A Hybrid Regression Technique for House Prices Prediction

Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh
Institute of High Performance Computing (IHPC), Agency for Science Technology and Research (A*STAR), Singapore
(lus@ihpc.a-star.edu.sg, liz@ihpc.a-star.edu.sg, qinz@ihpc.a-star.edu.sg, yangx@ihpc.a-star.edu.sg, gohsm@ihpc.a-star.edu.sg)

Abstract – **Usually, House price index represents the summarized price changes of residential housing. While for a single family house price prediction, it needs more accurate method based on location, house type, size, build year, local amenities, and some other factors which could affect house demand and supply. With limited dataset and data features, a practical and composite data pre-processing, creative feature engineering method is examined in this paper. The paper also proposes a hybrid Lasso and Gradient boosting regression model to predict individual house price. The proposed approach has recently been deployed as the key kernel for Kaggle Challenge "House Prices: Advanced Regression Techniques". The performance is promising as our latest score was ranked top 1% out of all competition teams and individuals.**

Keywords - **Feature Engineering, Machine Learning, Gradient Boosting, Lasso (Least Absolute Shrinkage and Selection Operator)**

I. INTRODUCTION

Machine learning develops algorithms and builds models from data, and uses them to predict on new data. The main difference with traditional algorithm is that a model is built from inputs data rather than just execute a series of instructions. Supervised learning uses data with result labeled, while unsupervised learning using unlabeled data. There are a few common machine learning algorithms, such as regression, classification, neural network and deep learning. Reinforcement learning and representation learning are heavily used for deep learning.

How to use machine learning algorithms to predict house price? It is a challenge to get as closely as possible result based on the model built. For a specific house price it is determined by location, size, house type, city, country, tax rules, economic cycle, population movement, interest rate, and many other factors which could affect demand and supply.

For local house price prediction, there are many useful regression algorithms to use. For example, support vector machines (SVM), Lasso (least absolute shrinkage and selection operator) [2], Gradient boosting [3], Ridge, Random forest. We will investigate and explore them in Part III.

After examining data, we find that the data quality is a key factor to predict the house prices. Data input feature density estimation is important for regression. Hence, normality test for each feature is to confirm whether it is well-modeled by a normal distribution and to explore possible transformation to a normal distribution. Homoscedasticity verification are also considered, hence regression algorithms with parameter more than 10000 iterations are applied. But the result is determined by the homoscedasticity between training data and test data. Linearity of each feature is the statistic fundamental of regression algorithm, therefore, many transformation are applied to enhance the linearity of input features.

Kaggle organizes a house prices competition [1], it provides data with 79 explanatory variables for part of residential home transactions in Ames, Iowa, and opens to all to predict price of each covered home transaction SalePrice.

This paper is organized as follows, it reviews related previous work in part II, and illustrates the details of methodology used in part III, then compares the test result of different algorithms in part IV, finally discusses the result and makes a conclusion in part V.

II. RELATED WORK

House price index (HPI) is used to measure price changes of residential housing in many countries, for example, the US Federal Housing Finance Agency HPI, S&P/Case-Shiller price index, UK National Statistics HPI, UK Land Registry's HPI, UK Halifax HPI, UK Rightmove HPI and Singapore URA HPI. Since HPI is a summary figure for all transactions, it isn't enough to use it to predict a specific house price.

A few documents explore the correlation among house price and local amenities, local area and renovation. Beracha et al [4] investigate the correlation between house price volatility, returns and local amenities, and proves that high amenity areas experience greater price volatility. Stephen Law [5] 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. Alexander and William investigates the result of property improvements in wide-scale US geographies [6], the result shows that the price could be increased 15% in the central districts of large cities, while less distortionary effect outside of downtown areas or in smaller cities.

For no fundamental factors, David et al [7] 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.

Daniel et al [8] explore nonparametric estimation for spatial data analysis, and find the feasibility for large datasets. For example, due to prices vary across neighborhoods within Chicago, the distance to the nearest rapid transit line is a misspecification,

Binbin et al [9] build Geographically Weighted Regression (GWR) model to study London house prices, they consider Euclidean distance (ED), road network distance and travel time metrics.

