Comparing Linear Regression and Decision Trees for Housing Price Prediction

Xiang LI*

[1] Fuzhou Pingdong Middle School of Fujian Province, Fuzhou, Fujian, 350000, China
*xil470@pitt.edu

Abstract. As artificial intelligence and machine learning become more and more advanced nowadays, they have been used in vast fields. As a fundamental and mature topic, housing price prediction remains popular among machine learning workers and researchers. Housing price prediction can contribute a lot to the real estate market and global economy as well as making it much more effective for investors to make decisions. There is a great variety of algorithms in machine learning, and algorithms are still updating as time passes. Housing price prediction applies a reasonable background for researchers conducting machine learning research. Linear regression and decision trees are two popular algorithms in machine learning, which are both possible for housing price prediction. Linear regression can fit a "line" that follows how housing prices change as variables change, while decision trees can also forecast house prices as their trees become deeper and deeper. In this research, the author will compare the performance and accuracy of linear regression and decision trees when used to predict house prices.

Keywords: Housing Price Prediction, Linear Regression, Decision Trees.

1 Introduction

Housing price prediction plays an important role due to its far-reaching implications on both individual households and the broader economy. The real estate market, where these prices fluctuate, has a great influence on economic stability and growth. For many citizens, homes represent the most significant financial asset, and accurate prediction of housing prices can provide buyers, sellers, and investors with more choices. Moreover, the real estate sector substantially contributes to economic activity, encompassing construction, mortgage lending, and related industries, making it a crucial driver of gross domestic product (GDP).

With the development of AI and machine learning, methods predicting housing prices have become more and more diversified. Traditional methods often lack accuracy in capturing the complex web of variables that influence housing prices—factors like location, property characteristics, economic indicators, and market trends. Machine learning models, particularly algorithms like linear regression, decision trees, and neural networks can easily process vast datasets and identify complex

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_9

patterns that human analysis often makes mistakes. By learning from historical price data and identifying correlations between variables and housing prices, these models offer more accurate forecasts. This technology not only aids individuals in making informed real estate decisions but also enables financial institutions, investors, and policymakers to better anticipate market trends and design more effective strategies.

Essentially, machine learning has made great advances in the field of housing price prediction and the real estate market, and machine learning has made a great contribution to providing data-driven insights for economic and personal gain. In this paper, two machine learning models, linear regression and decision trees, will be compared to the accuracy in predicting housing prices by comparing the two models' mean square error and r2 value. However, with vast data and numerous factors involved, predicting housing prices was not easy to conduct.

2 Literature Review

Research has been done to determine the possible factors that can affect housing prices by combining Spearman correlation with multiple linear regression [1]. Dr. M. Thamarai and Dr. S. P. Malarvizhi [2] have conducted research comparing the performances of multiple linear regression and decision trees. A mobile application was made by Aaron Ng [3] to make housing price predictions within London. Research has also been done to predict housing prices using stochastic gradient descent (SGD), random forest, and gradient boosting machine (GBM) [4]. Garcia et al. used the machine learning method to predict housing prices during the COVID-19 period [5]. Adetunji et al. conducted research using random forest to forecast house prices [6]. Deep learning and the ARIMA method were used to predict house prices by the Institute of Electrical and Electronics Engineers(IEEE) [7], while IEEE also conducted research using regression models to predict housing prices [8], but failed to forecast house prices using non-regression models. In 2022, Soltani et al. incorporated machine learning with spatio-temporal dependency [9], which was also used to forecast house prices.

3 Methodology

3.1 Data Processing

With the dataset imported in Python, it underwent several processions. First, it was checked whether it contained any missing or NaN values to guarantee that all the values stayed valid. Next, all the non-numbered data should be converted into numbers so that the computer can better understand what they stand for. Specifically, columns "mainroad" "guestroom" "basement" "hotwaterheating" "airconditioning" and "prefarea" all consisted of whether" yes" or "no". Therefore, all the yeses were converted into "1s" and all the noes were converted into "0s". In particular, the column "furnishingstatus" contains three different statuses, they were: furnished,

semi-furnished, and unfurnished. In this column, data can not simply be converted into a single number. "Furnished" is converted into "0 0", "Semi-furnished" is converted into "1 0", and "Unfurnished" is converted into "0 1".

