AI in Real Estate: Forecasting House Prices with Advanced Machine Learning Models

Jiashuo Cui[1,*][a], Zhitong Liu[2][b] and Yinghan Ma[3][c]
[1]*Beijing Royal Foreign Language School, Beijing, China*
[2]*Innovation Tianhui High School, Hebei, China*
[3]*Beijing Haidian Foreign Language Academy, Beijing, China*
*

Keywords: Machine Learning, House Price Prediction, Artificial Intelligence.

Abstract: Homes are essential to our lives, providing a comfortable space for living, working, and learning. Therefore, everyone needs a place to call home. This article describes how to use different machine learning algorithms (one of the core techniques of artificial intelligence, including linear regression, decision trees, random forests, and XGBoost) to complete the task of predicting house price trends at a time when house price trends are becoming the focus of attention. The performance of each model is compared to evaluate its different advantages and disadvantages in predicting housing prices. The results show that XGBoost model is the most effective in predicting house prices and becomes the best choice, followed by random forest model linear regression model, decision tree model can play a role in some cases. The results of the experiment can not only be convenient for home buyers or real estate developers to get help, but also can be used as a reference for government decision-making.

1 INTRODUCTION

Houses are the basic needs of people's lives. Houses allow us to have a comfortable place to live, work or go to school, so everyone has a requirement of houses. But in today's society, there is a factor that will first affect the house when people's income is higher. When they buy a house, this will cause the supply to be lower than the demand or demand needs which will cause house prices to change, the second war, the war will cause the price of the house to rise. But these factors often take a lot of time to calculate this data, so Artificial Intelligence (AI) can be considered for assisting us in this case since it has strong feature extraction and prediction capabilities, which can process more data faster, and is usually more efficient than human calculations. In addition, this can be conveniently modified by home buyers, real estate developers, and can be used as a reference by the government when making decisions.

Generally, AI enables computers to emulate human intelligence. In more detail, AI involves creating systems that possess the ability to learn, comprehend, strategize, perceive, reason, and engage with their surroundings.

Since the summer of 1956, McCarthy et al. at Dartmouth College in the United States determined the goal of "using machines to simulate human intelligence", how to achieve this goal has become an important issue for scientists and mathematicians. Since the release of ChatGPT, the AI field has ushered in a new wave of development. In just over a year, a variety of AI large models have sprung up, and their capabilities have also been rapidly upgraded and iterated, further stimulating people's discussion of AI progress and future applications.

Machine learning stands out as a sophisticated technology capable of recognizing, understanding, and analyzing highly intricate data structures and patterns. These intelligent behaviors give machines unprecedented capabilities such as speech and image recognition, Natural Language Processing (NLP), problem solving, and more. Prusty et al. discuss how different machine learning techniques can be used to

[a] https://orcid.org/0009-0005-9637-2638
[b] https://orcid.org/0009-0000-7223-3880
[c] https://orcid.org/0009-0000-2874-3330

Cui, J., Liu, Z., Ma and Y.
AI in Real Estate: Forecasting House Prices with Advanced Machine Learning Models.
DOI: 10.5220/0013511000004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 144-148
ISBN: 978-989-758-754-2

predict hypothyroidism and compare them to find the best predictive model (Prusty, 2022). Khan et al. also co-authored a paper discussing comparative studies of machine learning algorithms for detecting breast cancer and affirming the potential of machine algorithms such as XGBoost (Khan, 2021).

Therefore, based on the fact that AI has demonstrated extraordinary predictive performance on various tasks, this paper intends to consider also utilizing different machine learning algorithms for house price forecasting and analysis.

To cope with the issue of housing price prediction, this study utilized datasets sourced from Kaggle. The study implemented several machine learning models. The performance of each model was compared to evaluate their effectiveness in predicting housing prices. The results demonstrate the efficacy of machine learning models in forecasting, highlighting the strengths and limitations of each approach in the context of real estate analysis.

2 METHOD

2.1 Dataset Preparation

The dataset used in this study is sourced from Kaggle and contains 1,460 data points with 81 features related to housing prices. The features include details such as street, area, number of rooms, and other characteristics of the houses. The primary goal of this dataset is to predict house prices, making this a regression task.

The steps involved in data preprocessing include separating 35 numerical and 43 non-numerical columns, handling missing values, applying one-hot encoding, normalizing numerical features. In addition, the proportions for the train-test split were carefully chosen to ensure a balanced and accurate model evaluation. The specific methods and code implementations used for data preprocessing are based on industry-standard practices and tailored to the requirements of this study.

2.2 Linear Regression-Based Prediction

Linear regression is a method used to analyze the relationship between one dependent variable and one or more independent variables (Su, 2012; Montgomery, 2021; James, 2023), often used for prediction. Its principle is to find a line that best fits the dependent and independent variables, allowing us to use this equation to calculate new values for the dependent and independent variables. This method enables us to predict house prices.

