Real Estate Price Prediction based on Supervised Machine Learning Scenarios

Zhihang Liu

Rutgers, the state university of New Jersey, New Brunswick NJ08901, US

zl400@scarletmail.rutgers.edu

Abstract. House price prediction is one of the most common supervised learning tasks in the machine learning field, which makes it a perfect criterion for the effectiveness of different learning models. From basic regression models to neural networks, countless methods have been proposed to solve the house price prediction problem. In this paper, the focus is the performance of three regression models, linear, LASSO, and ridge. There will be a selected dataset of sold houses from the open-source website. The data will be explored and visualized for a better understanding and then implement the regression models for further testing. According to the analysis, the LASSO regression model can yield the most accurate prediction with 90.15% accuracy but need a specific λ value. The linear and ridge regression yields similar predictions with close to 90% accuracy. Therefore, the most effective model for the house price prediction problem is the LASSO regression model. Overall, these results shed light on guiding further exploration of the performance of different machine learning models.

Keywords: Machine Learning; Linear Regression; LASSO; Ridge.

1. Introduction

House price prediction is not only a fundamental and popular topic in machine learning but also has significant meaning in both economic and academic aspects. The price of real estate varies each day in a very opaque situation [1]. If the price is well predicted, it will give the edge to the consumers and the whole industry. Meanwhile, the change in house prices can also reflect on the social insurance policy, individuals' income, and the political economy [2]. Some of the research indicates that the factors that affect housing prices include but are not limited to residential locations, housing market policy, filtering, trapping, cycles of urban deprivation, and wealth distribution [3]. Thus, it's very difficult to accurately predict the real price change considering these non-numeric data.

Real estate is the foundation of a city's development, from material, political, and economic aspects. It formed a particular building space for a city and constructed a network of legal, political, and economic capitals that operate nationally and internationally [4]. There are some decent models and solutions for real estate price prediction. Two models named back propagation neural network (BPN) and radial basis function neural network (RBF) include the macroeconomic parameters in house price prediction. It built a model of non-linear prediction of change in house price [5]. There is another team has proposed a model called fuzz least-square regression (FLSR) and makes comparisons to some of the existing models such as Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference Systems. The outcome of the trivial shows that FLSR has a superior prediction function in capturing the relationship between independent and dependent variables in the house price datasets. The computational complexity of FLSR is also the lowest [6]. Furthermore, the support vector machine is proven effective in-house price prediction. The research team took the dataset quarterly from 1998 to 2008 in China and measure the performance of the SVR model in mean absolute error, root means square error, and mean absolute percentage error. Compare with the BPNN model, it's reasonable to conclude the support vector regression model is an excellent tool to forecast the real estate price in the future [7]. A neural network is also a solution for price forecasting. More precisely, the BP neural network and Elman neural network. It's convinced that Elman neural network has better performance for predicting price change [8]. A team in the UK gave a Gaussian Process Regression model which is claimed to be 96.6% accurate. Its score beat the regression-Kriging, random forests, and an M5Pdecision-tree [9]. Meanwhile, data mining can also be a way to forecast real estate prices. It requires a preprocessed dataset which means the data need to be normalized first. Researchers implement several data mining models, LASSO, SVM, XGBoost, kernel ridge random forest, LGBM, gradient boosting, and elastic net. The best model that yields 94% accuracy is the elastic net [10].

The purpose of the study is to sift through traditional and advanced models in machine learning and find out a generally effective model by applying these methods to real estate price prediction, thereby reveal an appropriate model to predict the house price. The Sec. 2 will introduce the dataset and models used for machine learning. The Sec. 3 will demonstrate the outcomes of different models and compare metrics and give an explicit discussion of the results. Finally, the overall computation will be examined and focus on the limitations of the models and datasets and the effectiveness of the models. Furthermore, improvement will also be discussed for future work and conclude the whole paper.

