Comparison Of Linear and Nonlinear Models in Predicting House Prices

Zuyu Lin

School of Economics and Management, Southeast University, Nanjing, Jiangsu, China

213240748@seu.edu.cn

Abstract. Nowadays, the increase in housing prices has led to a heavier financial burden on ordinary citizens, affecting their quality of life. This research used linear regression, random forest, and Support Vector Machine (SVM) to predict housing prices using the Boston housing price dataset, and compared the advantages and disadvantages of them. The paper compares the performance of linear and nonlinear models in predicting housing prices to reflect the importance and magnitude of various factors that affect housing prices. The overall research demonstrates that nonlinear models have unique advantages over linear models in multi-factor impact problems such as predicting housing prices. This is due to the comprehensive examination of multiple factors by nonlinear models, which linear models do not possess. The project still has many shortcomings in many aspects, such as not tuning the parameters of the SVM. There is also no addition of cross-validation, and a lack of visualization of the results.

Keywords: Linear and Nonlinear Models, Predicting House Prices, Support Vector Machine.

1. Introduction

Although the rise in housing prices has indeed driven the entire economic development, it is also a significant social problem, because housing prices are also very closely related to the cycle of the financial crisis [1]. As a key industry highly associated with the financial system, real estate is known as the mother of finance [2]. Historically, the subprime mortgage crisis in the United States originated from the bursting of the foam in real estate market bubble and eventually triggered a global economic recession. This reality fully demonstrates the importance of accurately predicting the trend of housing price changes. So, the research wants to use this report to discuss the gains and losses of using three models to predict housing prices. Through machine learning methods, we can use some models to explore how to predict housing prices, and explore their future trends, future development, and some corresponding countermeasures.

The main theme of the project this time is to compare the models. In order to better reflect the different roles of various models in prediction, three models are used: linear regression, random forest, and SVM, and the classic Boston housing price dataset is selected for training. The models are compared to better predict future housing price changes and better grasp future economic trends.

Housing price prediction is a critical task in real estate and finance, impacting decisions for buyers, sellers, investors, and policymakers. Machine learning (ML) models have increasingly been applied to this domain due to their ability to handle complex datasets and capture non-linear relationships. Support Vector Machines (SVM) and linear regression models have demonstrated

Excellent accuracy in previous studies, A 2024 study using Kaggle's King County dataset found XGBoost outperformed Support Vector Machines (SVM) and Linear Regression, achieving the highest accuracy (91%) and lowest error rates [3]. However, in the research, the linear regression model did not perform well [4]. Another study in Adelaide, Australia, spanning 32 years of housing data, reported that ensemble methods (GBM, Random Forest) better captured spatio-temporal dependencies than linear models. This result is similar to the findings presented in the research. Neural networks, including Multilayer Perceptrons (MLPs) and Long Short-Term Memory (LSTM) networks, show promise but exhibit instability [5]. A 2022 study noted MLPs underperformed (R^2=0.584) compared to the tree-based model. While LSTMs achieved moderate success (R^2=0.863) in time-series forecasting. Structural Time Series Analysis (STSA) was proposed for German real

estate markets, explicitly modeling cyclical components and non-stationary trends [6]. This viewpoint has also been referenced in the research [7]. A 2020 study on commodity futures predicted 3.74% sample-out-of-sample returns using decision trees, ten times that of linear models. This viewpoint leads me to speculate that non-linear models such as random forests perform better than linear regression models.

2. Methodology

2.1. Dataset

The Boston dataset has a wide range of data and multiple influencing factors, covering various factors. It can better explore the role of random forests in reflecting the importance of various factors and the fitting degree of various models. At the same time, it can be compared with many previous studies on the Boston housing price dataset. The problem encountered was whether to handle missing values during data preprocessing, but the Boston dataset is usually complete, so confirmation is sufficient. The Boston dataset and sample size are not large, and random forests may have the risk of overfitting, but the default parameters perform relatively well on the Boston dataset.