Moreover, after the research was conducted, it was found that the slope would be extremely huge if the data was not rescaled. The reason was that after all the data was converted into numbers, except for the column "area", all other data was small integers. Therefore, data was rescaled by using MinMaxScaler in sklearn. After the rescale, all the data shared a close value.

Next, the dataset was split into a trainset and a test set. After the data splitting, the train set occupied 80% of the data and the test set occupied 20% of the data.

3.2 Linear Regression

Linear Regression is a foundational technique in statistical modeling and machine learning and operates as a method to establish relationships between variables by fitting a linear equation to observed data. The fundamental principle of Linear Regression is to generate a line that best represents the underlying trend within the data points. This line is characterized by its slope, which signifies the magnitude and direction of the relationship between the independent variables and the dependent variable. Through a process of minimizing the sum of squared differences between the observed data points and the predicted values on the line, Linear Regression calculates the optimal coefficients for the equation. This allows for the estimation of the dependent variable's value based on given independent variables. Renowned for its simplicity and interpretability, Linear Regression serves as a crucial tool in various fields, aiding in predictive modeling, trend analysis, and uncovering causal relationships between variables.

When using linear regression to predict housing prices, a dataset should consist of all the basic characteristics. For example, the dataset used in this research, contains a property's area, bedrooms, bathrooms, stories, location(whether the property is on the main road(Yes/No) and whether the house is located in a preferred area (Yes/No)), number of guestroom, basement(Yes/No), hot water heating(Yes/No), air-conditioning(Yes/No) and furnishing status(furnished, semi-furnished, or unfurnished). What linear regression does in housing price prediction is to generate a best-fitting line between these characteristics and the housing price.

3.3 Decision Trees

Decision Trees is a powerful algorithm in machine learning, operated by recursively partitioning the dataset into subsets based on the values of input features. At each step, the algorithm selects the feature that best separates the data, aiming to maximize the homogeneity of the target variable within each subset. This process forms a tree-like structure where internal nodes represent feature tests, branches correspond to possible feature outcomes, and leaf nodes hold the final predictions. Decision Trees excel in handling complex, nonlinear relationships and are adept at capturing

interactions between variables. They can accommodate both categorical and numerical data, making them versatile for various applications.

The dataset used in decision trees is as same as the one used in linear regression, which is the dataset processed through several steps. When using decision trees to predict house prices, the dataset was also split into a train set(80%) and a test set(20%). Importantly, it should be treated by the DecisionTreeRegressor() method in sklearn. Next, use .fit() and .predict() methods to realize prediction.

3.4 Evaluation

In the research, root mean square error(RMSE) and R-square value are used to evaluate the accuracy and performance of linear regression and decision trees in predicting housing prices.

4 Experiment

4.1 Dataset

The dataset was chosen from Kaggle [10]. The dataset consisted of 13 columns, including price, which is also the value the research trying to predict, area, bedrooms, bathrooms, stories, location (whether the property is on the main road(Yes/No), and whether the house is located in a preferred area (Yes/No)), the number of guestroom, basement(Yes/No), hot water heating(Yes/No), air-conditioning(Yes/No) and furnishing status(furnished, semi-furnished, or unfurnished). Moreover, to make sure that the result was accurate, the dataset consisted of 545 groups of value. The dataset contained values like area, which probably highly corresponded to the housing price, and whether the house had hot water heating, which was probably not so corresponding to the housing price. By doing so, the experiment can test the accuracy of different models.

4.2 Configuration and Parameter

Choosing proper data and parameters in using linear regression and decision trees is vital. For linear regression, data should not be extremely big or small. Therefore, in the research, data was standardized by using MinMaxScaler() in the sklearn library in Python, so that all the values are close. For decision trees, the maximum depth of the tree should be set to avoid excessive branching, thereby mitigating overfitting. Moreover, minimum samples per split should also be set to prevent data from splitting further, thereby enhancing generalization. Therefore, in the research, the parameter "depth" was set to 10, and the parameter "min_samples_split" was set to 5. Moreover, in the research, the train set occupied 80% of the data, and the test set occupied 20% of the data.