2. Data & Method

The training and testing dataset is acquired from Kaggle "House price prediction". The dataset contains 81 attributes and 1259 data samples. The Sale Price is the independent variable and the rest are dependent variables. Fig. 1 shows the number of the different sale prices in the training dataset, where the houses sold in price from 0.1 million to 0.2 million dollars. The skew and kurtosis of the sale price are 1.96 and 6.99.

In this paper, linear regression, Ridge regression, and LASSO regression will be applied to predict price from the scikit-learn library in python. Linear regression is a simple but solid model in machine learning. It finds a linear relationship between the dependent variables and other variables, thereby revealing the variables which have stronger relations and deducing an appropriate result. Its equation is simply:

$$Y = a + bX \tag{1}$$

Here, X is the explanatory variable and Y is the dependent variable, b is the slope of the line and a is the intercept.

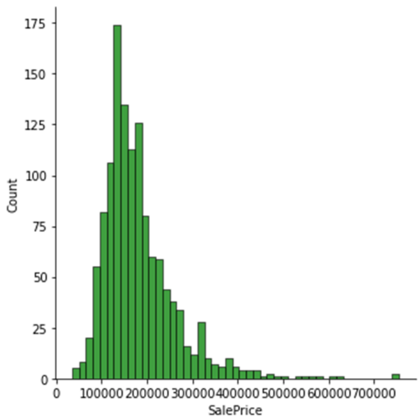

Fig 1. Distribution of Sale Price

To find the relations more accurately, outliers which are the points that are too far away from other clusters of points need to be removed. The LASSO regression is similar to the linear regression in the format but more complicated. The result is the residual sum of the square plus λ times the sum of the square of weights. Here, λ is the amount of shrinkage. If λ is equal to zero, all the features will be considered, and LASSO regression will behave like linear regression. As the λ approach to infinity, none of the features will be considered. LASSO regression is a regularization technique. Thus, regularizing the dataset first is essential. The principle of regularization is adding a boundary to the train data to produce less variance, which can be described as:

$$MIN\left(\left|\left|Y - X\theta\right|\right|^2 + \lambda\left|\left|\theta\right|\right|^2\right) \tag{2}$$

Ridge regression is a perfect tool for a dataset that possesses multicollinearity. Similar to LASSO regression, Ridge regression has a boundary or penalty term. By changing the λ value, the penalty term can be directly computed. It also shrinks the parameters to avoid multicollinearity. The shrinkage can also lead to less computational complexity. To apply ridge regression, the dataset must be standardized, which means that every value needs to subtract the corresponding mean and then be divided by the corresponding standard deviation. However, the bias and variance will change if λ is changed. If λ increases, bias will increase, and variance decrease, vice cetera. Therefore, it depends on the operator to apply the appropriate λ value. To evaluate the models, the metrics root of mean squared error and r2_score functions are utilized based on the API offered in the scikit-learn library.

3. Results & Discussion

To find the correlation in the data to examine the most related variables to the sale price, the heatmap of the correlation coefficients is given in Fig. 2. It's obvious that the most related variable is overall quality. The second place is GrLiveArea which stands for the living area above the ground in square feet according to the data description. Such an occasion is very reasonable since the quality of the house is essential for the real estate price and normally the bigger the house is, the higher the price is. Fig. 2 excludes the variables that have a correlation of less than 0.5 since it's less relevant to the sale price. Explore the relationship between other independent variables and the sale price.

