The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China

Chengke Zou *

University California Santa Barbara, Jinan, China

* Corresponding author email: chengke_zou@ucsb.edu

Abstract. House prices increase substantially in China from 1998. Because of expensive house prices, most Chinese people have only one chance to select suitable houses. Therefore, building a house price prediction model based on housing conditions is significant for customers to make decisions. This paper collects the estate market data of Jinan city from the HomeLink website and performs several feature selection algorithms to get critical features for house price prediction. The paper compares the classical machine learning methods for the problem, including Multiple Linear Regression, Random Forest, and Catboost. After cross-validation tests, the CatBoost, algorithm with the lowest Mean Square Error (MSE) is regarded as the most accurate algorithm to predict house prices. The analytic results show that the house price is dominated by the location features such as area and block.

Keywords: House Price; Machine Learning; Feature Selection; Regression.

1. Introduction

House is the foundation of human activities. According to the theory of humans by Abraham H. Maslow [1], physiological and safety needs are the essential demands people should satisfy before seeking spirit satisfaction. Houses not only provide a place to rest but a site with privacy to meet security requirements. Therefore, having a house becomes a primary demand to satisfy these needs. However, it is increasingly difficult for Chinese people to buy a house today. Since 2006, Chinese real estate prices have increased at an average rate of 10% (Wu, Gyourko, Deng (2016)) [2]. After the subprime mortgage crisis in 2008, there was excess volatility in the real estate market. House price was stimulated to skyrocket by the economic environment at that time. Finally, Chinese house prices have become a heavy burden for most families. Because of exorbitant house prices, most civilians have only one chance to buy their houses. To avoid possible discontent and inconvenience after moving into new houses, they should consider more factors about house conditions which may affect their livelihoods. Thus, building a model to predict the house price based on house conditions is significant. People can use the model for reference to select houses suitable to their demands and financial situation.

Lancaster's (1966) Hedonic pricing model [3] provides theoretical support for the house price prediction model based on house conditions. According to Lancaster's model, the price of products reflects their characteristics. In other words, people pay money for their demands of characteristics. Shanaka Herath and Gunther Maier apply this model to the real estate market and prove that house conditions and their surrounding environment can be considered features that may influence the house price [4]. From this conclusion, a model including all houses' features can be built by machine learning algorithms to predict house prices.

Although there are already several house price prediction models like the house price prediction models in Australia, the US, and Indonesia [5],[6].[7], the features used in these models, such as the garage, fireplaces, lot size, and basement type, are not suitable for building the Chinese house price prediction model. Most Chinese houses are apartments instead of detached houses in other countries. Besides, the Chinese government's macroeconomic control policies in the real estate market, like purchase restrictions, have tremendous influences on real estate prices [8]. Previous studies on housing prices in China mainly focus on the macro level, especially the changes in house prices across China [9]. Therefore, the house price prediction model based on microlevel in line with Chinese

economic characteristics is still blank. However, Chinese house prices continue to change rapidly. The model totally based on micro factors may not be accurate enough after house prices fluctuate wildly. In addition, the heterogeneity across real estate markets in different cities makes the barrier to building a universal and stable house price prediction model nationwide. Thus, this research selects Jinan, a city with a relatively low house price growth rate and the lowest 2.6% heterogeneity in this country, as the case for building the model.

This research uses three kinds of machine learning algorithms to build models. They are linear regression, random forest, and CatBoost, an algorithm using boosting gradient. Machine learning algorithms train the model on existing selected input and output data of predicted future according to given conditions. Using these algorithms, the researcher can express the relationships among features and prices perspicuously by fitting them together. The algorithms can also reflect the importance of each feature in the model to help people define their impact on the price. Three models' RMSE will be tested and used to calculate cross-value scores to increase the model's accuracy. The model with the highest cross-value score will be selected as the final house price prediction model for analysis..

