The real estate price prediction of US prediction based on multi-factorial linear regression models

Xiaoyu Xu*

School of Mathematics, University of Bristol, Bristol, UK

*Corresponding author: mf19117@bristol.ac.uk

Abstract. In every period, housing price prediction has always been a fascinating topic. Fluctuations in housing price are not only relevant to each individual resident but also to the politics and economy of the country. This essence of this research project is the usage of some real influencing factors to predict housing prices. In the Ames Housing dataset from Kaggle.com, five real factors that have a relatively strong correlation with housing prices are the overall material and finish quality, the above ground living area, the size of garage in car capacity, the garage area, and the total basement area. Based on these five real factors, two multiple linear regression models are constructed for predicting residential prices in Ames, Lowa, US. According to the analysis, when two independent variables are closely related, removing one of them does not necessarily reduce the fit of the model significantly, even if both independent variables are closely related to housing price. Therefore, choosing more appropriate variables is very important to increase the fit of the model. These results shed light on guiding further exploration of using more significant variables to find more accurate models to fit actual housing prices.

Keywords: Multiple Linear Regression; Housing price prediction; Variable Selection.

1. Introduction

Housing is an essential provision of life and it is probably one of the crucial investments that many people will make in their lifetime. Hence, people are more expected to spend their wealth on their dream house at the most opportune time by observing the fluctuation in housing prices. Furthermore, housing price is also associated with the national economic and social stability [1]. For instance, the sharp decline in US house prices in 2008 dampened the US and the rest of the world economy and led to a deterioration in the world economic outlook [2]. In addition, the COVID-19, which emerged in 2019, has had a considerable impact on housing prices. In the research of Bhat et al., they found that an increase in new COVID-19 cases led to an increase in housing prices based on the data on home value in major Texas cities [3]. Therefore, understanding the current situation of housing prices in the country has benefits for the formulation of relevant national policies as well as the financial management of ordinary families.

In many countries, House Price Index (HPI) is commonly used to calculate house price inflation for residential properties [4, 5]. One of the basic methods of constructing HPI is to refer to the median price for each period [6-8]. This type of index is easy to construct, but it has the disadvantage of having little or no control over quality [9, 10]. However, many factors in real life can lead to changes in housing prices, e.g., city, location, living area, the overall material and finish quality, the size of garage, number of bathrooms, and the original construction date. These factors can be broadly classified into three categories, namely house information, built environment, and surroundings of the community. Researchers from a variety of disciplines have long studied topics related to housing prices in order to understand the impact of property values in various social environments [11 -13].

This article will use multiple linear regression models to select the five factors in the Ames Housing dataset that are most strongly associated with housing prices for prediction. The rest part of the paper is organized as follows. The Sec. II will introduce the process of variable selection and the fundamental structure of multiple linear models. The Sec. III will compare the fitting degree of the two distinct multiple linear models. Eventually, a brief summary will be given in Sec. IV.

2. Methodology

2.1 Data

The Ames Housing dataset used in this project is compiled by Cock and published on Kaggle.com. There are 79 explanatory variables describing almost every aspect of housing in Ames, Lowa. For example, the style of dwelling, the electrical system and the original construction date. There are two categories contained in the Ames Housing dataset. The training set consists of 1,460 housing records with 81 columns of data, the first of which is ID, and the last is SalePrice, which is the target value to predict. In the testing set, there are 1,459 housing records with 80 columns of data, and there are a lot of missing values in the testing set. Hence, using Python to remove columns containing missing values is more applicable for this research. Correlation analysis of the remaining variables reveal that the five variables most strongly correlated with SalePrice (with correlation coefficients higher than 0.6) are the overall material and finish quality (OverallQual), the above grade (ground) living area square feet (GrLiveArea), the size of garage in car capacity (GarageCars), the size of garage in square feet (GarageArea), and the total square feet of basement area (TotalBsmtSF). The overall material and finish quality for each house ranges from one to ten, with one being very poor and ten being very excellent. Therefore, using these five variables to predict SalePrice is a practical approach to predict hosing prices. From Figure 1, OverallQual, GrLiveArea, and SalePrice are the most strongly correlated, followed by GarageCars and GarageAreaTotalBsmtSF has the weakest correlation with SalePrice of the five variables.

