House Price Prediction Based on Machine Learning --Taking Beijing House Price Data as an Example

Yuhong Xu*

{*Corresponding author: centralparkxia@gmail.com}

Dundee International Institution of Central South University, Changsha Hunan, China

Abstract: House price prediction has significant practical value in the real estate market, as it provides critical decision support for specific needs of buyers, sellers, and investors. However, accurate forecasting remains challenging due to the complex and multi-factorial nature of housing prices. This study investigates the application of machine learning algorithms for house price forecasting to enhance predictive accuracy. To achieve this aim, we examine several classical machine learning models including decision trees, linear regression, k-nearest neighbors, random forests, AdaBoost, Bagging, gradient boosting, and Extra Tree. Using housing price data from Beijing containing attributes such as transaction time, age of the house, and the average community price, we employ feature engineering to process the data. Specifically, we extract quarterly information from transaction times and calculate house age based on the difference between transaction time and construction time. We also generate difference features to increase model expressivity. Experimental results demonstrate that feature engineering significantly improves model performance. Furthermore, comparing RMSE values before and after feature engineering confirms its efficacy in enhancing baseline algorithms, which further confirms the effectiveness of feature engineering. In conclusion, this study highlights the potential of machine learning and feature engineering to advance house price prediction, offering practical value for real estate market participants while provides insights for further research on machine learning-based housing price forecasting.

Keywords: Machine Learning, Feature engineering, House Prices, Prediction.

1. Introduction

Accurate prediction of house prices has become an issue of profound socioeconomic interest in the real estate market. This study aims to investigate and compare various methods for house price prediction through the application of machine learning algorithms, as well as examine the influence of feature engineering on prediction performance.

This study utilizes housing price data from Beijing. During data preprocessing, we cleaned the data and processed date-time information. For feature engineering, we introduced quarterly information, house age, and neighborhood house price differences to better capture factors influencing house prices. We then constructed house price prediction models using machine learning algorithms including decision trees, linear regression, and random forests. By evaluating different algorithms, we aim to determine the optimal method for prediction. Moreover, we emphasize the vital role of feature engineering in enhancing model performance.

ICEMME 2023, November 17-19, Beijing, People's Republic of China

DOI 10.4108/eai.17-11-2023.2342672

The results can furnish investors and policy makers with more accurate house price forecasts while providing valuable references for future research.

Next, we will detail the steps of data preprocessing and feature engineering, as well as the prediction performance of different algorithms. Through these experiments, we comprehensively assess house price prediction models and deliver critical insights for research and practical applications in this domain.

2. Related works

Many researchers have proposed various methods for house price prediction. In [1], authors present a random forest regression model for house price prediction using the Boston dataset and R. They optimize parameters through grid search and cross-validation, achieving a test RMSE of 0.0478, and compare it to linear regression, SVM, and neural networks. In [2], authors introduce a machine learning model for COVID-19-era house price prediction in Spanish cities. They employ ensemble algorithms like boosting and bagging, comparing them to linear regression models. Their systematic approach ensures an efficient and interpretable model. In [3], the authors introduce the incorporation of time and space features, including seasonality and geographical information, to enhance real estate price predictions. Their model showcases improved performance in accounting for seasonal variations in the real estate market. In [4], authors suggest a stacking ensemble method for house price prediction, integrating various base models. While enhancing prediction accuracy and robustness, it demands significant data and computation. The approach stands out for its innovative integration of multiple models, fostering cross-modal synergy and collaboration. [5] proposed a method using Bayesian, back propagation neural network and support vector machine and other technologies and uses mean square error and mean absolute error as evaluation indicators. This method can deal with nonlinear and high-dimensional data and improve the accuracy and stability of prediction. [6] employs machine learning and neural network techniques to predict house prices and assesses model performance. Methods include linear regression (LR), support vector machine (SVM), backpropagation neural network (BP neural network), random forest (RF), and deep neural network (DNN). Evaluation metrics include mean squared error (MSE) and mean absolute error (MAE). Results indicate DNN performs best, followed by BP neural network and RF, then LR and SVM.

