House Price Prediction Based on Machine Learning Model

Haojie Chen

Sydney Institute of Language and Commerce, Shanghai University, Shanghai, China
706410415@qq.com

Abstract. This paper explores the challenging task of housing price prediction using machine learning algorithms. Leveraging a dataset of Beijing housing prices from 2011 to 2017, various preprocessing techniques, including handling missing values and feature extraction, were employed. Attributes were selected based on Pearson correlation coefficient, covariance, and principal component analysis (PCA) to improve prediction accuracy. The performance of different models was evaluated using root-mean-square error (RMSE), with RandomForest demonstrating the best performance initially. However, through attribute selection and model optimization, notably using Pearson correlation coefficient and covariance, significant improvements were observed, particularly in GradientBoost and ExtraTree models. Additionally, PCA enhanced the performance of Linear Regression. The combination of covariance and PCA further optimized model performance, underscoring the importance of attribute selection and model optimization in housing price prediction.

Keywords: house price, prediction, machine learning

1. Introduction

With the development of economy and the acceleration of urbanization, the housing price has become a hot issue that people pay attention to. The prediction of housing price trend is of great significance in investment decision, loan decision, urban planning and development, risk management, market analysis and policy making.

Housing prices are affected by a variety of factors, such as economic conditions, geographical location, and policy regulation. As a result, forecasting house prices has become a challenging problem. Forecasting house prices is a complex task that involves a variety of difficulties and challenges, including: House prices are affected by a variety of factors, including economic conditions, the job market, interest rates, supply and demand, government policies, geographical location, and so on. These factors interact with each other in a complex way, making accurate predictions difficult. Real estate market data can be volatile and affected by seasonal, cyclical and regional factors. Therefore, the volatility of these data needs to be processed and corrected. Unpredictable external factors: House prices may also be affected by unpredictable external factors, such as natural disasters, political events, financial crises, etc. These factors can suddenly change market conditions and make forecasting more

M. R. Mohyuddin and N. A. D. IDE (eds.), *Proceeding of the 2024 International Conference on Diversified Education and Social Development (DESD 2024)*, Advances in Social Science, Education and Humanities Research 899,
https://doi.org/10.2991/978-2-38476-346-7_18

difficult. Long-term and short-term factors: Housing price forecasting needs to consider both long-term and short-term factors. Long-term factors may include population growth and urban development, while short-term factors may include changes in interest rates and market sentiment.

Taking these factors together, house price forecasting is a challenging task that requires the use of multiple methods and data sources and requires careful handling of uncertainties. Even experienced analysts may not be able to predict the future direction of house prices with complete accuracy. Therefore, investment decisions and policy making should be based on a combination of multiple factors and multiple forecasting models.

2. Related Works

In the academic circle, many scholars have carried out the research of housing price forecasting and achieved certain results.

The paper [1] compares the performance of various machine learning technologies in housing price prediction, including linear regression, decision tree, random forest, etc., and provides performance comparison and recommendation. In this paper [2], time series analysis is combined with machine learning to predict housing prices, taking into account the seasonal and cyclical factors of the real estate market. In the paper [3], the author discussed the effect of deep learning models in housing price prediction, including convolutional neural networks (CNN) and recurrent neural networks (RNN). The paper [4] studies the use of satellite images and deep learning technology to predict housing prices in urban areas, taking into account the impact of geographic information on housing prices.

The paper [5] proposes a hybrid approach that combines machine learning and econometric models to improve the accuracy of house price forecasts. In the paper [6], the author studied the use of stacking method to integrate the prediction results of multiple machine learning models to improve the stability and performance of housing price prediction. [7] Methods of using mobile phone data for house price forecasting in developing countries are explored to address the problem of insufficient data. The paper [8] studied the impact of sentiment analysis in housing price forecasting to understand the impact of market sentiment on prices. The paper [9] combines spatial autoregressive models and machine learning methods to consider the impact of geographical location on housing prices.

