Decision Support System Technical Report
Applying Regularized Linear Model in Predicting house price

NGUYEN Duc Thien VU Hoang Duc Hieu LE Duc Thang TRINH Ngoc Khang PHAM Van Hai[*]
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam

Abstract

This paper presents our approaching model to solve the task of *House Prices - Advanced Regression Techniques* which was public on (Kaggle). We implemented the . Our model used ... to . Experiments on the test show that our model achieves 0.12299 Root Mean Square Logarithmic Error (RMSLE) score.

1 Introduction

Nowadays, house resale is an important field in every country. A wide variety of factors affects the price of a house, ranging from the house's location, its features, as well as the property demand and supply in the real estate market. The housing market is also one essential component of every nations' economy. Therefore, forecasting housing values is not only beneficial for buyers, but also for real estate agents and economic professionals. Studies on housing market forecasting investigate the house values, growth trend, and its relationships with various factors.

The improvement of machine learning techniques and the proliferation of data or big data available have paved the way for real estate studies in recent years. There is a variety of research leveraging statistical learning methods to investigate the housing market. The goal of this paper is to analyse and design a system that helps users to predict house price would address a huge number of troubles for people who want to buy a house or even the one who sells it. Moreover, it could benefit the projection of future house prices and the policymaking process for the real estate market.

The rest of this paper is organized as follows. The next section, Section 2, provides an overview about others' work. Section 3 presents our algorithms and solutions. The experiments of the dataset and results are shown in Section 4. Finally, Section 5 concludes the paper and gives some perspectives for the work.

2 Related Works

Many previous studies have been done in different countries, applying a variety of models. Beracha et al. (2018) investigate the correlation between house price volatility, returns and local amenities, and proves that high amenity areas experience greater price volatility. Law (2017) finds that the strong links between Street-based local area with house price and it shows that using Street-based local is better than using region-based local area. For no fundamental factors, Ling et al. (2015) find that the sentiment of home buyers, home builders and lenders are related to real house price appreciation over the next two quarters during booms and busts. Lu et al. (2014) built a Geographically Weighted Regression (GWR) model to study London house prices.

Comparing the 79 variables provided in Kaggle competition, we indicate that the information is incomplete, and a few features are unnecessary for prediction, there are implicit conditions in house price. We identify those implicit characteristics among those features, applying normal distribution and transforming for increasing linearity. We use regression algorithms with parameters adjustment and consider the coupling effect of different algo-

[*]Corresponding author

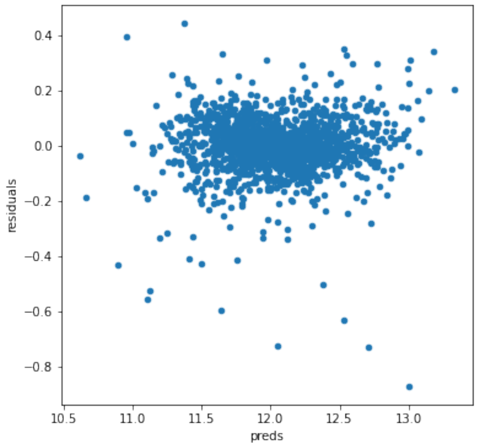

Figure 1: Data residual

rithms to achieve better test results.

3 Methodology

In this part, we will refer to the features engineering and describe how to apply multi-regression algorithms. There are many regression algorithms that can be used to build models and predict house prices. After investigation, we find that Ridge (Rifkin and Lippert, 2007), Lasso (Friedman et al., 2010) from Scikit-learn (Pedregosa et al., 2011) and Gradient boosting (Chen and Guestrin, 2016) are more useful. Finally, the efficiency and effectiveness of Lasso and Gradient will be shown in the final sub section.

3.1 Data pre-processing

We investigate the value distribution and correlation of SalePrice for each variable and introduce many new variables. For example, Figure [2] shows log transformation SalePrice distribution for each neighborhood. There are significantly different SalePrices among different neighborhoods. Details of feature engineering are listed in the following paragraphs.

