Research on House Price Prediction based on Machine Learning

Xiangjun Yang

Department of Computer Science, Gonzaga University, 99258, United States

Abstract. Accurately predicting house prices is of vital importance to individual home buyers and investment groups, which not only profoundly affects the formulation of home-buying strategies, but also is closely related to the smooth operation of the economy and the overall development of the society. In recent years, machine learning techniques have shown remarkable potential in house price prediction, as these models can mine the complex nonlinear correlations in large amounts of historical data to produce more detailed and accurate predictions. This study aims to evaluate and compare the performance of various machine learning models on the task of house price prediction. For the house price prediction task, Random Forests generally perform better than Linear Regression and Single Decision Tree because they can better capture complex patterns in the data and reduce the risk of overfitting. Linear regression models are simple and easy to interpret, but may not be accurate enough when dealing with nonlinear relationships and outliers. The advantages of random forests are reflected in higher predictive accuracy, robustness to outliers, and the ability to handle interactions between variables automatically.

1.Introduction

Housing prices are an important indicator of the well-being of urban residents, which is directly related to people's financial behavior and quality of life. Timely and accurate prediction of real estate prices can help reduce people's economic losses, enhance people's sense of well-being, and strengthen social stability. Machine learning has become an important tool for real estate price prediction because it can handle massive amounts of data, portray nonlinear relationships, and constantly revise the model. This project proposes to predict housing prices using several machine learning methods and compare and analyze their performance.

Housing price forecasting is a key link in real estate market analysis, and is of great guiding significance to investors, government policymakers, and ordinary homebuyers. When discussing the importance of housing price forecasting and its methodology, we first need to recognize the complexity and variability of the real estate market[1]. Real estate prices are affected by factors such as economy, policy and society, and have obvious cyclical and geographical characteristics. In addition, the problems of information asymmetry and data availability increase the difficulty of house price forecasting. However, with the

Corresponding author: xyang@zagmail.gonzaga.edu

development of big data and artificial intelligence technology, researchers have begun to try to utilize advanced methods such as machine learning and deep learning to improve the accuracy and efficiency of house price prediction.

In recent years, research on real estate price prediction has grown significantly. People have been looking for more efficient prediction models from traditional statistics to machine learning. Machine learning methods such as multiple linear regression, decision tree, and random forest have been widely used in real estate price prediction because they effectively deal with complex nonlinear relationships and massive feature information. Meanwhile, deep learning algorithms represented by convolutional neural networks and recurrent neural networks also show good application prospects in the field of real estate price prediction. This project intends to combine the above theories and methods to establish a housing price prediction model that can fully take into account the characteristics of China's real estate market and data characteristics, so as to provide more accurate prediction and decision-making support for the behavior of China's real estate market players.

2. Data and Methods

2.1 Data Collection

The data used for the house price prediction model in this study comes from the Kaggle website, an online platform that provides a wide range of datasets, including detailed real estate information and house price data. The community has validated these datasets to be of high quality and utility, providing a reliable data base for the house price prediction study in this paper.

The dataset mainly contains the following structural variables.The LotFrontage attribute indicates the street frontage width of the parcel on which the house is located. Larger frontage widths may increase the value of a home.LotArea is also one of the most important factors in determining home prices. Generally, the larger the lot, the higher the price of the home. OverallQual ratings reflect the overall construction quality and maintenance of the home. Higher quality homes usually have a higher value. Ground floor living area is the sum of the living area on a home's first and second floors. More living space usually means a higher price. The number of vehicles that can be accommodated in a garage is also an important consideration. More garage space usually increases the value of a home.

2.2 Data Pre-processing

When delving into the real estate price forecasting research domain, the importance of data preprocessing as the first building block for constructing accurate and reliable forecasting models cannot be overstated. This phase begins with an exhaustive review of the dataset, aiming to identify and respond to missing value issues, and to maintain data integrity and accuracy by strategically filling in or excluding records. Further, multivariate statistical techniques such as box-and-whisker plot analysis and standard deviation tests are employed to accurately identify and deal with outliers, to minimize their potential interference with the model's performance.