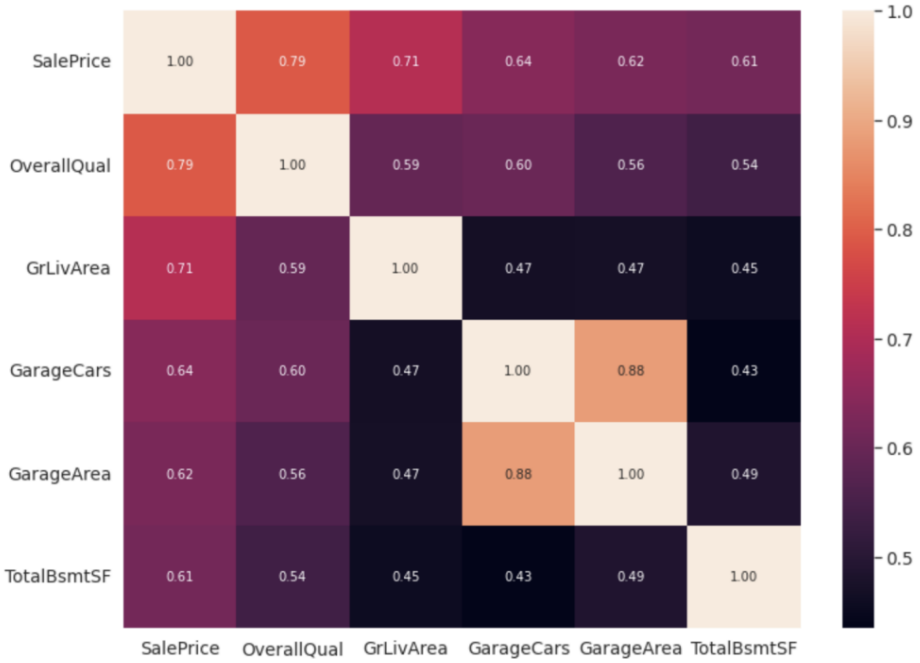

Figure 1. Correlation Analysis between 5 variables and SalePrice.

Table I. Least squares coefficient estimates of the multiple Linear Regression of SalePrice on OverallQual, GrLivArea, GarageCars, GarageArea and TotalBsmtSF.

	Coefficient	Std.error	t-statistic	p-value
Intercept	-99072.05	4638.45	-21.359	<2E-16
OverallQual	23635.007	1072.532	22.037	<2E-16
GrLivArea	45.346	2.489	18.218	<2E-16
GarageCars	14544.315	3022.681	4.812	1.65E-06
GarageArea	17.133	10.468	1.637	0.102
TotalBsmtSF	31.501	2.904	10.848	<2E-16

2.2 Multiple linear regression

Linear regression is an important statistical learning method, which can be used in a wide range of quantitative disciplines and this model is also the basis for a larger number of more complex models. Multiple linear regression is a further enhancement of simple linear regression and is also a straightforward approach for predicting a quantitative response Y based on predictors X_i. It assumes that there is approximately a linear relationship between X_i and Y. Mathematically, the model is

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_k X_k + \epsilon \quad (1)$$

Where β_0 is a constant indicating intercept and β_i are the regression coefficients for X_i. β_i quantifies the association between the predictor X_i and the response variable Y. ε stands for the error term which is independent of X_i typically. Estimating the regression coefficients, one derives the formula:

$$\hat{Y} = \widehat{\beta_0} + \sum_{i=1}^{k} \widehat{\beta_k} X_k \quad (2)$$

In the end, the values of the root mean squared error (RMSE) and R^2 statistic will be used to test the fit of the model.