A new method [10] is proposed to predict house prices using recurrent neural networks (RNN) and short term memory (LSTM) to account for the effects of time series data. The paper [11] uses geographic information data and random forest methods to predict housing prices in urban renewal areas to aid urban planning and development. The paper [12] uses a Bayesian method to forecast housing prices and estimates the uncertainty of the forecast to provide more comprehensive information.

The paper [13] uses reinforcement learning methods to predict housing prices, taking into account the dynamic changes of the market and the decision-making process. The paper [14] proposes a hybrid deep learning model for real-time housing price

forecasting to meet the timeliness of market demand. The paper [15] uses transfer learning methods to apply existing market data to emerging markets in order to improve the accuracy of housing price forecasting.

3. Our Method

Aiming at the need of accurate prediction of housing price, this paper proposes a method of predicting housing price by using machine learning algorithm. By preprocessing and reintegrating data, the prediction accuracy of machine learning model is improved. The following is a detailed introduction to the method in this paper.

A. *Introduction to Datasets*

The Beijing housing price dataset contains 318,851 pieces of data from 2011 to 2017, with a total of 22 characteristics. This data set is used to analyze the trends of Beijing housing prices and forecast cyclical changes in the market.

B. *Dataset Preprocessing*

1)Handling missing values

In the process of feature processing, the missing values in the data set are processed, and the rows or columns containing the missing values are deleted to ensure the integrity of the data. The data set used in this paper contains more than 300,000 real estate transaction data samples, and deleting a small number of samples with missing information will not affect the overall characteristics of the data.

2)Date feature processing

The date feature is processed to extract the attributes of year, month, day and whether it is peak season. The aim is to better understand the overall trend of house prices over time and capture possible seasonal effects.

Fisrtly, extracting year information helps to understand the overall trend of house prices over time. This can help determine whether the market is in an upward, downward, or relatively stable phase.

Secondly, extracting month information helps to identify seasonal trends. Prices can be affected by seasonal changes, for example in summer or winter when people are more likely to trade properties.

Thirdly, 12, January and February are the eve of the Spring Festival, and this period is defined as the peak season for housing transactions. This property can be used to capture whether there is a greater impact on house prices during certain time periods (e.g. holidays, special seasons).

C. *Selecting Attributes*

Since each attribute in the data set has a different degree of impact on the housing price, it is decided to screen out the attributes that have a greater impact on the housing price, and then use the model to forecast, so as to reduce the interference of the attributes with less impact on the model. The following are the basis for the three filter attributes:

1)Pearson coefficient

The Pearson correlation coefficient is a statistic that measures the strength and direction of a linear relationship between two variables. It is often used in statistics and data analysis to measure the degree of linear correlation between two continuous variables. The Pearson coefficient is calculated as follows:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

Where, X_i and Y_i are the observed values of the two variables respectively, $\bar{X}$ and $\bar{Y}$ are the mean values of the two variables respectively.

The Pearson coefficient ranges from -1 to 1 and has the following meaning. When the coefficient is 1, it means that the two variables are completely positively correlated, that is, when one variable increases, the other variable also increases. When the coefficient is -1, it means that the two variables are completely negatively correlated, that is, one variable increases and the other variable decreases. When the coefficient is 0, there is no linear relationship between the two variables.

2)Covariance to filter attributes

Covariance is a statistic that measures the strength and direction of a linear relationship between two random variables. Covariance describes the tendency of these two variables to deviate from the mean at the same time, i.e. an increase in the value of one variable is accompanied by an increase or decrease in the value of the other variable. The formula for calculating covariance is as follows:

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

Where, X_i and Y_i are the observed values of the two variables respectively, $\bar{X}$ and $\bar{Y}$ are the mean values of the two variables respectively, N are the number of samples.

The covariance can range from negative infinity to positive infinity. The specific meanings are as follows: when the covariance is positive, it means that the two variables are positively correlated, that is, the increase of one variable is accompanied by the increase of the other variable. When the covariance is negative, it means that the two variables are negatively correlated, that is, the increase of one variable is accompanied by the decrease of the other variable. When the covariance is close to zero, there is no linear relationship between the two variables.