3. Methods and experimental design

This study leveraged Beijing house price data from Kaggle, most data is traded in 2011-2017, some of them is traded in Jan,2018, and some is even earlier (2010,2009), comprising attributes including geographic location, house size, and average house price in the neighborhood totally 22 columns and about 300,000 rows in all, for an example see Figure 1. Data cleaning and preprocessing removed missing values and outliers. We then conducted feature engineering, extracting date information, and generating different features. For model selection, eight algorithms were evaluated: Decision Tree[7], Linear Regression[8], K Nearest Neighbor[9], Random Forest[10], AdaBoost[11], Bagging[12], Gradient Boosting[13], and Extra Trees[14].

Cross-validation calculated the RMSE value for each algorithm as a performance metric. The processing of each feature is detailed separately below.

	Lng	Lat	tradeTime	...	subway	district	communityAverage
0	116.475489	40.019520	2016/8/9	...	1.0	7.0	56021.0
1	116.453917	39.881534	2016/7/28	...	0.0	7.0	71539.0
2	116.561978	39.877145	2016/12/11	...	0.0	7.0	48160.0
3	116.438010	40.076114	2016/9/30	...	0.0	6.0	51238.0
4	116.428392	39.886229	2016/8/28	...	1.0	1.0	62588.0

Fig. 1. The samples of Beijing House trading data

3.1. Feature 1: Quarter

3.1.1 Description

The "quarter" feature represents the quarter of the year in which a real estate transaction took place. This valuable temporal attribute can capture the seasonal patterns and trends in the housing market. For this feature engineering, we extracted and leveraged the quarter information from the original "tradeTime" column, initially formatted as datetime.

3.1.2 Feature Engineering

To extract the "quarter" feature, we first converted the "tradeTime" column to datetime format. With this conversion completed, we leveraged datetime functionality to extract the quarter information for each data point. Specifically, we utilized the ‘dt.quarter’ method to obtain the corresponding quarter for each transaction entry. This newly created "quarter" feature is then incorporated as a predictive attributes in our machine learning models.

3.1.3 Feature processing effect

Incorporating the "quarter" feature has proven vital for our housing price prediction task. By considering the transaction quarter, our models can better capture seasonal house prices fluctuations. This feature engineering enhanced model performance, enabling more accurate price predictions based on temporal patterns. Notably, the Random Forest model demonstrated significant improvements with the addition of this feature. The comparison of these data is all reflected in Figure 2.

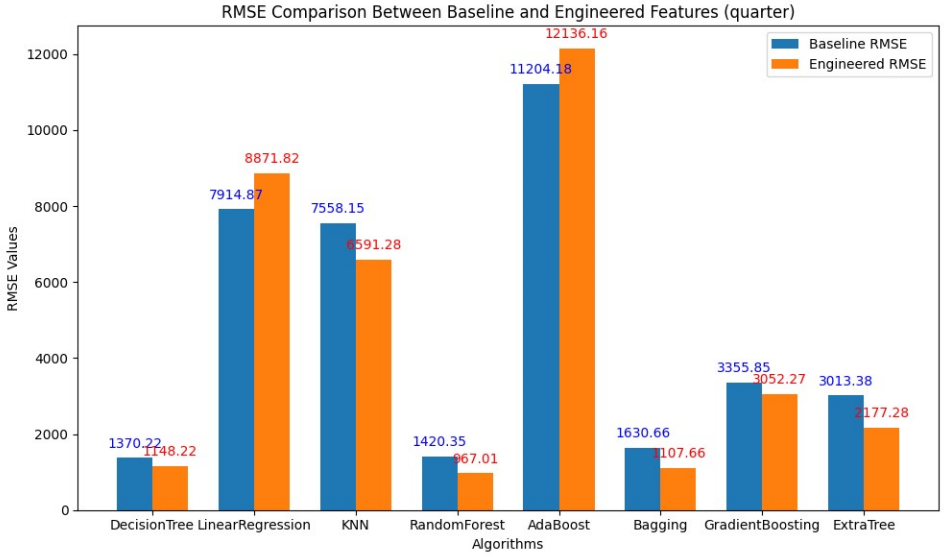

Fig.2. The RMSE comparison between baseline and feature “quarter”