978-1-5386-0948-4/17/$31.00 ©2017 IEEE

Marco et al [10] emphasize the importance and complex of Spatial Heterogeneity in Austria, and proposed a Mixed geographically weighted regression (MGWR) model to reduce prediction errors.

Using State level data in USA, Sean et al [11] examine the correlation among house prices and real per capita disposable income, common shocks, macroeconomic and local disturbances, net borrowing cost, state level population growth and spatial factors.

Based on the administrative data from the Netherlands, Joep et al [12] find that higher income and wealth buyer leads to higher purchase price, while higher income and wealth seller leads to lower selling price.

In Singapore property market, the main factors are private or HDB property, new or resale, freehold, potential enbloc (collective sale), distance to MRT, location, size, and high level floors, and economic cycle.

Comparing with the 79 variables provided in Kaggle competition, we find that the information is incomplete, and a few features are unnecessary for prediction, there are implicit conditions in house price. We try to find those implicit characteristics among those features, then transform those features to normal distribution and transform for increasing linearity. We also explore regression algorithms with parameters adjustment and consider the coupling effect of different algorithms to achieve better test result.

III. METHODOLOGY

In this part, we describe the details of creative feature engineering, and describe how to apply multiple regression algorithms. Finally, we explain the coupling effect of Lasso and Gradient boosting algorithms.

A. Creative Feature Engineering

We investigating the value distribution and correlation of SalePrice for each variables and introduce many new variables. For example, Fig. 1 shows log transformation SalePrice distribution for each neighborhood. There are significant different SalePrices among different neighborhoods. Details of feature engineering are listed in following paragraphs.

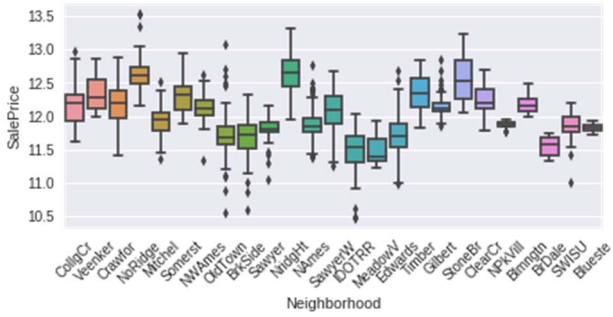

Fig. 1 Neighborhood Log Transformation of SalePrice

- Changing numerical type values to string category, and introducing some quality level numerical value.
- Changing few string category types to numerical types based on average SalePrice.
- Using mode to fill some missing values, for example MSZoning, SaleType; If too many missing values in a feature, we introduce NoValue type, for example, Alley; Replacing some missing values with 0, for example BsmtFullBath, and replacing some missing values to median values, for example GarageArea.
- Introducing sale price group predicted with SVM with few input features.
- Adding new features, we multiply of Lot Area, GrLiveArea, TotalBsmtSF, etc. with OverallQual, ExterQual, and KitchenQual etc. to add new features.
- 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 left side, the log transformation of SalePrice distribution, while in right side it is SalePrice distribution.

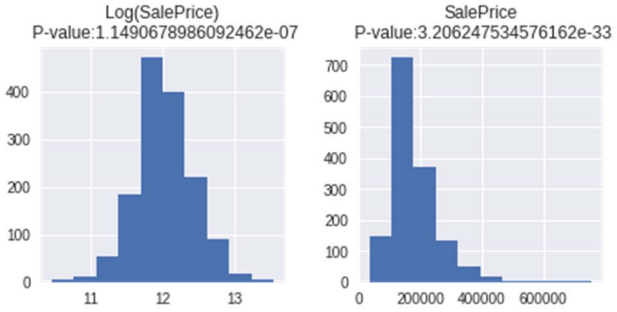

Fig. 2 Log transformation SalePrice and SalePrice distribution.