4.3 Performance

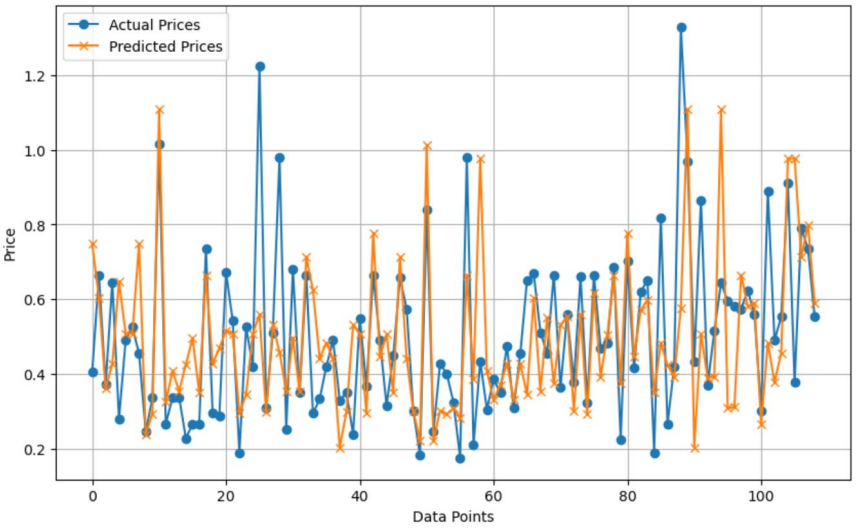

Fig 1. Actual and Predicted Housing Prices(Decision Trees)(Picture credit: Original)

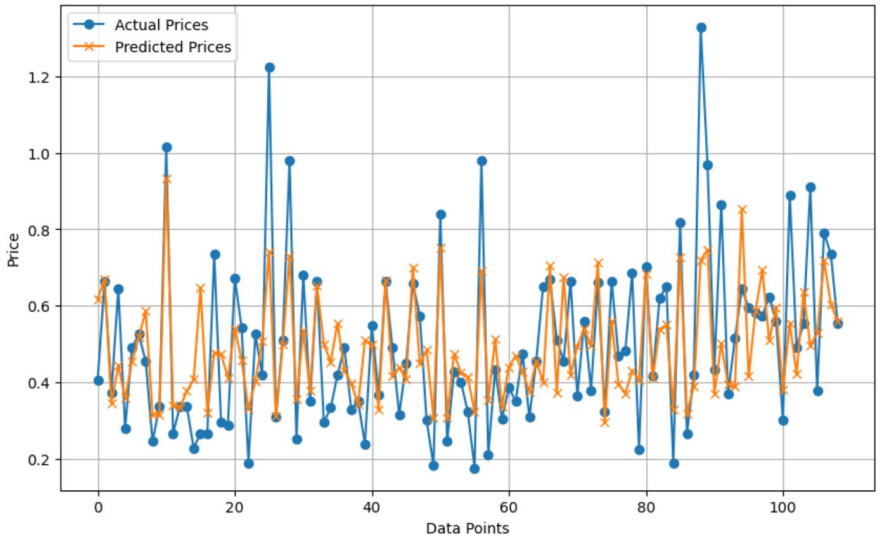

Fig 2. Actual and Predicted Housing Prices(Linear Regression) (Picture credit: Original)

As is shown in both figures, the x-axis stands for different data points, and the y-axis stands for the housing price. Moreover, the blue-dot line stands for the actual prices, which is also the y-test data when conducting the research. Conversely, the orange-cross line stands for the predicted prices, which is also the data that the machine generated. The distance between two dots within one column is the difference between the actual price and the predicted price. As shown in Figure 1, the maximum value of the predicted housing price by decision trees is 11083333.333333334, while the minimum of the predicted price by decision trees is 2030000.0. On the other hand, the maximum value of the predicted housing price by linear regressions is 9338793.18082853, while the minimum of the predicted price by linear regression is 2963780.6043453473. As the variables change, the predicted price fluctuates.

It also calculated the root mean square error(RMSE) and R-square of the two models by using sklearn. It turned out that the RMSE for decision trees was 2017096.3374790787 and the R-square for decision trees was 0.19504973986626317. Conversely, the RMSE for linear regression was 1567718.610600157 and the R-square for decision trees was 0.5137585349037072. As the result is shown, in the research, the RMSE of linear regression is smaller than that of decision trees, R-square of linear regression is greater than that of decision trees. Therefore, a conclusion can be drawn that in this research, linear regression performed better than decision trees in housing price prediction.

4.4 Pros and Cons

For linear regression, it has several advantages:
1. Simple: Linear regression is a straightforward algorithm with a simple principle. It is easy to comprehend and conduct.
2. Feature-sensitive: Linear regression allows users to identify the importance of different variables and features based on their coefficients.
3. Linear fit: When the situation comes to a linear or near-linear model, linear regression can well fit the situation and make reasonable and close predictions.
However, linear regression also has disadvantages:
1. Inflexible: While linear fit becomes its advantage, it also becomes linear regression's disadvantage. When the situation becomes more complex and not linear, linear regression can not make good predictions.
2. Outlier-sensitive: In linear regression, outliers can contribute huge errors to the result.
For decision trees, it has several advantages:
1. Flexibility: Compared to linear regression, decision trees can better handle more complex and non-linear situations. However, in housing price prediction, linear regression performed better than decision trees (because housing price prediction is linear).

2. Robust to outliers: Compared to linear regression, decision trees showed better robust ability to outliers. Outliers contribute less damage to the overall structure of decision trees.
It also has disadvantages:
1. Overfitting: Decision trees are easy to become overfitting, especially when it becomes deeper and more complex.
2. Instability: Small changes in the data can change tree structures, making decision trees less stable for making consistent predictions.

5 Conclusion

In conclusion, linear regression performed better than decision trees when predicting housing prices. Linear regression has a smaller RMSE than decision trees. Linear regression also generated an R-square that was closer to 1 compared to that of decision trees. The result is reasonable because housing price prediction is more of a linear model so linear regression showed better performance. Linear regression made better predictions in housing prices. However, decision trees also showed well predictions.
The research has been done to compare the accuracy between linear regression and decision trees when predicting housing prices, providing reasonable results to the field of machine learning in housing price prediction. Linear regression remains a better performance than decision trees in linear situations.

However, more methods can still be used to predict house prices, including neural networks and random forest. More work should also be done to predict house prices using these and other methods. Housing price prediction in machine learning is still a complex job to conduct. Except for simply using numbers to make predictions, other factors like citizens' demand, economic conditions, government policy, etc. Future research should focus more on comprehensive factors that can affect housing prices by using more complex models. While housing price permeates into vast fields, housing price prediction remains important nowadays.

References

1. Zhang, Q. (2021). Housing price prediction based on multiple linear regression. Scientific Programming, 2021, 1–9. https://doi.org/10.1155/2021/7678931Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
2. Thamarai, M., Malarvizhi, S. P., House Price Prediction Modeling Using Machine Learning. MECS (1999).
3. Aaron, Ng.: Machine Learning for a London Housing Price Prediction Mobile Application. Imperial College, London (2015).
4. Ho, W. K., Tang, B., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

5. García, R. T. M., Céspedes-López, M. F., & Perez-Sanchez, V. R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. Land, 11(11), 2100. https://doi.org/10.3390/land11112100
6. Adetunji, A., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science, 199, 806–813. https://doi.org/10.1016/j.procs.2022.01.100
7. House Price Prediction Approach based on Deep Learning and ARIMA Model. (2019, October 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8962443
8. Machine Learning based Predicting House Prices using Regression Techniques. (2020, March 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9074952 Soltani, A., Heydari, M.,
9. Aghaei, F., & Pettit, C. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 131, 103941. https://doi.org/10.1016/j.cities.2022.103941
10. Housing price prediction. (2023, July 7). Kaggle. https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction