Machine Learning for House Price Prediction

Zixian Ning *

School of Myddelton College, Jinhua, China

* Corresponding Author Email: student_zixian.ning@myddeltoncollege.cn

Abstract. Housing stands as a core human necessity, and residential property prices—shaped by elements such as local crime levels and air quality—hold great significance for all stakeholders in the real estate sector. A dependable price prediction system plays a key role in supporting well-informed decisions related to property transactions. This research utilized the Boston housing dataset sourced from Kaggle and carried out all-round data preprocessing: it addressed missing or incorrect values, removed outlier data points, applied Z-score normalization to reduce data complexity, and conducted data transformation as part of the data wrangling process. Three machine learning models were evaluated in the study, namely Support Vector Regression (SVR) with a radial basis function kernel, a five-layer Artificial Neural Network (ANN) consisting of three dense layers and two dropout layers, and Extreme Gradient Boosting (XGBoost). Additionally, five-fold cross-validation was employed to select regularization parameters, and principal component analysis was used to enhance model performance. Experimental results revealed that the Artificial Neural Network outperformed the other two models, achieving the highest R-squared value of 0.86, the lowest Mean Squared Error of 0.0046, and the smallest Mean Absolute Error of 0.047. Support Vector Regression also delivered strong performance, with an R-squared of 0.83 and a Mean Squared Error of 0.0054, while XGBoost exhibited relatively higher errors, with a Mean Squared Error of 0.1214.

Keywords: Machine learning, house price prediction, artificial intelligence.

1. Introduction

A home is among the most essential elements of human life, alongside food, water, and various other necessities. Generally, housing prices are influenced by a range of factors. The interest in understanding housing price trends and the factors that determine them extends beyond tenants; it also includes homeowners, real estate professionals and policymakers, as well as urban and regional planning authorities. With the assistance of a computer-based prediction system, these groups can make informed decisions about whether to purchase a property and when to do so.

Machine learning whose primary objective is to enable computer systems to automatically detect patterns in data and generate predictions or make decisions without the need for explicit programming [1]. Unlike traditional programs, machine learning models can extract meaningful patterns from historical data and apply this knowledge to new, previously unseen data. This characteristic enables them to achieve remarkable success in challenging tasks such as financial forecasting, image recognition, and natural language processing. Over the past decade, machine learning has made significant progress, moving from academic and experimental environments to widespread applications in various industries. From financial forecasting and medical diagnosis to autonomous driving and smart manufacturing, AI technologies are fundamentally transforming the ways in which societal production operates. Looking forward, as computing capabilities continue to advance and algorithmic innovations emerge, the scope and impact of machine learning applications are expected to expand even further.

In recent years, machine learning has emerged as a crucial prediction method because it can forecast property values more accurately based on their characteristics, without relying on data from previous years. Numerous studies have explored this subject and confirmed the effectiveness of machine learning approaches [2-4]. To obtain more accurate final predictions, Random Forest, a type of ensemble model, combines the predictions of multiple decision trees. Previous research has proven that Random Forest is a powerful tool [5]. Raschka and Mirjalili [6] have outlined the steps that constitute the Random Forest algorithm: 1. Generate a bootstrap sample of size n by randomly

selecting n samples from the training set with replacement. 2. Using the bootstrap sample, construct a decision tree at each node: (a) Randomly select d characteristics without replacement. (b) Based on the feature that maximizes information gain, split the node using the optimal split according to the objective function.

This report reviews and examines the current research on the key factors affecting housing prices, as well as the data mining techniques used to predict housing values. Theoretically, houses in prime locations—such as those near shopping malls or other facilities—are usually more expensive than those in rural areas with fewer amenities. An accurate prediction model would allow both housing developers and investors to determine the affordable price of a home and the reasonable market price for a property. This study discusses the characteristics that previous researchers used to predict housing prices through different prediction models. All the findings of the survey indicate that Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Support Vector Regression (SVR) are capable of estimating housing prices. These models are developed based on a variety of input characteristics and have a significant positive impact on housing price prediction. In summary, the purpose of this study is to support and assist future researchers in developing a practical model that can easily and accurately predict housing values. The results of this research should be used to validate these models in further work on real-world models.

2. Method

2.1. Dataset Preparation

This study obtained the Boston housing dataset from Kaggle [7], which contains information on various features that affect housing prices, including Crime Rate (CRIM), Business Proportion (INDUS), Proximity to River (CHAS), Air Quality (NOX), and Age of Houses (AGE).

Since housing prices are continuous variables, regression analysis seems more logical. However, classifying housing prices into specific price ranges can also provide valuable insights for users; it also helps to explore different techniques that may be specific to either regression or classification. Given that the dataset contains 288 features, regularization is necessary to avoid overfitting. To determine the regularization parameter, during both the classification and regression phases of the project, this study first performed K-fold cross-validation with $k = 5$ on a broad range of regularization parameter options. This facilitated the selection of the optimal regularization settings during the training phase. Additionally, this study implemented principal component analysis pipelines for all models to further improve their performance and cross-validated the number of components suitable for each model to achieve optimal results.

Before data is input into a model, it undergoes a refinement process known as data preprocessing. The four stages of data preprocessing can be roughly categorized as follows: a) data cleansing b) data editing c) data reduction d) data wrangling. For inaccurate data or empty data fields, the solution involves either removing the entire record from the dataset or filling the empty field with the mean or median value. These problems often occur when data is recorded manually. The mean value is calculated by considering the number of records in the dataset and the values of the characteristics. Data editing refers to the process of identifying and removing outliers from the data. Outliers in data are mainly caused by experimental errors, such as those resulting from faulty equipment or other factors. Data reduction is the technique of simplifying data by applying some form of normalization to make data processing more manageable. One commonly used method for normalization is the Z-score. Data wrangling is the process of transforming or mapping data, which includes data aggregation, data visualization, and data munging. Data visualization is the technique of creating graphs using statistical data. Data aggregation involves filtering data before it is fed into a model.

2.2. Machine Learning Models

In this study, all models were trained and optimized using a portion of the dataset. The same dataset was divided into training and testing sets, with 75% used for training and 25% selected randomly for

testing. Linear regression is a statistical method used to predict a dependent variable by establishing a relationship between the dependent variable and one or more independent variables [8]. This method assumes a linear relationship between the independent and dependent variables. In this research, it is assumed that all features in the dataset have a linear relationship with housing prices. For the linear regression model, the sklearn. linear_model Python library was used.

2.2.1 SVR

Support Vector Regression (SVR) is a variation of the Support Vector Machine (SVM) algorithm and is applied to regression tasks rather than classification tasks. Similar to the support vector machine, this method finds the hyperplane in the high-dimensional feature space that best fits the data points, while ensuring that the maximum number of points lie within a specific hyperplane margin. Several types of kernel functions are commonly used in this method, including linear, polynomial, and radial basis functions. This allows SVR to represent features that have non-linear relationships with the target variable and enables the algorithm to map the input features into a higher-dimensional space. In this study, after comparing the performance of all available kernels, the radial basis function was selected as the kernel function because it clearly demonstrated higher accuracy compared to the other two kernels. To build this model, the sklearn.svm library was used, and the kernel of the Support Vector Regression was set to the radial basis function [9].

2.2.2 ANN

The ANN model is a multi-layer perceptron with ReLU and linear activation functions in the output layer, enabling it to learn complex non-linear relationships. This study used TensorFlow to construct the ANN model, utilizing Dense, Batch Normalization, and Dropout layers from the TensorFlow Keras library [10]. The five-layer Artificial Neural Network (ANN) model consists of three "dense" layers and two "dropout" layers. The output shape of the first "dense" layer is (None, 128), which means it contains 128 neurons and 1792 parameters. It is followed by a "dropout" layer, which has no parameters and also has an output shape of (None, 128). The second "dense" layer has an output shape of (None, 128), with 128 neurons and a total of 15512 parameters. The next "dropout" layer has no parameters and an output shape of (None, 128). The output shape of the final "dense" layer is (None, 1), indicating that it has only one neuron and 129 parameters.

2.2.3 XGBoost

XGBoost is the abbreviation for extreme gradient boost regression. Compared to other machine learning algorithms, it performs exceptionally well. Based on its workflow, XGBoost is one of the top-performing supervised learning algorithms; it consists of base learners and an objective function. XGBoost's ensemble learning approach considers a large number of models, known as base learners, to predict a single value. It is not expected that all base learners will produce poor predictions; instead, when the total number of predictions is summed up, the good predictions will offset the bad ones. When provided with features, a regressor fits a model and predicts an unknown output value.