Data normalization, as the core link in the preprocessing process, aims to unify the scales of feature values of different dimensions to ensure that the contribution of each feature to the prediction results reaches a balanced state during model training, a move that not only consolidates the robustness of the model, but also promotes the fairness of the model's assessment of the sensitivity of different features.

In feature engineering, we focus on core features closely related to property prices, such as OverallQual, GrLivArea, etc., which significantly impact house price fluctuations. At the same time, we are also committed to innovatively constructing new features to uncover underlying trends and deeper correlations in the data to improve the predictive accuracy of our models further.

In addition, to ensure a high degree of fit between the data format and the model requirements, we implemented a rigorous data format conversion strategy, including the use of a solo thermal coding technique to transform the categorical data into numerical data so that the model can perform in-depth analysis more efficiently. This step is crucial to maintain the uniformity of the dataset and the accuracy of the model inputs.

2.3 Model

In this in-depth study of house price prediction, we adopt three distinctive machine learning architectures: linear regression, random forest algorithm, and decision tree model. Among them, linear regression is known for its simplicity in model structure and directness in output, providing users with a clear and intuitive view of the linear dependencies between housing prices and their influencing factors. However, it is worth noting that the applicability of this method is somewhat limited when faced with complex and variable nonlinear relationships. The generalization ability of linear regression models may be limited by the model assumptions, especially when dealing with complex real estate market data, which may not be flexible enough[1]. The random forest model is an integrated learning algorithm that incorporates multiple decision trees, which can effectively enhance the model's generalisation performance. Random forests usually have better generalization capabilities due to their integrated learning properties[2]. Several studies have shown that Random Forest maintains high prediction performance on different datasets. By virtue of its unique mechanism-i.e., independently predicting each decision tree and processing its output through integrated voting or averaging, the Random Forest algorithm effectively cuts down the risk of overfitting and demonstrates excellent adaptability in dealing with the complex and variable real estate price prediction challenges. In contrast, the decision tree method stands out for its intuitive and easy-to-understand advantages, and has been prominent in studies exploring the correlation between house prices and their features. The method carefully segments the dataset through a recursive strategy and constructs a prediction model based on the principle of feature selection, paving a well-organized logical path for the house price prediction process. The usefulness of the technique in studying the relationship between house prices and housing characteristics, identifying important determinants of house prices, and predicting house prices is demonstrated[3].

By combining these three models, our forecasting system is able to comprehensively capture the factors influencing house prices and improve the accuracy of the forecasts through the complementary strengths of the models. This multi-model fusion strategy provides real estate market participants with an accurate and reliable house price forecasting tool.

3. Analysis of Results

3.1 Data Visualization

The importance of data visualization as a central strategy in exploring in-depth research in the field of real estate price forecasting cannot be overstated. This technique skillfully transforms complex data sets into intuitive diagrams and charts, greatly facilitating the

comprehensibility of information. Using various visualization techniques, such as line graphs, box plots, and bar charts, we conducted a detailed exploratory analysis of the research data, revealing the distribution patterns of the data and accurately identifying potential outliers. Further, with the help of advanced visualization tools such as Matplotlib, Seaborn, and Plotly, we analyzed the complex correlations between real estate characteristic variables and house prices. This process deepens the understanding of market dynamics and lays a solid foundation for subsequent analysis.

As shown in Fig.1., the average selling price of these properties is pegged at $180,921,000, with a wide range of prices stretching from a low of $34,900 to a high of $755,000 This data range is important for quickly assessing whether the home price distribution exhibits symmetry and its similar characteristics. If the distribution is observed to deviate significantly from the norm or does not follow a normal distribution pattern, it is critical and necessary to take timely pre-processing measures.