- Appling log transformation for other skewed numeric features.

After listing top positive and negative correlation features with log SalePrice, we add new features with square, cube, square root transformations of top correlation features.

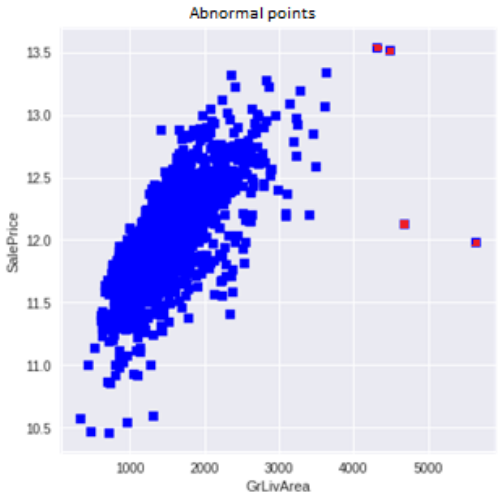

Fig. 3 Abnormal Points.

We also apply get_dummies method in Python Pandas module. It converts categorical variables into dummy/indicator variables. Then we remove generated new features with very few non zero values to avoid overfitting.

In order to improve the prediction accuracy, based on Lasso result, we drop more than 260 low correlation features from 490 features.

Finally, we remove 4 abnormal points as depicted in the right part of Fig. 3.

B. Regression Algorithms

There are many regression algorithms that can be used to build models and predict house prices. After investigation, we find that Ridge, Lasso from sklearn [2] and Gradient boosting [3] are more useful.

Ridge and Lasso regressions are used to model cases with large number of features. In especial, Lasso regression could model cases with a million features. In order to avoid overfitting, Ridge regression performs L_2 regularization and Lasso regression performs L_1 regularization. In the following part, we investigate the Ridge, Lasso and Gradient Boosting regression algorithms.

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. In Fig. 4, it shows the Ridge prediction in X, and SalePrice in Y for training data.

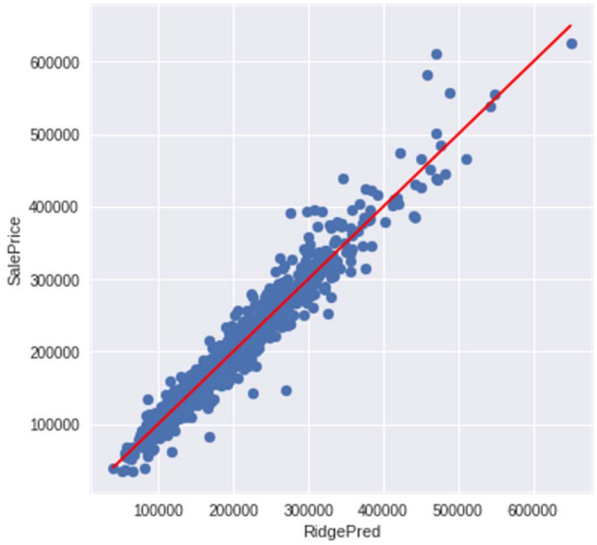

Fig. 4 Ridge Prediction for Training Data.

Lasso regression there is also one parameter alpha for Lasso regression to choose. Lasso is useful to perform feature selection. After selecting the alpha parameter, and building Lasso model, Lasso is able to list coefficients zero features, those features could be dropped late. In Fig. 5, it shows the Lasso prediction in X, and SalePrice in Y for training data.

Gradient boosting there are few parameters to choose for gradient boosting. You may do grid search for parameter selection. In order to keep dynamic balance of overfitting and regularity, we follow the following steps to choose parameters.

