Write A Code Using Linear Regression and Neural Layered Structure To Predict The House Price

Kehan Chen [1] and Wanting Lu [2] and Yicheng Yan [3,*]

[1] Nanjing Foreign Language School, Nanjing, Jiangsu, China
[2] Shanghai Experimental Foreign Language School, Shanghai, China
[3] YiChang International School Longpanhu, Yichang, Hubei, China
* 2016123749@jou.edu.cn

Abstract. House price prediction is a challenging task, and for home buyers, it is difficult to accurately predict house prices due to the complexity and dynamics of the real estate market. Secondly, as far as the data is concerned, house price prediction is affected by several indicators and it has a great deal of randomness, so this is not easy for a machine to predict. The essence of house price prediction is to analyze and process the text, i.e. to do regression tasks. Therefore, in this essay, we propose a method for house price prediction by using linear regression and a neural layered structure. We demonstrate the effectiveness of these techniques on a dataset of 506 records from house price reports in Boston, Massachusetts, USA. Linear regression models provide an initial understanding of data trends, while neural network models use the power of deep learning to capture more complex patterns and relationships. Linear regression is a supervised learning algorithm used for predicting a continuous output based on input features. It assumes a linear relationship between input variables and the target variable. It's a suitable choice when you have a dataset with numerical features and a continuous target variable. The neural network refers to the human brain neuron network and forms different networks according to different connection methods to complete information processing and establish a certain model.

Key words: Linear Regression, neural layered structure, predict the house price.

1 Introduction

House price prediction is closely related to our daily lives, and for people who want to buy a house, a house price prediction system can help them to have a good perception of the price of the house. It has great socio-economic value.

The traditional methods of predicting housing prices are limited by the ability to obtain data and predict algorithms, resulting in a lack of timeliness and accuracy, which is not satisfactory. Finding new data construction indicators to improve the model's effectiveness is one of the problems that needs to be solved [1].

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_70

The rapid development of big data and the gradual maturity of machine learning and natural language processing technologies have made it possible to analyze real estate prices based on network search data, which can compensate for the subjectivity of indicator selection in traditional prediction model construction and obtain more comprehensive and abundant influencing factors, thereby better reflecting the formation mechanism of housing prices and improving prediction accuracy [2].

So, house price predictions can be made very effectively with neural networks. A typical neural network consists of an input layer, a hidden layer, and an output layer. Each layer contains several data-storing nodes, called neurons, and each neuron is connected to all neurons in the next layer, with each pair of connections having a corresponding parameter. This connection maintains the transfer of information between neurons. Because the neurons inside the neural network can transfer data, in the process the neural network can extract information about the features in the text. Therefore, this type of model is very suitable for the prediction of house price text.

The remainder of this paper is organized as follows. Section 2, explains the related work of the house price prediction. Part 3 explains the methodology used for house price forecasting. Part 4 explains the model theory and model structure. And finally, there is a summary of the whole experiment.

2 Related Research

2.1 The Method We Use In This Experiment

For this experiment we use linear regression for house price prediction, here are some studies by others.

The accuracy rate of four different machine learning algorithms—C4.5, RIPPER, Naive Bayesian, and AdaBoost—is analyzed to determine which algorithm produces the highest accuracy rate [3]. We use machine learning methods including Leaner Regression, Decision Tree, k-Means, and Random Forest to forecast the price of homes in this suggested system [4]. It was demonstrated through research that combining visual and textual data produced better estimation accuracy than textual features alone. Furthermore, given the same dataset, NN produced better results than SVM [5]. To summarise, the above three papers use C4.5, RIPPER, Naïve Bayesian, AdaBoost, and Leaner Regression, and in the third article, we can see that better results are obtained using neural networks than SVMs.

In addition, we have used neural networks for house price prediction, so we need to compare the utility of other arithmetic. Our findings suggest that if the performance indicator is the percentage of subprime loans correctly identified, then tailored neural networks represent a viable direction. However, logistic regression models and the neural networks technique are comparable if the performance metric is the percentage of good and bad loans properly categorized [6]. Unlike NNs, GPs can anticipate results as well as outcomes, and permitting an investigation of the generated genetic programming may also help with comprehending the medical diagnosis. Compared to previously proposed LDFs [7], neural network approaches were more effective in differentiating between glaucomatous and healthy eyes. This development shows that

linear discriminant approaches and neural network techniques have roughly equal potential for use in glaucoma diagnosis [8]. To summarise , neural networks are more dominant than logistic regression in lending, but GP is better in healthcare.