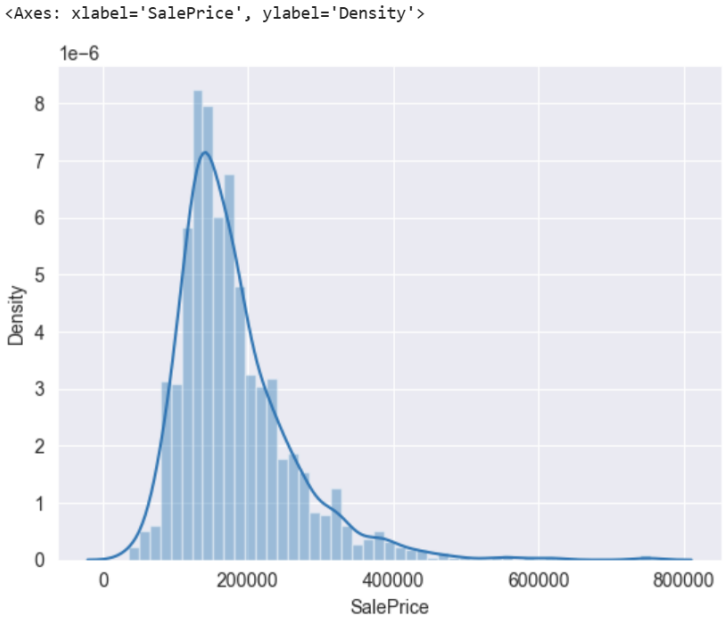

Fig.1. Normal distribution of labeled house prices (Photo/Picture credit : Original)

By visually analyzing the correlation coefficient plot (Fig.2.), this paper finds that the ten characteristics of OverallQual, GrLivArea, TotalBsmtSF, 1stFlrSF, GarageCars, GarageArea, FullBath, TotRmsAbvGrd, and YearBuilt, as influencing the house price key factors. The identification of these features is essential for constructing accurate house price prediction models

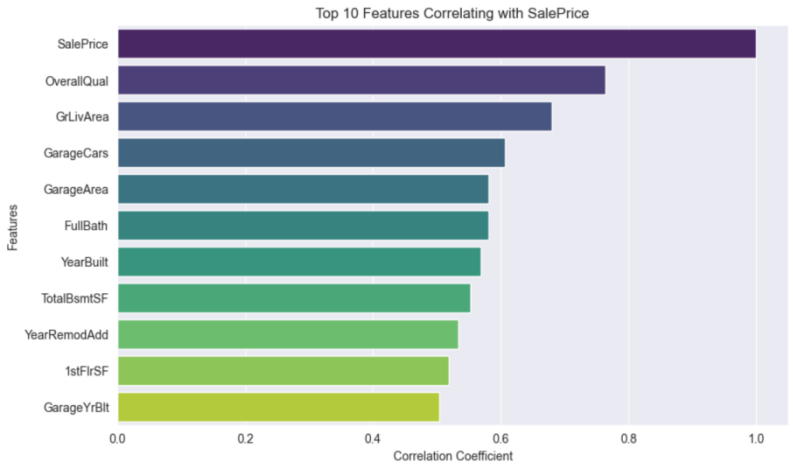

Fig.2. Top 10 features with the highest correlations (Photo/Picture credit : Original)

Merging the training and test sets for feature processing is to ensure consistency in data preprocessing. After merging 2,919 pieces of data, we perform uniform processing steps such as missing value filling, outlier correction and data normalization on 79 features. After the processing, the dataset is split into separate training and test sets for subsequent model training and evaluation to ensure accurate assessment of model performance.

3.2 Forecasting

In modeling housing price forecasts, linear regression was first applied, with the introduction of the necessary third-party libraries as an aid. Fifty different alpha parameters were used, and a cross-confirmation method was used to evaluate the performance of each parameter. The average variance information collected during the cross-confirmation period showed that the minimum error of the model was 0.140 for alpha values between 10-20, indicating that this interval is the optimal range of parameters selected for this study. Finally, based on the cross-checking, the optimal alpha is filtered to ensure the accuracy and robustness of the model in predicting housing prices.

