# Analysis of Various Models for House Price Prediction based on Machine Learning

Yuhan Wu

Ohio State University, Columbus OH, 43212, USA

wu.3966@osu.edu

**Abstract.** The real estate prices prediction is a relatively popular data analysis problem. With the gradual increase in the competitiveness of the housing trading platform in recent years, and the basic model of house prices predicted is different from the factual transaction price. Therefore, it is necessary to design more accurate house prices prediction models. This article will adopt different linear regression models and machine learning models. By training these models of training Set, through the accuracy algorithm of Kaggle, the accuracy of each model is obtained. According to the analysis, this article finally discovered that the accuracy of the new model is improved than the accuracy of the basic model (i.e., new model predicts house price more precision), but it can be improved in the case of less data in the real environment. The reason may be related to economic and public opinion effects. Overall, these results shed light on guiding further exploration of house price.

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

## 1. Introduction

With the diversification of housing demand and the development of the Internet, real estate online trading platform development is faster, and the user income in the platform has gradually increased [1]. Under the influence of epidemic viruses in recent years, the competitiveness of online housing trading platforms has increased significantly. Therefore, it is necessary to meet the needs of users to maintain the competitive advantage.

The more accurate predicting the price of houses is an important part. There are many studies of house price prediction models so far [2], but there are many basic models that are still improved under certain circumstancesfor instance, linear regression models [3], KNN based lasso and ridge regression models [4], support vector regression, and boosting algorithms such as Extreme Gradient Boost Regression [5], every model has its own advantages and disadvantages, need to be considered based on the objective of the case study [6]. This article will list three different models and two linear models of mechanical learning and try to integrate new models that can improve the accuracy of sales predictions through these models.

## 2. Data & Method

### 2.1 Data Gathering

The dataset is directly from a Kaggle competition: “House Prices - Advanced Regression Techniques” from url https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data, the dataset contains 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. The reason to use this data set is that it's not just the number of bedrooms or white fences that can influence price negotiations, but also minor factors, such as basement ceiling heights or the distance of the house near the east-west railroad.

### 2.2 Modeling

Scikit-Learn (Sklearn) is an important Python library that can provide machine learning algorithms. SCIKIT-Learn can provide the indispensable auxiliary functions of machine learning, such as classification, regression, and clustering [7]. Kernel ridge regression (KRR) is a combination of Ridge

regression and classification with kernel trick. Thus, it learns linear function and non-linear function in correspond space induced by the respective kernels. Kernel ridge regression is identical to support vector regression (SVR) [8]. Kernel ridge regression uses squared error loss while support vector regression uses insensitive loss. Kernel Ridge Regression guarantees it does not relate to Bayesian assumptions [8]. eXtreme Gradient Boosting is an effective and widely used machine learning algorithm based on tree boosting system. XGBoost is a type of scalable end-to-end tree boosting system used in data science [9]. Compared to tree boosting system, it develops a novel sparsity-aware algorithm for sparse data [10]. It was invented by developer Tianqi Chen and used as a tookit in Distributed Machine Learning Community. The algorithm can be used for regression with large and complex datasets. LightGBM is a novel gradient boosting cision tree algorithm which contains two techniques, Gradient-based One-Side Sampling and Exclusive Feature Bundling to deal with large tasks and data sets [11, 12].

## 3. Results & Discussion

After importing datasets and libraries, a heatmap is plotted to show relationships between numerical variables as shown in Fig. 1. Since 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 illustrates the result after filtration and arrangement based on correlation coefficient with feature SalePrice, which are sale price, ground living area, garage cars, garage area, etc.

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. Take ground living area (named as GrLivArea in Fig. 3) as an example, the left panel of Fig. 3 shows the distribution of data before elimination of outliers, and the right panel shows the distribution of data after elimination of outliers. Comparing the results, two data points on the right lower side on left panel are regarded as outliers and removed.