2.2. Models

Linear Regression. Linear regression maintains a foundational role in machine learning as a transparent and computationally efficient method for modeling linear relationships. Its primary strength lies in interpretability—coefficients explicitly quantify the directional impact and magnitude of each feature on the target variable, a critical advantage in domains requiring auditable decision-making, such as healthcare diagnostics or financial risk assessment. The model's simplicity enables rapid implementation, making it a practical baseline for benchmarking more complex algorithms. Diagnostic metrics like residual plots and R-squared values also provide immediate insights into data quality issues, such as heteroscedasticity or nonlinear patterns, which often persist even when using advanced models.

However, the method's rigid assumptions frequently constrain its applicability. The presumption of linearity rarely aligns with complex real-world systems, while sensitivity to outliers and multicollinearity can produce unstable coefficient estimates. Although techniques like polynomial feature engineering or regularization (e.g., Lasso, Ridge) partially address these limitations, they introduce trade-offs in model complexity without guaranteeing robust performance. The assumption of independent, homoscedastic errors further falters in scenarios involving temporal or spatial data dependencies.

Despite these constraints, linear regression remains widely deployed due to its computational lightness and pedagogical value. It serves as an introductory framework for understanding core machine learning concepts like loss minimization and gradient descent. In industrial applications, regularized variants often feature in high-dimensional prediction tasks, leveraging sparsity-inducing penalties for feature selection. While overshadowed by nonlinear models in tasks involving intricate patterns, linear regression persists as a pragmatic choice when data scarcity, interpretability requirements, or operational efficiency outweigh the need for higher predictive accuracy. Its enduring relevance underscores a key machine learning principle: model complexity should align with problem constraints rather than defaulting to algorithmic sophistication (Fig. 1.).

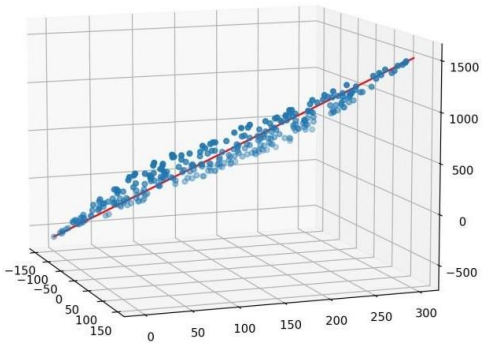

Fig. 1. Presentation of Multiple Linear Regression (Photo/Picture credit: Original).

Random Forest. Random Forest is a robust ensemble learning method that excels in handling diverse machine learning tasks through its combination of decision trees and randomness. Its primary strength lies in its ability to mitigate overfitting by aggregating predictions from multiple uncorrelated trees, trained on bootstrapped data subsets and random feature selections. This design typically yields high accuracy across classification and regression problems, particularly with high-dimensional datasets containing non-linear relationships or noisy features. The model inherently manages missing values and maintains resilience to outliers, reducing the need for extensive data preprocessing compared to more sensitive algorithms like linear regression. Feature importance scores generated during training also provide actionable insights into variable contributions, aiding interpretability in complex scenarios.

However, the method has notable limitations. Computational costs escalate with tree quantity and data size, making it less practical for real-time applications or resource-constrained environments. While feature importance metrics offer partial interpretability, the "black-box" nature of aggregated tree predictions complicates model debugging and stakeholder communication, a critical drawback in regulated industries like healthcare or finance. Performance may also plateau or degrade in tasks requiring extrapolation beyond training data ranges or when dealing with strictly linear relationships, where simpler models often outperform with lower computational overhead.

In practice, Random Forest serves as a versatile default choice for exploratory modeling and baseline establishment, especially when data patterns are unclear or relationships involve complex interactions. It frequently underpins applications like recommendation systems, genomic data analysis, and fraud detection, where prediction accuracy outweighs interpretability needs. While newer algorithms like gradient-boosted trees or neural networks may achieve superior performance in specific contexts, Random Forest's balance of reliability, ease of use, and built-in diagnostics ensures its continued relevance. Its effectiveness ultimately depends on aligning model strengths, such as handling heterogeneous features and non-linear patterns, with problem-specific requirements, while acknowledging trade-offs in speed and transparency.

