

Multiple Machine Learning Models in House Price Prediction: Performance Evaluation and Comparison

Jiahao Wu

Department of Electronics and Information Engineering, Hebei Normal University of Science & Technology, Qinhuangdao, China

laifu547@tzc.edu.cn

Abstract. House pricing is one of the most critical concerns for people. Accurately predicting house prices is essential for participants and investors in the real estate market. This study delves into the realm of predictive modeling for house prices using a multimodal approach in machine learning. The research methodology spans various stages including data collection, preprocessing, feature engineering, and model selection. Based on a dataset sourced from Kaggle, housing data from Iowa comprising 79 pertinent features influencing property prices is utilized. Performance evaluation is conducted using the root mean square error (RMSE), comparing the efficacy of linear regression models such as Lasso regression, Kernel Ridge regression, and ElasticNet regression against tree-based models including XGBoost, Gradient Boosting, and LightGBM. Additionally, the study implements two stacked ensemble models to identify the most optimal predictive model for house prices. Experimental procedures involve training and evaluating these models on specific datasets to gauge their predictive accuracy. Findings and discussions underscore the comparative analysis of diverse machine learning models within the context of house price prediction. Moreover, the study employs the Gradient Boosting method in machine learning to pinpoint the ten most influential features impacting house prices. Recommendations for future research are provided to enhance the precision and robustness of house price prediction models. This research contributes to the advancement of predictive modeling in the real estate domain, offering insights into the efficacy of various machine learning techniques for forecasting property prices.

Keywords: House price, Machine Learning, Stacking Model, Feature Importance.

1. Introduction

As economic growth accelerates and urbanization intensifies, the real estate market experiences a continuous surge, with housing price predictions emerging as a matter of concern. The price of a house is the paramount concern for most individuals and holds significant weight in their decision-making process. Accurately forecasting housing prices is of paramount importance for participants and investors in the real estate market when it comes to house price prediction. With the rapid advancements in machine learning and data science witnessed over the past decades, machine learning-based approaches have emerged as a prominent research area for house price forecasting. By harnessing extensive real estate data alongside sophisticated machine learning algorithms, precise models can be constructed to predict housing prices [1].

Machine learning techniques develop rapidly in recent years due to its ability to automatically learn and improve from data without being explicitly programmed, making it widely applied in fields like image recognition, language processing, and recommendation systems [2-4]. Top of form in machine learning-based house price predicting, previous studies have tried a variety of algorithms, including linear regression which is the most commonly used statistical model for data analysis [5], decision trees, random forests, neural networks etc. Among them, models about regression algorithms are more commonly used and convenient to make predictions. For example, the regression algorithm model used in this article, the Lasso regression, it was first introduced in [6], it provides a theoretical basis for house price forecasting.

Moreover, the use of ensemble methods, such as stacking multiple models, has been documented as an effective strategy to boost predictive power and address the limitations of individual algorithms.

Some studies have utilized ensemble learning methods, such as gradient lift trees and XGBoost to further improve prediction performance [7].

In order to evaluate the performance of housing price forecasting models, current studies usually use indicators such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The model evaluation criterion chosen in this paper is RMSE. In Chen's work [8], the performance of RMSE index in evaluating the prediction accuracy of regression models is described. Not only the definition and calculation methods of RMSE are elaborated, but also the advantages and disadvantages of RMSE which is a widely used evaluation index of regression models are discussed in detail.

Furthermore, the identification of key features influencing house prices has been a focal point in many studies, with researchers employing techniques like Gradient Boosting to extract important variables. Inspired by Cieslak's work [9], which is a study on the influence of feature importance on housing prices, this article adds feature importance analysis, using Gradient Boosting.

This article compares a large number of basic machine learning models in the sklearn package to find the most suitable predictive model. Additionally, through the ensemble learning technique of stacking [10], different types of basic learners complement each other, so as to enhance the overall model's generalization ability and performance. Subsequently, based on weight redistribution, the model with the lowest final RMSE value was chosen as the ultimate model.