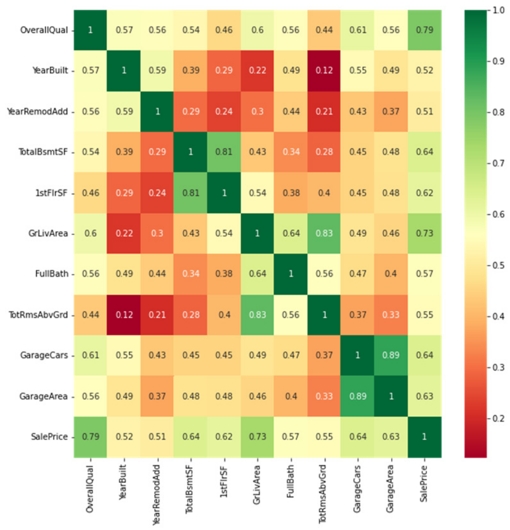

Fig 2. Heat map of correlation of different variables

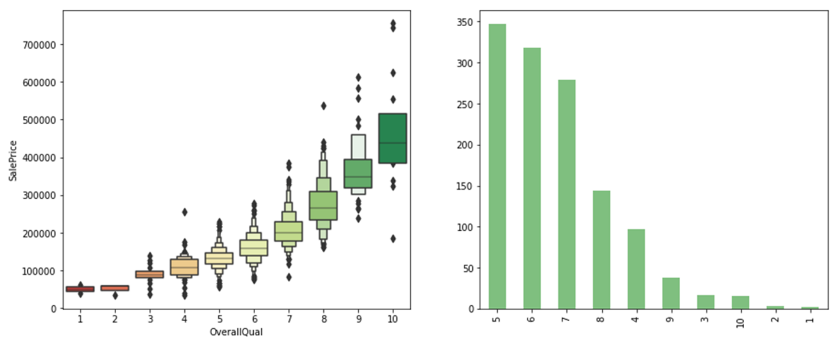

Fig 3. Box plot overallQual versus SalePrice and sale distribution of overallQual

The left panel of Fig. 3 demonstrates the relationship between the overall quality and sale price, it's roughly linear. As the overall quality grows, the sale price goes higher. The right panel shows the number of sales for different quality indexes. The medium-quality houses are sold the most and the best and worst-quality houses are sold the least.

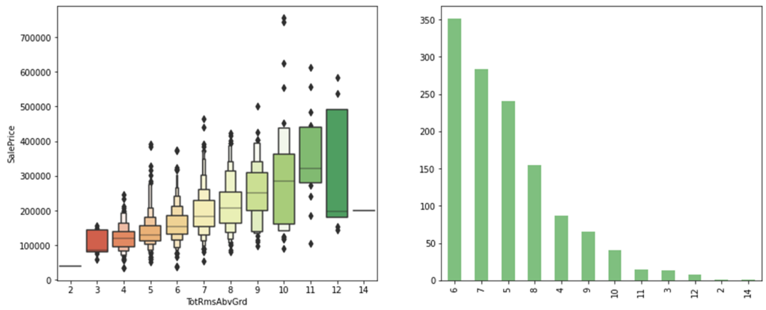

Fig 4. Box plot TotRmsAbvGrd versus SalePrice and sale distribution of TotRmsAbvGrd

The left panel of Fig. 4 shows the sale price for the houses which possess different numbers of residential rooms above the ground level. Somehow, some houses that have 10 rooms have the highest sale price. When the number of rooms is above 10 rooms, the highest sale price seems to drop a bit. Similarly, the right panel shows the number of sales for houses has different numbers of rooms. The 6-room houses are sold most. Its relationship is similar to the number of sales for different quality houses.

