Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220593

A Polynomial Linear Prediction Model for Housing Price in the USA

Yuyao He

University College London, London, WC1E 6BT, UK

hecaroline@yeah.net

Abstract. The purpose of this research was to build a linear model for predicting the price of houses. The price of the house could be approximated without knowing the price of every house. In the process of the experiment, the data from real estate markets would be analyzed for the supervised study. A linear model would be utilized to predict the price. Different values of learning rate would be compared, and the most efficient value according to the cost function would be chosen. Finally, the prediction model with learning rate 2 would be chosen and used by people who would like to know the price of houses without spending a long time. The price of the house can be successfully accessed by inputting the values of features -numbers of bedrooms, bathrooms, area, and floors- to the model.

Keywords: supervised learning, linear regression, cost function, house price prediction, gradient descent

1. Introduction

According to the data provided by the University College of London, the number of housing units in the United States in the year 2021 was 141.95 million [1]. The great number of houses leads to the inconvenience of collecting the prices of all houses in each city. A method is required to figure out an efficient way to predict the house price.

Under the environment of current society, time is a treasure. For people who tend to sell, rent, or buy a house, budget is the most important factor. In the past, market research was essential to be done if a person needs to do a price prediction of a house. However, this method was inefficient. The first reason is that the time required for market research is extremely long. Secondly, a person's energy is limited, which leads to the limited number of houses being investigated. An insufficient number of samples would cause an inaccurate prediction of the price of the house.

Machine learning is a way to let the computer study the samples of given houses and build a model to predict the price of other houses. Compared to the number of houses a person could investigate, computers could learn more than a million samples of houses in a day, which is hard to achieve. In addition, the time for people who want to know the price of a house could be decreased greatly if the computer can help with the calculation of the price of a house. Therefore, using computers to predict the house price is the purpose of this research.

In order to figure out the prediction model, the poly linear regression was used. Therefore, we would utilize the knowledge under supervised learning to build a linear model to do the price prediction for the

DOI: 10.54254/2755-2721/2/20220593

houses in Washington. Supervised learning [2] means using the data sets which were classified to analyze the matched relationships between inputs and output. Then, linear regression was used to build the model. The main idea of linear regression was to use a linear function to predict the output with all features as variables. In other words, the influence levels of each feature decided the parameters before each variable, and all variables were the factors that affected the output. In this report, the process of choosing each parameter would be clarified.

This research would give a solution to house prediction by utilizing a linear regression model with the most efficient learning rate. The house price could be predicted with the given features.

2. Literature reviews

Nowadays, in the existed models utilized to predict the prices of houses used by most agents and users, the area of the house and the location would be the decisive factors that affect the final prices. However, more factors should be included in the model as the price of a house is a result of multiple influences.

In the beginning, the development of a house price prediction started with a simple linear regression with one variable. In particular, the linear regression method predicted the output values by examining the linear relationship between the output variable Y and the input variable X. Corresponding with the problem of houses prices, one dependent variable was not sufficient to support the estimation of the final prices although the mathematic theories required for linear regression was not complex. The linear regression method was a suitable start for analyzing the prices of houses because the concept of the linear regression was a solid base for the further research extension. The idea that making a hypothesis to illustrate the dependent variable by analyzing the effects of changes in values of the independent input variable was extremely useful to develop a better model to predict the changes in the prices of houses. The idea should be kept; however, more dependent input variables should be considered to build the relationship between the input variables and the final output value. The main factor can give an approximation of the prices of houses in a roughly right direction. Nevertheless, ignoring the effects of factors that are seemingly unimportant would lead to an inaccurate approximation in each particular situation.

Therefore, it was essential to analyze the prices of houses in a comprehensive view. In this project, the characteristics of houses would be included as the factors which relate to the final price to obtain a more accurate house price prediction model.

Then, considering the paper "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization" [3] published by Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, and Wayan Firdaus Mahmudy in 2017, more factors were included in their model building. The factors they used were the area of the house and the situation of transportation. The area of the house was a continuous variable which was a continuous number. The situation of public transportation around the building was a binary variable that only contained two situations: there was a public transportation nearby the house or no public transportation was set nearby the house. To analyze the discontinuous variable, the value for there was a public transportation nearby was 1, and the value for there was no public transportation nearby was 0, so it was easier to represent the values of this variable in the equation of the hypothesis. Thus, this research provided a good indication that the continuous variables and the discontinuous variables could be utilized simultaneously when the house price prediction model was built.

