# House Price Prediction and Analysis Based on Random Forest and XGBoost Models

Han Li*

Department of Applied Mathematics, Capital University of Economics and Business, Beijing, China

*Corresponding author: 568230251@qq.com

**Abstract.** Accurate prediction of house price is important in housing market. It's difficult to forecast housing price because it's influenced by many factors. There has been many discussions on housing price prediction by kinds of machine learning algorithm. This paper attempt to predict housing price by Random Forest and XGBoost models, and compares the performance between them. In this paper, missing values processing, correlation analysis and standardization of samples are carried on the initial data at first, then two machine learning models are constructed, trained and test on the same dataset. The Kaggle house price dataset ("House Prices-Advanced Regression Techniques") is used in the paper. The dataset contains 1420 samples with 79 features that represent almost all characteristics of house. The results show that XGBoost algorithm achieve higher R square score of 89% , indicating that XGBoost model can be more efficient and accurate on measurement and prediction of housing sales prices.

**Keywords:** House price prediction; Random Forest; XGBoost.

## 1. Introduction

The housing is a major livelihood issue in society and related to the vital interests of the people. Housing price prediction can not only benefit sellers and buyers in sales transaction, but also has a positive effect on state regulating policies. Housing price prediction is a challenging task because the price is restricted by many factors. HPI(Housing Price Index) is the relative number reflecting the trend and changes of housing price in many nations. However, it's difficult to predict price of individual house accurately by HPI[1].

Machine learning algorithms develop quickly in 21st century and have been widely used in many fields such as image recognition, object detection, medical diagnosis, financial risk management, natural language understanding and auto polit.[2]. It's important to compare the performance of various kinds of ML method and select appropriate models for housing prediction. Adetunji and Akande develop RF model on Boston housing dataset and achieved $R^2$ of 90%[1]. Xiaotong Li and Xuan Guo compare various kinds of machine learnging algorithms on the price of second-hand houses in Beijing. The result shows that RF is the best algorithm and SVM follows it[3]. The research of José-María Montero indicates the semi-parametric regression algorithm outperforms the traditional parametric regressions[4]. Wang Dongxue and Guo Xiujuan trained housing price prediction model based on XGBoost and use RMSLE as evaluation method[5]. Tao Ran optimized XGBoost method for housing price prediction and the final regression $R^2$ was 87%[6]. In research of B.Vijay Kumar, a Gradient Boosting model with depth 6 and 1000 trees is developed to forecast housing price[7]. Chen Lei compare performance of prediction based on ARIMA Model and SARIMA Model[8].

In this paper, Random forest and XGBoost models are trained on a Kaggle dataset with 79 features. The test result of two algorithms are evaluated and compared to find the best selection.

## 2. Methodology

### 2.1. Data preprocessing

There are 79 feature columns and 1460 samples in the dataset. The house sale price is indicated by the label column named SalePrice. Data types include float64, int64 and object(discrete string type). Some features containing too many null or duplicate values are dropped, including 'Alley ',

'YearRemodAdd', 'FireplaceQu', 'PoolQC', 'Fence' , 'MasVnrType' and 'MiscFeature'. Feature 'ID' is the identity of sample and is dropped too.

For continuous features, we use mean value of featrure to fill null values and for categorical features, we use value of 'NONE' to replace 'NaN'.

We use method of hard encoding to transform categorical features. The class of OrdinalEncoder in sklearn package is responsible for the transformation.

The StanardScaler is used to normalize the dataset, the formula is:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{1}$$

Where μ means the mean of feature, σ means the standard deviation of feature.

### 2.2. Correlation Analysis

Pearson correlation is the statistical indicators expressing correlation degree between variables.

Mathematical formulation:

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{2}$$

$\overline{X}$ and $\overline{Y}$ are samples mean value, n is number of samples.

We use function corr() of Pandas.DataFrame class to calculate Pearson correlation between SalePrice and other 71 attributes.

### 2.3. RF Regression

Random forest regression algorithm is a classifier that contains multiple individual decision trees.[9]. Multiple CART decision trees are integrated to obtain high accuracy. For regression problem, the final prediction result is decided by aggression of model output.