2.3 Decision Tree-based Prediction

A decision tree is a common machine learning model used for classification problems (De, 2013; Song, 2015). It represents the process using a tree-like structure, where each node in the structure corresponds to a feature that helps determine the strategy most likely to achieve the goal.

The decision tree algorithm utilizes a hierarchical structure to perform classifications. It consists of several essential components: the root node that holds all the samples, internal nodes that test specific feature attributes, and leaf nodes that provide the decision outcomes. In the prediction phase, the algorithm examines an internal node's attribute value to decide the path towards a leaf node, where it delivers the final classification result. This supervised learning method operates on if-then-else logic, with the decision rules derived from data training, instead of manual construction.

Indeed, the decision tree is one of the most straightforward machine learning algorithms, known for its ease of implementation, clarity, and alignment with human reasoning. It is widely applicable across various fields. However, the inherent nature of decision trees can lead to the creation of overly complex models. This complexity often results in poor data generalization, commonly referred to as overfitting.

2.4 XGBoost-Based Prediction

XGBoost is a highly efficient and scalable machine learning algorithm that implements gradient boosting for decision trees (Chen, 2016; Nielsen, 2016; Torlay, 2017). It’s optimized for performance, supporting parallel and distributed computing, making it ideal for handling large datasets. It is designed to be exceptionally fast, scalable, and portable, making it a powerful tool for machine learning tasks, particularly in distributed computing environments.

XGBoost is widely used in data science for tasks like classification, regression, and ranking. Its key features include handling missing data, regularization techniques to prevent overfitting, and working seamlessly in environments like Hadoop and MPI.

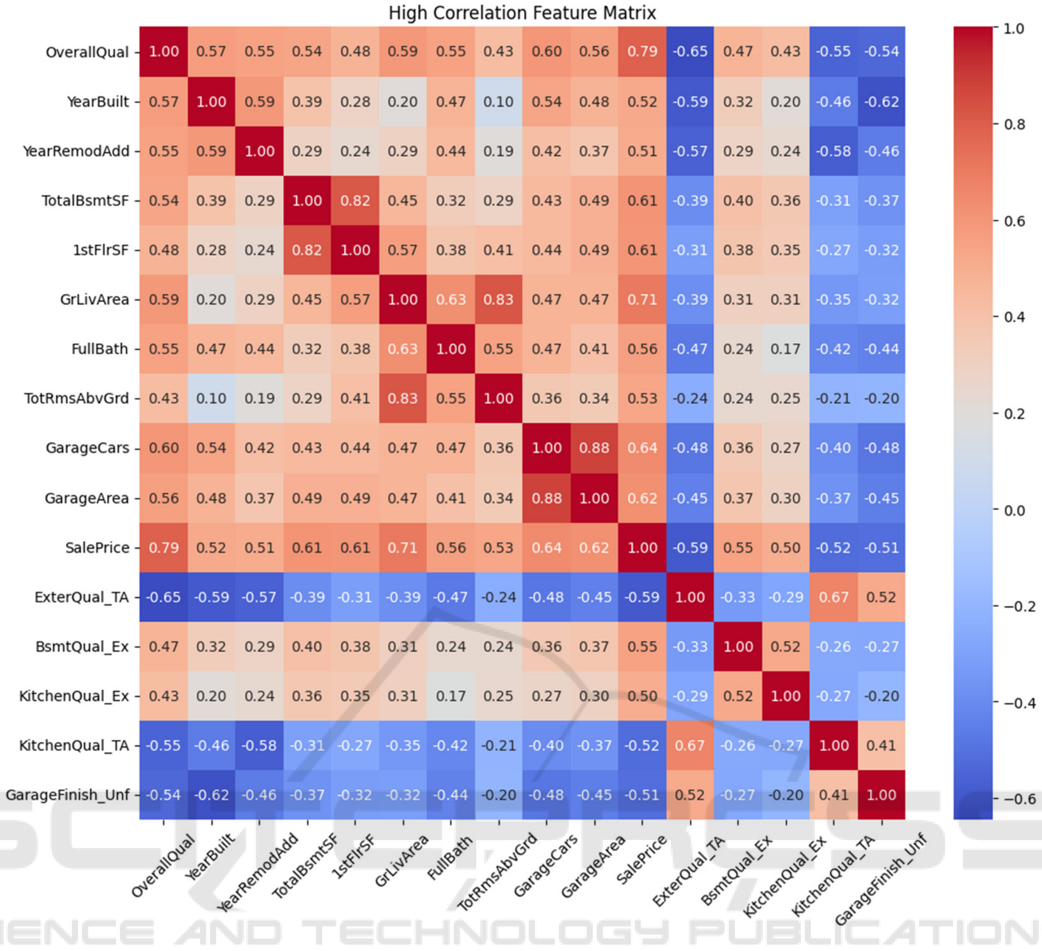

Figure 1: Correlation matrix (Photo/Picture credit: Original).