Combined the advantage in the paper "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization" that the discontinuous variables could also be considered and correspond to the numbers and the disadvantage of the use of one dependent variable which led to an inaccurate approximation result, in this project, more factors would be used as the features to do the prediction of the prices of the houses in Washington.

3. Methods

Based on the idea of linear regression, the main purpose of the research is to find a linear relationship between the input independent variables and the output dependent variable. However, to improve the accuracy of the linear regression, more factors would be considered in this research. The dependent

DOI: 10.54254/2755-2721/2/20220593

variables are chosen if the variables lead to a change in the output values. Relating to the context of this research, the characteristics of a house would be analyzed to predict the price of the house if there is a strong correlation between the characteristics of the house and the price of the house.

Several steps were needed to complete this research.

First and foremost, the data set was downloaded from Kaggle [4], which was a well-known website that provided a wide range of data sets. The data set was the basis of this research. In this research, the data from Kaggle was qualified to be used since the features provided in this data source were valuable and credible. The number of samples in this database is enough to use in this program. In addition, the features included in this dataset contain the features that would be chosen for the building of the prediction model of the house prices in this project. The dataset described the house price in Sydney and Melbourne in 2014. The whole datasheet contained 6400 information on houses. However, in this research, 500 pieces of information were picked as the material for supervised learning.

The second step was to decide the features that should be included in this research. The main factors that affected the price of houses should be chosen. After the consideration, the number of bedrooms, the number of bathrooms, the area of the house, and the number of floors were picked to evaluate the price of houses. The area of the house is the main factor that most people consider first when making the hypothesis of the price of a house. Because normally, in nowadays society, the total price of a house is determined by the product of the price of a unit area of the house and the total area of the house. The number of bedrooms is included as a feature because the number of bedrooms in the houses is highly related to the need of people who are going to purchase or rent a house. Although the area is large enough, the lack of bedrooms which cannot meet the requirement of people would also lead to a decrease in the price of the house. Similarly, the number of bathrooms is another feature that is an essential requirement for people. The number of floors contributes to the preference of people who would rent or buy a house. For example, people who care about the convenience of doing cleaning would prefer to choose a smaller number of floors.

Thirdly, the feature scaling [5] was an important step to avoid the effect of the huge difference in original values of different features. For instance, the value of the area of the house could be 1500 square feet; nevertheless, the number of bedrooms was only 3. The huge gap between these values would result in the low efficiency of gradient descent because a longer time was needed to calculate the parameters. The parameters of each variable are decided by running gradient descent to compute the cost function. Gradient descent is an algorithm [6] by trying different parameters for each variable starting from initial parameter values. The start parameters are set randomly, a large difference in values of different features would need to run a significantly large number of gradient descent algorithms to compute the appropriate parameters of all features. Therefore, normalizing all features in a certain range could effectively decrease the amount of time that was required to calculate parameters. In this research, feature scaling was achieved following the equation: $X_{norm}=(X-X_{mean})/(X_{max}-X_{min})$. All values of features should be a range from -1 to 1.

After that, gradient descent was done by giving the initial learning rate and the number of iterations. The initial values were set with the following values: the learning rate was 0.2 and the number of iterations was 4000. The initial values did not need to be precise. However, with more experience in machine learning, the initial values given would be closer to the final values.

Drawing the cost function [7] was the fifth step to ensure the accuracy of the model with the calculated parameters. The cost function represented the difference between the predicted prices of houses and the real predicted prices of houses. Therefore, a minimum was the thing that the gradient descent algorithm was looking for. The value of the cost function can indicate the accuracy of the prediction model. Therefore, at minima, the house price prediction model can be proved to be accurate and with fewer mistakes in prediction. The prices of houses were considered to be close to the actual prices of houses with similar features. The result of gradient descent can be utilized to draw the cost function. According to the diagram of the cost function, the value of the learning rate and the number of iterations should be adjusted to acquire the parameters with the lowest cost function.

DOI: 10.54254/2755-2721/2/20220593

Finally, the parameters for each feature were calculated and the final prediction model was built. The price of houses could be predicted when the values of features were given to the model.

4. Results and discussion

During the process of experiments, several results were printed using MATLAB. Figure1 showed the first 10 examples of datasheets imported were right. This output was to check whether the data was imported successfully and in the right pattern. Comparing the first 10 examples to the original dataset, all values were matched.

```
First 10 examples from the dataset:
 x = [5.000 2.500 3650.000 2.000], y = 2384000.000
 x = [3.000 2.000 1930.000 1.000], y = 342000.000
 x = [3.000 2.250 2000.000 1.000], y = 420000.000
 x = [4.000 2.500 1940.000 1.000], y = 550000.000
 x = [2.000 1.000 880.000 1.000], y = 490000.000
 x = [2.000 2.000 1350.000 1.000], y = 335000.000
 x = [4.000 2.500 2710.000 2.000], y = 482000.000
 x = [3.000 2.500 2430.000 1.000], y = 452500.000
 x = [4.000 2.000 1520.000 1.500], y = 640000.000
 x = [3.000 1.750 1710.000 1.000], y = 463000.000
```

Figure 1. samples of dataset.