In the visual presentation of the linear regression, we observe a series of trends in the cross-validation error (i.e., cv Error) modulated by different values of Alpha (Fig. 3).The Alpha value, as the key coefficient of the regularization term, is set to balance the model complexity with the risk of overfitting. The graph clearly shows that with the decreasing trend of the Alpha value, the cross-validation error exhibits a gradually shrinking trajectory until the Alpha value drops to a significant error trough at 0.140. This phenomenon strongly suggests that, in the context of the application of this particular dataset, adopting a lower Alpha value setting can significantly enhance the model's generalization ability. However, it is worth noting that the decreasing trend in error does not remain significant as the Alpha value further approaches zero, which may allude to the fact that the move to further weaken the strength of regularization has tended to have limited marginal enhancement in model performance (Fig.4).

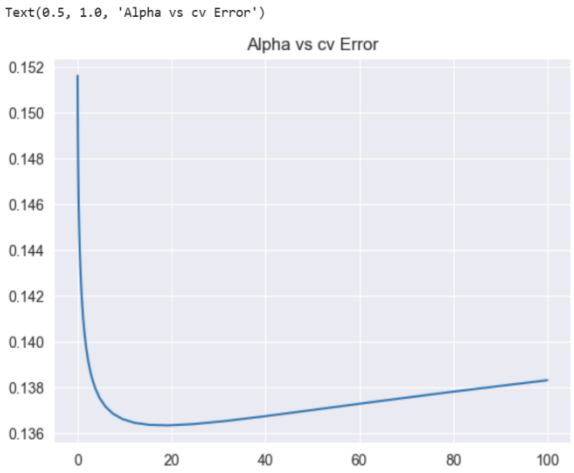

Fig.3. Alpha vs cv Error (Photo/Picture credit : Original)

We then use the Random Forest algorithm and focus on its key parameters: n_estimators (the number of trees in the forest) and max_features (the number of randomly selected features used to split the nodes). We set max_features to 0.3 times the total number of features and then explored the optimal number of trees by varying the value of n_estimators. Using a cross-validation approach, we evaluated the model performance under different values of n_estimators. Through graphical analysis, we found that when n_estimators is set to 150, the model's mean square error reaches the lowest at 0.1372, which indicates that the random forest model performs best under this parameter setting.

In the graphical presentation of the random forest, we explored the dynamic correlation between the number of estimators (i.e., the N_estimators parameter) and the cross-validation error in the model (Fig.4.). The number of estimators here refers to the total number of decision trees within the random forest architecture. The graph clearly reveals a trend: as the number of estimators grows steadily, the cross-validation error shows a gradual decrease, a change that clearly indicates a positive increase in the generalization ability of the model. However, when the number of estimators reaches a certain threshold, the rate of error reduction slows down, which may be interpreted as a sign that the marginal improvement of the model performance by increasing the number of estimators is weakening, or a hint that we are approaching the boundary of the optimal estimator allocation.

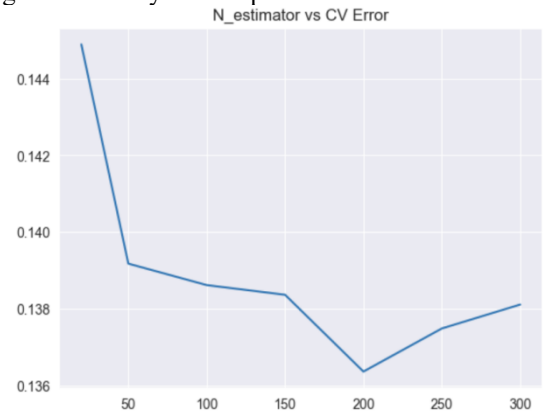

Fig.4. N_estimator vs CV Error (Photo/Picture credit : Original)

A decision tree regression method was used to predict house prices and optimize its parameters. In this study, we propose to evaluate the performance of each depth by iterating the data with different depths, using a 10-fold cross-test method, and determine the optimal tree depth by calculating the mean mean square error (RMSE). This method not only enhances the generalization ability of the model, but also intuitively understands the intrinsic connection between the tree's depth and the model's performance through visual analysis. Although the decision tree method has the advantages of intuition and good interpretability, it is often combined with integrated models to improve the overall prediction accuracy and robustness.