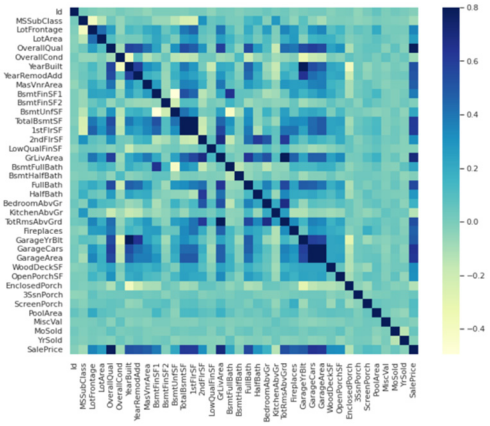


**Fig 1.** An overview result of heatmap

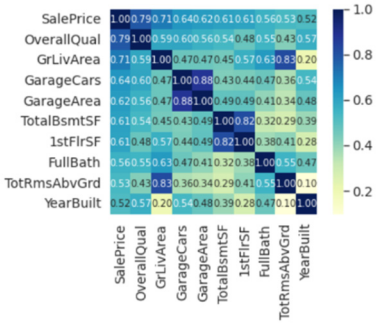



**Fig 2.** A heatmap after filtration of variables

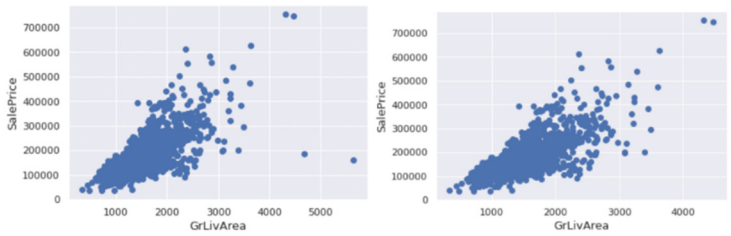



**Fig 3.** Datasets of GrLivArea feature before and after elimination of outliers.

In this step, transformation of a dependent feature to a normal shape using box-cox transformation is produced. Take the example of the feature sale price (named as SalePrice in Fig. 4, distribution and probability plot of the feature is firstly generated, the distribution plot shows a non-normal distribution of the feature. Fig. 4 depicts distribution of sale price feature follows normal distribution after box-cox transformation.

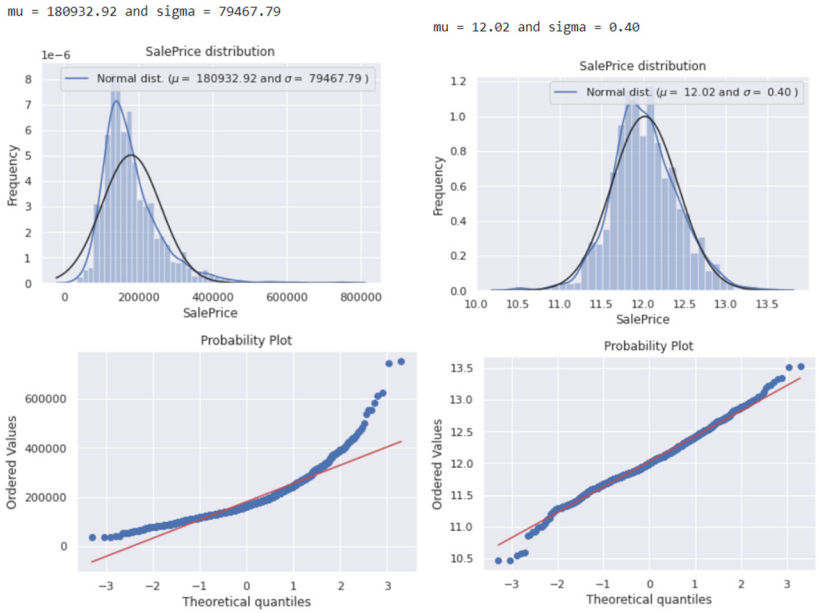



**Fig 4.** Distribution and probability plots of Saleprice feature before and after box-cox transformation

Missing data of all features are searched and filled with corresponding default values. Fig. 5 presents 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.

This paper uses root square errors as scoring standards for each model. Root mean square error (RMSE) is a measurement between the model prediction or estimation value (sample value or overall value) and the observed values. The root mean square error represents the square root of the second -order sample torque (sample standard deviation) between the predicted value and the observation value, or the square average of the difference. The root mean square error can summarize the amplitude of the error in each data point prediction into prediction capabilities. It is related to the proportion. It is generally used to compare the predictive error of different models of the given data set. Generally speaking, the lower the RMSE, the more accurate the prediction. The formula of RMSE is:

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

As shown in Table. 1, the result of the basic models shows that the accuracy of 5 models are arranged between 0.112 to 0.117. The stack model do the predictions of all base models on the test

data and use the average predictions as meta-features to done by the meta-model and get the final predictions. This study selects Kernel Ridge, XG Boost and Gradient Boosting as the base models and lasso as the meta model. The result shows its accuracy is better than the single base model.