Support Vector Machine. SVM is an important model in machine learning, which mainly includes the following three types: 1. Linear Separable Support Vector Machine: When the training data is linearly separable, this type of SVM can perform at its best. It divides data points of different categories by finding a hyperplane. Two Linear support vector machine: For linearly inseparable data, linear support vector machine can handle some special cases by introducing relaxation variables, thereby performing linear partitioning to a certain extent. Three Nonlinear support vector machine: When the dataset cannot be partitioned by linear functions, a nonlinear support vector machine comes in handy. It maps data to a high-dimensional space through kernel techniques, making originally linearly inseparable data separable. This optimal hyperplane is called the decision boundary. Specifically, SVM aims to find a hyperplane that maximizes the distance between sample points closest to the hyperplane. Therefore, the decision boundary of SVM is entirely determined by these support vectors, which is why SVM is named a support vector classifier.

2.3. Evaluation Methods

Mean Squared Error (MSE) is a metric used to measure the difference between a model's predicted value and its true value, quantifying the model's prediction accuracy by calculating the average of the squared prediction errors. Its core function is to evaluate the performance of the regression model, with smaller values indicating that the prediction is closer to the true value. The mathematical definition of MSE is the average of the squared prediction errors of all samples. The characteristics of MSE include: simple calculation: the calculation formula of MSE is intuitive and easy to understand, and is easy to program and implement. High sensitivity: sensitive to outliers, able to reflect the predictive ability of the model in extreme situations. Wide applicability: MSE is one of the standard methods for evaluating the accuracy of model predictions in regression problems. However, MSE also has some drawbacks, such as being overly sensitive to outliers. In extreme cases, individual outliers may greatly affect the value of MSE, leading to less robust model evaluation results. Meanwhile, MSE itself is only a numerical value and difficult to directly interpret as an intuitive description of model performance. MSE is widely used in model evaluation of various regression problems, including but not limited to housing price prediction, stock trend analysis, weather prediction, and other fields. In these scenarios, MSE can help us quantify the accuracy of model predictions, thereby guiding us in model selection and tuning. Share and answer again.

R-squared (R2) is an important indicator for measuring the fit of a linear regression model. It represents the percentage by which the independent variables can collectively explain the variance of the dependent variable. However, it is worth noting that a higher R-squared value is not always better. Sometimes, excessively high R-squared values may actually mask other potential issues with the model. Therefore, when evaluating the goodness of fit of a regression model, we need to consider multiple factors comprehensively, rather than just relying on the high or low R-squared value.

The goal of linear regression is to find an equation that minimizes this gap, specifically, to find an equation that minimizes the sum of squared residuals. In statistics, if the difference between observed values and predicted values is small and unbiased, we say that the regression model can fit the data well.

However, before relying solely on numerical indicators such as R-squared to evaluate the goodness of fit, we should first examine the residual plot. Residual plots can more effectively discover potential biases by revealing patterns in residuals. If the model has bias, its results may not be reliable. We should only further evaluate R-squared and other statistical data when the residual plot shows good patterns.

3. Result

Table 1. MSE and R-squared values predicted using three models

Model	Mean Squared Error (MSE)	Coefficient of Determination (R^2)
Linear regression	24.29	0.67
Random forest	7.91	0.89
Support Vector Machine (SVM)	13.00	0.82

3.1. Results Analysis

Random forests perform the best due to their ability to handle nonlinear relationships, with the smallest MSE value of 7.91 and the closest R-squared value of 0.89 (Table 1).

SVM performance performs second and may require parameter tuning (such as using grid search to adjust C and gamma parameters).

Linear regression performs poorly in all aspects, with MSE values reaching 24.29 and R-squared reaching 0.67, the lowest among the three, indicating that there is a certain linear relationship in the data, but it is not strong.

Linear regression models have always been considered a good choice for predicting housing prices, but it has been proven that random forest models can reflect the importance and magnitude of various factors that affect housing prices, as well as more accurate predictions of housing price trends and better models for housing price trends. SVM did not perform well, possibly due to the lack of parameter adjustment. However, the overall research still demonstrates that nonlinear models have unique advantages over linear models in multi-factor impact problems such as predicting housing prices.

