# Comparative Analysis of Machine Learning Performance in House Price Prediction for Ames Iowa

Xingchen Zhou
Miami University, Oxford, OH, 45056, USA
zhoux45@miamioh.edu

**Abstract.** The price of housing is closely related to the quality of people's lives, which is reflected in various aspects (e.g., finding a place to live and managing their finances). On this basis, being able to predict house prices would be a great convenience, and this study examines the accuracy of multiple mechanistic learning regression models in predicting future house price changes. To be specific, this paper downloaded the dataset from Kaggle which includes the training and testing csv files. and ran many different regression techniques including lasso and more models on google Colab. Furthermore, a comparison between the results of different methods is carried out by estimating the mean square error. Then, this research combined some of the best yielded results of the model to combine into a final model and got a result with a high score. According to the analysis and evaluations, the study concluded that the hybrid model resulted in a more accurate prediction than the single model. These results shed light on guiding further exploration of real estate prediction.

**Keywords:** Data Processing; Machine Learning; House Price Prediction; Regression.

## 1. Introduction

In real life, housing prices are closely related to human beings lives. By the anatomy of subject well-being, residential satisfaction is one of six essential correlates of people's measure of supervisory well-being [1-3]. Many people are busy for half their lives finding a place to live, whether renting or buying a home. At the same time, housing transactions have created a booming industrial environment. Many people enter the real estate industry to work. Furthermore, during the golden age of the U.S. real estate boom, many American consumers put their saved values into real estate purchases to get a return on their investment. Ultimately these investors lost their money in the subprime mortgage crisis of 2007-2008 [4,5]. Fluctuations in house prices have an impact on all aspects of life, work, and investment, so for these sellers, buyers, and speculators, it would be very convenient for them to be able to predict house prices.

In a previous study, Dr. Thamarai and Dr. Malarvizhi used decision tree regression and multiple linear regression to predict house prices for small towns in West Godavari district of Andhra Pradesh and concluded that multiple linear regression had better performance [6]. In addition, by estimating house prices in Arlington County, Virginia through random forest, the random forest model captures the nonlinear hidden relationship between house prices and house locations, which would provide better results relative to the baseline linear regression [7]. In a research paper by Vineeth, Ayyappa and Bharathi, three different models were used to predict house prices in king county of USA, namely simple linear regression, multiple linear regression and neural networks, and it was concluded that neural networks would produce better prediction results [8]. In the paper by Ja'afar, Mohamad and Ismail, the study summarizes the previous mechanical learning models and provides five algorithms including Support Vector Machine (SVM) [9]. In a study by Kang, et al, they introduced multi-source big geo-data in a machine learning framework and concluded that properties with low house prices and small square footage have higher appreciation potential in Boston [10].

# 2. Data & Method

## 2.1 Data

The dataset is from the Kaggle website of Housing Price-Advanced Regression Technique [11]. This dataset contains a large number of attributes that could comprehensively include the overall features of a single house. In this dataset, the explanatory variables cover almost any aspect of a house that can be involved in Ames, Iowa. This covers both continuous and discrete variables, for a total of 79 features across 1420 properties, including SalePrice, MSSubClass, MSZoning, etc.

## 2.2 Model

In the whole project, this study selects google colab as the IDE to write the code [12]. In this paper, the Root-mean-square error (RMSE) is used as the scoring criteria for each model. In this model it is the train and test file. When these deviations are calculated from the sample of data itself used to calculate the estimate, the difference is usually called residual. When the difference is not based on the estimate derived from the sample, it is usually called error or prediction errors. Root means square error (RMSE) is a measure of accuracy and is used to compare the prediction errors of different models for a given data set, but not between data sets because it is scale dependent. In general, the lower the RMSE the better, where the mathematical description is given as follows:

$$RMSE = \sqrt{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2 / N} \quad (1)$$

## 2.3 Model Selection

In the model selection part, this research imported some regression models from scikit-learn, and this study will use the following models as the basic models for prediction. The Lasso regression is the traditional linear regression objective function followed by an L1-parametric, which makes the coefficient of unimportant features compressed to 0 to achieve a more accurate parameter. However, this model is very sensitive to outliers and there may still be some certain outliers in the data set. Hence, one uses RobustScaler to normalize the data set. This paper also used two other models such as elastic network and gradient boosting regression. Extreme Gradient Boosting (XGBoost) is an improved algorithm based on GBRT, which improves the computational efficiency of the algorithm by introducing model complexity to reconstruct the objective function and minimizing it in the process of iterating each tree, while minimizing the loss function of the model LightGBM is an optimization improvement based on the GBRT algorithm.