Displayed equations are centered and set on a separate line. Therefore, the setup of the project that the data used to be analyzed and figured out the contributions of the prices of houses were effectively finished. After importing data successfully, the correct data set can be used to do the feature scaling.

```
Feature Scaling ...
mu = [3.412 2.163 2126.478 1.502]
max_min = [8.000 7.000 12820.000 2.000]
```

Figure 2. mean and the difference between the maximum and minimum values.

From Figure 2, the values of the mean values of each feature were printed, and the range of values of each feature was also calculated. The output matrix of mu had a length of 5 and the output matrix of the difference between the maximum values and the minimum values of each feature was computed correctly by comparing the values calculated by the computer program and the values calculated by hand.

Then, the result after feature scaling improved that the values of each feature were efficiently set in the range from -1 to 1. Figure 3 could approve. Similarly, in figure 1, the first ten samples of features were printed. By observing the output, all feature values were correct after feature scaling. Thus, the purpose of feature scaling was achieved, and further steps could be processed.

```
after feature scaling
x = [0.199 -0.051 -0.051 0.074]
x = [-0.176 -0.176 0.074 -0.051]
x = [0.074 -0.051 0.048 -0.023]
x = [0.012 0.048 -0.166 -0.023]
x = [0.048 0.048 -0.023 -0.059]
x = [0.119 -0.015 -0.010 -0.015]
x = [-0.097 -0.061 0.046 0.024]
x = [-0.047 -0.032 0.249 -0.251]
x = [-0.251 -0.251 -0.251 -0.251]
x = [0.249 -0.251 -0.001 -0.251]
```

Figure 3. feature scaling.

DOI: 10.54254/2755-2721/2/20220593

Figure 4 could show the process of gradient descent. When the line "Gradient Descent…" was printed on the computer, some time would need to get the final result of values of theta. This was because of the quality of gradient descent, which was to try different parameters from the initial values until the cost function reached a local minimum.

```
Gradient Descent ...
Theta computed from gradient descent:
 550790.138000
-113051.683919
-56751.216971
 3050771.226826
 83742.583723
```

Figure 4. gradient descent.

According to the figure5, the cost function was plotted with different values of the learning rate[8].

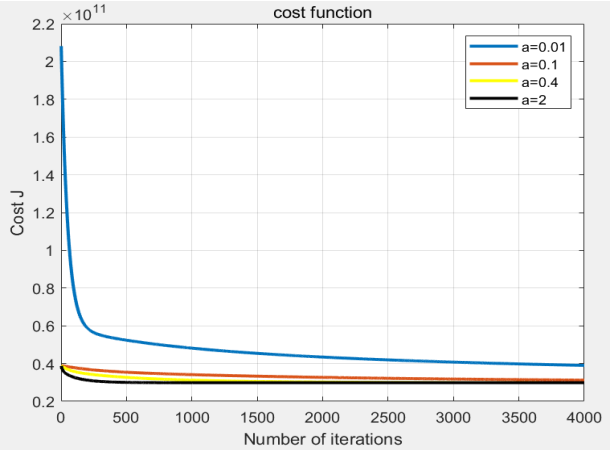

Figure 5. cost function.

The learning rate of 0.01 was too small to reach the local minimum of the cost function with 4000 iterations. Comparing the learning rates 0.1, 0.4, and 2 could all reach the local minimum with 4000 iterations; nevertheless, when the learning rate was equal to 2, the cost function could be at the local minimum the fastest. Thus, the learning rate 2 was selected as the step of gradient descent. However, 0.1 and 0.4 can also reach the final result. The choice of 2 as the learning rate was a method to increase the effectiveness of this computer program. Therefore, the time spent to find all suitable parameters would decrease sharply compared to the time needed when the learning rate is 0.1 and 0.4 because the small step of changing would allow more values of parameters to be tried during the process of gradient descent.