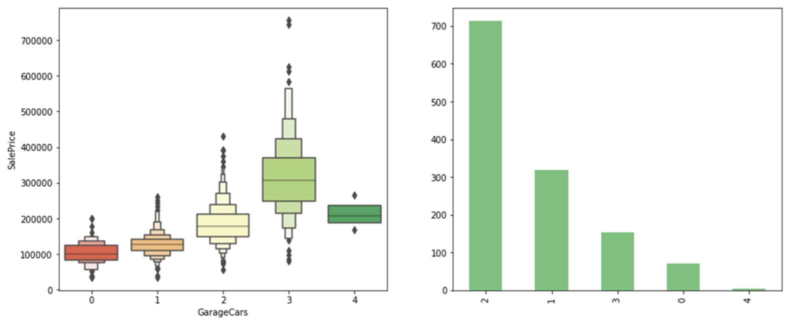

Fig 5. Box plot GarageCars versus SalePrice and sale distribution of GarageCars

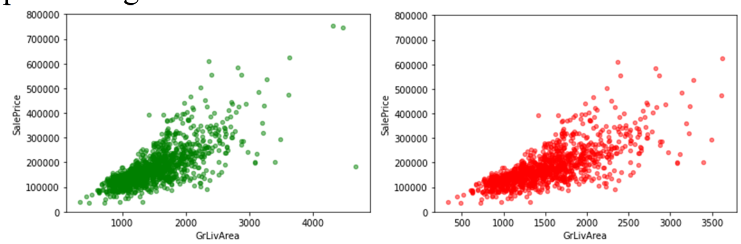

Fig 6. Scatter plot of GrLiveArea between SalePrice before and after outliers' removal.

The left panel of Fig. 5 shows the houses with different numbers of garages, from 0 to 4, and their corresponding sale price. It's clear that houses with 3 garages can have the highest price. Nevertheless, the prices of the 4-garages house are much lower than the 3-garages house prices. It's a very strange phenomenon since more garages usually mean higher prices in common sense. Meanwhile, the right panel clearly shows that houses of 2 garages are sold the most and the houses with 4 garages are sold the least. The left panel of Fig. 6 is the relationship between the living area above the ground of the houses which is the second most correlated variable with the sale price and sale price. Most of the points cluster around 1000 to 2000 in the living area and there are some outliers spotted in the plot. The right panel of Fig. 6 shows the same content as the previous plot but removed outliers. The same procedure can be applied to other variables.

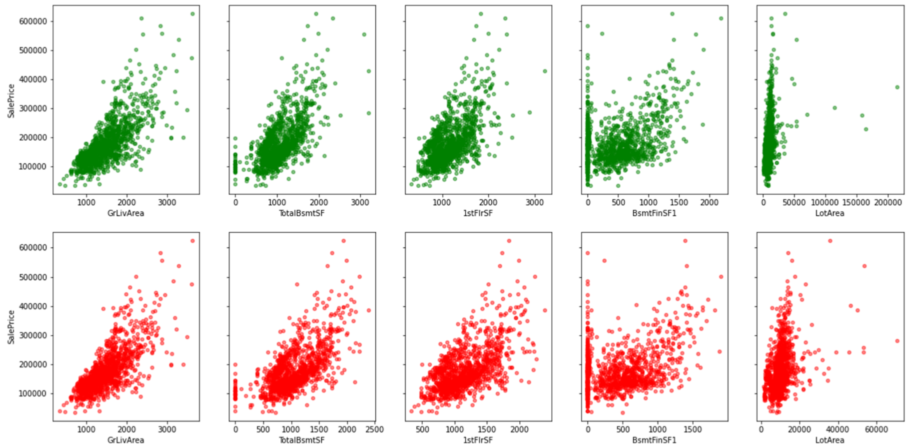

Fig 7. Scatter plot of different variables between SalePrice before and after outliers' removal

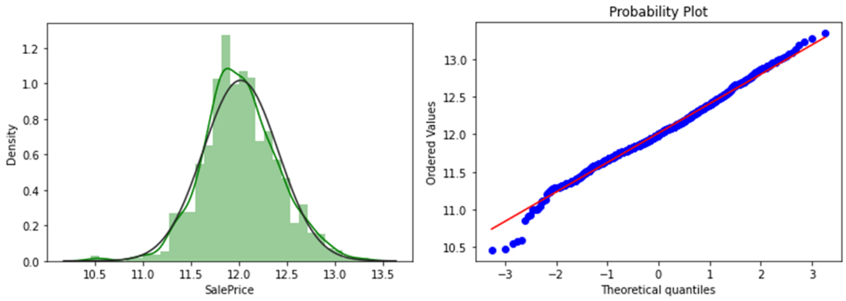

Fig 8. Density distribution of SalePrice and QQ plot.