2.2 Vector

The data set, $D = \{(X_1, Y_1), (X_2, Y_2), \cdots\cdots (X_2, Y_2)\}$, there are a total of n data samples, each of which contains d feature attributes and 1 label value, denoted as $X_1, X_2, \cdots\cdots, X_d$, and Y. The i-th data sample is represented as $t = (X_i, Y_i) = (X_{i1}, X_{i2}, \cdots, X_{id}, Y_i)$. For convenience, this article represents dataset D as a combination of attribute set X and label set y, i.e. $D = \{X, Y\} = \{(X_1, Y_1), (X_2 Y_2), \cdots, (X_n, X_n)$. Properties to quantity $X_{iT} = (X_{i1}, X_{i2}, d, X_{id})$ and satisfy the ■ j ≤ 1, $y = (Y_1, Y_2, d, Y_n)$ T was a signal to a corresponding symbol with a d-Viet characteristic [9].

2.3 Artificial Neurons And Perception Machines

$X_i(i = 1,2, \cdots, n)$ represents the input of neurons, $w_i(i = 1,2, \cdots, n)$ represents the weight values between each neuron, and the weighting of the input values and weight values yields the output y. The specific neural network structure is shown in Figure 1, where the circles represent neurons. In this lead, the concept of perceptron and the perceptron model are two two-layer artificial neural networks[10].

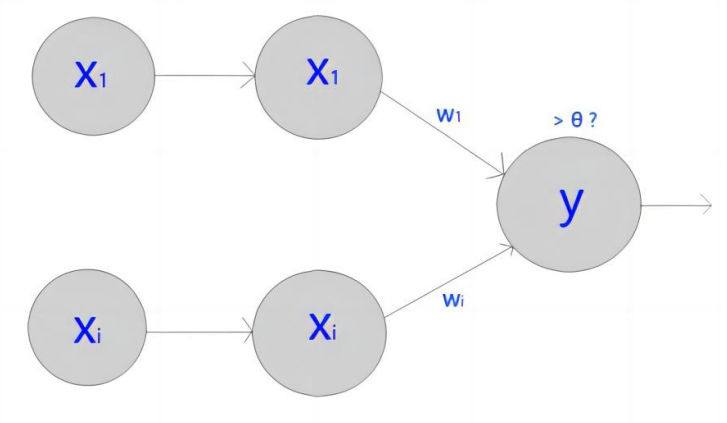

Fig. 1. Work of layers [9]

Here, we take the perceptron model at i=2 as an example. xi(i=1,2) represents the input signal, y is the output signal, and wi(i=1,2) is the weight, without considering the bias term. When each neuron receives the input signal, it is multiplied by the weight w and added up. If the set threshold is exceeded, the lower neural element performs an output operation, which is also known as the "activation" of the neuron. The working principle of the perceptron can be expressed as:

$$\sum_{i=1}^{n} x_i w_i \leq \theta, y = 0 \qquad (1)$$

$$\sum_{i=1}^{n} x_i w_i > \theta, y = 1 \quad (2)$$

3 Neural layered Structure

Biological neurons use electricity to transmit signals. The dendritic part receives signals. The cell body of the neuron processes the signals, and then the signals are transmitted to the next neuron from the axon and Axon terminal. This pattern can be abstracted into a three-layer model: the first layer is the input layer, the second layer is the processing layer, and the third layer is the output layer. So it does three things: inputting data, processing data, and outputting data[9].

3.1 Activation Function

The h(x) function can convert the weighted sum of input signals into output signals. This function is also known as the activation function, which determines how to activate the weighted sum of input signals in neural networks. The input signal is weighted and calculated to obtain node a, which is then converted into node y through the activation function h(a), where "node" and "neuron" represent as[9]:

$$y = h\left(b + \sum_{i=1}^{n} x_i w_i\right) \quad (3)$$

$$h(x) = 0, (x \leq 0) \quad (4)$$

$$h(x) = 1, (x > 0) \quad (5)$$

3.2 Principles Of Neural Networks

Neural networks are divided into forward neural network structures and feedback neural network structures based on the direction of information transmission. The information transmission process of forward neural networks is as follows: input layer → hidden layer → output layer. Each neuron has only one output signal, and it can only be used as an input signal for the next layer and cannot be transmitted back. The characteristic of feedback neural networks is to transmit the output signal of a certain neuron as the input signal of itself or other neurons, which makes the learning process of neural networks more complex based on the feedback of output information. The training and testing of the model also take longer. The structure of this type of neural network is shown in Figure 2.