In contrast to previous simple house price predictions, this article also conducts feature importance analysis. It analyzes 79 different features in the data, selects the features that have the greatest impact on house prices, and visually presents those important features. Through the research on the housing price prediction model in machine learning, this paper can enable housing related personnel to better predict the housing price, and through the visual display of the most important features affecting the housing price, people who have a demand for housing purchase can have a better understanding of the housing price.

2. Method

2.1. Dataset Description

This article uses a dataset sourced from Kaggle [11], which includes 79 features related to factors influencing the sale prices of houses in Iowa. These features include the total area of the house, the street on which the house is located, and others. The dataset consists of a training set and a test set, with the training set comprising 1,460 rows and 81 columns, including the target variable SalePrice.

2.2. Dataset Preprocessing

a). Outlier Handling

Outliers may have a negative effect on the performance of the model and make the prediction results inaccurate [12]. By plotting scatter plots, it is possible to visually identify outliers in the features of the training set. The removal of outliers can improve the predictive performance of the model, but this study has retained some outliers to enhance robustness.

b). Target Variable Analysis

SalePrice is the target variable that this study aims to predict. Firstly, a distribution plot and a Quantile Quantile Plot (QQ plot) of SalePrice are generated. If the points on the QQ plot lie close to a straight line, it indicates that the data is approximately normally distributed. From Fig. 1, it can be observed that the distribution of this data is right-skewed. Therefore, this study applies a logarithmic transformation to SalePrice using the log1p function from the numpy package to make the data closer to a normal distribution. As shown in Fig. 2, the fitted curve aligns closely with the reference normal distribution curve, and most of the points on the QQ plot are clustered around the straight line.

c). Missing Value Handling

Initially, this article examines the presence of missing values in the given dataset features and identifies a total of 34 features with missing values.

The methods applied in this article for filling missing values are all common practices, such as using the median, mode, inputting 'None', and deletion.

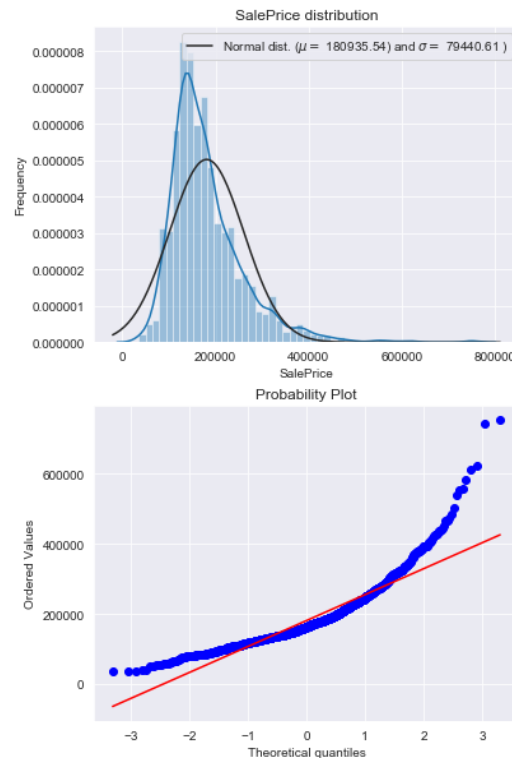


Figure 1. Non-normal distribution (Photo/Picture credit: Original)