3. Results & Discussion

In this housing price prediction model, the predictors x_1 to x_5 are OverallQual, GrLivArea, GarageCars, GarageArea and TotalBsmtSF and $\widehat{\beta}_1$ to $\widehat{\beta}_5$ are the regression coefficients for x_i. Table I gives the model coefficient values for various attributes, where the intercept $\widehat{\beta}_0$ is -99072.050, the coefficient on the OverallQual variable is $\widehat{\beta}_1 = 23635.007$, which means that one additional grade point of the overall material and finish quality is associated with 23635.007 dollars higher in SalePrice. The coefficient on GrLivArea variable is $\widehat{\beta}_2 = 45.346$. This means that on average, for every additional square foot of above grade living area, the SalePrice increases by 45.346 dollars. For GarageArea, each additional vehicle accommodated may result in an increase in the SalePrice of 14544.315 dollars, as the coefficient on GarageArea variable is $\widehat{\beta}_3 = 14544.315$. Since the coefficients on GarageArea and TotalBsmtSF variables are $\widehat{\beta}_4 = 17.133$ and $\widehat{\beta}_5 = 31.501$, each additional square foot in their area increases the SalePrice by 17.133 dollars and 31.501 dollars respectively.

Table II. Predicted SalePrice values.

ID	Actual SalePrice	Predicted SalePrice	Difference
6	143000	143251.503	251.503
28	306000	277814.915	-28185.085
45	141000	127161.25	-13838.75
211	98000	85498.793	-12501.207
477	380000	356837.391	-23162.609
516	402861	351120.591	-51740.409
666	230500	259811.93	29311.93

Table III. Predicted SalePrice values.

Quantity	Value
Residual Standard error	38900
Adjusted R	0.7603
F-statistic	926.5
RMSE	38815.45

Table II lists the predicted price using the multiple linear regression model, the fourth column is the result of subtracting the real value from the predicted vale. Prices of ID numbers 6 is predicted more correctly, however, there are also relatively large differences for prices of ID numbers 28, 45, 211, 477, 516, and 666.

The R^2 in the Table III is the square of the correction between the response and the fitted linear model. The R^2 is 0.7603 is close to 1, hence, this multiple linear explains a large portion of the variance in the response variable.

3.1 Variable Selection

From Table I, GarageArea has the largest p-value compared to the other predictors, while the other variables have extremely very small p-values. Hence, GarageArea may be less statistically significant. Moreover, the Figure1 illustrates that the correlation coefficient between GarageArea and GarageCars is 0.88, which indicates a strong correlation between them. To test whether the addition of GarageArea factor would improve the fit of the model, a second multiple linear regression model will use only the other four factors for prediction.

3.2 Second Multiple Linear Regression Model

Removing GarageArea from the predictors list, the newly updated predictors list will include OverallQual, GrLivArea, GarageCars and TotalBsmtSF.

Table IV and Table V present the information for the multiple regression on OverallQual, GrLivArea, GarageCars and TotalBsmtSF. According to the results, the model that uses all five features to predict SalePrice has an R^2 of 0.7603 and the model that uses the features excluding GarageArea to predict SalePrice has an R^2 value of 0.76. The Addition of the GarageArea variable to the model containing the other four predictors resulted in only a tiny increase in R^2. Comparing their values of RMSE, the first multiple regression model fits the real data better, but the addition of the GarageArea variable does not improve the model significantly.

A Comparison of Table II and Table VI shows that the difference values with IDs 28, 211 and 477 are reduced by removing the GarageArea factor from the model. In other words, as the factor of GarageCars is already included in the independent variable, adding GarageArea as an independent variable would only slightly improve the fit of model.

Table IV. Least squares coefficient estimates of the second multiple Linear Regression of SALEPRICE ON OverallQual, GrLivArea, GarageCars and TotalBsmtSF.

	Coefficient	Std.error	t-statistic	p-value
Intercept	-99248.853	4639.866	-21.39	<2E-16
OverallQual	23572.236	1072.465	21.98	<2E-16
GrLivArea	45.643	2.484	18.38	<2E-16
GarageCars	18582.209	1747.412	10.63	<2E-16
TotalBsmtSF	32.52	2.838	11.46	<2E-16

Table V. More information about the least squares model for the second multiple regressions.