2. Data

2.1 House Pricing Factors

Lancaster's (1966) characteristics model is a neoclassical economic model proposed by Kelvin J. Lancaster in his research on consumer theory [3]. It assumes that commodities are the sum of characteristics that cannot be divided individually. Commodities themselves only have characteristics, and these characteristics provide utilities to consumers. In the premise that consumers will desire the best collection of characteristics within the extent of affordability, Lancaster's (1966) characteristics model detects the effect of consumers' preferences and products price on consumer behaviors. The product price in this framework reflects a series of characteristics, but consumers cannot know the specific price of an individual characteristic. For instance, the price of a vehicle reflects the combination of its characteristics. It may correlate with the engine, automotive interiors, exteriors, type of vehicle-mounted system, brands, and some consumers' subjective perceptions like driving feel. Based on Lancaster's (1996) characteristics theory, a rough hedonic pricing model that can parse the weight of each feature and predict the price by the change of characteristics is built. This model can be formulated as follows:

$$p = f(c_1, c_2, c_3, c_4, \dots, c_n) \quad (1)$$

Where p is the product's price, and c represents the product's characteristics.

House is a kind of good with complex characteristics. Its individual characteristics' impact on price cannot be detected directly. Thus, the house is suitable to be analyzed by this hedonic model. Furthermore, real estate possesses perdurability characteristics, structural design comfort, and urban social space [10]. In other words, house price variables are location, structure, and time. The process building models should collect and consider all attributes related to these three concepts of house.

2.2 Data Collection

We use the XML Path Language and Python to gather the Jinan real estate market data from the HomeLink website, the biggest real estate information platform. First, we select seven central administrative districts of Jinan. They are LiXia, HuaiYin, TianQiao, ShiZhong, LiCheng, GaoXin, ChangQing. In each section of districts, we crawled 50 pages of data. Each page contains 30 houses' market prices and other detailed information. Finally, we collected 9875 houses' information as samples used to build models.

We collect 15 features from the HomeLink website. These house features are housing estate, block, district, house type, area, orientation of the room, degree of decoration, floor, build times, price, the unit price per square meter, introduction, ladder household ratio, housing property, and elevators. Among these 15 features, quantitative features are area, floor, build times, price, and unit price per square meter. Rest features are unmeasurable. Algorithms are not able to utilize unmeasurable

features in the model-building process. Unmeasurable features should be transformed into measurable features as more as possible after the data processing to make models closer to reality.

The location of a house can be expressed not only by residential community and neighborhood but longitude and latitude. Using the Baidu map application program interface, we checked the latitude and longitude of the house according to its address. Then, we transformed the house type into two measurable features room and a bathroom. Because the data of house type is presented in the data set as a string containing the number of rooms and halls, we extracted the numbers from the string and added them to two new columns of rooms and halls. Afterward, houses with blank information are deleted from the dataset. Finally, the dataset has 17 pieces of feature information for 9483 houses.



Fig 1. Word cloud

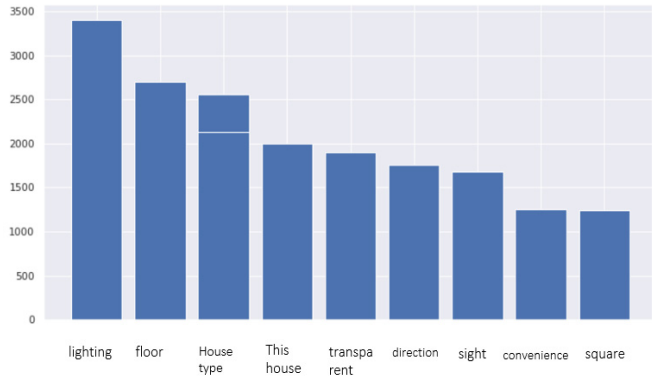


Fig 2. Frequency statistics

Among these 17 features, the introduction is the most special one. It records the specific introduction to the selling points of the house. Therefore, it contains much-complicated information related to houses, including characteristics we did not gather. The introduction is just the same thing as advertising words and cannot affect the price as a feature. However, it can help us to discover the relationship between other features and price and make feature selections if we use it properly. Fig. 1 is the word cloud made by the Jieba program according to the data in the introduction. Then, we count the word frequency in the word cloud and use it to make a bar statistics graph. Fig. 2 is the bar graph of the top ten most frequent words in the word cloud, following descending order. In this graph, the word "lighting" appears more than 3000 times in 9483 houses' introduction, meaning that good lighting conditions must be attractive to buyers. Continuing analysis with this logic, buyers can find attractive house features: lighting, intermediate floor, house type, north-south transparency, sight, convenient transportation, and house square. From this conclusion, we hypothesize that these features have a more significant impact on the price than other features and decide to test this hypothesis using feature selection, as described in the following section.