After scoring the basic model, research wanted to get a better result, so the average model and the stacked model are designed. For the four models from scikit-learn, calculate the average of their scores to get a more accurate result. Stacked Models are applied to do the predictions of all base models on the test data and use the average predictions as meta-features to be done by the meta-model and get the final predictions. At the least, this study fuse XGBoost, LightGBM and StackedRegressor and AverageRegressor together in an average manner to build the final prediction model. For each model takes 25% of the weight

## 2.4 Model Parameters

In the lasso model the alpha is set as 0.0005 and the random state is set the number of 1. For the ENet model the alpha value is 0.0005, the l1_ratio is .9 and the random state is set 3. For the kernel Ridge the alpha is 0.6, the kernel is polynomial, the degree is 2 and the coef0 is 2.5. For GradientBoostingRegressor model the n_estimators are 3000, learning rate is 0.05, max depth is 4, max feature set as sqrt, min samples leaf is 15, min samples split is 10, the loss is 'Huber' and the random state is 5. For xgb.XGBRegressor mode the colsample bytree is 0.4603, the gamma value is 0.0468, the learning rate is 0.05, the max depth is 3, the min child weight is 1.7817, the n_estimators is 2200, the reg alpha is 0.4640, the rag lambda is 0.8571, the subsample is 0.5213, the silent value is 1 and the nth read is -1. For lgb.LGBMRegressor model the objective is set 'regression', the number

of leaves is 5, the learning rate is 0.05, the n_estimators is 720, the max bin is 55, the bagging fraction is 0.8, the bagging freq is 5, the feature fraction is 0.2, the feature fraction seed is 9, the bagging seed is 9, the min data in leaf is 6, the min sum hessian in leaf is 11 and the verbose is -1.

## 3. Results & Discussion

### 3.1 Feature Engineering

After importing datasets and libraries, a heatmap is plotted to show relationships between numerical variables as shown in Fig. 1. It does not give closely related variables directly. A filtration of variables is generated and 9 mostly related features based on the correlation against sale price are chosen. Fig. 2 shows the result after filtration and arrangement based on correlation coefficient with feature Sale Price, which are sale price, ground living area, garage cars, garage area, etc.