In the illustration of the decision tree model, we delve into the subtle connection between MaxDepth and cross-validation error (Fig.5.). MaxDepth, a core measure of decision tree complexity, quantifies the stretch length of the longest path in the tree. Analyzing the graph, it is easy to find that as the maximum depth increments, the cross-validation error shows a trend of decreasing and then increasing. This trend suggests that increasing the depth of the decision tree within a reasonable range can significantly improve the generalization effectiveness of the model; however, once the depth exceeds a certain ideal threshold, the model may fall into the dilemma of overfitting, resulting in the error climbing again. In particular, the lowest point of the error in the illustration, which corresponds to a maximum depth of about 10.0, is a finding that provides strong data support for us to determine the optimal depth of the decision tree model.

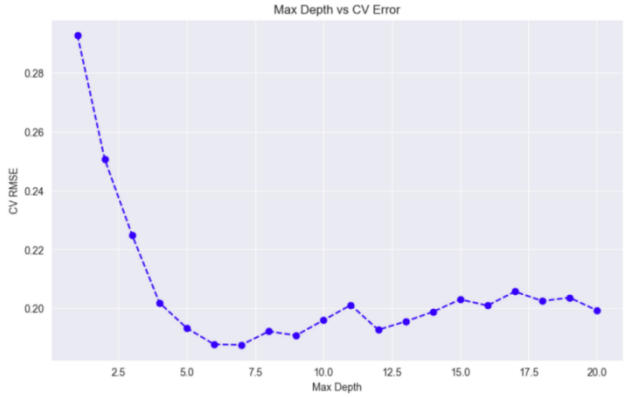

Fig.5. Max Depth vs CV Error (Photo/Picture credit : Original)

4. Discussion

4.1Model Performance Comparison

In the house price prediction study, we compared three models: linear regression, random forest and decision tree. The linear regression model is easy to explain the economic meaning represented by each coefficient due to its intuitive mathematical form, which is advantageous in certain application scenarios[4]. It is suitable for scenarios where data relationships are more linear, although it is difficult to capture complex nonlinear relationships, but it demonstrates its effectiveness in this study with an MSE of 0.140. Random forest, as a model of integrated learning, significantly improves the generalization ability by constructing multiple decision trees, effectively suppresses overfitting, and slightly outperforms linear regression with an MSE of 0.1372, especially when dealing with complex house price data. Although random forests are usually better than linear regression in terms of prediction performance, they may not be as good as the linear regression model in terms of providing

direct economic explanations, whereas the decision tree model is known for its intuition and ease of comprehension[5]. decision-making process and ease of comprehension are effective in identifying key factors affecting house prices and are used to predict house prices. However, they are prone to overfitting when used independently, which affects the generalization ability[6].

In summary, each model has its own characteristics: linear regression is suitable for scenarios with linear trends and requiring high interpretability; random forests excel in handling complex nonlinearities and improving generalization ability; and decision trees are suitable for scenarios requiring intuitive understanding of the model's decisions due to high transparency. In the practice of house price prediction, flexible selection of models according to data characteristics and prediction needs is the key to realize accurate prediction.