2.3 Feature Selection

Feature selection is a process of deleting unrelated or redundant and preserving important features so that features that remain can describe more accurate models. Redundant and unrelated features are noise in machine learning models because they will waste the computational performance of the model and make the model overfit to lose accuracy [11]. There are three feature selection methods: filter, package, and embedded selection. Since package selection and embedded selection directly targets the learner algorithms for feature optimization, they contradict the step of this experiment to determine data first and then determine the optimal machine learning algorithms based on the results. In contrast, filter selection completes the feature selection of datasets before training them in learner algorithms. Therefore, the research chooses filter selection. Specifically, the research uses mutual information algorithm and z-score test to select features

2.3.1 Mutual Information Score

Mutual information is a kind of information measure in information theory [12]. It tests shared information of two random variables and reflects reduced uncertainty of a random variable caused by another. A high mutual information score means that there are strong relations between two variables. The Mutual information score can be calculated as follows:

$$I(X;Y) = H(X) - H(X|Y) \tag{2}$$

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{3}$$

Where H is entropy, the measure of uncertainty; $P(x,y)$ is the joint probability density function of X and Y; $P(x)$ and $P(y)$ are the marginal probability density functions of X and Y, respectively.

In this research, variable Y is price, and each feature other than unit price and introduction is evaluated once as variable X. Since price is equal to area times unit price, unit price and area essentially represent the same kind of feature. The model mainly focuses on the effect of characteristics on house prices. Thus, the unit price is dropped when calculating the mutual information score.

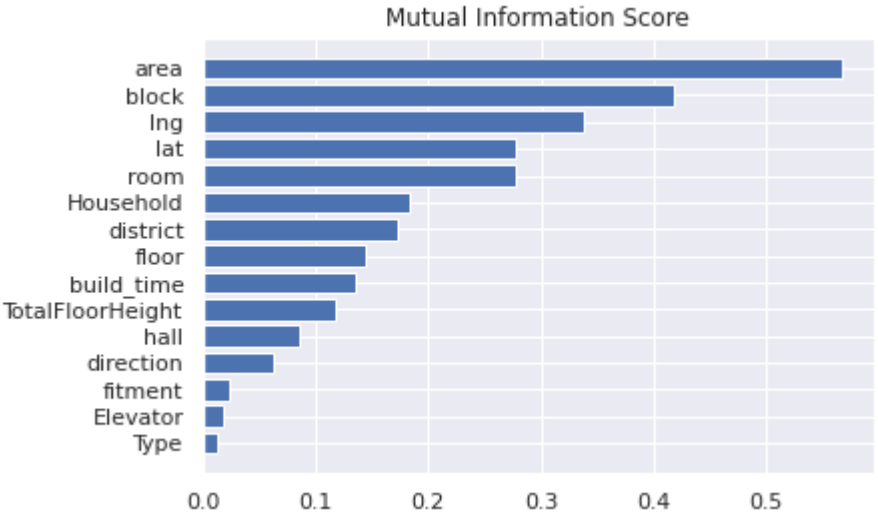


Fig 3. The rank of mutual information score

Fig. 3 shows each variable's mutual information score to house price. The Mutual information score positively correlates with certainty between two variables [12]. When the mutual information score of a variable and price is at 0 or very low, this variable has minimal or no impact on the house price. In Fig. 3, the area and the variables representing the house's location have the highest score. The change in these variables will lead to more dramatic price swings than the change in others. Decoration, elevators, and housing properties have relatively lowest mutual information scores, less than 0.05. They are almost unrelated to the house price, consistent with the research result of the house introduction. Therefore, these three variables are dropped in the dataset. The dataset preserves the remaining 11 features to build the model.

2.3.2 Z-score

Z-score is a significant measure of statistics as a deviation from the mean in standard deviation. Z-score can test the divergence of an observation point from other points in the dataset [13]. If the Z-score of an observation point is bigger than threshold value three or less than threshold value minus three, this point is an outlier point. It is a valuable tool for testing and removing outliers in the dataset. We cannot ensure that the seller inputs data align with real estate market rules. To avoid outlier points diverting the performance and accuracy of the model, removing them is necessary. The formula of the Z-score is as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{4}$$

Where X is primary data, μ is the mean value, and σ is the standard deviation.