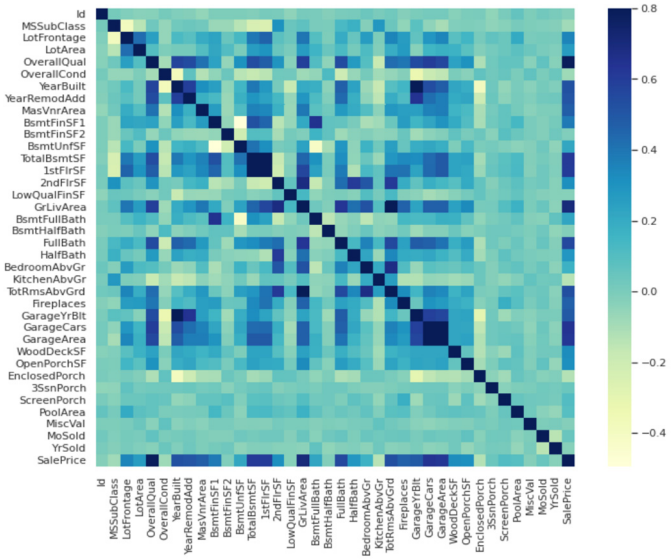



**Fig 1.** An overview result of heatmap

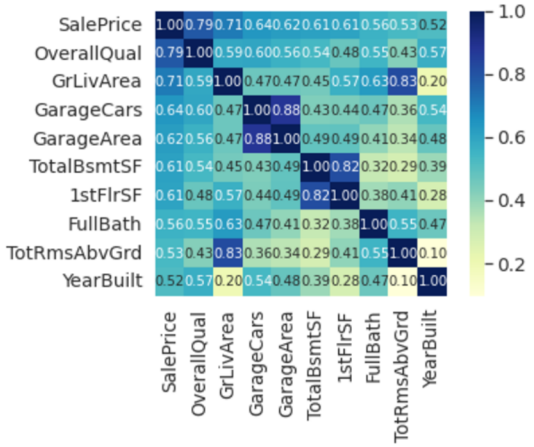



**Fig 2.** A heatmap after filtration of variables

### 3.2 Elimination of Outliers

In elimination of outliers of the 9 features, graphs of all variables against sale price are produced and outliers which obviously affect the regression are removed. Taking ground living area (named as GrLivArea in Fig. 3 and Fig. 4) as an example, Fig. 3 illustrates the distribution of data before elimination of outliers, and Fig. 4 presents the distribution of data after elimination of outliers. Comparing Fig. 3 and Fig. 4, two data points on the right lower side on Fig. 3 are regarded as outliers and removed.




**Fig 3.** Datasets of GrLivArea feature before elimination of outliers.

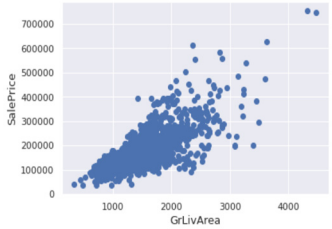



**Fig 4.** Datasets of GrLivArea feature after elimination of outliers

### 3.3 Box-cox Transformation of Data

In this step, transformation of a dependent feature to a normal shape using box-cox transformation is produced. Taking the example of the feature sale price (named as SalePrice) in Fig. 5 and Fig. 6, distribution and probability plot of the feature is first generated, the distribution plot shows a non-normal distribution of the feature. Fig. 6 shows distribution of sale price features follows normal distribution after box-cox transformation.

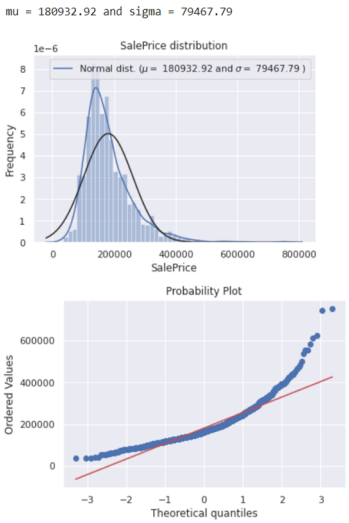



**Fig 5.** Distribution and probability plot of SalePrice feature before box-cox transformation

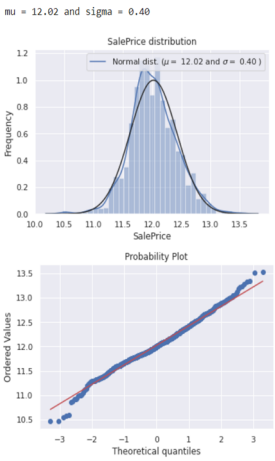



**Fig 6.** Distribution and probability plot of SalePrice feature after box-cox transformation

### 3.4 Search are Filling Missing Data

In this part, missing data of all features are searched and filled with corresponding default values. Fig. 7 illustrates the percentage of missing value of all features. These missing data are filled by two criteria based on data type, which is 0 for numerical values and NAN for strings.