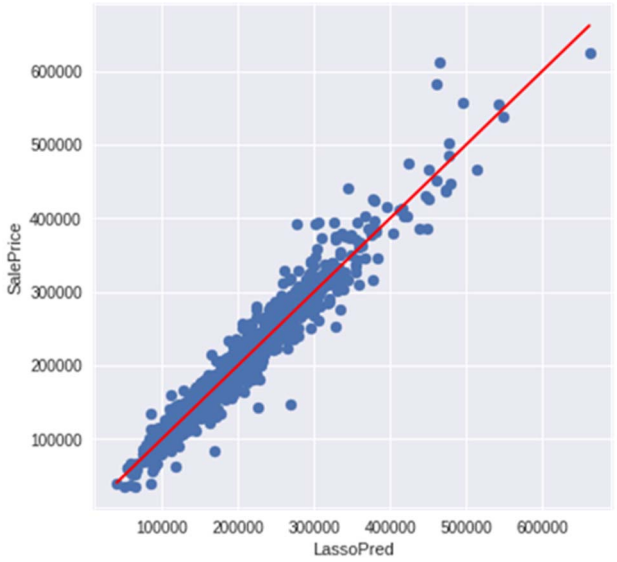

Fig. 5 Lasso Prediction for Training Data.

Fixing learning rate = 0.1, no of estimators = 10000, and determine the optimum number of tree specific parameters (max_depth, min_child_weight, gamma, subsample, colsample_bytree) for decided learning rate and number of trees.

- Setting regularization parameters (lambda, alpha).
- Reducing learning rate and decide those optimal parameters again.

In Fig. 6, it shows the Gradient boosting prediction in X, and SalePrice in Y for training data.

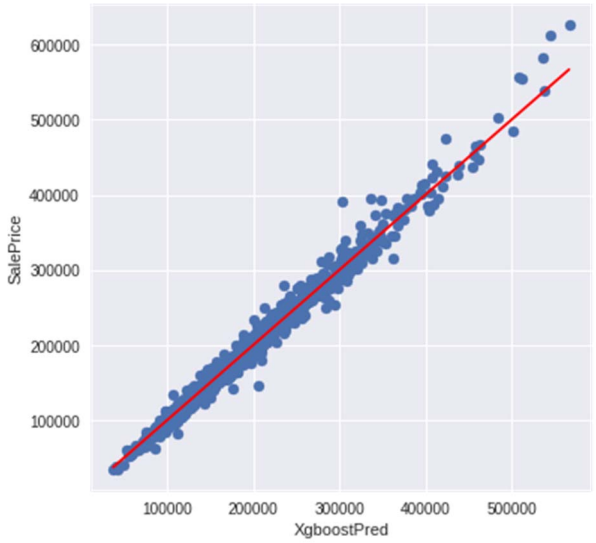

Fig. 6 Gradient Boosting Prediction for Training Data.

C. Hybrid Regression

Since Kaggle House Prices competition [1] covers the SalePrice field for test data, users only can get score after submission. We find the coupling effect of multiple regression algorithms. The hybrid regressions result is better than one specific regression algorithm. There is a

need to verify different hybrid combination to get the best score. In Fig. 7, it shows Lasso and Gradient boosting hybrid prediction in X, and SalePrice in Y for training data.

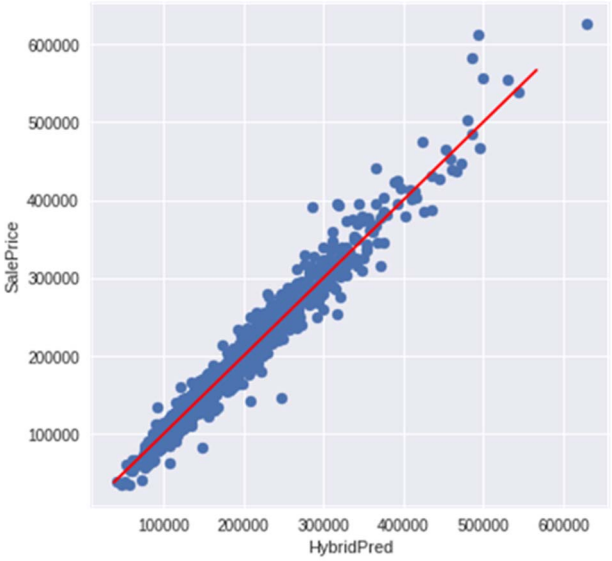

Fig. 7 Lasso and Gradient Boosting Hybrid Prediction.