Fig. 7 is a series of plots of sale prices versus different variables. The green plot shows the outliers before removal, the red plot shows the outliers after removal. The log transformation needs to be applied to the sale price variable and make it more linearly distributed. Fig. 8 illustrates the result of log transformation and redistribution. After this step, the missing values in the dataset will be taken care of, since there are very few houses that have every feature described in the dataset. Seen from Fig. 9, it's easy to observe that most houses don't have pools, miscellaneous features, alley access, fences, and fireplaces. At first, encode the categorical variables. Subsequently, one needs to get rid of unwanted features and variables, which are less relevant and useful. If the missing value of a feature is higher than 0.7, that feature will be discarded.

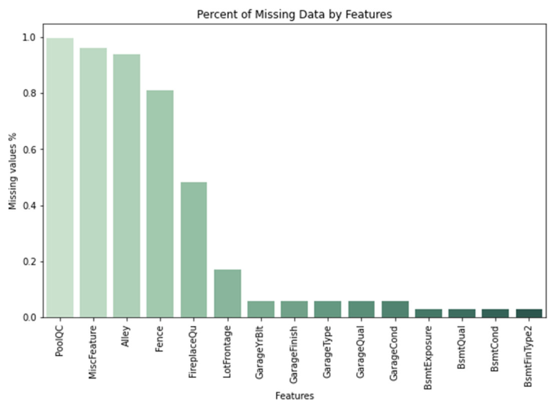

Fig 9. Percent of Missing Data by Features.

Table 1. Title

λ	0.0001	0.001	0.01	0.05	0.1	0.3	1
Accuracy (%)	90.01	90.15	89.74	85.14	83.29	83.37	81.87
RMSE	0.013	0.012	0.013	0.019	0.021	0.021	0.023

Eventually, the three regression models can be implemented. For the linear regression, the accuracy of the prediction reaches 89.96%, and the root of the mean squared error is 0.0125. The result of linear regression is excellent, the accuracy can reach nearly 90% which is very accurate among all the machine learning techniques. Regarding to the LASSO regression, the situation will be different. The λ value needs to be decided first, there are 7 different values in total chosen for the LASSO regression model, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.3, 1. Table. 1 summarizes the corresponding accuracy and root of mean squared error. According to the results, when the model is the most accurate, the λ value is 0.001 and the root of the mean squared error is the smallest. When the model is the least accurate, the λ value is 1 and the root of mean squared error is the largest. The accuracy of ridge regression is 89.98%, very close to 90%. The root of mean squared error is 0.0125. All the regression models have been implemented. The LASSO regression can produce the most accurate result, however, require careful selection for λ value, otherwise, the LASSO regression can't predict the correct result. The ridge regression and linear regression yield similar results.

4. Limitations & Prospects

The models chosen for this research only include regression models, which are very basic solutions for machine learning problems. In fact, there are many distinct solutions for house price prediction. For example, the previously mentioned neural network. It contains many advanced types of solutions (e.g., artificial neural networks), which have high performance and predict accurate results. In the meantime, the dataset selected for the research is too small compared to real-world data. The dataset contains 1259 sold real estate with 81 features, which limited the performance of the different models since the dataset is relatively small and simple. In the previous section, the 3 models have very diminutive differences. The accuracy of the 3 models is around 90% except for the LASSO regression model with an inappropriate λ value. In addition, the way to deal with the missing value is too simple, most of the less relevant features are simply erased without a complex method to fill in the missing values. The metric of assessing the performance of the models is limited as well. It only contains the root of mean squared errors and the in-library judgment tool. The removal of outliers is not as good as expected, as there are still very visible outliers existing in the graph. This research is only based on one question to find the best model for prediction in the supervised learning task, thus there is no guarantee that the best model found in real estate price problems can also be the best for other problems.