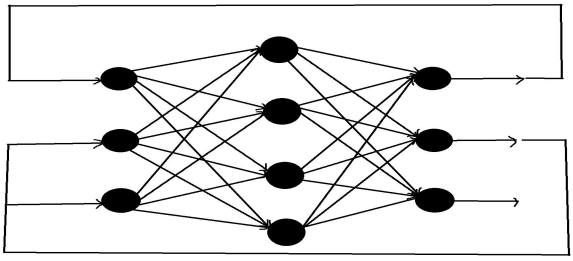

Fig. 2. Neural Networks(Picture credit: Original).

3.3 BP Neural Networks

Back Propagation Neural Network, was established in 1986 mentioned by a research team led by Rumelhart and McClelland proposed in the journal Nature [6]. Its principle is to ensure input update parameters based on error backpropagation while the signal propagates forward. During forward propagation, the input signal of the input layer passes through the hidden pass layer processing to the output layer. When backpropagation occurs, the gradient descent method is used to reduce the error between the actual output and the expected output the new weight value achieves the effect of minimizing error by assigning errors to neurons at each layer. Learning of BP Neural Networks It is necessary to continuously update the connection weights and topology structure between neurons based on the training errors of each batch, to determine the parameters of the neural network model.

4 Multiple Regression Prediction Model

The process of establishing a multiple regression model is to find the regression coefficient w based on the known y and the corresponding y. It is generally believed that the error is the difference between the expected output y value and the actual output y value. To avoid the positive and negative differences of the error canceling out each other, the model error is in the form of a square error.

4.1 The Boundedness Of The Multiple Regression Prediction Model

If the data to be fitted is complex, we blindly pursue multiple regression. When the error of the model reaches its minimum, there is a possibility of overfitting. Overfitting refers to the situation where the model's prediction performance is very good on the training set, but it performs generally or even poorly on the test set. The occurrence of overfitting indicates that the model only considers known sample situations, while its predictive ability for unknown data is average, and the model's generalization ability is poor.

5 Experiment

5.1 Something About the Data Sets

The data sets are taken from the StatLib library maintained by Carnegie Mellon University which is about house prices in Boston, Massachusetts. There are 506 rows in this data set. Each row represents a town in Boston, and each column represents an attribute (CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT), and the housing price we need to focus on MEDV, the median value of owner-occupied homes in $1000's.

5.2 Data Preprocessing

After unzipping the file, we get a plain text file. Then, load the data using the read.csv function, which generates data. frame object and adds the feature names to the data frame. And our target of prediction is the Median value of owner-occupied homes in $1000's(MEDV).

5.3 Feature Selection

It can be seen from the data in the sample that the value range varies greatly. If the network is directly input, the network may automatically adapt to this value range, but learning will be more difficult. Therefore, we will standardize each feature by subtracting the feature mean from each input feature and dividing it by the standard deviation. The resulting feature mean is 0 and the standard deviation is 1. Then, find the relevance between the features and draw a graph of them. Next, split the target variable and independent variables into training data and test data.

5.4 Experimental Setup and Anvironment

In House Price Prediction Based on Densely Connected Neural Networks, three fully connected layers are used. Among them, the DenseNet is a stack of Dense layers, which is used to process vector data. The activation function of the first two layers is ‘real'. The last layer has no activation layer, which allows predicting any range of values.

5.5 The Linear Regression and Neural Layered Structure

Biological neurons use electricity to transmit signals. The dendritic part receives signals. The cell body of the neuron processes the signals, and then the signals are transmitted to the next neuron from the axon and Axon terminal. Simply put, it involves doing a few things: inputting data, processing data, and outputting data. Linear regression, Polynomial regression, Logistic regression, etc. can also be regarded as a neural network. So Linear regression is a part of the Neural layered structure.

5.6 Analysis of Combined Methods

House Price Prediction Based on Densely Connected Neural Networks, uses K-fold cross-validation for model tuning to find the hyperparameter values that optimize the model's generalization performance. Once found, the model is retrained on the entire training set and an independent test set is used to make a final evaluation of the model performance. During each iteration, K-fold cross-validation has only one chance for each sample to be included in the training or test set.

In Linear regression, use cross-validation. This is a technique for evaluating model performance, estimating the true prediction error of models, and tuning model parameters. It helps estimate how well a model generalizes to unseen data and aids in selecting optimal hyperparameters improving the reliability of linear regression predictions.

