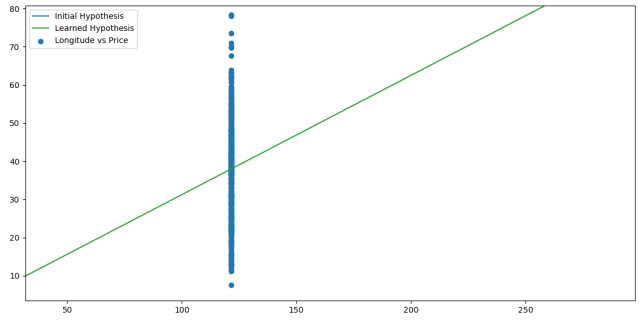

Figure 7. Graph for Longitude vs Price

The scatterplot does not match the optimized regression line. The optimized cost is 92.311.

3. ANALYSIS

3.1 Initial Thetas and Learning Rate

In terms of the initial thetas, all pairs started with (0, 0).

In terms of the learning rate, Transaction Date vs Price, Distance to the nearest MRT station vs Price, and Longitude vs Price had the smallest learning rate. Then, Latitude vs Price had the next smallest learning rate which was closer to the previous learning rate. Small learning rates were used to reach an optimized cost to create the optimized regression line and to avoid overshooting the optimized cost. Then, House Price vs Age and Number of Convenience Stores vs Age had a big gap from the previous learning rates. Both pairs had the biggest learning rate.

3.2 Optimized Cost

Before reaching the final optimized cost, the cost underwent 10,000 computations.

For the Transaction Date vs Price, its optimized cost was the fifth lowest value and had the smallest learning rate.

For the House Age vs Price, its optimized cost was the second lowest value with the biggest learning rate.

For the nearest MRT station vs Price, its optimized cost was the sixth lowest value and had the smallest learning rate.

For the Number of Convenience Stores vs Price, its optimized cost was the lowest value and had the biggest learning rate.

For the Latitude vs Price, its optimized cost was the third lowest value and had the middle learning rate.

For the Longitude vs Price, its optimized cost was the fourth lowest value and had the smallest learning rate.

In consideration of the numerical values, the Number of Convenience Stores had the lowest optimized cost (in respect to the consideration for the other features' learning rates).

3.3 Visual Representations

For the Transaction Date vs Price, the optimized regression line shows that the transaction date is directionally proportional to the price of the houses. However, the optimized regression line is far from the scatterplot.

For the House Age vs Price, the optimized regression line shows that the houses become cheaper as they become older. However,

the scatterplot is near to a small segment of the optimized regression line.

For the Distance to the nearest MRT station vs Price, the optimized regression line shows that the houses become more expensive as the distance of the nearest MRT station increases. Even if the line matches the scatterplot, the line is far from the scatterplot.

For the Number of Convenience Stores vs Price, the optimized regression line shows that the houses become more expensive as the number of convenience stores increases. The scatterplot's majority is near the line.

For the Latitude vs Price, the optimized regression line shows that the houses become more expensive as the latitude increases. However, the line does not represent the plots. The scatterplot was vertically stacked while the line was diagonal.

For the Longitude vs Price, the optimized regression line shows that the houses become more expensive as the longitude increases. However, the line does not represent the plots. The scatterplot was vertically stacked while the line was diagonal.

In consideration of the visual representations, the Number of Convenience Stores vs Price had the best match for the optimized regression line and the scatterplot.

3.4 Synthesis

For the Transaction Date vs Price, even if the pair had the fifth smallest optimized cost, the scatterplot is distant from the optimized regression line.

For the House Age vs Price, even if the pair had the second lowest optimized cost, a huge segment of the optimized regression line is far from the scatterplot.

For the Distance to the nearest MRT station vs Price, even if the optimized regression line's implication matches the scatterplot, the optimized cost is large (a larger learning rate results in an infinity error), and the line is far from the scatterplot.

For the Number of Convenience Stores vs Price, the pair had the lowest optimize cost, and its optimized regression line is near the scatterplot.

For the Latitude vs Price, even if the pair had the third smallest optimized cost, the optimized regression line did not represent the scatterplot.

For the Longitude vs Price, even if the pair had the fourth smallest optimized cost, the optimized regression line did not represent the scatterplot.

4. CONCLUSION

Number of Convenience Stores vs Price had the lowest optimized cost with a relatively fair learning rate (in consideration of the other features). Moreover, the pair's optimized regression line is the nearest to the scatterplot.

As a result, Number of Convenience Stores is the best feature to predict the House Price.

5. REFERENCE

[1] Bruce. 2019. Real estate price prediction. Retrieved from https://www.kaggle.com/