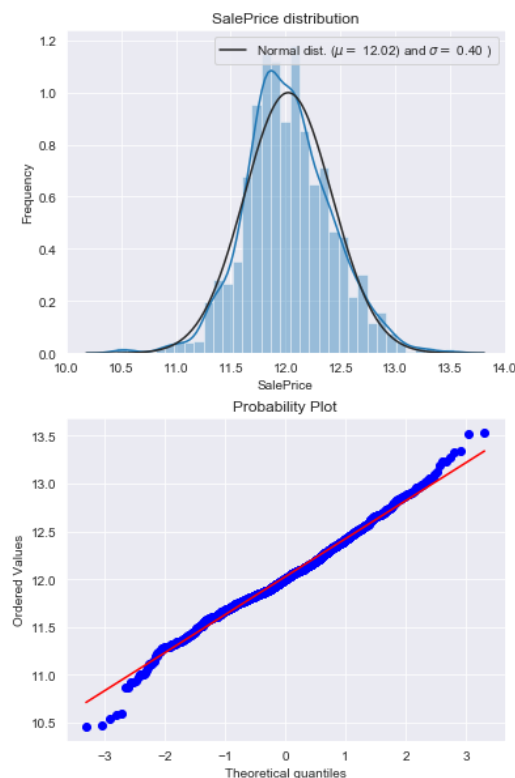


Figure 2. Normal distribution (Photo/Picture credit: Original)

2.3. Feature Engineering

a). The methods applied in this article for feature processing are also commonly known, involving the transformation between numerical and categorical features based on different situations.

b). Feature Relationships and Creating New Features: By plotting the feature correlation matrix heatmap, as shown in Fig. 3, this study selected important features such as Total square feet of basement area (TotalBsmtSF), 1stFirSF, and 2ndFirSF, which have a significant correlation with SalePrice to construct a new feature TotalSF, which representing the total area of the house.

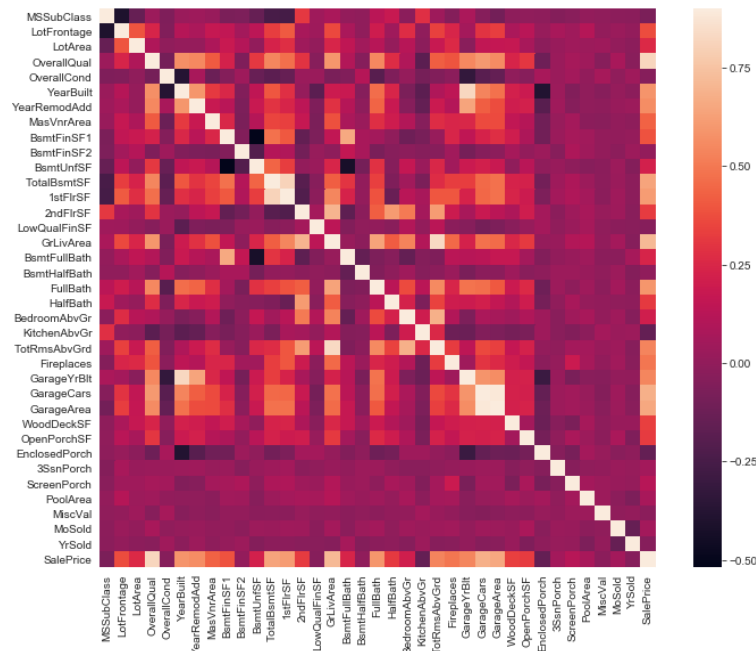


Figure 3. Feature Correlation Matrix Heatmap (Photo/Picture credit: Original)

c). To make numerical features more closely follow a normal distribution, this study performed Box-Cox transformations on features with high skewness (skewness greater than 0.75), smoothing the skewness to 0.15. For categorical features, one-hot encoding was applied.

2.4. Machine Learning Models

This study utilizes four fundamental regression models in machine learning, namely Lasso Regression, Kernel Ridge Regression and ElasticNet Regression, along with three decision tree-based ensemble models, GBoost, XGBoost and LightGBM.

The study also constructs a stacking model based on the results of the base models.