Since there are many limitations existed, many improvements can be done in the future. More topics can be included to prove the effectiveness of different models. For example, future research can include image recognition and cancer prediction. More models can be selected to solve machine learning problems including but not limited to the neural network, regression, data mining, and inference system. Future work can also choose multiple larger and more complex real-world datasets. One possible way is to select several different cities, counties, and even countries' real estate selling data from different time intervals and may include some significant incidents (e.g., economic crises), which allow the models to react to the accidental factors. Therefore, the performance of different models can be visualized more clearly. Simultaneously, there are more solid data that prove the accuracy of models in different scenarios. When introducing more complex models and datasets, more plots will be presented to demonstrate the relationships of variables and principles of each model to give a better understanding of the structure of the datasets and models. For more accurate prediction, the less relevant variables will also be included in the future. The missing values in the dataset won't be simply discarded, for the numeric variables, mean or median values will be inserted according to the dataset description. For the categorical variables, the logical method will be determined by the description since the result can be sensitive to some categorical variables. When evaluating the performance of different models, more exhaustive and advanced metrics will be implemented to comprehensively judge the effectiveness of models.

5. Conclusion

In summary, the most accurate model to predict house price is the LASSO regression model with a λ value equal to 0.001. It reaches 90.15% of accuracy and 0.0123, the smallest error. The accuracy of linear regression and ridge regression is nearly 90%. This number indicates that linear regression and ridge regression are excellent choices for real estate price prediction as well. However, there are many problems as well. The dataset is too small and simple. The number of chosen models is only three. Real estate price prediction is a basic problem in the machine learning field, no further complex problems are covered. These flaws also limit the visibility of differences in models. In addition, complex metrics can be used for evaluating models in the future. To improve further research, more complex datasets can be used and more topics in machine learning can be included. Apparently, more advanced learning models will be included as well to expand the scope of the study. Nevertheless, this paper focuses solely on real estate price prediction, LASSO regression model is the most accurate one but needs to manually determine the λ value. The wrong value will lower the accuracy severely. Linear and ridge regression are suitable for house price prediction problems if one wants to avoid manual parts in code. In the future, more methods and models will be implemented. Overall, these results offer a guideline for future performance measurement of different machine learning models.

References

[1] Varma, A., Sarma, A., Doshi, S. et al.: House Price Prediction Using Machine Learning and Neural Networks. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 1936-1939 (2018).

[2] Ansell, B.: The Political Economy of Ownership: Housing Markets and the Welfare State. American Political Science Review, 108(2), 383-402 (2014).

[3] Maclennan, D.: Some Thoughts on the Nature and Purpose of House Price Studies. Urban Studies, 14(1), 59–71 (1977).

[4] Yates, A.: Real Estate and Global Urban History (Elements in Global Urban History). Cambridge: Cambridge University Press, 2021.

[5] Li, L., Chu, K. H.: Prediction of real estate price variation based on economic parameters. 2017 International Conference on Applied System Innovation (ICASI), 87-90 (2017).

[6] Sarip, A. G., Hafez, M. B., Daud, M. N.: Application of Fuzzy Regression Model for Real Estate Price Prediction. Malaysian Journal of Computer Science, 29(1), 15–27 (2016).

[7] Li, D. Y., Xu, W., Zhao, H., and Chen, R. Q.: A SVR based forecasting approach for real estate price prediction. 2009 International Conference on Machine Learning and Cybernetics, 970-974, (2009).

[8] Hu, X., Zhong, M.: Applied research on real estate price prediction by the neural network. 2010 The 2nd Conference on Environmental Science and Information Application Technology, 384-386 (2010).

[9] Henry, C., Paul, D., Theo, D., Stephen A. J.: A spatio-temporal, Gaussian process regression, real-estate price predictor. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPACIAL '16). Association for Computing Machinery, New York, NY, USA, Article 68, 1–4 (2016).

[10] Dey, S. K., Urolagin, S.: Real Estate Price Prediction using Data Mining Techniques. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 1-4 (2021).