3.2. Feature 2: House Age

3.2.1 Description:

The "houseAge" feature denotes the age of a property in years, calculated as the difference between the transaction year (refers to the calendar year in which the property sale transaction occurred) and the construction year (refers to the calendar year in which the property was built and completed). This feature provides valuable insights into the property's age, which can substantially influence its market value.

3.2.2 Feature Engineering:

For the "houseAge" feature engineering, we first extracted year information from the "tradeTime" (transaction year) and "constructionTime" (construction year) columns. The "tradeTime" column was first converted to a numeric format to facilitate the calculation. We then derived each property’s age by subtracting the construction year from the transaction year. This generated a new feature, "houseAge", which represents each property’s age in years.

3.2.3 Feature processing effect:

The inclusion of the "houseAge" feature has proven to be essential in our housing price prediction task. As property age substantially impacts prices, our models benefit from this attribute's inclusion. By integrating "houseAge", our machine learning models can capture the age influence on prices, enabling more precise predictions. This feature engineering process markedly enhanced models predictive performance, improving property value estimation based on age. In particular, the Random Forest model exhibited notable improvements with the addition of this engineered feature. The comparison of these data is all reflected in Figure 3..

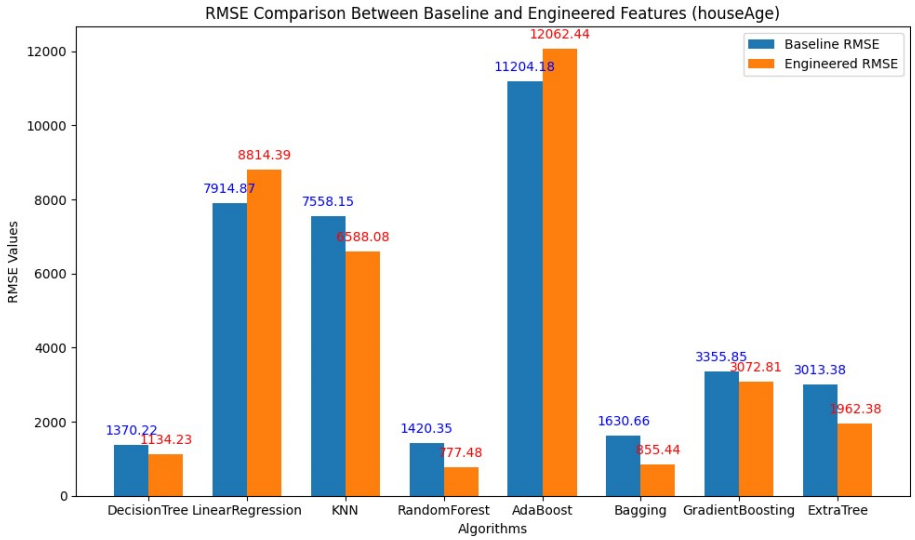

Fig.3. The RMSE comparison between baseline and feature “houseAge”

3.3 Feature 3: Community Average Price Difference

3.3.1 Description:

The "communityPriceDiff" feature denotes the difference between the community's average property price and the actual price of a specific property. This feature captures how an individual property’s price compares to the community's average price, providing valuable insights into its relative market position.

3.3.2 Feature engineering:

In the feature engineering process for "communityPriceDiff," we calculated the price difference between the "communityAverage" and each property’s actual price. The "communityAverage" price reflects the typical property price within a given community. By computing the difference between the community's average price and each property's actual price, we derived the "communityPriceDiff" feature, quantifying each property's price deviation from the community norm.

3.3.3 Feature processing effect:

The inclusion of the "communityPriceDiff" feature has proven to be highly effective in our housing price prediction task. This feature enables models to consider both inherent property characteristics and relative community price. Capturing significant pricing deviations from community averages is crucial, as such variances can indicate unique attributes influencing market values. Integrating "communityPriceDiff" allowed our machine learning models to better capture such variations, improving predictive accuracy. Except for KNN, all algorithms exhibited varying degrees of enhancement with this feature. The comparison of these data is shown in Figure 4

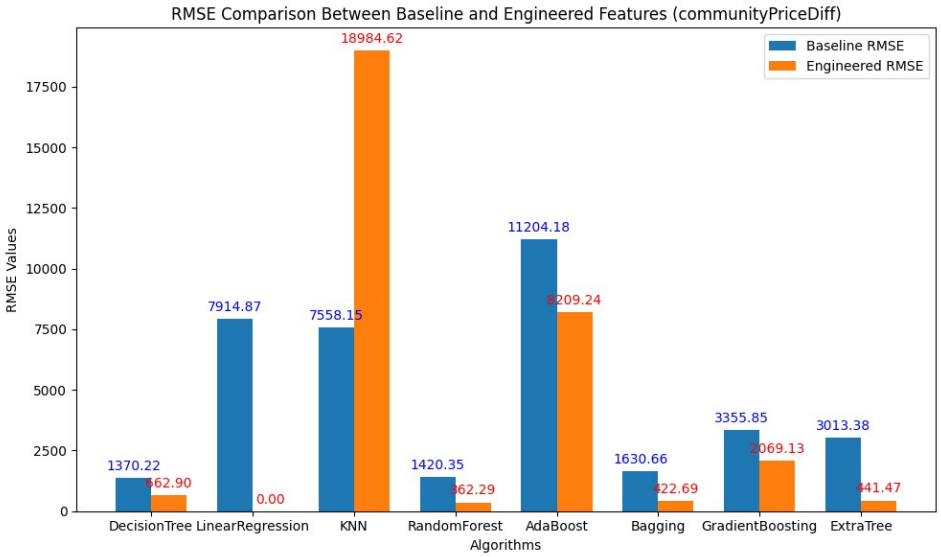

Fig. 4. The RMSE comparison between baseline and feature "communityPriceDiff"

4. Overall result analysis

This study undertakes property price prediction utilizing machine learning algorithms, with an emphasis on the Beijing housing market. Our analysis and feature engineering efforts aimed to improve model predictive performance. This section summarizes the results obtained from different algorithmic approaches and feature engineering techniques, emphasizing salient findings and their implications.

4.1. Baseline Model

Our baseline DecisionTree model achieved a RMSE of approximately 1370.22, establishing a starting point to evaluate feature engineering and algorithm enhancements.

4.2. Feature Engineering: Quarter

The introduction of the "quarter" feature denoting the transaction quarter significantly improved accuracy. The Random Forest algorithm emerged as the top performer, reducing RMSE to approximately 967.01. This highlights the importance of transaction timing, as seasonal trends substantially impact housing prices.

4.3. Feature Engineering: House Age

The "houseAge" feature, derived from the difference between the year of the transaction and the year of construction, further improved predictive accuracy. Once again, the Random Forest algorithm outperformed others, achieving an RMSE of approximately 777.48. This result suggests that the age of a property is a crucial factor in determining its market value, with newer properties often commanding higher prices.

4.4. Feature Engineering: Community Average Price Difference

Our final feature, "communityPriceDiff," quantified the difference between a property's price and its community average price. Interestingly, Linear Regression proved most effective here, remarkably reducing RMSE to 0.00177. This result underscores the significance of understanding how a property's price compares to the community average, as even minor deviations can indicate unique attributes or local market dynamics.

4.5. Overall Findings

Our study reveals several key insights. First, transaction timing and property age significantly influence market prices. Second, a property’s community relative pricing is essential for accurate predictions. Lastly, algorithm selection also plays a critical role, with different algorithms excelling under varying feature sets.