Quantity	Value
Residual Standard error	38920
Adjusted R	0.76
F-statistic	1156
RMSE	38851.19

Table VI. Predicted SalePrice valus using the second model.

ID	Actual SalePrice	Predicted SalePrice	Difference
6	143000	143828.431	828.431
28	306000	278265.414	-27734.586
45	141000	127081.986	-13918.014
211	98000	86145.159	-11854.841
477	208900	195558.54	-13341.46
516	402861	349654.114	-53206.886
666	230500	260581.421	30081.421

3.3 Limitation

It should be noted that this paper has some defects and limitations. Specifically, these test IDs are selected at random and are not representative of the overall data. Further, in order to further improve the accuracy and credibility of the prediction, some features that are less correlated with SalePrice need to be included in the consideration of the independent variables. Nevertheless, the inclusion of factors that are not correlated with house prices may lead to a decrease in the fit of the model due to over-fitting. Moreover, some features are directly related to each other, i.e., their interaction terms are also need to be test at this point. Therefore, the process of variable selection is constructive to improve the accuracy of predictions. Having more data is also likely to result in better prediction.

4. Conclusion

In summary, this article develops multiple factorial linear regression models to predict housing price. By comparing the information from the two multiple linear regression models, it was found that their R^2 values were approximately equal and that the RMSE value of the model without the GarageArea factor only increased by 35.74. Therefore, adding features that are more correlated with housing price as independent variables to predict SalePrice does not necessary improve the fit of the model significantly. On this basis, determining the important variables is very significant to the statistical models. In the future, dataset could cover a wider range of information and advanced regression techniques could be used to build housing price prediction models. Overall, these results offer a guideline for real estate price prediction using multi-factorial linear regression models.

References

[1] M. Chen, W. Liu, and D. Lu. "Challenges and the way forward in China’s new-type urbanization." Land use policy, vol. 55, 2016, pp. 334-339.

[2] G. P. Kouretas, and A. P. Papadopoulos, eds. Macroeconomic analysis and international finance. Emerald Group Publishing, vol. 1, 2014, pp. 79-102.

[3] M. R. Bhat, J. Jiao, and A. Azimian, "The impact of COVID-19 on home value in major Texas cities," International Journal of Housing Markets and Analysis, 2021.

[4] A. B. Adetunji, et al. "House Price Prediction using Random Forest Machine Learning Technique." Procedia Computer Science vol. 199, 2022, pp. 806-813.

[5] X. Wang, K. Li, and J. Wu. "House price index based on online listing information: the case of China." Journal of Housing Economics vol. 50, 2020, 101715.

[6] T. M. Crone, and R. P. Voith. "Estimating house price appreciation: a comparison of methods." Journal of Housing Economics vol. 2.4, 1992, pp. 324-338.

[7] D. H. Gatzlaff, and D. C. Ling. "Measuring changes in local house prices: an empirical investigation of alternative methodologies." Journal of Urban Economics vol. 35.2, 1994, pp. 221-244.

[8] F. T. Wang, and P. M. Zorn. "Estimating house price growth with repeat sales data: what's the aim of the game?" Journal of Housing Economics vol. 6.2, 1997, pp. 93-118.

[9] S. C. Bourassa, M. Hoesli, and J. Sun. "A simple alternative house price index method." Journal of Housing Economics vol. 15.1, 2006, pp. 80-97.

[10] K. E. Case, and R. J. Shiller. "Prices of single-family homes since 1970: New indexes for four cities.", 1987.

[11] A. R. Archer, D. H. Gatzlaff, and D. C. Ling. "Measuring the importance of location in house price appreciation." Journal of Urban Economics vol. 40.3, 1996, pp. 334-353.

[12] K. Cao, M. Diao, and B. Wu. "A big data–based geographically weighted regression model for public housing prices: A case study in Singapore." Annals of the American Association of Geographers vol. 109.1 2019, pp. 173-186.

[13] Y. Kang, et al. "Understanding house price appreciation using multi-source big geo-data and machine learning." Land Use Policy vol. 111 2021, 104919.