3)Principal component analysis

Principal component analysis (PCA) is a commonly used dimensionality reduction technique for converting high-dimensional data to low-dimensional data while preserving the main information in the data. The goal of PCA is to project the data into a new coordinate system by finding the principal components in the data so that the variance of the data in the new coordinate system is as large as possible. The steps of principal component analysis are as follows: First, the data is centralized, that is, the mean of each feature is subtracted to ensure that the mean of the data is zero. Next, calculate the covariance matrix of the data. The covariance matrix describes the relationship between different features in the data. Eigenvalue decomposition of covariance matrix. The eigenvalue obtained from the eigenvalue decomposition represents the variance of the data in the new coordinate system, and the corresponding eigen-

vector represents the direction of the new coordinate system. Select the eigenvector corresponding to the largest k eigenvalues as the principal components, where k is the dimension after the desired reduction. The data is projected onto the selected principal component to obtain the data after dimensionality reduction.

Advantages of using principal component analysis: Firstly, PCA can reduce the dimensions of data, improve computational efficiency, and retain the main information of the data. Secondly, PCA can remove the correlation in the data and obtain mutually independent features by selecting the principal component. Thirdly, through dimensionality reduction, high-dimensional data can be visualized in two-dimensional or three-dimensional space for easy understanding and observation. Fourthly, PCA helps to filter out noise in the data and improve the generalization ability of the model.

4. Experiment

After processing the data set, a variety of models including LinearRegression, KNN, GradientBoost, ExtraTree, AdaBoost, Bagging, RandomForest, and DecisionTree are used to predict house prices. For the evaluation of the model, the root-mean-square error (RMSE) was used for comparison in this experiment. For each observed value, the difference between the predicted value and the actual value, known as the Residual, is calculated, and the square of all the residual is summed to obtain the mean square error (MSE), and the square root of the mean square error (RMSE) is obtained.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where, n is the number of samples, y_i is the actual value of the first observation, $\hat{y}_i$ is the corresponding predicted value.

In the housing price prediction problem, RMSE can be used to measure the prediction accuracy of the model to the housing price. The smaller the value, the smaller the difference between the housing price predicted by the model and the actual housing price.

A. *Raw Data*

The properties in the dataset were not screened, and all the data in the dataset were used for prediction. Among all the prediction results, the RMSE value of the RandomForest model was the smallest, indicating that the RandomForest model had the best effect on housing price prediction in this experiment.

TABLE I TEST RESULTS TABLE WITH RAW DATA

Model name	Attribute number	RMSE
DecisionTree	22	1093.257
LinearRegression	22	8877.554
KNN	22	6592.807
RandomForest	22	789.624
AdaBoost	22	12029.284
Bagging	22	946.815

GradientBoosting	22	3072.807
ExtraTree	22	2054.304

B.*Use Pearson Correlation Coefficient to Filter Attributes*

The first 6-25 attributes with the largest absolute value of Pearson correlation coefficient between housing prices are selected and predicted by the model. It can be concluded that when 11 attributes are selected, the effect of KNN, GradientBoost and ExtraTree models is optimal; When 12 attributes are selected, the AdaBoost and Bagging models are optimized. When 14 attributes are selected, RandomForest and DecisionTree models achieve the best effect. Among all the models, the RandomForest model with the smallest RMSE indicates that RandomForest model has the best effect on housing price prediction in this experiment. Pearson correlation coefficient was used to screen the properties and then forecast the housing price. Only two models failed to improve the RMSE value, and six models were able to achieve the minimum RMSE and all had different degrees of reduction, showing obvious improvement effect. The RMSE values for the GradientBoost and ExtraTree models have been reduced by 1168 and 1045.