These variables, which served as features of the dataset, were then used to predict the average price per square meter of each house.The next step was to investigate missing data. Variables with more than 50% missing data would be removed from the dataset. Any observations which had missing values were also removed from the dataset. Below are a few feature engineering processes which were done to cleanse the dataset.

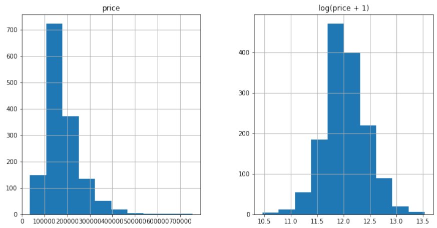

Figure 2: Log Sale Price

Log transformation, in order to approximate normal distribution, log transformation has been applied for SalePrice, LotArea and LotFrontage etc. The Shapiro-Wilk test for normality is depicted in Fig. 2, in the left side, the log transformation of SalePrice distribution, while in right side it is SalePrice distribution.

We transform the skewed numeric features by taking $log(feature + 1)$ - this will make the features more normal. Create Dummy variables for the categorical features. Replace the numeric missing values ($NaN's$[1]) with the mean of their respective columns. Finally, the residual of dataset was shown in Figure [1]

3.2 Ridge Linear Regression

Ridge regression there is one parameter alpha to choose for Ridge regression. Based on mean squared error scorer, we choose the alpha to get maximum score. Ridge regression give a penalty of L2 norm to the loss function, as follow:

$$L = ||Y - X(\theta)||_2^2 - \alpha(||\theta||_2^2) \tag{1}$$

When $\alpha = 0$, the penalty term has no effect, and the estimates produced by Ridge regression will be equal to Linear Regression. However, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. As can be seen, selecting a good value of α is critical. Cross validation comes in handy for

[1] Not an Number

this purpose. The coefficient estimates produced by this method are also known as the L2 norm.

3.3 Lasso Linear Regression

The Lasso model, as known as Least Absolute Shrinkage Selector Operator, is quite similar to ridge, but lets understand the difference them by implementing it in our big mart problem. Lasso model using L1 norm instead of L2 norm in the loss function, as follow:

$$L = ||Y - X(\theta)||_2^2 - \alpha(||\theta||_2) \tag{2}$$

This model is the same as the one above, but with a slight modification. Instead of penalty by L2 norm, this model uses the L1 one.

3.4 Gradient Boosting

XGBoost is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by Chen and Guestrin (2016). The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking. The package is made to be extendible, so that users are also allowed to define their own objectives easily.

4 Experiments

4.1 Evaluation

Kaggle evaluation standard is Root-Mean-Squared-Logarithmic-Error (RMSLE) between the logarithm of the predicted value and the logarithm of the observed sales price.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(log(y_i + 1) - log(\hat{y}_i + 1))^2} \tag{3}$$

Where y_i stands for $log(SalePrice_i)$ and $\hat{y}_i$ stands for $log(PredictSalePrice_i)$. The core part equates to $log(SalePricei/PredictSalePricei)$. Using prediction SalePrice ratio could avoid the heavy weight age for expensive houses. In other words, prediction of cheap house prices is more important here. Kaggle house prices competition Public Leaderboard only shows the score for half of test data, it could avoid overfitting by many times of tries. Since prediction results need to be verified via Kaggle website to get the test score, and there is a limitation of no of submission per day, only selected test results are submitted.

Method	RMSLE
Ridge	0.13042
Lasso	0.12455
XGBoost	0.13278
Hybrid	**0.12342**

Table 1: Comparasion of RMSLE.

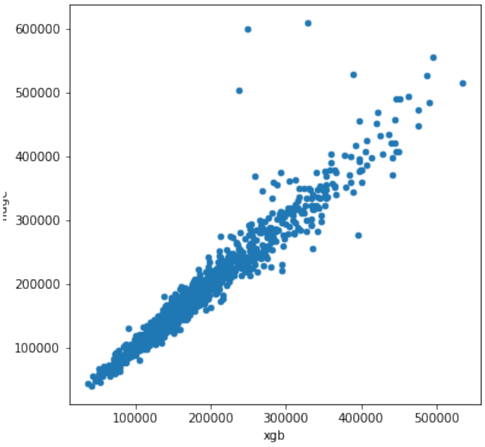

Figure 3: Ridge Prediction

4.2 Results

After applying Ridge model to real estate price prediction, we obtained the result as shown in Figure [3]. Then we did the same with Lasso model, the results performed better (Figure [4]). For the best results, we aggregate them into a hybrid model between Ridge and Lasso as follow equation:

$$H = 0.5 * Ridge + 0.5 * Lasso \tag{4}$$

as the following update equation

$$L = ||Y - X(\theta)||_2^2 - 0.5 * \alpha(||\theta||_2) - 0.5 * \alpha(||\theta||_2) \tag{5}$$