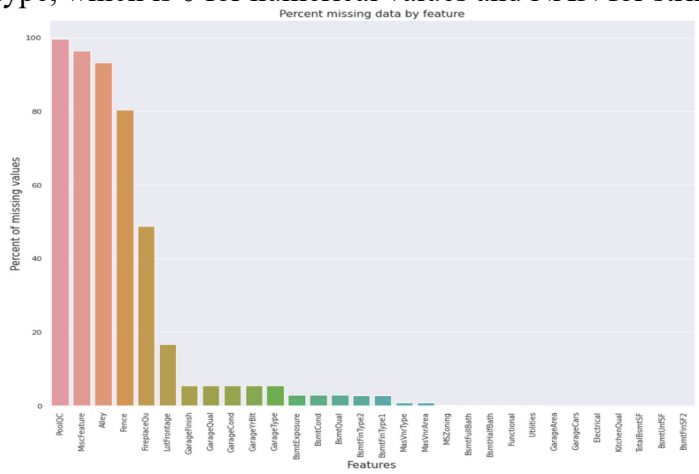



**Fig 7.** Percentage of missing data for all features

### 3.5 Evaluations of Different Models

By using these base models to run the already processed housing information data, one can see that the elastic network has the Highest score as summarized in Table. 1. By calculating the average of lasso, Kernel Ridge, ElasticNet and Gradient Boosting, the score of the average model is 0.1089. As for stacking averaged model, this study chose Kernel Ridge, ElasticNet and Gradient Boosting as the base models and Lasso as the meta model in the case of this study. The final prediction score is 0.1083 which is better than the single base model. After a series of calculations, the result is 0.074. Then, by trying for each model, one came up with a relatively high-quality score. Compared with the highest score of 0.1119 obtained by a single model, the composite model can obtain a higher score of 0.0748. In summary, this study processed the data by censoring the discrete values, correcting the normal distribution, and filling in the missing values, and scored the model by RMSE to finally obtain a house price prediction result that is relatively close to the test level.

**Table 1.** The Scores of different models.

| Model | Score |
|---|---|
| ElasticNet | 0.1119 |
| Lasso | 0.1120 |
| Kernel Ridge | 0.1145 |
| Xgboost | 0.1158 |
| Lgboost | 0.1166 |
| Gradient Boosting | 0.1167 |

## 4. Conclusion

In summary, this paper investigates the application of mechanical learning models for predicting housing prices in Ames, Iowa. Specifically, Lasso, elastic Net and ridge regression models were used to predict house prices in Ames, Iowa and the results were scored using RMSE. According to the analysis, elastic Net has the best results among the individual models. In addition, the composite model yielded a relatively higher score than the single regression model. However, this study has some drawbacks and shortcomings. To be specific, in terms of data the competition data was used as the dataset for training. In addition, the final model used in this study could perhaps be further improved based on other algorithms. In the future, more housing information data and other models

could be used for training to obtain a more accurate result. Overall, these results provide a guide for future mechanical learning training based on house price prediction.

## References

[1] Van Praag, B. M., Frijters, P., Ferreri, C. A.: The anatomy of subjective well-being. Journal of Economic Behavior & Organization, 51(1), 29-49 (2003).

[2] Andrews, F. M., Inglehart, R. F.: The structure of subjective well-being in nine western societies. Social indicators research, 6(1), 73-90 (1979).

[3] Weidemann, S., Anderson, J. R.: A conceptual framework for residential satisfaction. In Home environments (pp. 153-182). Springer, Boston, MA 1985.

[4] Barberis, N.: Psychology and the Financial Crisis of 2007-2008. Financial innovation: too much or too little, 15-28 (2013).

[5] Reinhart, C. M., Rogoff, K. S.: Is the 2007 US sub-prime financial crisis so different? An international historical comparison. American Economic Review, 98(2), 339-44 (2008).

[6] Thamarai, M., Malarvizhi, S. P. House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering & Electronic Business, 12(2), (2020).

[7] Wang, C., Wu, H.: A new machine learning approach to house price estimation. New Trends in Mathematical Sciences, 6(4), 165-171 (2018).

[8] Vineeth, N., Ayyappa, M., Bharathi, B.: House price prediction using machine learning algorithms. In International Conference on Soft Computing Systems (pp. 425-433). Springer, Singapore (2018, April).

[9] Ja'afar, N. S., Mohamad, J., Ismail, S.: Machine learning for property price prediction and price valuation: a systematic literature review. Planning Malaysia, 19 (2021).

[10] Kang, Y., Zhang, F., Peng, W., et al.: Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, 111, 104919 (2021).

[11] Kaggle datasets for house prices. Retrieved from: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

[12] Colab official website. Retrieved from: https://colab.research.google.com/.