TABLE II TEST RESULTS TABLE WITH PEARSON CORRELATION

Model name	Attribute selection criteria	Attribute number	RMSE
DecisionTree	pearson	14	626.289
LinearRegression	pearson	13	9256.239
KNN	pearson	11	8659.349
RandomForest	pearson	14	423.222
AdaBoost	pearson	12	11742.580
Bagging	pearson	12	443.021
GradientBoosting	pearson	11	1904.353
ExtraTree	pearson	11	1009.7462

C.*Using Covariance to Filter Attributes*

The first 6-12 properties with the largest absolute value of covariance between DecisionTree and housing price are selected, and the model is used for prediction. It can be concluded that when 8 properties are selected, the effect of Decisiontree, GradientBoost and ExtraTree models is optimal. When 9 attributes are selected, RandomForest and Bagging model achieve the best results. When 10 attributes are selected, the KNN model achieves the optimal effect. The LinearRegression model is optimized when 12 attributes are selected. When 14 attributes are selected, the effect of AdaBoost model is optimized. Among all the models, the RandomForest model with the smallest RMSE indicates that RandomForest model has the best effect on housing price prediction in this experiment. Using the covariance screening attribute to predict the housing price, only the RMSE value of two models is not improved, there are 6 models that can reach the minimum RMSE have different degrees of reduction, there is a significant improvement effect, in which the RMSE value of the GradientBoost and ExtraTree models is even reduced by 1409 and 1234.

TABLE III TEST RESULTS TABLE WITH COVARIANCE

Model name	Attribute selection criteria	Attribute number	RMSE
DecisionTree	cov	8	569.621
LinearRegression	cov	12	9365.000
KNN	cov	10	8584.153
RandomForest	cov	9	368.122
AdaBoost	cov	12	11672.880
Bagging	cov	9	414.916
GradientBoosting	cov	8	1838.467
ExtraTree	cov	8	645.289

D.*Use Principal Component Analysis to Filter Attributes*

After PCA operation on the original data, the data will be obtained and the model will be used for prediction. Among all the models, the RMSE value of the LinearRegression model is the smallest, indicating that the LinearRegression model has the best effect on housing price prediction in this experiment. Using PCA method to reduce the dimensionality of the data and then predict the housing price, all the eight models have different degrees of reduction in the minimum RMSE, which has a significant improvement effect. Among them, the RMSE value of LinearRegression decreases by 8877.

TABLE IV TEST RESULTS TABLE WITH PRINCIPAL COMPONENT ANALYSIS

Model name	Attribute selection criteria	Attribute number	RMSE
DecisionTree	PCA	22	335.095
LinearRegression	PCA	22	0.000
KNN	PCA	22	128.677
RandomForest	PCA	22	213.601
AdaBoost	PCA	22	5703.472
Bagging	PCA	22	249.278
GradientBoosting	PCA	22	249.278
ExtraTree	PCA	22	177.777

E.*Use Covariance and PCA combination to Filter Attributes*

By comparing the results of selecting attributes based on covariance with those based on Pearson correlation coefficient, it can be found that the minimum RMSE value that can be achieved by all models based on covariance is less than the minimum RMSE value that can be achieved based on Pearson correlation coefficient. Therefore, the covariance and PCA were combined to optimize the model. Among all the models, the RMSE value of the LinearRegression model is the smallest, indicating

that the LinearRegression model has the best effect on housing price prediction in this experiment. Among the 8 models, the minimum RMSE value that can be achieved by using the combination of covariance and PCA to screen the number of attributes is smaller than that achieved by only writing covariance, but larger than that achieved by only PCA, namely DecisionTree, Bagging and GradientBoosting model. For these models, PCA method is better than covariance and PCA combined method to filter the number of attributes. For the other five models, the optimal RMSE value achieved by the combination of covariance and PCA in selecting the number of attributes is better than that achieved by PCA and covariance. For LinearRegression, KNN, RandomForest, AdaBoost and Bagging models, the combination of covariance and PCA is the best method to filter the number of attributes.