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} w_i h_i(x) \tag{3}$$

Where $h_i(x)$ is output of base learner $h_i$, $w_i$ is the weight of $h_i$

One of the advantages is that importance of features can be output, which can help to optimize all algorithms. The 1000 samples in dataset are used to model training and the rest 460 samples are used to test. X array is composed of 30 features and SalePrice feature acts as predicted y. The model is developed based on class RandomForestRegressor in sklearn package and RMSE is used to estimate the result. GridSearchCV is choosen to be parameter search method because our dataset is small.

### 2.4. XGBoost

XGBoost is a machine learning algorithm based on Gradient Boosting, which uses the Newton method as loss function and calculates second derivative resulting in faster fitting and higher accuracy. [10]. The objective function comprehension:

$$obj^t = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^t}\right) + \sum_{k=1}^{t} \Omega(f_k) \tag{4}$$

Where *l* is prediction of the t th tree, $\Omega(f_k)$ is used to control complexity

We build the model based on class XGBRegressor in xgboost package. The train and test set are same as that of random tree regression. Table 1 shows main hyperparameters of the model.

**Table 1** The hyperparameters for XGB regressor

| max_depth | learning_rate | n_estimators | objective | booster |
|---|---|---|---|---|
| 3 | 0.1 | 100 | squarederror | gbtree |

## 3. Result

Table 2 lists the features with absolute value above 0.6 which indicates high correlation. We choose top 30 features as input of learning algorithms.

**Table 2**. Pearson correlation coefficient with 'SalePrice'

| 1stFlrSF | GarageArea | GarageCars | GrLivArea | OverallQual |
|---|---|---|---|---|
| 0.6047 | 0.6075 | 0.64015 | 0.7117 | 0.7835 |

In Random forest regression algorithm, The most important hyperparameter of is n_estimators. Through searching in scope of 1-100, we find the optimal RMSE is 0.41 and it is achieved when n_estimators equals 95. Figure 1 shows that training curve with the value of n_estimators parameter from 1 to 100.

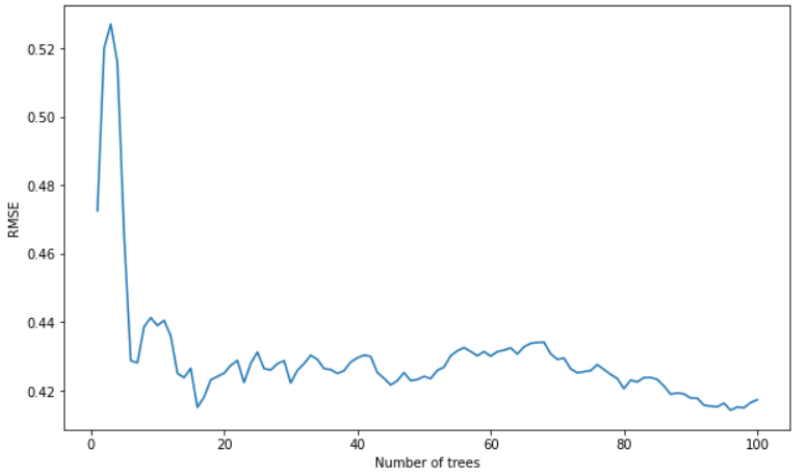


**Fig. 1** RMSE of random tree regression

Table 3 shows statistic result of Random forest regression prediction. 82% of $R^2$ indicates that the model should be optimized to improve accuracy of prediction.

**Table 3**. The hyperparameters for RF regressor

| explained variance | mean absolute error | R square | mean squared error |
|---|---|---|---|
| 0.82 | 0.24 | 0.82 | 0.17 |

We develop XGBoost model and train it on same dataset with Random forest model. The actual value and predicted value for test samples are contrasted in Figure 2.