2.4.1 Basic Regression Models

a). Lasso Regression: Lasso regression achieves feature selection and model simplification by incorporating L1 regularization, making it particularly useful for high-dimensional data and cases where a sparse model is desired [13].

b). Kernel Ridge Regression: Kernel Ridge Regression leverages the kernel trick to handle nonlinear relationships in the data by mapping the features into a higher-dimensional space, allowing for more flexible modeling while controlling overfitting through regularization [14].

c). Elastic Net Regression: Elastic Net Regression combines the advantages of Lasso and Ridge regression by incorporating both L1 and L2 regularization, making it effective in addressing issues like feature selection, multicollinearity, and overfitting. This method is widely used in practice due to its ability to automatically select features, handle multicollinearity, and mitigate the risk of overfitting.

2.4.2 Ensemble Models Based on Decision Trees

a). GBoost: It is an ensemble learning method that iteratively trains multiple weak learners (typically decision trees), with each iteration aimed at reducing the residual errors of the previous model. By iteratively training models and integrating their prediction results, Gradient Boosting can

establish a robust overall model that is suitable for various machine learning tasks and has been widely applied in practice.

b). XGBoost: XGBoost is an ensemble learning algorithm that trains multiple decision tree models to make predictions. In each iteration, XGBoost optimizes the model using gradient boosting, minimizing the loss function to improve model performance. XGBoost introduces innovative techniques such as sparsity-aware split finding, weighted quantile sketch, cache awareness, and approximate tree learning to enhance training speed and model performance.

c). LightGBM: LightGBM is based on the gradient boosting framework and uses a novel technique called Gradient-based One-Side Sampling (GOSS) to filter out the data instances during the training process, reducing the computational resources required for training while maintaining accuracy. LightGBM introduces techniques like exclusive feature bundling, histogram-based decision tree learning, and leaf-wise tree growth to optimize the training process and improve performance. It also supports GPU acceleration for faster computations [15].

2.4.3 Stacking Model

Stacking is a meta-learner that combines the predictions of these basic learners to improve overall prediction performance [16]. The study utilizes five-fold cross-validation for model evaluation and constructs a two-layer stacking model. Based on the training results of the base models, the KRR model, ENet model, and GBoost model are used as the first layer, while the Lasso model is utilized as the second layer. Fig. 4 illustrates the general principle of a two-layer stacking model with five-fold cross-validation.

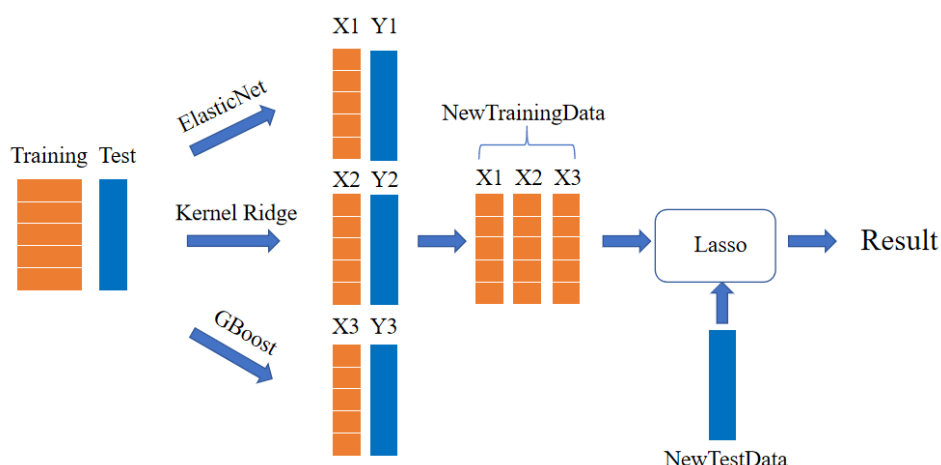


Figure 4. Two-layer stacking model (Photo/Picture credit: Original)

2.4.4 Analysis of Feature Importance

Gradient Boosting helps improve the generalization ability of the model and reduce the risk of overfitting [17]. Gradient Boosting is a popular machine learning technique used for regression and classification tasks. When it comes to predicting house prices, Gradient Boosting models can be particularly effective due to their ability to capture complex relationships in the data.