We performed a Z-score test for each measurable feature, including longitude, latitude, area, room, hall numbers, build times, and floor height. If any house feature has an absolute Z-score value more significant than threshold value 3, that house will be removed from the dataset. A total of 349 houses are removed from the dataset as outliers. After the Z-score test, the dataset has 9134 houses to build the model.

3. Method

3.1 Linear Regression

Linear regression is a frequently-used statistics model to predict dependent variables through the input of multiple independent variables. Linear regression uses least squares to fit a linear equation [14]. This linear equation explains the relation between independent and dependent variables, and the prediction model is built by this linear equation. For instance, if multiple independent variables are having an impact on the dependent variable, their linear equation will look like follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \varepsilon \quad (5)$$

Where x_1, x_2, x_3 ..., and x_n are features, and β and ε are the best parameters of the fit.

3.2 Random Forest

The random forest algorithm is an efficient classifier using multiple decision trees [15]. The decision tree is a tree structure classification. In the decision tree, each non-leaf node tests a feature property; each branch represents the output of this property in codomain; each leaf node stores a result of classification. After filtering, the items to be classified pass through branches to reach the leaf node, and the leaf node is responsible for storing them as outcomes of decision-making. The random forest algorithm creates multiple decision trees and averages their outcomes of decision-making to produce a prediction model with relatively high accuracy.

The advantage and characteristic of the random forest algorithm are that it selects training datasets and features randomly. In other words, every decision tree is different, even unique. Therefore, all these trees are biased and generate different outcomes. Random forest algorithms seek common points of these outcomes to ensure accuracy. In this mode, it can handle plenty of variables without a dimension reduction process and have a better ability to reduce errors.

3.3 Catboost

Catboost is a complex algorithm combining gradient boosting and categorical features. The gradient boosting algorithm is a forward distribution algorithm approaching the final optimized formula through learning new primary functions and parameters step by step [16]. It fits residuals from the last calculation to minimize losses. In this process, residuals will shrink step by step until reaching a sufficiently acceptable range.

However, because of using average to calculate, traditional gradient boosting algorithm leads to overfitting easily sometimes. Catboost reduces overfitting by adopting a new method to process categorical features and optimizing the traditional gradient boosting algorithm.

Data processing disorganizes categorical features randomly, calculates the frequency of categories, and generates new numerical features to describe categorical features. Assuming that the feature σ=(σ1, σ2,…,σn) is arranged randomly, the formula to transform it is as follows:

$$X_{\sigma P,k} = \frac{\sum_{j=1}^{p-1}[X_{\sigma j,k}=X_{\sigma P,k}]*Y_{\sigma j}+\alpha*P}{\sum_{j=1}^{p-1}[X_{\sigma j,k}=X_{\sigma P,k}]*Y_{\sigma j}+\alpha} \quad (6)$$

Where P is the prior value equal to the mean of label value, and α is the weighting coefficient.

As for the computing mode difference between Catboost and gradient boosting, gradient boosting adds nodes layer by layer to optimize the results, whereas Catboost uses oblivious trees, a complete

binary tree, to create mirrored nodes. With mirrored nodes for comparison, Catboost can avoid overfitting and increase its prediction efficiency.

4. Experiments

The research builds house prediction models by the scikit-learn library from Python. We create a training dataset. The training dataset randomly selects 20% of data from the house information dataset. The rest data in the dataset is used as the test set to do cross-validation. This paper uses mean absolute error, mean squared error, and R-squre score of models to compare the most accurate model. The model with the lowest RMSE will be chosen as the final model because low RMSE will lead to high accuracy. Moreover, each model's mean cross-validation score is used to show the specific accuracy to make models more intuitive.

Table 1. Prediction Result

	RMSE	R-square	MAE	Accuracy
Linear Regression	1973.054	0.779	30.681	78.45 %
Random Forest	948.810	0.893	18.129	89.77 %
Catboost	772.408	0.913	17.506	91.39 %

It can be seen in the table that the Catboost algorithm has the lowest RMSE and highest accuracy among the three algorithms. The accuracy of the random forest model is slightly less than Catboost's, but there is a big gap between the accuracy of the linear regression model and the other two. The RMSE of linear regression is even two times greater than the random forest's RMSE. The reason is that simple linear regression is challenging to handle datasets including a number of categorical features. Although many features are transformed into numerical features in the data processing step, three significant categorical features (district, direction, block) that still cannot be transformed participated in the fitting process. Instead, both random forest and Catboost models are models based on trees. Trees can accurately partition data based on categorical variables [17].