TABLE V TEST RESULTS TABLE WITH COVARIANCE AND PCA COMBINATION

Model name	Attribute selection criteria	Attribute number	RMSE
DecisionTree	cov+PCA	5	372.784
LinearRegression	cov+PCA	5	5.42E-12
KNN	cov+PCA	4	119.042
RandomForest	cov+PCA	6	290.161
AdaBoost	cov+PCA	10	3895.458
Bagging	cov+PCA	5	298.932
GradientBoosting	cov+PCA	5	741.337
ExtraTree	cov+PCA	6	184.369

5. Conclusion

In the analysis of Beijing housing price data set, the RandomForest model has the best performance in the original data, but through attribute screening and model optimization, the prediction effect of multiple models has been significantly improved, especially when Pearson correlation coefficient and covariance are used to screen properties. The performance of the GradientBoost and ExtraTree models is significantly improved. Principal component analysis (PCA) has also successfully improved the performance of the LinearRegression model. It is worth noting that the combination of covariance and PCA method further optimizes the forecasting performance of the model, which provides a new idea for the study of housing price forecasting. These results emphasize the importance of property selection and model optimization to improve the accuracy of housing price prediction, and provide a useful reference for future housing price prediction research and practical application.

References

1. Li, Y., Zhang, L., & Li, X. (2017). A Comparative Study of Machine Learning Techniques for Housing Price Prediction. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM) (pp. 581-586).
2. Gao, X., & Sun, Y. (2019). Predicting Housing Prices with Machine Learning Using Time Series Analysis. International Journal of Environmental Research and Public Health, 16(23), 4782.
3. Chen, C., & Wu, L. (2020). Deep Learning Models for Housing Price Prediction: A Comparative Study. IEEE Access, 8, 155758-155769.
4. Chen, L., & Li, X. (2018). Predicting Housing Prices in Urban Areas Using Satellite Imagery and Deep Learning. Computers, Environment and Urban Systems, 71, 101-110.
5. Wang, J., & Liu, Y. (2019). A Hybrid Approach for Housing Price Prediction: Integrating Machine Learning and Econometric Models. Applied Soft Computing, 84, 105738.
6. Zhang, H., & Yu, L. (2017). Predicting Housing Prices with Ensemble Learning: A Stacking Approach. Neurocomputing, 226, 190-197.
7. Doan, T. T., & Le, T. D. (2018). Housing Price Prediction in Developing Countries Using Mobile Phone Data. In Proceedings of the 10th International Conference on Mobile Computing, Applications and Services (pp. 118-132).
8. Fu, X., & Zhang, Y. (2020). Exploring the Impact of Sentiment Analysis on Housing Price Prediction. Expert Systems with Applications, 142, 112977.
9. Yang, J., & Zhang, S. (2019). Forecasting Housing Prices with Spatial Autoregressive Models and Machine Learning. Journal of Real Estate Research, 41(2), 227-262.
10. He, X., & Wu, T. (2019). A Novel Approach to Housing Price Prediction Using Recurrent Neural Networks and Long Short-Term Memory. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM) (pp. 957-962).
11. Li, L., & Zhang, W. (2018). Predicting Housing Prices in Urban Renewal Areas Using Geospatial Data and Random Forest. Journal of Geographical Sciences, 28(11), 1689-1704.
12. Kim, Y., & Kang, J. H. (2019). A Bayesian Approach to Housing Price Prediction with Uncertainty Estimation. IEEE Transactions on Neural Networks and Learning Systems, 31(5), 1618-1630.
13. Zhang, M., & Wang, Y. (2020). Predicting Housing Prices in a Dynamic Market Using Reinforcement Learning. Expert Systems with Applications, 141, 112984.
14. Wu, H., & Zhang, J. (2020). Hybrid Deep Learning Models for Real-time Housing Price Prediction. Neurocomputing, 396, 398-406.
15. Hu, W., & Lee, D. H. (2017). Predicting Housing Prices in Emerging Markets: A Transfer Learning Approach. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM) (pp. 1167-1172).