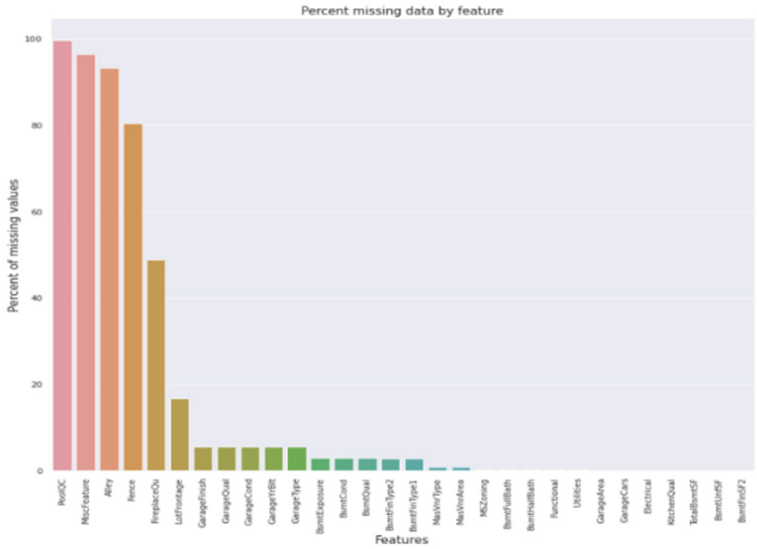



**Fig 5.** Percentage of missing data for all features

**Table 1.** RMSE of different models

| Model | RMSE |
|---|---|
| Lasso | 0.1120 |
| Kernel Ridge | 0.1145 |
| XG Boost | 0.1158 |
| LightGBM | 0.1166 |
| Gradient Boosting | 0.1167 |
| Stacked Models | 0.1028 |

## 4. Conclusion

In summary, this paper investigates accuracy of different linear, machine learning models and stacked model for house price prediction to see the improvence of the accuracy of stack model compare to basic models. Specifically, different basic models are selected from recent paper in house price prediction, accuracy of different models is calculated from root mean square error. According to the analysis, the accuracy of stacked model shows lower root mean square error, which indicates high accuracy compared to 5 basic models for predicting house price. Nevertheless, there are still spaces of improvence for prediction of house price, the stack model cannot predict house price in react to views prevalent among the general public. In the future, there may exist an algorithm shows possibilities of trend of house price considering different economical effect. Overall, these results offer a guideline for predicting house prices.

## References

[1] Varma, A., Sarma, A., Doshi, S., & Nair, R.: House price prediction using machine learning and neural networks. In 2018 second international conference on inventive communication and computational technologies (ICICCT) pp. 1936-1939 (2018, April).

[2] Bisong, E.: Introduction to Scikit-learn. Building machine learning and deep learning models on Google cloud platform. Apress, Berkeley, CA, 2019.

[3] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.

[4] Hujia Yu, and Jiafu Wu. "Real estate price prediction with regression and classification." CS229 (Machine Learning) Final Project Reports (2016).

[5] Manasa, J., Gupta, R., Narahari, N. S.: Machine Learning based Predicting House Prices using Regression Techniques," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 624-630 (2020).

[6] Ke, G., Meng, Q., Finley, T., et al.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30 (2017).

[7] Mohd, T., Jamil, N. S., Johari, N., Abdullah, L., Masrom, S.: An overview of real estate modelling techniques for house price prediction. Charting a Sustainable Future of ASEAN in Business and Social Sciences, 321-338 (2020).

[8] RH, N. S., Poojaa, R.: Price Prediction of House using KNN based Lasso and Ridge Model. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 1520-1527), (2022, April).

[9] N. N. Ghosalkar, Dhage, S. N..: Real Estate Value Prediction Using Linear Regression. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1-5 (2018).

[10] Thamarai, M., Malarvizhi, S. P.: House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering & Electronic Business 12.2 (2020).

[11] Vovk, V: Kernel ridge regression. Empirical inference. Springer, Berlin, Heidelberg, 105-116 (2013).

[12] Xiao, Y.: Hedonic housing price theory review. Urban morphology and housing market. Springer, Singapore, pp. 11-40 (2017).