3. Results and Discussion

All trained models were successfully trained and evaluated using the same testing dataset. The ability of each model to predict housing prices with different levels of accuracy is presented in Table 1 below:

Table 1. Model performance comparison

Algorithm	MSE	R-squared	MAE
XGBoost	0.1214	0.85	0.2420
ANN	0.0046	0.86	0.047
SVR	0.0054	0.83	0.056

In this performance evaluation, the Artificial Neural Network outperformed the other two algorithms in all metrics. The SVR with the RBF kernel also performed well, showing a lower MSE than XGBoost.

R-squared is a statistical metric that indicates the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. A higher R-squared value indicates a better fit of the model to the data. The ANN's R-squared value of 0.86 means that the model can explain 86% of the variance in the target variable. This is the highest score among all the algorithms evaluated. XGBoost also achieved a high R-squared value, followed by the SVR with the RBF kernel. The R-squared values of these three algorithms are relatively close to each other.

The results from the evaluation metrics consistently show that the Artificial Neural Network (ANN) model is superior, as it outperforms the other regression algorithms. It demonstrated a better fit to the data and higher predictive accuracy by achieving the lowest MAE, the highest R-squared value, and the lowest MSE.

The primary goal for future research is to expand the scope of the study while addressing its current limitations. This research intends to collect a large amount of data to improve the accuracy of the results for future enhancements. Furthermore, this study aims to extend the data coverage to larger regions, such as entire states or countries. It is essential to use information from reliable and trustworthy sources. Additionally, this study can utilize multiple datasets to conduct comparative analyses of results, which will help to improve the prediction accuracy and the generalization ability of the model.

However, there are challenges associated with acquiring large amounts of data, such as the requirement for significant computing time. In this study, hyperparameter tuning for a single model—with fewer than 3,000 observations—took approximately eight hours. This time could be much longer when dealing with millions of observations. Two potential solutions are to optimize the settings to reduce computing time and to invest in high-performance workstations. Another aspect that should be considered is spatial interpolation, which takes into account the locations of amenities such as parks, hospitals, and train stations—all of which have a significant impact on housing values [11].

4. Conclusion

This study aimed to apply machine learning to predict housing prices, a topic of great importance to all individuals and entities involved in the real estate industry. It used the Boston housing dataset from Kaggle, which includes details such as crime rate and air quality that influence housing prices. Before training the models, the study conducted necessary data preprocessing: in the data cleansing stage, it fixed incorrect or missing data; in the data editing stage, it removed abnormal data points; in the data reduction stage, it simplified the data through normalization; and in the data wrangling stage, it transformed the data format. Then it tested three models. The test results showed ANN worked better than the other two. It had the smallest Mean Squared Error (0.0046), the highest R-squared (0.86), and the smallest Mean Absolute Error (0.047). Right now, the study's data size is not big enough. Later, it will use more data and spatial interpolation methods to make price predictions more accurate.

References

[1] Pandey D, Niwaria K, Chourasia B. Machine learning algorithms: a review. Mach Learn. 2019;6(2).

[2] Fan C, Cui Z, Zhong X. House prices prediction with machine learning algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018). 2018. doi:10.1145/3195106.3195133.

[3] Phan TD. Housing price prediction using machine learning algorithms: the case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE). 2018. doi:10.1109/ICMLDE.2018.00017.

[4] Mu J, Wu F, Zhang A. Housing value forecasting based on machine learning methods. Abstr Appl Anal. 2014; 2014:1–7. doi:10.1155/2014/648047.

[5] Breiman L. Random forests. Mach Learn. 2001; 45:5–32. doi:10.1023/A:1010933404324.

[6] Raschka S, Mirjalili V. Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow. 2nd ed. Birmingham: Packt Publishing; 2017.

[7] Kaggle. The Boston housing dataset. Available from: https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset [accessed 2025 Sep 11].

[8] Parekh J. House price prediction using linear regression model. Int J Multidiscip Res. 2023.

[9] Karlik B. Development of optimum ANN structures for house price estimation. Int J Comput Appl. 2017; 1:21–32.

[10] Matey V, Chauhan N, Mahale A, Bhistannavar V, Shitole A. Real estate price prediction using supervised learning. In: 2022 IEEE Pune Section International Conference (PuneCon). IEEE; 2022. p. 1–5.

[11] Viana D, Barbosa L. Attention-based spatial interpolation for house price prediction. In: Proceedings of the 29th International Conference on Advances in Geographic Information Systems. Beijing, China; 2021 Nov 2–5. p. 540–9.