Feature importance in Gradient Boosting models is determined by how much each feature contributes to reducing the loss function during the training process, which is applied in this study. The higher the reduction in loss attributed to a feature, the more important that feature is considered.

3. Result and Discussion

3.1. Performance of Different Models

The presentation of the results primarily focuses on the RMSE values of the six base models and the stacking model. Choose the model with the smallest RMSE as the final prediction model for

housing price prediction. A bar chart is created to visualize the top ten features influencing house prices, based on the results obtained through Gradient Boosting analysis.

Table 1. The performance of the model used in this article

Model Type	RMSE
Lasso	0.1173
ElasticNet	0.1281
Kernel Ridge	0.1227
Gradient Boosting	0.1210
XGBoost	0.1219
LightGBM	0.1220
Stacking	0.1148

The results presented in Table 1 serve as the basis for building the two-layer stacking model in this study. According to the results, the Lasso model performed the best among these six base models, with the lowest RMSE value. Therefore, when stacking models, using the Lasso regression model as the second layer is the most reasonable choice, as it can help further filter features, enhance the overall model stability, and improve generalization ability.

The RMSE value of the stacking model constructed based on the base models is 0.1148. It can be observed that this model's RMSE value is lower than those of the individual base models, indicating that the newly created stacking model provides better predictive performance for house prices. Therefore, this stacking model will be selected as the final model for predicting house prices in the test dataset in this study.

3.2. Feature Importance

The purpose is to combine the final predicted house prices with the features in the test set and utilize the most appropriate model to analyze feature importance, aiming to identify the most significant features among the 79 features.

Based on the results shown above, the Gradient Boosting model exhibits relatively stable prediction performance with low RMSE values after training. Therefore, this study selects this model for feature importance analysis. A bar chart (Fig. 5) is plotted in this article for visual representation. To provide a more intuitive understanding, the top ten important features are listed from high to low.

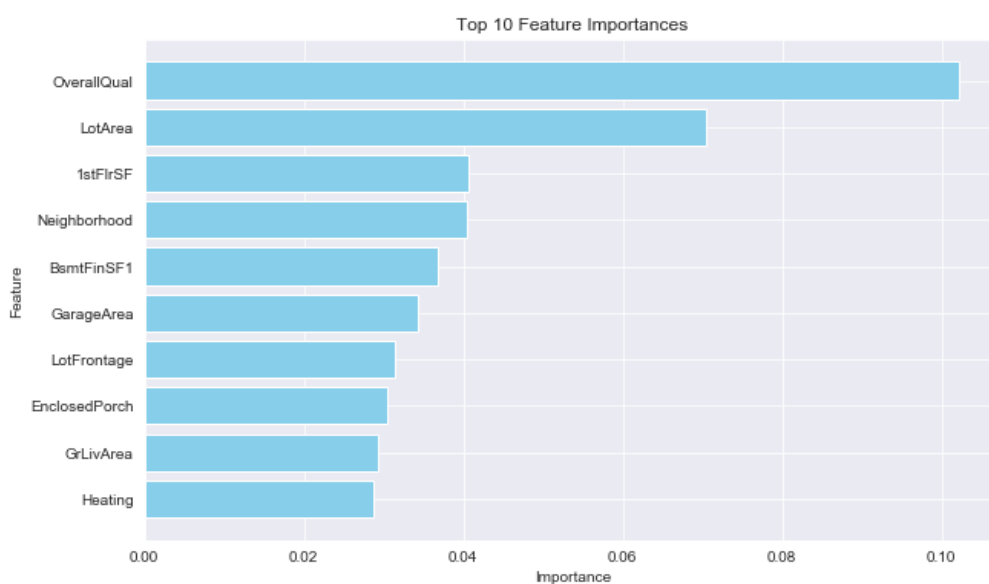


Figure 5. Top ten important features bar chart (Photo/Picture credit: Original)

From the bar chart, it can be observed that the feature OverallQual is the most prominent among the 79 features, being 1.5 times that of LotArea and more than twice the value of the other features.