3 RESULTS AND DISCUSSION

3.1 Exploratory Data Analysis

This study's initial step involved an exploratory data analysis to understand the relationships between various features considered for house price prediction. This analysis was crucial in identifying potential correlations between features, which could influence the performance of the machine learning models.

A heatmap shown in Figure 1 was generated to visualize these relationships, displaying the selected features' correlation matrix. For instance, the strong positive correlation between "OverallQual" and "SalePrice" (0.79) suggests that the overall quality of a house significantly impacts its selling price. Conversely, a notable negative correlation between "ExterQual_TA" and "OverallQual" (-0.65) implies that houses with a "TA" exterior quality rating tend to have lower overall quality.

3.2 Feature Selection

Based on the results of the correlation analysis, certain features were selected or excluded to mitigate the risk of multicollinearity, which could adversely affect the model's predictive accuracy. Features exhibiting high correlations with each other were carefully evaluated, and in cases where the correlation was deemed too strong (e.g., above 0.8), only one feature was retained to prevent redundancy and overfitting.

3.3 Model Performance

The performance of four models was evaluated using the selected features. Each model was assessed based on its Mean Squared Error (MSE) and R^2 score on the test dataset. The results are summarized in Table 1 below.

Table 1: Summary of Model Performance Metrics

Model	MSE	R^2Score
Linear Regression	873794723.7045932	0.8860811522089055
Decision Tree	1858389542.921233	0.7577170132373361
Random Forest	850798837.9281938	0.8890791845161693
XGBoost	685000624.3803617	0.9106947183609009

3.4 DISCUSSION

Linear Regression showed a strong performance with an MSE of 873,794,723.70 and an R^2 score of 0.8861. This indicates that the model captured the underlying patterns in the data fairly well, though it leaves some room for improvement in reducing prediction errors.

Decision Tree performed less effectively compared to the other models, with an MSE of 1,858,389,542.92 and an R^2 score of 0.7577. The model appears to have overfitted the training data, resulting in a less generalized performance on the test set. This overfitting is evident from the significantly higher MSE and lower R^2 score, suggesting that the Decision Tree model is less reliable for making accurate predictions in this context.

Random Forest improved upon the Decision Tree results, with an MSE of 850,798,837.93 and an R^2 score of 0.8891. The ensemble nature of the Random Forest model likely contributed to its better generalization ability, reducing the overfitting issue seen with the Decision Tree. This model performed closely to the Linear Regression model, with a slightly better MSE, indicating that it is a strong contender in predictive accuracy.

XGBoost outperformed all other models, achieving the lowest MSE of 657,230,838.02 and the highest R^2 score of 0.9143. This model's superior performance suggests that it effectively captured the complexities of the data, providing the most accurate and reliable predictions. The combination of gradient boosting with decision trees in XGBoost likely contributed to its ability to handle intricate patterns within the dataset, resulting in better overall performance.

4 CONCLUSIONS

This study assessed the performance of four models. The analysis revealed that XGBoost is the most effective model, offering the highest accuracy and best generalization capability. This finding highlights the value of advanced techniques like gradient boosting in capturing complex patterns within housing data.

While simpler models such as Linear Regression and Random Forest also showed reasonable performance, they were surpassed by XGBoost in terms of predictive accuracy. These results suggest that adopting more sophisticated models can lead to better predictions in real estate markets.

For future work, there is potential to further improve model performance by incorporating additional features or exploring other advanced algorithms. The insights gained from this study provide a foundation for more accurate and informed decision-making in real estate pricing and market analysis.

AUTHORS CONTRIBUTION

All the authors contributed equally, and their names were listed in alphabetical order.

REFERENCES

Chen, T., & Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

De Ville, B. 2013. Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), 448-455.

Inan, M. S. K., Hasan, R., & Alam, F. I. 2021. A hybrid probabilistic ensemble based extreme gradient boosting approach for breast cancer diagnosis. In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1029-1035). IEEE.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. 2023. Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Cham: Springer International Publishing.

Montgomery, D. C., Peck, E. A., & Vining, G. G. 2021. Introduction to linear regression analysis. John Wiley & Sons.

Ngiam, K. Y., & Khor, W. 2019. Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262-e273.

Nielsen, D. 2016. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? (Master's thesis, NTNU).

Prusty, S., Patnaik, S., Dash, S. K., & Prusty, S. G. P. 2022. Designing a Pipeline for Predicting Hypothyroidism with Different Machine Learning Classifiers. In 2022 2nd Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON) (pp. 1-5). IEEE.

Song, Y. Y., & Ying, L. U. 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

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

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. 2017. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain informatics, 4, 159-169.