5.7 Model Running Results

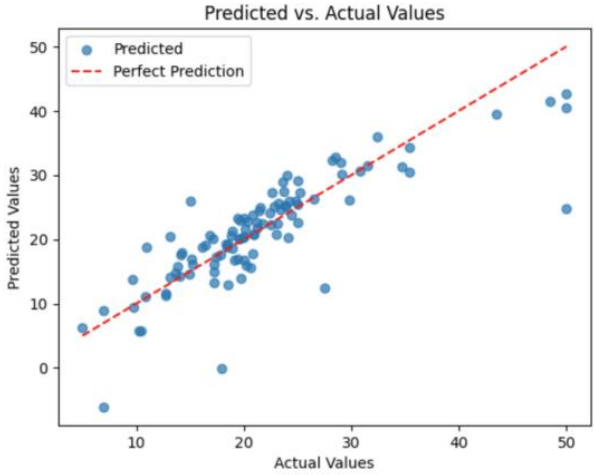

Fig. 3. Linear regression (Picture credit: Original).

As shown in Figure 3, this graph shows the difference between the actual house price and the predicted house price. And overall, the Linear regression model of the predicted value of the house price in Boston is successful. The relationship between the predicted house price and the actual house price can be approximated as a function image of x=y, and the points composed of the predicted and actual values are roughly near the function image, but there are still outliers. Therefore, linear regression still has a certain reference value.

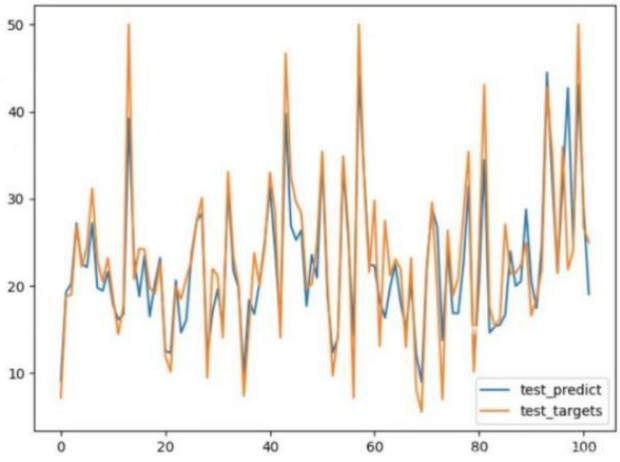

Fig. 4. Neutral network(Picture credit: Original).

As shown in Figure 4, in this line chart, the predicted value fits the target value to a high degree, so the project of using a neural network to predict house prices in Boston is valuable. In the chart, the blue line represents the predicted house price and the yellow chart represents the target house price. We can see from the figure that the target house price fits the predicted house price to a high degree, and the trend and important features are the same, except for some extreme values, such as the cases where the target house price is particularly high and particularly low. All in all, the accuracy of using neural networks to predict housing prices is still very high.

6 Conclusion

In this research, we use the Linear Regression and Neural Network Structure to predict the house price in Boston, Massachusetts, USA. But it's just a small place in the world and the result also has some errors because the house price is always changed, and affected by the amount of things-which include the people's purchasing desire and the guideline of the government-so the graph drawn by Linear Regression is not correct so we need further research in the future.

Authors Contribution

All the authors contributed equally, and their names were listed alphabetically.

References

1. Qiu Shi Predicting Real Estate Price Index Using Internet Search Index [D]. Xiamen University, 2023. DOI: 10.27424/d.cnki.gxmdu.2020.003320
2. Liu Min Analysis of House Price Prediction Based on Network Search Data [D]. Shandong University, 2018
3. Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems With Applications, 42(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040
4. Rawool, A. G., Rogye, D. V., Rane, S. G., & Bharadi, V. A. (2021). House price prediction using machine learning. Int. J. Res. Appl. Sci. Eng. Technol, 9, 686-692.
5. Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In Economic analysis of the digital economy (pp. 89-118). University of Chicago Press.
6. Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. IEEE Transactions on Evolutionary Computation, 5(1), 17-26.
7. J. P. S. Ngan, M. L. Wong, K. S. Leung, and J. C. Y. Cheng, "Using grammar-based genetic programming for data mining of medical knowledge," in Genetic Programming 1998: Proc. 3rd Annu. Conf., 1998
8. Bowd, C., Chan, K., Zangwill, L. M., Goldbaum, M. H., Lee, T. W., Sejnowski, T. J., & Weinreb, R. N. (2002). Comparing neural networks and linear discriminant functions for

glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc. Investigative ophthalmology & visual science, 43(11), 3444-3454.

9. Li Kejia, Hu Xuexian, Chen Yue, et al. Differential Privacy Linear Regression Algorithm Based on Principal Component Analysis and Function Mechanism [J]. Computer Science, 2023,50 (08): 342-351
10. Huang Sheng Analysis of the Neural Network Prediction Model for Real Estate Prices in Xi'an City [D]. Xi'an University of Finance and Economics, 2021. DOI: 10.27706/dcnki.gxacj.2021.000084