In conclusion, our research demonstrates that feature engineering and appropriate algorithm selection can substantially enhance the accuracy of housing price predictions. Incorporating seasonality, property age, and community price differentials provides a holistic perspective on influential factors These findings offer practical applications in real estate valuation and investment decision-making, ultimately benefiting both buyers and sellers in the Beijing housing market.

5. Conclusion

In this study, we implemented machine learning techniques to forecast property values in Beijing's highly volatile real estate market. Through comprehensive exploration of feature engineering approaches and evaluation of diverse algorithms, our research sought to optimize predictive accuracy. The results demonstrate that integrating temporal factors, property age, and community price dynamics significantly improves price prediction. In particular, incorporating quarterly transaction and age attributes enabled random forest models to substantially reduce RMSE. Meanwhile, quantifying community average price deviations allowed linear regression to achieve remarkably low error rates.

These findings offer valuable practical insights for real estate appraisal, investment analysis, and market research in Beijing. By accounting for timing, age, and relative community pricing, stakeholders can make better-informed decisions despite constant housing market fluctuations.

In summary, strategic feature engineering and algorithm selection considerably enhance machine learning-based property price forecasting. Our study takes a holistic perspective of influential drivers like seasonality and localized market dynamics. The dramatically improved predictive capabilities provide data-driven decision support amidst Beijing's volatile real estate landscape. Further research can build upon these techniques to create widely generalizable and comprehensive pricing models.

References

[1] Mora-Garcia, R.-T., Cespedes-Lopez, M.-F. and Perez-Sanchez, V.R. (2022) Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land, 11, 21001

[2] Vineeth, N., Ayyappa, M. and Bharathi, B. (2018) House Price Prediction Using Machine Learning Algorithms. In: Satapathy S.C., Bhateja V., Das S., Raju K.S. (eds) Soft Computing Systems. ICSCS 2018. Communications in Computer and Information Science, vol 837. Springer, Singapore

[3] Chen, L., Zheng, B., Cao, X., and Leung, H. (2018) Forecasting Real Estate Prices with Multiple Model Fusion. In Proceedings of the International Joint Conference on Artificial Intelligence

[4] Wu, Y., and Zhang, Y. (2022). Spatial and Temporal House Price Prediction: A Machine Learning Approach. arXiv preprint arXiv:2204.09050.

[5] Chen, Y., Xue, R., and Zhang, Y. (2021) House price prediction based on machine learning and deep learning methods. In 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), Changchun, China.

[6] Varma, A., Sarma, A., Doshi, S., and Nair, R. (2018) House Price Prediction Using Machine Learning and Neutral Networks. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India.

[7] Safavian, S.R., and Landgrebe, D. (1991) A survey of decision tree classifier methodology. In IEEE Transactions on Systems, Man, and Cybernetics, 21. 660-674.

[8] Naseem, I., Togneri, R., and Bennamoun, M. (2010) Linear Regression for Face Recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 32. 2106-2112.

[9] Keller, J., Gray, M., and Givens, J. (1985) A fuzzy K-nearest neighbor algorithm. In IEEE Transactions on Systems, Man, and Cybernetics, SMC-15. 580-285.

[10] Belgiu, M., and Drăguţ, L. (2016) Random forest in remote sensing: A review of applications and future directions. In ISPRS Journal of Photogrammetry and Remote Sensing, 114. 24-31.

[11] Schapire, R.E. (2013) Explaining AdaBoost. In: Schölkopf, B., Luo, Z., Vovk, V. (eds) Empirical Inference. Springer, Berlin, Heidelberg

[12] Breiman, L. (1996) Bagging predictors. Mach Learn. 24. 123-140.

[13] Friedman, J. (2002) Stochastic gradient boosting. In Computational Statistic & Data Analysis, 38. 367-378.

[14] Geurts, P., Ernst, D., and Wehenkel, L. (2006) Extremely randomized trees. In Mach Learn, 63. 3-42.