Predicting House Prices with Machine Learning: A Comparative Study of Linear, Lasso, and Ridge Regression

Hanrun Hu *

Guanghua Cambridge International School, Shanghai, China

* Corresponding Author Email: harryrun@126.com

Abstract. Machine Learning has seen rapid advancements this decade, marked by daily innovations in applications and algorithms. One significant application is in the real estate sector, particularly in predicting house prices—a crucial task given the continual rise in property values. This project employs Linear Regression, Lasso Regression, and Ridge Regression, utilizing the Scikit-Learn tool to explore the potential of these models. By considering various factors such as the number of bedrooms, the age of the house, proximity to transportation hubs, local schools, and shopping centers, these models aim to assist potential buyers in identifying homes that best meet their requirements. Among the tested methods, the Lasso Regression emerged as the most effective, boasting the lowest root-mean-square error of 190,124.751, which notably outperformed both Ridge and Linear Regression. This outcome not only underscores the efficacy of Lasso Regression but also highlights the overall utility of regression models in accurately predicting house prices.

Keywords: Machine Learning, Linear Regression, Lasso Regression.

1. Introduction

House price prediction is achieved based on the importance of different factors influencing the price of houses [1-3]. Real estate markets reflect economic vitality, with housing prices influenced by supply and demand, interest rates, inflation, and growth. In such situations, the prediction based on traditional models e.g. financial equations is challenging. Technological advancements, especially in big data and Artificial Intelligence (AI), have enhanced forecasting accuracy [4-6]. Accurate housing price predictions are crucial for investment decisions, meeting residential needs, urban planning, and policymaking.

In the last decades, many related studies have been carried out in terms of price prediction. For instance, Lu et al. utilized Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and other algorithms to predict stock prices with minimal error, and these models have also shown promising results in other areas such as gold price forecasting [7]. Among the various tasks, housing price prediction holds significant importance. Naga Satish et al. employed Linear regression, Gradient boosting regression, and Lasso regression for housing price forecasting. The Gradient boosting regression and Lasso regression models performed exceptionally well, achieving accuracies of 76% and 91%, respectively. Thamarai et al. also leveraged AI to predict housing prices using the decision tree method, achieving a final mean absolute error (MAE) of 0.5, mean squared error (MSE) of 0.5, and root mean square error (RMSE) of 0.7. They also reported a Mean Absolute Error of 2.1, Mean Squared Error of 6.6, and Root Mean Squared Error of 2.5, indicating very small errors and a high level of predictive accuracy. As their second predictive model, Multiple Linear Regression yielded performs better with a mean absolute error of 1.9, mean squared error of 6.1, and root mean squared error of 2.4, suggesting that the performance of multiple linear regressions is superior to decision tree regression for predicting house prices.

Nor Zulkifley et al. considered Locational, Structural, and Neighborhood factors to be the most significant in determining house prices in their study [8]. They utilized Multiple Regression Analysis, Support Vector Regression, Artificial Neural Networks, and XGBoost, achieving RMSE values of 0.1261, 0.2362, 0.5155, and 0.1212, respectively. These results demonstrate that the models, developed based on various input attributes, have a significantly positive impact on house prices.

To evaluate the performance of different models and find the superior model, this paper aims to utilize data from Kaggle and employs three models: Linear Regression, Lasso Regression, and Ridge Regression for comparison. Additionally, the paper analyzes the weight coefficients of the models for determining the importance of each feature, such as the higher weight of geographical location and the lower weight of the number of toilets.

2. Method

2.1. Dataset Preparation

The dataset for this study was sourced from Kaggle, comprising 4,601 records and 18 attributes [9]. The findings indicate that including too many features in a model can be detrimental, as it may lead to overfitting and poorer performance on new datasets. Therefore, it may be more beneficial to exclude features that do not significantly enhance the model. Additionally, high correlation between features can also be problematic. When two features are highly correlated, retaining both can contribute to overfitting. For example, if overfitting occurs, features like sqt_above and sqt_living—due to their high correlation—might be removed to improve model performance.

2.2. Machine Learning Models-based Prediction

This study used linear regression, lasso regression and ridge regression models， and uses sklearn in python code. These models use RMSE and Loss (actual magnitude- predicted magnitude) to evaluate the efficiency and accuracy of each of these three models.

2.2.1. Linear Regression

Linear regression represents a fundamental statistical tool employed to establish the correlation between a response variable and one or more explanatory variables [10]. The objective is to pinpoint the most precise linear formula that encapsulates this correlation. Critical performance indicators in linear regression encompass the R-squared value, also known as the Coefficient of Determination, which measures the proportion of the response variable's variability that is attributable to the explanatory variables, with scores that scale from 0 to 1, where higher scores indicate a more robust fit. Additionally, the F-statistic is a pivotal metric that gauges the model's overall significance by contrasting it against a baseline model devoid of predictors, aside from the constant term, to ascertain whether the predictors' inclusion substantially enhances the model's predictive power.

2.2.2. Ridge Regression

Ridge regression serves as a method for determining the coefficients in multiple regression models, particularly when there is a high degree of correlation among the independent variables. This technique, alternatively referred to as Tikhonov regularization in honor of its developer Andrey Tikhonov, is utilized to tackle multicollinearity within linear regression models, a common challenge in scenarios with numerous parameters. It is extensively applied across various disciplines, including econometrics, chemistry, and engineering. By incorporating a regularization term, Ridge regression enhances the stability of parameter estimates. This is achieved by slightly increasing the bias while significantly reducing the variance. It was developed to counteract the imprecision of least squares estimators when dealing with multicollinear variables, providing a ridge estimator that often yields more reliable parameter estimates with lower variance and mean squared error compared to traditional least squares methods.

2.2.3. Lasso Regression

Lasso Regression is an approach that applies regularization to models to mitigate overfitting, akin to Ridge Regression but with a unique penalty mechanism. It emphasizes the selection of significant variables and their coefficients by striving to minimize prediction errors while enforcing a constraint that pushes the regression coefficients toward zero. This is achieved by constraining the sum of the absolute values of the coefficients to a predetermined threshold (λ). Consequently, the complexity of

the model is managed, and variables with coefficients that are reduced to zero are effectively disregarded in the model. The most effective value for λ is commonly identified through k-fold cross-validation, which involves partitioning the dataset into k equally sized portions. This method aids in curbing overfitting and ensures that every segment of the dataset is utilized in the validation process, eliminating the necessity for a separate subset dedicated to internal validation.

Table 1. The results of methods

Model	Details	Root Mean Squared Error	R-squared	Adjusted R-squared
Lasso regression	Alpha = 1	190124.751	0.694	0.7
Ridge regression	Alpha = 1	190947.879	0.694	0.7
Linear regression	-	234987.298	0.389	-

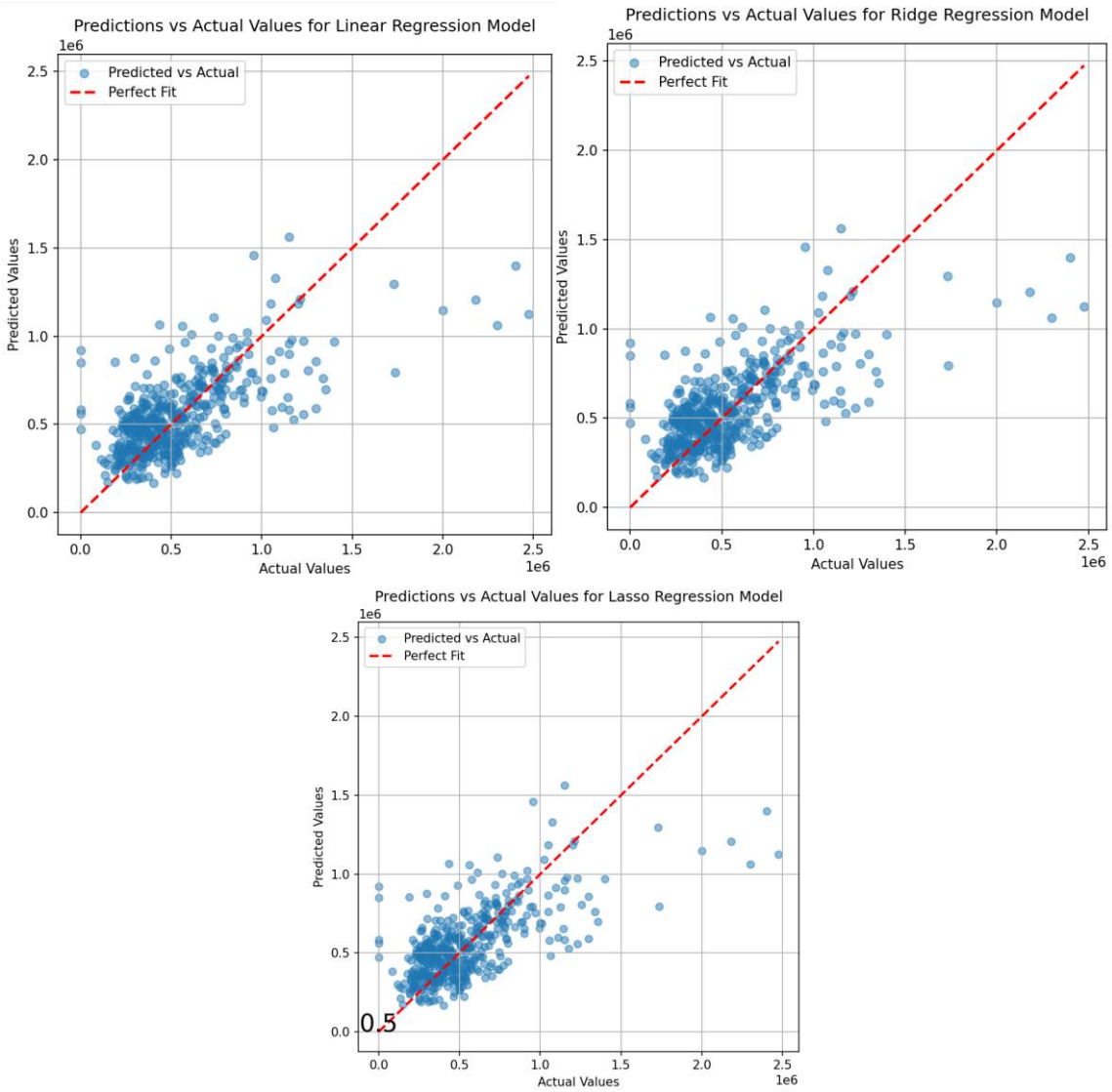


Figure 1. The prediction performance based on different models (Photo/Picture credit: Original).

3. Results and Discussion

3.1. Model Performance Comparison

The visual representations, including the graph and the chart, demonstrate that both the Ridge and Lasso regression models outperform the Linear regression model as depicted in Table 1 and Figure 1. Specifically, the Lasso regression model exhibits a marginally superior performance over the Ridge regression model, with an RMSE that is 23 units lower, and more accurate than linear regression with

linear regression's RMSE 62385 higher than Ridge regression, which means that Lasso regression performs the best and Linear regression is less accurate.

Linear regression serves as a prevalent and potent technique for forecasting outcomes and deciphering the connections among different variables. It accomplishes this by adjusting a linear equation to align with the empirical data, which facilitates the prediction of future trends and the evaluation of the relevance and influence of each variable. Gaining a deep comprehension of its foundational parameters and the presuppositions it relies on is essential for constructing dependable and precise linear regression models.

Ridge, Lasso regression are variations of linear regression that are specifically crafted to handle challenges such as multicollinearity and overfitting. Ridge regression excels at mitigating multicollinearity, a scenario where predictors exhibit strong intercorrelations. Conversely, Lasso regression stands out for its feature selection capabilities, as it can reduce certain coefficients to zero, thereby effectively eliminating those features from the predictive model. Both methods incorporate regularization to introduce a controlled bias, which serves to lower variance and enhance the model's predictive performance on unseen data.

In the conducted research, Lasso regression demonstrated the highest level of accuracy, registering a root-mean-square error (RMSE) of 0.1121. This outcome aligns with expectations given the high level of multicollinearity among the variables, which makes Lasso's feature selection capability particularly advantageous in this context.

3.2. Feature Importance Analysis

In this study, parameters within the dataset were examined using various charts, including the correlation matrix depicted in Figure 2. This matrix reveals the relationships among selected features. For instance, a correlation coefficient of 0.20 between house prices and the quantity of bedrooms indicates a modest positive link. This implies that although homes with more bedrooms tend to have higher prices, the impact is not significant. The feature with the second highest correlation to Sales Price is the Above Grade Living Area, which is consistent with the notion that larger homes typically have higher prices. Furthermore, factors such as the quality of the view, the condition of the property, and the year it was built also play significant roles in determining house prices. Buyers often seek homes with attractive views for relaxation and generally prefer newer properties for their cleaner appearance and reduced safety risks.