The matrix of theta could therefore be calculated. The final model was price=[1 X1norm X2norm X3norm X4norm].*[theta0 theta1 theta2 theta3 theta4]. This model showed the linear regression equation. Each feature had its relationship with the output value which is the house prices. The values of parameters indicated the strength of correlation between each variable and the final prices. From the result of theta which can be observed in figure 4, the value of theta3 is the largest. Theta3 is the parameter of x3 which is the area of the houses. Because the feature scaling is done, the original large value of the area of houses would not affect the parameter. The large value of the parameter showed that the area of houses was the most important factor that would decide the prices of houses. This conclusion was matched with the assumption made before that the area of houses was the main factor that contributed to the prices of houses.

DOI: 10.54254/2755-2721/2/20220593

The results of cost functions with different learning rates were plotted in the same figure to have a stronger contrast.

```
Predicted price of a 3bedrooms 2.5bathrooms 2600sqft 2floors house (using gradient descent):
 $687419.967195
```

Figure 6. Prediction result.

Figure 6 demonstrated the function of the prediction model. All four cost functions were decreasing with 4000 iterations. However, when the learning rate was 0.01, the final lowest cost function value was higher than others， which means within 4000 iterations, the learning rate of 0.01 cannot compute the values of theta that can predict the prices of houses most accurately and with the lowest error. The difference between remains learning rates was the number of iterations they needed to reach the lowest value of the cost function. Learning rate 2 was the fastest to reach the lowest cost. The learning rate was the step for the algorithm to try different values of parameters. With a small step, more numbers would be tried for the parameters. Therefore, the learning rate of 0.1 and the learning rate of 0.4 needed more time to reach the final values of theta. The learning rate of 0.1 which was the smallest step spent the longest time to reach the final values of theta.

The predicted price of the house of 3 bedrooms, 2.5 bathrooms, 2600 square feet, and 2 floors was 687419.967195 dollars, which was a reasonable value for the price of a house. To further check the accuracy of the result, the house with the same features from the datasheet was found. The price of the house in the datasheet was close to the price predicted using this house price prediction, model. Therefore, the model could be used as a tool to predict the prices of houses with given values of features.

5. Conclusion

In conclusion, in this research, a linear model for the prediction of the house price was built successfully because the final predicted price of a house of a given number of bedrooms, bathrooms, area of the house, and floors was close to the actual price of the house with the same features in the datasheet. Several values of learning rate were utilized in the experiment; however, the learning rate 2 was the most suitable step to do the gradient descent to reach the local minimum of the cost function. The reason for picking the learning rate 2 was that this learning rate can successfully reach the lowest cost function value, and the time needed to reach the local minima was the shortest. By considering both the final result and the effectivity, learning rate 2 was the most suitable step for the algorithm to do the gradient descent.

Finally, the Washington house price prediction model using the knowledge of polynomial linear regression model was

Price (in dollars) = 550790-113051*number_bedrooms(normalized)-56751*number_bathrooms(normalized)+3050771*area(normalized)+83742*number_floors(normalized).

In this research, a linear regression model was successfully built to predict the house price utilizing the dataset which included the information on the real estate market in the United States in 2014. The appropriate value of the learning rate ensured the efficiency of the model. The cost function could reach the local minimum at the fastest speed. The model could give a clue to all customers and sales about the price of houses in the United States in 2014 with a known number of bedrooms, bathrooms, areas, and floors. This function of the model could effectively reduce the time needed to estimate the price of a house.

Nevertheless, there were improvements to increase the accuracy of the model. More features could be included to build the model because the number of features used in this research was limited, which led to the disadvantage that some factors that affected the price of houses were ignored when the model predicted the price.

DOI: 10.54254/2755-2721/2/20220593

References

[1] University College of London, "Number of housing units in the United States", in 2021, Available: https://www.statista.com/statistics/240267/number-of-housing-units-in-the-united-states/
[2] ScienceDirect, "Supervised Learning", Available: https://www.sciencedirect.com/topics/computer-science/supervised-learning
[3] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, and Wayan Firdaus Mahmudy, in 2017,
[4] "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization" Available:https: //www.researchgate.net/profile/Wayan-Mahmudy-2
[5] Shree, updated in 2019, "House data", Available: https://www.kaggle.com/shree1992/housedata
[6] Baijayanta Roy, published in 2020, "All about Feature Scaling", Towards Data Science, Available: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35
[7] D.P.Mandic, Feb. 2004, "A generalized normalized gradient descent algorithm", published in "IEEE Signal processing letters"
[8] Hang Zhang. In 2020, "House Price Prediction with An Improved Stack Approach", available in: https://iopscience.iop.org/article/10.1088/1742-6596/1693/1/012062/meta
[9] Matthew D. Zeiler, 2012, "An Adaptive Learning Rate Method", Available: https://arxiv.org/abs/1212.5701