Comparative result of each model was shown on Table [1]. The hybrid model of Ridge, Lasso

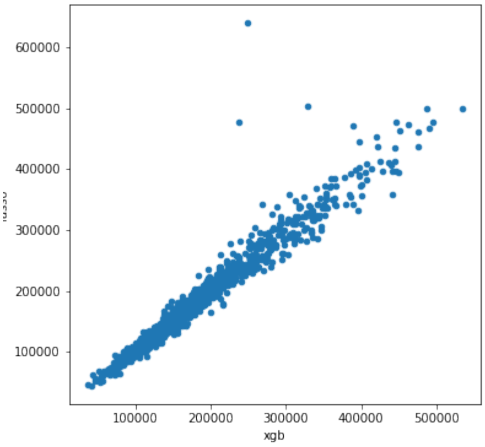

Figure 4: Lasso Prediction

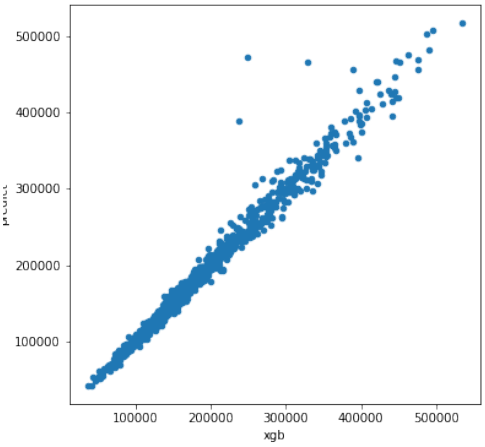

Figure 5: Ridge combine with Lasso Prediction

and XGB achieved the highest results in all models with minimum RMSLE value, 0.12342

The data distribution of three models, Ridge, Lasso and Hybrid , are shown on Figure [3], Figure [4], Figure [5] respectively.

5 Conclusion

The test result indicates the usefulness of creating more new features. For those missing values the program needs to generate default values with different methods statistically. For example, mode method, noValue, zero value or median value. There are a lot of key variables that affect house prices. If data are available ,a good idea is to introduce more features, for example income, salary, population, local amenities, cost of living, annual property tax, school, crime, marketing data.

Furthermore, Neural Networks are state-of-the-art models; it may help to improve prediction accuracy. We plan to build a Neural Network model to detect and predict house prices in Vietnamese, specially in Hanoi.

Acknowledgements

The authors wish to thank Assoc. Prof. PHAM Van Hai for his reviews and encouragements; thank (Kaggle) to organise this greatful competition; thank De Cock (2011) for the dataset of Ames House Price.

References

Eli Beracha, Ben T Gilbert, Tyler Kjorstad, and Kiplan Womack. 2018. On the relation between local amenities and house price dynamics. *Real Estate Economics*, 46(3):612–654.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Dean De Cock. 2011. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Kaggle. House prices - advanced regression techniques.

Stephen Law. 2017. Defining street-based local area and measuring its effect on house price using a hedonic price approach: The case study of metropolitan london. *Cities*, 60:166–179.

David C Ling, Joseph TL Ooi, and Thao TT Le. 2015. Explaining house price dynamics: Isolating the role of nonfundamentals. *Journal of Money, Credit and Banking*, 47(S1):87–125.

Binbin Lu, Martin Charlton, Paul Harris, and A Stewart Fotheringham. 2014. Geographically weighted regression with a non-euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science*, 28(4):660–681.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ryan M Rifkin and Ross A Lippert. 2007. Notes on regularized least squares.