4.2 Factors Affecting Accuracy

Data quality is the cornerstone of improving prediction accuracy in house price prediction research. Data cleaning and pre-processing aim to eliminate noise and inconsistency to avoid interference with model performance. For missing data, fill-in or exclusion strategies are adopted. Selecting the right features is essential to improve the prediction accuracy of the model. The Random Forest model evaluates the importance of each feature to help identify the factors that impact house prices most [7]. We have selected core features such as “OverallQual” and “GrLivArea”, which have a significant impact on house prices.

Optimizing model parameters is the core of improving prediction accuracy, such as setting regularization coefficients in linear regression to prevent overfitting, and adjusting the number of trees in random forests to increase accuracy and diversity. Meanwhile, limiting the tree's maximum depth simplify the model and enhance the generalization ability.

In addition, external factors such as economic cycles, interest rates, and regional characteristics. For example, the number of cab traffic around the household area, public facilities, schools, shopping services, subway lines, and living services are factors that significantly impact housing prices in the household area[8]. It also has an important impact on house price prediction. The cyclical nature of China's real estate market makes house price fluctuations closely linked to the external economic environment [9,10]. Therefore, forecasting requires comprehensive consideration of internal and external factors, the use of integration strategies, deep learning and other advanced models, and deepening of feature engineering to explore the value of data. Continuous tracking and real-time analysis of model performance is the key to guaranteeing its long-term stability [11].

5. Conclusion

This paper used machine learning techniques such as linear regression, random forests and decision trees to provide insights into house price prediction. The data were acquired through Kaggle, and after cleaning, feature engineering and visualization, the prediction model was constructed and cross-validation was used to optimize the performance. It is found that linear regression is suitable for dealing with linear relationships, while random forest performs better in complex scenarios by virtue of its integration advantage, and decision tree is easy to understand but prone to overfitting. House price prediction is affected by multiple factors such as data quality, model parameters, economic cycles and policies. Future research can optimize model parameters, innovate algorithms to handle complexity, mine new features, integrate model strengths, and develop real-time prediction systems. This study provides market participants with data-driven forecasting tools that help make informed decisions and provide references for forecasting problems in other fields. With the development of big data and AI technology, house price prediction is expected to achieve higher accuracy and wide

application, bringing more accurate insights and decision support to the market. This study is not only an exploration of the application of technology, but also an important step in grasping future market trends.

References

1. B.W. Luo, Z.Y. Hong, J.Y. Wang, Application of multiple linear regression statistical modeling in house price prediction. Computer Age , (6), 51-54 (2020).
2. Y.L. Gong, Y.H. Yang, Construction of automatic real estate appraisal model of random forest and its comparative study. China Asset Appraisal, (1), 32-41(2022).
3. G.-Z. Fan, S.E. Ong, H.C. Koh, Determinants of House Price: A Decision Tree Approach. Urban Studies 2006, 43(12), 2301-2315.
4. S.A. Septianingrum, M.A. Dzikri, M.A. Soeleman, P. Pujiono, M. Muslih, Performance Analysis of Multiple Linear Regression and Random Forest for an Estimate of the Price of a House, in Proceedings of the 2022 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2022, pp. 415-418.
5. Lei Gan, Research on second-hand house valuation model in Chongqing based on random forest model, Ph.D. thesis, Chongqing University of Technology, (2020).
6. G.Z. Fan, S.E. Ong, H.C. Koh, Determinants of House Price: A Decision Tree Approach. Urban Studies, 43(12), 2301-2315 (2006).
7. L. Zhang, M. Xie, Exploration of the relationship between internal attributes of houses and house prices: based on random forest method. Modern Business, (22), 59-61(2019).
8. Z. Zhang, D. Zheng, Mining analysis of objective factors affecting regional house price. Computer Application and Software 36(11), 32-38, 85 (2019).
9. P.-F. Pai, W.-C. Wang, Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. Appl. Sci. 10, 5832 (2020).
10. Train, House Rates - Advanced Regression Techniques. Kaggle (2024, August 15). https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview.
11. Test, House Rates - Advanced Regression Techniques. Kaggle (2024, August 15). https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview.