Boston House Price Prediction Based on Machine Leaning Methods

Ze Li

College of Arts and Science, Boston University, 233 Bay State Road, Boston, MA02215

lize83@bu.edu

Abstract. A significant part of the U.S. economy is real estate and the housing industry. The level of housing prices influences the market's total price level to some extent. Data from Kaggle website is evaluated by using linear regression, random forest regressor and SVM regressor. There are 505 samples and the relationships between aim price with 13 different features. In terms of accuracy of the prediction, the method random forest is best method to predict house price. However, better model like XGboost could be chosen to improve prediction results.

Keywords: Boston house price; Linear regression; Random Forest regressor; SVM regressor.

1. Introduction

For the past 10 years, the global policy-making and financial communities have been paying special attention to the housing market. If the any one of countries housing market economy collapses, it might have a domino effect on other nations, such as slowing the US and European economies [1-3]. A significant part of the U.S. economy is the housing industry and real estate. In the United States, residences are frequently a significant source of household wealth, and housing development generates a significant amount of employment at the individual level since the majority of occupied housing units are owner-occupied. A sizable portion of the economy is devoted to the housing sector, therefore changes in this sector might have a significant impact on the economy as a whole.

First, as it is the basic price, the level of housing prices influences the market's total price level to some extent. Housing costs influence both the material cost of producing commodities and the cost of labor as price-of-production considerations. Due to the high value of housing, the reasonableness of housing prices not only influences the real degree of production costs and market prices of all commodities, but it also impacts household consumption expenditure. In terms of the consumer price of the entire society, the cost of housing contributes significantly, and to some extent, it also determines the consumer price of the entire market.

Secondly, housing costs serve as a significant consumer good and play a significant purpose in determining how inhabitants' living standards are regulated. If the housing price is high, the affordability of the residents will be low, and accordingly the living standard and living quality will be reduced. Consequently, the level of housing costs has grown to be a significant economic and social concern pertaining to the urgent needs of inhabitants.

In this paper, we focus on analyze Boston housing data by using scikit-learn's Boston dataset. To predict future Boston house price, we facilitate the best measurement according to different models. In section 2, we introduce the data description and three different models' functions and operations, explaining the process of evaluating given data and calculate the prediction of test data.

2. Data Research

2.1 Data Description

There are 505 samples and the aim predictable price with 13 different characteristics from www.kaggle.com is evaluated.

There are many influencing factors can alter the change of house price. In Fig.1, the information of 13 features in the dataset is applied to form a feature correlation heatmap. The variables have a perfect positive linear correlation when the value is 1, and a perfect negative linear correlation when

the value is -1. From the graph of feature correlation heatmap, TAX and RAD have the highest positively linearly correlated index 0.9, then is NOX and INDUS positively correlated at 0.8. However, DIS and NOX is the highest negatively correlated at -0.8. PRICE and LSTAT, DIS and INDUS, and DIS and AGE are all negatively correlated at -0.7.

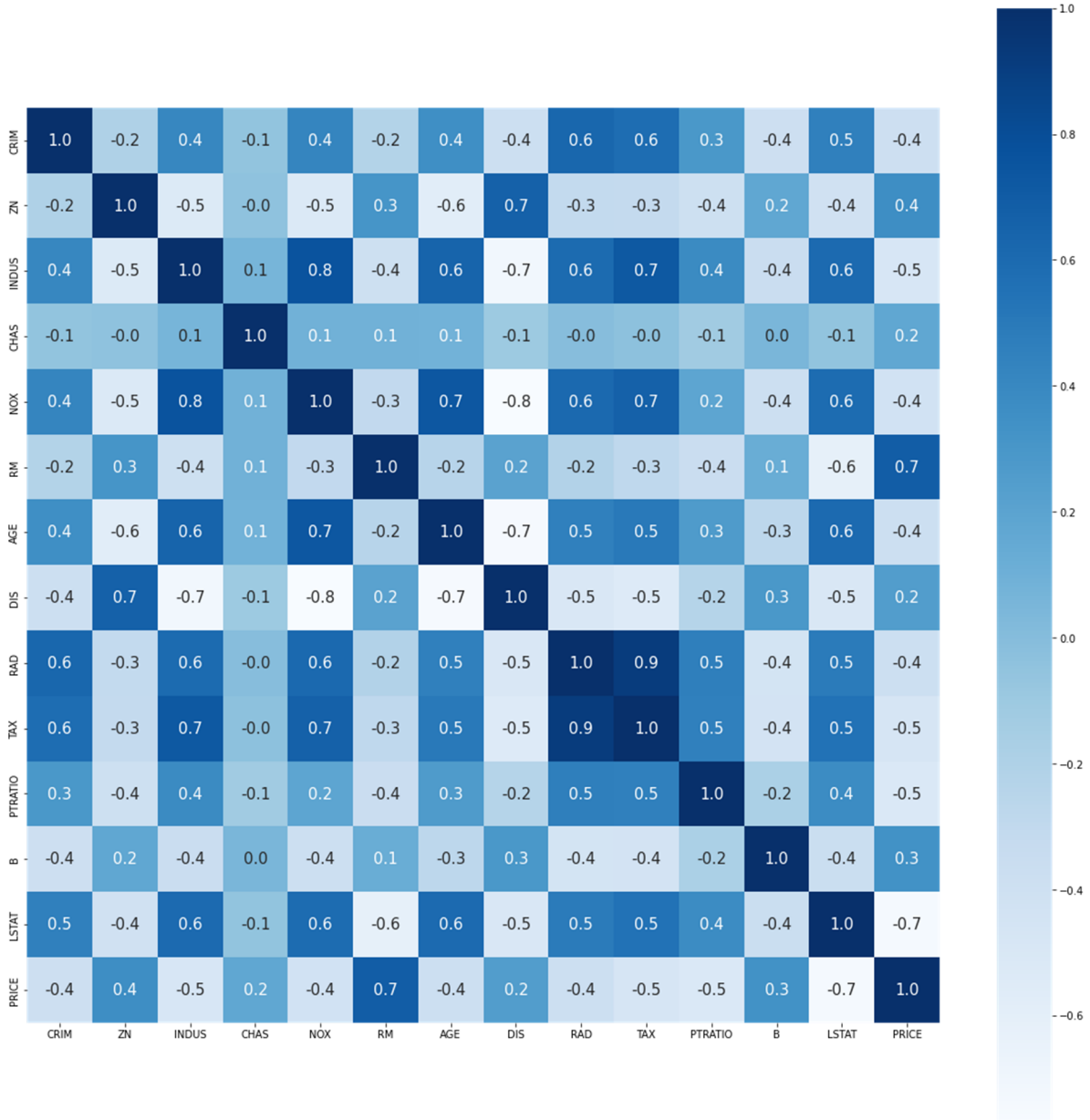

Fig. 1 Heatmap of feature correlation

In this section, the ratio of train data and test data is 8:2. In order to measure the fitness of different models, we compute these five terms for the further comparison. R squared is the ratio of the data's variance of target predictable price in terms of 13 characteristics in a regression model as a statistical aspect. Adjusted R squared, similarly, is used for linear models as a corrected goodness-of-fit for accuracy. It illustrates the percentage of variance in the target price in terms of 13 characteristics [8]. The mean of the absolute value of errors (MAE), the mean square error (MSE), and the mean square error (MSE) are the other three important standards to measure three models' fitness. All individual differences are equally weighted on the mean in a linear score called the MAE. Although standard deviation is also a good statistical measurement, the research objects and research purposes of standard deviation and root mean square are different. The root mean square error is used to account for the high dispersion of the sample. When the model is nonlinear fitting, it is preferred to apply the one has smaller RMSE.

2.2 Models

2.2.1 Linear Regression

Utilizing linear predictor functions, linear regression models the connections between variables [4-6] and the common expression is

$$y=wTx+e \quad (1)$$

and e is error having a mean value of 0 in a normal distribution.

According to the definition of linear regression, the coefficient values of linear regression model is

Price = -0.123CRIM + 0.056ZN - 0.009INDUS + 4.693CHAS -14.436NOX + 3.280RM – 0.003AGE – 1.552DIS + 0.326RAD – 0.014TAX – 0.803PTRATIO + 0.009B – 0.523LSTAT + 36.357.

To visualizing fitness of linear regression is good, the first graph in Fig. 2 shows the difference between actual prices and predicted values, since the dots can approximately scatter as a line. In order to keep the linear regression valid, checking residuals is necessary. In the second graph, the dots scatter randomly and distributed equally around 0. Therefore, the assumption of linearity is established. In the third graph, the residuals distributed normally, so it also satisfies the normality assumption.

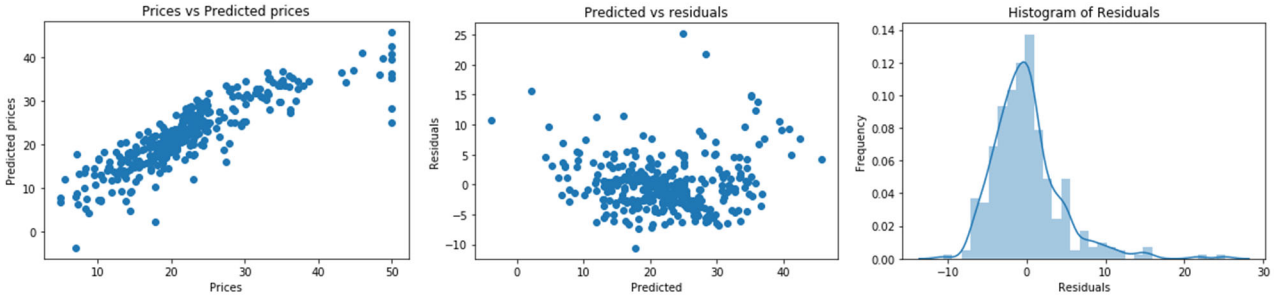

Fig. 2 Linear Regression Visualization and Normality of Residuals

After evaluating the test data by the five measurements mentioned above, these measurements are R squared is 74.66% with adjusted R squared is 73.69%. MAE, MSE, and RMSE are 3.09, 19.07, and 4.37.

2.2.2 Random Forest Regressor

A classifier known as a random forest employs several trees to train and predict samples, and the output categories are defined by the mode of the individual trees' output categories. The ability to deal with datasets with several predictor variables is a key advantage of utilizing random forests for prediction modeling; nevertheless, in reality, it is frequently preferable to minimize the number of predictors needed to produce result predictions in order to increase efficiency [9]. Because it is built on regression trees, random forest can also handle with nonlinear connections and the unsteadiness of variable effects over various segments. It gives a mechanism for detecting variable interactions through experimentation; it can balance the error for uneven classification datasets; and it computes the proximity in each example, which is highly useful for data mining, finding outliers, and displaying data.

The prediction on train data is visualized as Fig. 3. The first graph is predicted y linearly distributed with respect of prices. And the second graph is residuals have equally distributed around zero with respect of predicted value, which means that it satisfied the assumption of linearity. To evaluate the model, R squared is 96.61% with adjusted R squared is 96.48%. MAE, MSE, and RMSE are 0.94, 2.55, and 1.60.

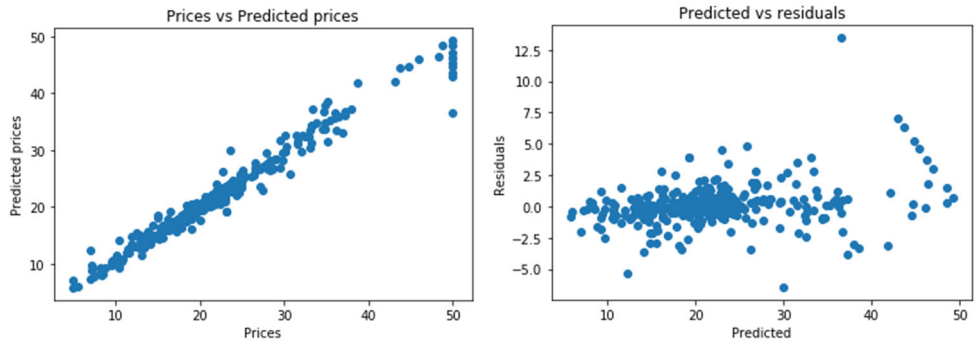

Fig. 3 Random Forest Model Visualization

For the prediction on test data, the five measurements of the random forest model is evaluated. R squared is 86.41% with adjusted R squared is 85.13%. MAE, MSE, and RMSE are 2.57, 14.19, and 3.77.

2.2.3 SVM Regressor

In machine learning, the maximum margin hyperplane serves as the decision boundary for the Support Vector Machine (SVM), a generalized linear classifier that uses binary classification to perform supervised learning on data. Since both structural and empirical risk reduction are taken into account in the SVM optimization problem, it is stable. From a geometric standpoint, the stability of the SVM is demonstrated by the fact that it needs the most space between the interval bounds to accommodate the test samples while building the hyperplane decision boundary. As a proxy loss, SVM employs the hinge loss function. Because of the value of the hinge loss function, the SVM is sparse, the support vector alone determines its decision border, and the remaining sample points are not used in the empirical risk minimization. The resilience and sparsity of SVM in nonlinear learning with kernel techniques decreases the computational and memory cost of kernel matrices while maintaining accurate solution outcomes [10].

In the train model, the SVM regressor is coded by control cache_size, gamma, and kernel. Cache size is the size of kernel cache which is 200 MB. Gamma is 'auto_deprecated' as 1/n features. Kernel is set in default of 'rbf' because the type of hyperplane 'rbf' for separating the data. The other elements are all in defaults. The coef0 is zero since it is not 'poly' or 'sigmoid'. Degree is in default of 3 since it is in 'rbf', which representing for the polynomial kernel function. Epsilon is in default of 0.1. Shrinking is a boolean set in default of True in order to determine whether to use the shrinking heuristic.

In the evaluation of train set, SVM model is visualized in Fig. 4. Similarly, just like the other two models, the first graph shows the linear relationship between predicted prices and prices. In the second graph, dots are equally distributed around 0, satisfying the assumption of linearity. R squared is 64.19% with adjusted R squared is 62.82%. MAE, MSE, and RMSE are 2.94, 26.95, and 5.19.