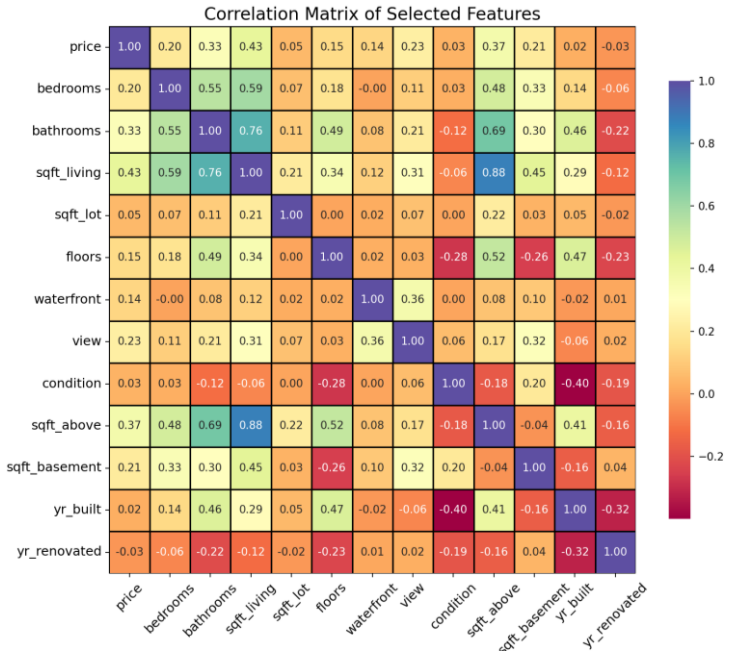


Figure 2. The correlation between variables (Photo/Picture credit: Original).

4. Conclusions

This initiative utilizes core machine learning techniques, such as linear regression, Ridge regression, and Lasso regression, within its framework. with implementation carried out using the Scikit-Learn tool. The primary aim of this work is to assist users in predicting both the availability and prices of houses in the city. The project has demonstrated that these regression methods are effective for house price prediction, with Lasso regression showing superior performance compared to the other two methods. Despite this, the accuracy of the predictions remains limited, as indicated by a high root-mean-square error (RMSE). In order to achieve better predictive outcomes, it's advisable to augment the dataset with a richer set of attributes and delve into more complex machine learning algorithms to refine the accuracy of the house price forecasting model. This approach could lead to more nuanced insights and improved predictive capabilities.

References

[1] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. March 2019. House price prediction using regression techniques: A comparative study. In 2019 International conference on smart structures and systems (ICSSS) (pp. 1-5). IEEE.

[2] Mohd, T., Jamil, N. S., Johari, N., Abdullah, L., & Masrom, S. 2020. An overview of real estate modelling techniques for house price prediction. In Charting a Sustainable Future of ASEAN in Business and Social Sciences: Proceedings of the 3rd International Conference on the Future of ASEAN (ICoFA) 2019—Volume 1 (pp. 321-338). Springer Singapore.

[3] Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. 2019. House price prediction using machine learning. Journal of Innovative Technology and Exploring Engineering, 8(9), 717-722.

[4] Hunt, E. B. 2014. Artificial intelligence. Academic Press.

[5] Holmes, J., Sacchi, L., & Bellazzi, R. 2004. Artificial intelligence in medicine. Ann R Coll Surg Engl, 86, 334-8.

[6] Fetzer, J. H., & Fetzer, J. H. 1990. What is artificial intelligence? Springer Netherlands.

[7] Lu, W., Li, J., Wang, J., & Qin, L. 2021. A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 33(10), 4741-4753.

[8] Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. 2020. House price prediction using a machine learning model: a survey of literature. International Journal of Modern Education and Computer Science, 12(6), 46-54.

[9] Kaggle. 2018. House price prediction. Retrieved from https://www.kaggle.com/datasets/shree1992/housedata.

[10] Su, X., Yan, X., & Tsai, C. L. 2012. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.