D. Prediction Submission and Valuation

Kaggle house prices competition [1] evaluation standard is Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

Where y_i stands for log(SalePrice$_i$) and $\hat{y}_i$ stands for log(pred_SalePrice$_i$).

The core part equates to log(SalePrice$_i$/pred_SalePrice$_i$). Using prediction SalePrice ratio could avoid the heavy weightage 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.

IV. RESULTS

Many iterations of feature engineering have been done to find the optimal no of features. Based on training data, then we tried to find optimal alpha parameter for Ridge regression, optimal alpha parameter for Lasso regression, and a group of optimal parameters for Gradient boosting regression, for example colsample_bytree, gamma, max_depth, min_child_weight, subsample, reg_alpha, reg_lambda, learning_rate. Finally, considering the Coupling effect among regression algorithms, we have evaluated a couple of combinations of hybrid regression to get combination of 65% Lasso and 35 % Gradient boosting.

A. Creative Feature Engineering

After read kernels and discussion topics in Kaggle [1], we start to investigate features provided. As mentioned in Part III Methodology, we add 400 more new features. In Tab. 1, it illustrates that the more features we added, the better score from Kaggle evaluation for test data.

Regression Algorithms	No of Features	Score
lasso	150	0.12730
0.7Lasso+ 0.3 Xgb	320	0.12106
lasso	440	0.11918
lasso	490	**0.11520**

Tab. 1 Creative Feature Engineering

B. Ridge, Lasso and Gradient Boost result

As mentioned in Part III, we use Lasso for features selection to remove unused features. After investigated Ridge, Lasso and Gradient boosting regression algorithms, we find that the best score for test data is using 230 features. In Tab. 2 the minimum Root Mean Squared Error (RMSE) for training data is 160 features, while best score generated by Ridge regression using test data from Kaggle is 230 features.

No of Features	Alpha	RMSE	Score
160	13	**0.112276**	0.11638
230	18	0.113627	**0.11558**
280	20	0.114547	0.11583

Tab. 2 Feature Selection for Ridge

Using Lasso regression, we find same trend of no of features for test data. In Tab. 3, the minimum RMSE for training data is 160 features, while best score generated by Lasso regression using test data from Kaggle is 230 features.

No of Features	Alpha	RMSE	Score
160	1.55E-04	**0.113838**	0.11706
230	3.70E-04	0.114974	**0.11499**
280	5.40E-04	0.115464	0.11675

Tab. 3 Feature Selection for Lasso

Using Gradient boosting regression, same result happens again. The optimal parameters for training data is Subsample = 0.5, while Subsample = 0.6, we get better score from Kaggle for test data.

No of Features	Subsample	Score
160	0.6	0.12032
230	0.5	**0.11876**
230	0.6	**0.11843**
280	0.6	0.12057

Tab. 4 Feature Selection for Gradient Boosting

C. Hybrid Regression Result

After investigated few combination of hybrid prediction of test data, we find that the results for 230 features are better than that for 280 features. Ridge and Lasso prediction results are close. Hybrid Lasso and Gradient boosting achieves the best score, furthermore, the combination is 65% Lasso with 35 % Gradient boosting.

In Tab. 5, it shows the test data score of hybrid method of Ridge, Lasso, Gradient boosting regression, while Xgb is standard for Gradient boosting.