High-quality materials and decorations typically indicate a higher overall quality and more attractive appearance of a house. This can enhance the value and appeal of the property, thereby influencing its price. For homebuyers, the quality of materials and decorations is one of the important factors to consider when choosing a house. The overall quality of materials and decorations can directly impact the value and appeal of the property, thus affecting the housing market in Iowa.

The area, location, and street features of a house are well-known factors that influence its selling price, a phenomenon not unique to Iowa. A larger living area and a more comfortable and convenient neighborhood are certain to increase the selling price of a house. However, an unexpected feature that plays a significant role in determining the price of a house in Iowa is the Heating attribute, which ranks tenth in importance.

According to the report from the National Weather Service [18], in the state of Iowa, winter temperatures often drop below freezing, and extreme cold weather conditions may occur, making heating not only a matter of comfort but also a necessity of life. Insufficient heating can lead to issues such as frozen pipes and damage to the structure of homes. Therefore, maintaining good heating equipment and systems is crucial for residents of Iowa, ensuring that they stay warm and safe during the cold winter months.

4. Conclusion

In this study, various basic machine learning models were employed for training and testing to improve model accuracy and better fit real data. Building upon these base models, a stacking model was constructed to enhance house price prediction and feature importance analysis. The housing data from Iowa was sourced from the Kaggle website. Extensive preprocessing was conducted on the various features in the data to meet the model's standards. These models were evaluated on the training set using 5-fold cross-validation accuracy. Additionally, the evaluation metric used was RMSE. Experimental results demonstrate that using the stacking model leads to more accurate predictions. Among 79 features, "OverallQual" was identified as the most influential feature affecting house prices.

In the future, more attributes can be added to improve the accuracy of housing price prediction. Furthermore, a potential improvement method for this study could be to reweight and stack different ensemble models to reconstruct the final model. By adjusting the weights of various ensemble models and combining them to form a new ensemble model. For example, readjusting the weights of XGBoost, LightGBM, and the stacking model, it is possible to enhance prediction performance and further optimize results.

References

- [1] Dimitriadi Y, Tziralis G. A comparative study of machine learning methods for housing price prediction. *Expert Systems with Applications*, 2019, 129: 109-117.
- [2] Qiu Y, Wang J, Jin Z, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 2022, 72: 103323.
- [3] Sidorov G, Velasquez F, Stamatatos E, et al. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 2014, 41 (3): 853-860.
- [4] Roy PK, Chowdhary SS, Bhatia R. A Machine Learning approach for automation of Resume Recommendation system. *Procedia Computer Science*, 2020, 167: 2318-2327.
- [5] Gelman A, Hill J, Vehtari A. *Regression and Other Stories*. Cambridge University Press, 2020.
- [6] Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58 (1): 267-288.
- [7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 2016: 785-794.

- [8] Smith J, Johnson A, Brown C. Assessing the Predictive Accuracy of Regression Models: A Comparison of RMSE, MAE, and R-Squared. *Journal of Statistical Methods*, 2018, 45 (3): 267-288.
- [9] Cieslak D, Chmielowiec A, Tabor J. Feature selection for house price prediction: A comparative study. *Expert Systems with Applications*, 2019, 129: 170-182.
- [10] Breiman L. Stacked regressions. *Machine Learning*, 1996, 24 (1): 49-64.
- [11] Kaggle. House prices advanced regression techniques. <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>, 2024.
- [12] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017: 3146-3154.
- [13] Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 2008, 52 (3): 1694-1711.
- [14] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [15] Smola AJ, Schölkopf B. A tutorial on support vector regression. *Statistics and computing*, 2004, 14 (3): 199-222.
- [16] Wolpert DH. Stacked generalization. *Neural Networks*, 1992, 5 (2): 241-259.
- [17] Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 2002, 38 (4): 367-378.
- [18] Weather, National Weather Service, <https://www.weather.gov/wrh/Climate>, 2024.