4. Shortcomings and Prospects

4.1. Shortcomings

The project has many shortcomings in many aspects, such as not tuning the parameters of the SVM. There is also no addition of cross-validation, and a lack of visualization of the results. In addition, it lacks application to multiple linear regression models when it comes to linear regression models. The evaluation of linear regression models may be too arbitrary. Additionally, there is still room for improvement in feature standardization. Among the three models, the random forest model may have the risk of overfitting, and SVM requires too long a training time. Additionally, my code structure still needs improvement. I hope that in the future, there will be research on predicting housing prices, making more objective, scientific, and intuitive comparisons that include more models.

4.2. Prospective Comparison of Linear Regression, Random Forest, and SVM: Trends and Future Directions

Linear Regression. Linear regression remains a cornerstone for interpretable modeling, particularly in domains requiring causal inference and transparency. For instance, in osteoporosis prediction among elderly cardiovascular patients, logistic regression outperformed SVM and random forest (RF) with an AUC of 0.751, attributed to its ability to identify key predictors like low-sodium diets and genetic variants [8]. However, its limitations in handling multicollinearity and nonlinear patterns are evident in housing price prediction tasks, where SVM and neural networks achieve lower RMSE [9].

Random Forest. RF excels in high-dimensional and heterogeneous datasets. In cardiovascular disease prediction, RF achieved a macro-accuracy of 94% using 5-fold cross-validation, leveraging its ensemble structure to capture interactions between features like cholesterol levels and exercise-induced angina [10]. Despite its robustness, RF struggles with overfitting in low-dimensional data and lacks interpretability, as seen in regression tasks where linear models outperformed RF (RMSE 0.27 vs. 0.29) [11].

Support Vector Machine (SVM). SVM demonstrates superiority in small-sample, high-dimensional scenarios. For hyperspectral classification of dehydrated banana slices, SVM with polynomial kernels achieved 96.07% accuracy, surpassing convolutional neural networks (81.39%). Yet, its computational inefficiency and sensitivity to hyperparameters limit scalability in real-time applications like agricultural monitoring [12].

5. Conclusion

Overall, it processed the Boston housing price dataset using three models: linear regression, random forest, and SVM, and then calculated their MSE and R-squared scores. The paper first imported some necessary libraries, preprocessed some data, and defined some evaluation indicators. Then, it loaded the Boston dataset and performed some preprocessing to standardize features. Then it divided the data into a training set and a testing set, and trained these three models separately to predict the test results. Finally, it calculated MSE and R-squared and concluded.

Random forest performs the best, with the smallest MSE value of 7.91 and the closest R-squared value of 0.89. SVM performs second best and may require parameter tuning. Its MSE value is about

13.00, and R-squared is about 0.82, while linear regression performs poorly in all aspects, with MSE values reaching 24.29 and R-squared reaching 0.67, the lowest among the three, indicating that there is a certain linear relationship in the data, but it is not strong.

However, the overall research still demonstrates that nonlinear models have unique advantages over linear models in multi-factor impact problems such as predicting housing prices. Overall, nonlinear models perform better in predicting housing prices. This is due to the comprehensive examination of multiple factors by nonlinear models, which linear models do not possess.

References

[1] Lan, X.: Inside China's Development: The Role of the Government and the Economy. Shanghai People's Publishing House, Shanghai (2021)

[2] Ren, Z.: Real Estate Cycle. People's Publishing House, Beijing (2017)

[3] Wang, Y.: Research on machine learning based housing price prediction. Finance 14(4), 1552-1562 (2024)

[4] Soltani, A., et al.: Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 103941 (2025)

[5] Sahu, M.: House Price Prediction Using Machine Learning. IJRASET 4(3), 190 (2022)

[6] Wernecke, M.: Structural time series analysis for real estate forecasting. ERES (2024)

[7] Kintzel, J.D.: Price Prediction and Deep Learning in Real Estate Valuation Models. Harvard University (2025)

[8] Peng, L., et al.: Osteoporosis prediction in high-risk cardiovascular patients. BMC Geriatrics (2025)

[9] Liu, Y.: The principle of R language support vector classifier SVM and its application in predicting housing price data. Tencent Cloud Community (2025)

[10] Abdullah, A.: Comparative analysis of heart disease prediction. Scientific Reports 3(8) (2025)

[11] Dong, Y.: Comparison of RF and traditional methods. Statistics and Application (2023)

[12] CSDN Blog.: SVM vs. Random Forest: Advantages and limitations. (2024)