Features	Hybrid Method	Score
230	0.65Ridge+0.35Xgb	0.11318
230	0.70Lasso+0.30Xgb	0.11294
230	0.65Lasso+0.35Xgb	**0.11260**
230	0.60Lasso+0.40Xgb	0.11277
230	0.3Ridge+0.35Lasso+0.35Xgb	0.11285
230	0.25Ridge+0.40Lasso+0.35Xgb	0.11283
280	0.65Ridge+0.35Xgb	0.11458
280	0.65Lasso+0.35Xgb	0.11539

Tab. 5 Hybrid Combination of Regressions

V. DISCUSSION AND CONCLUSION

A. Importance of Creative Feature Engineering

In part IV, the test result shows that it is useful to create more new features. For those missing values, based on statistical result, the program needs to set default value with different method, for example, mode method, noValue, zero value or median value. For some skewed distribution features, log transformation is a very useful method.

An interesting finding is how to use Lasso to select features, there is a need to try and verify the features to be removed. For this example, it is about 230 features to remain.

The purpose of features engineering is to improve data normality and linearity, while set parameter of high iteration times is used to improve data homoscedasticity. Another interesting finding is the optimal no of features for training data may not be the best one for test data. The optimal group of parameters for Gradient boosting for training data, it may not be the best one for test data.

B. Hybrid Regression

In part IV, the result proves the coupling effect of multiple regression algorithms. Based on the result, the hybrid regressions are better than one from Ridge, Lasso or Gradient boosting regression. The best hybrid regression result for test data is 0.11260 with 65% Lasso and 35% Gradient boosting combination.

Lasso is very useful for feature selection, and one more benefit of removing useless features is it could decrease Ridge, Gradient boosting regression MSE result.

If we introduce more features, such as economic cycle, population movement, interest rate, the prediction SalePrice will be more accurate for future transactions.

The best position in Kaggle House Prices competition Public Leaderboard is top 1% among all teams and individuals.

Since it is a real and very fierce competition, our experience is that the program needs to be excellent in creative feature engineering, features selection, regression methods, parameter selections, and hybrid regression selection.

C. Future Work

As mentioned in Part II related work, there are a lot of key variables 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, Random forest is an advanced regression algorithm; it may help to improve prediction accuracy.

Finally, we suggest building a separate algorithm to detect and predict abnormal transactions SalePrice.

Kaggle is a good place to develop sharp tools for machine learning and test the result in blind mode. This paper demonstrates how to make machine learning more useful in normal life.

ACKNOWLEDGMENT

Thank Kaggle.com [1] for organizing this competition; thank Dean De Cock for compiling the Ames Housing dataset, and thank those contributors for providing ten thousands kernels and hundreds discussion topics.

REFERENCES

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques
[2] http://scikit-learn.org/stable/install.html
[3] https://github.com/dmlc/xgboost
[4] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016.
[5] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.
[6] Alexander N. Bogin , William M. Doerner, "Property Renovations and Their Impact on House Price Index Construction", https://www.fhfa.gov/PolicyProgramsResearch/Research/PaperDocuments/wp1702.pdf
[7] David C. Ling, Joseph T.L. Ooi and Thao T.T. Le, "Explaining house price dynamics: Isolating the role of nonfundamentals", Journal of Money, Credit and Banking, vol. 47, Issue S1, pp. 87-125, March/April 2015.
[8] Daniel P. McMillen, Christian L. Redfearn, "Estimation And Hypothesis Testing For Nonparametric Hedonic House Price Functions", Journal of Regional Science, vol. 50, Issue 3, pp. 712–733, Aug. 2010.
[9] Binbin Lu, Martin Charlton, Paul Harris & A. Stewart Fotheringham, "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data", International Journal of Geographical Information Science, pp. 660-681, Jan 2014.
[10] Marco Helbich, Wolfgang Brunauer, Eric Vaz, Peter Nijkamp, "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria", Urban Studies, vol. 51, Issue 2, Feb. 2014
[11] Sean Holly, M. Hashem Pesarana, Takashi Yamagata, "A spatio-temporal model of house prices in the USA", Journal of Econometrics, vol. 158, Issue 1, pp. 160–173, Sep. 2010.
[12] Joep Steegmans, Wolter Hassink, "Financial position and house price determination: An empirical study of income and wealth effects", Journal of Housing Economics, vol. 36, pp. 8-24, June 2017.