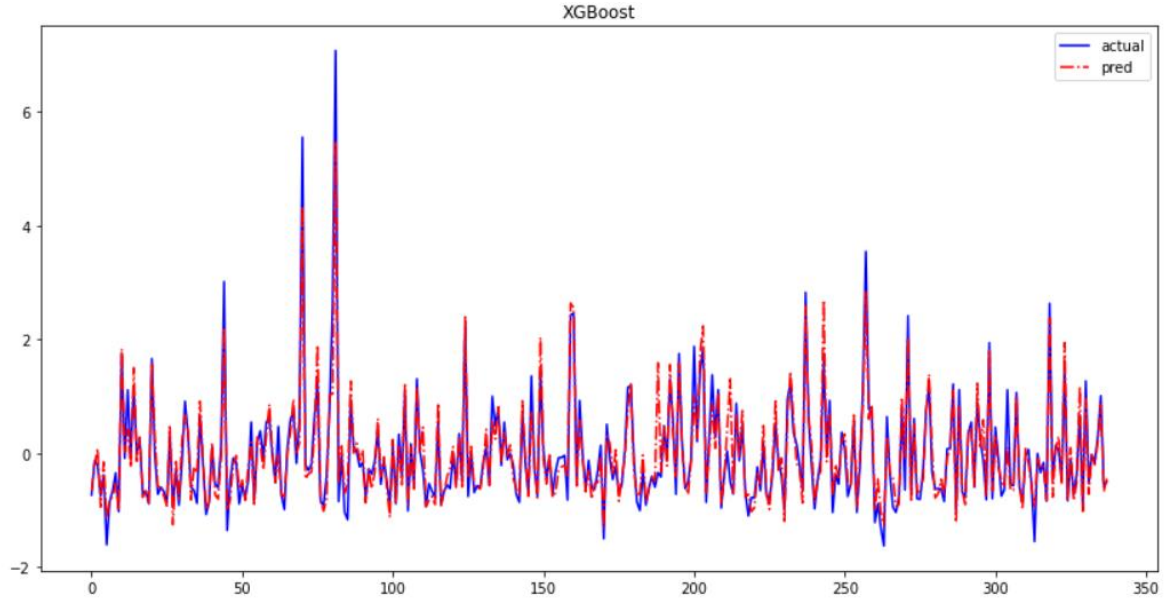


**Fig. 2** Acutal value and prediction in XGB

Table 4 shows statistic result of XGB prediction. Mean squared error fell to 0.1 and $R^2$ is improved to 89%, which indicates that the performance of XGB model is superior to Random Forest model.

**Table 4**. The hyperparameters for XGB regressor

| explained variance | mean absolute error | R square | mean squared error |
|---|---|---|---|
| 0.89 | 0.21 | 0.89 | 0.1 |

## 4. Conclusion

According to the result of validation, Random forest model and XGBoost model can both be competent for task of house price prediction. By comparing analysis, XGBoost model achives smaller prediction error and higher accrucy. At same time, the $R^2$ value of 2 models are both lower than 90% which indicates that they are not be able to used directly in practical application. In future work, Random Forest and XGBoost can be combined to improve performance of prediction, for example, Random Forest can be used to select the most important features as input of XGBoost. In addition, the hyperparameters of XGBoost can be futher optimized for higher accuracy.

## References

[1] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, etal. House Price Prediction using Random Forest Machine Learning Technique. The 8th International Conference on Information Technology and Quantitative Management, 2020&2021

[2] Akash Dagar, Shreya Kapoor. A Comparative Study on House Price Prediction. International Journal for Modern Trends in Science and Technology, 2020, 6(12):103-107.

[3] Xiaotong Li, Xuan Guo, Chengjie Wang. Analysis of Beijing Second-Hand House Price Based on Random Forest. Hans Journal of Data Mining, 2017, 7(2), 37-45

[4] José-María Montero, Román Mínguez and Gema Fernández-Avilés. Housing price prediction: parametric versus semi-parametric spatial hedonic models. Journal of Geographical Systems, 2018, 20, 27–55

[5] Wang Dongxue, Guo Xiujuan. Housing price prediction model based on XGBoost. Northern Architecture, 2021.6, 6(3): 79-82.

[6] Taoran. Optimized housing price prediction based on XGBoost. Journal of Sichuan University(Natural Science Edition), 2022.5, 59(3): (037001)1-18.

[7] B.Vijay Kumar, B.Ashritha, CH.Teja, M.Vineeth. House Price Prediction Using Gradient Boost Regression Model. International Journal of Research and Analytical Reviews, 2020, 7(1), 52-55

[8] Chen Lei. Forecast of Commercial Housing Prices in Nanjing—A Comparative Analysis Based on ARIMA Model and SARIMA Model. Statistics and Applications, 2022, 11(2): 280-287.

[9] Breiman I. Random forests. Machine Learning, 2001, 45:5.

[10] Chen T, Guestrin C. Xgboost: a scaled booting system. Proceedings of the 22nd acmsigkdd integrational conference on knowledge discovery and data mining, 2016.