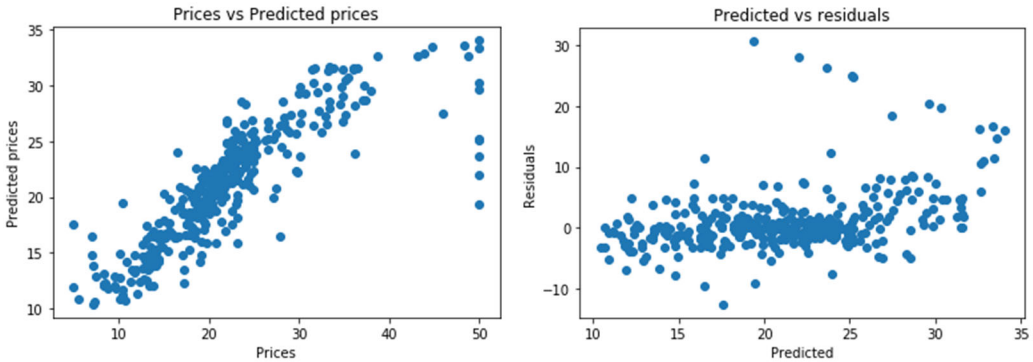

Fig. 4 SVM Model Visualization

However, In the test data, R squared is 59.00% with adjusted R squared is 55.14%. MAE, MSE, and RMSE are 3.76, 42.81, and 6.54.

2.3 Results

In this section, the comparison of three models is needed to evaluate the best fit model. The results of three method of prediction are summarized in Table 1. The R^2 of Random Forest is the highest among three methods. Therefore, the Random Forest Regression works the best for the prediction of Boston house price.

Table 1. Test Results of Three Methods Comparison

Name	R^2	Adjusted R^2
Linear Regression	74.66%	73.69%
Random Forest	86.41%	85.13%.
SVM-Regressor	59.00%	55.25%

3. Conclusion

In this article, linear regression, random forest, and support-vector machines are presented to analyze Boston house price. The helpful prediction of house price is derived from large datasets from scikit-learn's Boston dataset. In the Fig. 1, the feature correlation heatmap indicates that TAX is the highest positively linearly correlated with RAD, and NOX is the second highest positively linearly correlated with INDUS. However, DIS is the highest negatively linearly correlated with NOX.

To further analysis, three models are computed to predict the test data. From the aspect of accuracy, the random forest regression has the highest R squared. Nevertheless, we could choose better model like XGBoost to enhance the prediction result. XGBoost proposes an approximate algorithm that can greatly increase the learning rate when the data set is large, and it also introduces a regularized coefficient in the loss function in order to process pre-pruning on decision tree. In this case, we can get a more accurate prediction.

References

[1] M. S. M. Radzi, C. Muthuveerappan, N. Kamarudin, and I. S. Moham-mad, "Forecasting house price index using arificial neural network," International Journal of Real Estate Studies, vol. 7, pp. 43–48, 2012.

[2] Gupta, R., Kabundi, A., & Miller, S. M. (2011). Forecasting the US real house price index: Structural and non-structural models with and without fundamentals. Economic Modelling, 28(4), 2013-2021.

[3] Sampathkumar, V., Santhi, M. H., & Vanjinathan, J. 2015. Forecasting the land price using statistical.

[4] Wu, S., Huang, Z., Yang, X., Zhou, Y., Wang, A., Chen, L., ... & Cai, J. (2012). Prevalence of ideal cardiovascular health and its relationship with the 4-year cardiovascular events in a northern Chinese industrial city. Circulation: Cardiovascular Quality and Outcomes, 5(4), 487-493.

[5] Wu, J., Deng, Y., & Liu, H. (2012). Housing Price Index Construction in the Nascent Housing Market: The Case of China. National University of Singapore Institute of Real Estate Studies Working Paper IRES2011-017.

[6] Ye-Hui, W. A. N. G., & Xiao-Heng, X. U. 2020. Research on the Measurement of China's Adjusted House Price Index. DEStech Transactions on Social Science, Education and Human Science, (icesd).

[7] "Least Absolute Deviation Regression". The Concise Encyclopedia of Statistics. Springer. 2008. pp. 299–302. DOI= http://dx.doi.org/10.1007/978-0-387-32833-1_225. ISBN 9780387328331.

[8] https://www.ibm.com/docs/fi/cognos-analytics/11.1.0?topic=terms-adjusted-r-squared.

[9] Speiser, Jaime Lynn, et al. "A comparison of random forest variable selection methods for classification prediction modeling." Expert systems with applications 134 (2019): 93-101.

[10] Zhou Zhihua. Machine learning. Beijing: Tsinghua University Publication, 2016: pp.121-139, 298-300.