Compared with the random forest algorithm, Catboost has a better algorithm to handle categorical features and other features. Because of random selection, the random forest algorithm cannot handle imbalanced data well. Some rare categorical features are omitted by the selection, and this phenomenon will affect the stability of the algorithm [18]. Many features in the dataset are imbalanced, like districts. There are seven administrative districts. However, the number of the seventh district is far less than the other six districts with equal numbers. These imbalanced data impact the random forest algorithm and decrease its prediction accuracy.

5. Conclusion

This research builds three models to predict the house price of Jinan and choose the most accurate one as the final model. In an environment of relatively stable housing prices like Jinan, the Catboost model can predict the most accurate house price based on cross-sectional data of houses among the three models. This model has an advantage over the other two models in dealing with complex category features. Therefore, it has better performance than the other two models. The model can help Jinan's citizens and people living in a secondary city similar to Jinan to estimate the price of the houses they want to buy or sell. With this model, people have a helpful tool to assess the risk of buying or selling real estate to avoid or reduce unnecessary economic losses. However, the model still has room for improvement. This work uses the districts and blocks to replace concepts of points of interest and school district housing. Thus, some significant factors affecting house prices cannot be incorporated into the model in detail. In future research, there should be more commitment to refining the dataset to make the model more accurate and precise

References

[1] Maslow, A. A Theory of Human Motivation. Psychological Review, 50(4), 1943, pp.370-396.

[2] Jing Wu & Joseph Gyourko & Yongheng Deng. "Evaluating the risk of Chinese housing markets: What we know and what we need to know," China Economic Review, 2016, vol 39, pages 91-114.

[3] Lancaster, K. J. 'A new approach to consumer theory', The Journal of Political Economy, 1966, Vol. 74, No. 2, pp. 132- 157.

[4] Herath, Shanaka & Maier, Gunther. "The hedonic price method in real estate and housing market research. A review of the literature," SRE-Discussion Papers 2010/03, WU Vienna University of Economics and Business, 2010.

[5] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.

[6] Byeonghwa Park, Jae Kwon Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, Volume 42, Issue 6, 2015, Pages 2928-2934.

[7] M. Thamarai, S P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning", International Journal of Information Engineering and Electronic Business (IJIEEB), Vol.12, No.2, pp. 15-20, 2020.

[8] Jing Wu, Joseph Gyourko, Yongheng DEng, Evaluating the risk of Chinese housing markets: What we know and what we need to know, China Economic Review, Volume 39, 2016, Pages 91-114.

[9] Yongheng Deng, Joseph Gyourko and Jing Wu, "Land and House Price Measurement in China," in Property Markets and Financial Stability, A. Heath, F. Packer and C. Windsor (ed.), Reserve Bank of Australia, (2012), ISBN: 978-0-9873620-3-2.

[10] Chau, Kwong Wing and Chin, T. L., A Critical Review of Literature on the Hedonic Price Model (June 12, 2002). International Journal for Housing Science and Its Applications 27 (2), 145-165, 2003.

[11] Y. Cong, J. Liu, B. Fan, P. Zeng, H. Yu and J. Luo, "Online Similarity Learning for Big Data with Overfitting," in IEEE Transactions on Big Data, vol. 4, no. 1, pp. 78-89, 1 March 2018.

[12] Wren, J.D. Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics 5, 145 (2004).

[13] V. Aggarwal, V. Gupta, P. Singh, K. Sharma and N. Sharma, "Detection of Spatial Outlier by Using Improved Z-Score Test," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 788-790.

[14] Aiken, L.S., West, S.G. and Pitts, S.C. (2003). Multiple Linear Regression. In Handbook of Psychology, I.B. Weiner (Ed.), 2003.

[15] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).

[16] Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support, 2018.

[17] Huang Z. Clustering large data sets with mixed numeric and categorical values. InProceedings of the 1st pacific-asia conference on knowledge discovery and data mining, (PAKDD) 1997 Feb 23 (pp. 21-34).