Predictive modeling for house price prediction

Minsu Kim, Lingjie Qiao, Kevin Liao, Cheng Peng
University of California at Berkeley

***Abstract*— This paper summarizes the background, problem, methodology and results of our analysis on predicting house price in Ames, Iowa. To make full use of statistical models and machine learning techniques we have learned from Stats 159 Reproducible and Collaborative Statistical Data Science, we participate in Kaggle Competition, "House Price: Advanced Regression Techniques". We use the Ames Housing dataset compiled by Dean De Cock for use in data science education. While the competition only emphasizes the accuracy of predicted values, this paper elaborates thorough explanations on exploratory analysis, feature engineering and statistical modeling. Furthermore, we put our effort on reproducibility of this analysis, so that the reader can reproduce the exact same result from our code.**

I. INTRODUCTION

The House Price project thoroughly explores the predictive modeling process and advanced regression techniques. From previous study, in order to understand the relationship of one dependent variable with several independent variables, we fit a multiple linear regession with Ordinary Least Squares. However, since OLS may have high variance and include irrelevant variables, Predictive Modeling Process can improve the results in terms of Prediction Accuracy and Model Interpretability.

The competition sets the background of the project: Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition requires participants to predict the final price of each home. Our team therefore follows the idea of model prediction and tries to use different techniques in order to most accurately predict the final sales price of each house.

II. DATA DESCRIPTION

The datasets are obtained from the Kaggle Competition website. We have access to four files:

- *data description*, which provides the official definition for fields.
- *train.csv*, which provides 1459 real observations that can be used for model construction.
- *test.csv*, which is used to fit the predictive model and create submission entry for the final sales price of 1460 observations
- *sample submission*, which gives an example of how the fitted values should be submitted.

The train dataset has in total 80 variables, 79 predictors and 1 target variable called *SalePrice*. We observe both categorial predictors, such as *FireplaceQu*, *GarageCond* and *MasVnrType* as well as numerical predictors, such as *PoolArea*, *EnclosedPorch* and *YrSold*. Since we can potentially create a lot of different new variables, our goal is to understand the relationship between *SalePrice* and these predictors with statistical fitting procedures that minimizes Mean Square Error.

III. EXPLORATORY DATA ANALYSIS

In order to choose the model to accurately predict the housing price, we first need to understand the provided dataset by conducting exploratory data analysis (aka EDA). Separating categorical predictors and numerical predictors, we can explore in two perspectives: 1) the distribution of varible itself, and 2) the relationship between explanatory variable and response variable.

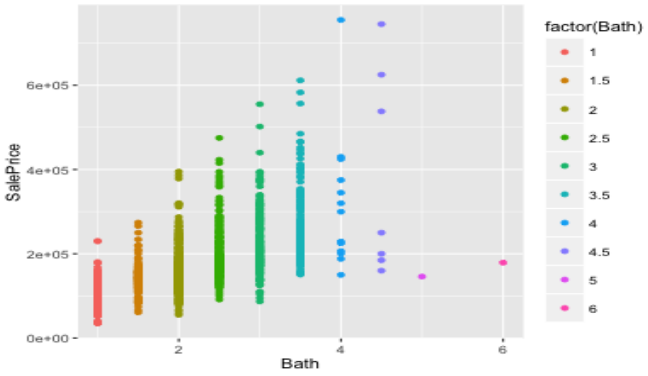

Fig. 1. Exploratory Data Analysis - Relationship Between Bath and SalePrice

Figure 1 is an example of explanatory data analysis, which presents the relationship between explanatory variable bath and the response variable SalePrice. Figure 2 is a more complicated yet informative plot, which shows the relationship between sales price and the difference of year sold and year built, namely the age of house after being built, as well as the difference of year sold and year remodeled, namely the age of house after being remodeled.

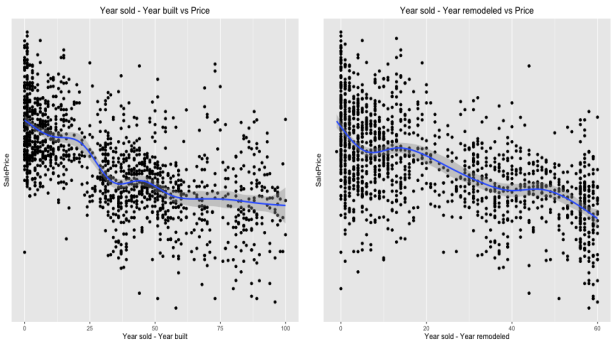

Fig. 2. Exploratory Data Analysis - scatter plot of year sold - year built vs. sale price as well as year sold - year remodeled vs sale price

Some interesting observations we made include: 1) While some variables are equally distributed in their own segments, some variables are skewed in certain directions (either to the left or to the right), which means some potential data transformation can be made to improve the predictive model; 2) We observe some linear relationship between variables, which can be used in the future model selection; and 3) While some variables have strong linear relationship with the response variable SALEPRICE, other variables do not display the similar pattern. This might encourage us to remove unnecessary predictors from the dataset in terms of model fitting and selecting.

IV. METHODOLOGY

The goal of this analysis is to accurately predict the final price of each home. Therefore, we frame this problem as a regression problem, and decide to use the L2 loss function [10] which is often used in regression problem. Taking this objective into account, we preprocess original dataset so that regression models can work well. Furthermore, we extract more features by involving feature engineering. Finally, we fit two shrinkage models and two ensemble models. The details are explained in the following subsections.

A. Evaluation and objective loss function

We specifically use root mean squared logarithmic error [5], which makes more sense in our problem setting because errors in predicting expensive houses and cheap houses should affect the result equally. The following is the formula of RMSLE.

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1) - log(a_i+1))^2}$$

Since L2 loss function minimizes the squared differences between the estimated and existing target values [10], L2 error will be much larger in the case of outlier compared to L1 and therefore L2 loss function is highly sensitive to outliers in the dataset. So, in the preprocessing step, we eliminate outliers to remedy this issue.

B. Preprocess

According to the exploratory data analysis, we find a lot of NA values in most of predictors. Since regression models cannot handle missing data, we need to either remove or impute data using appropriate methods [7]. The data description provided by client [6] indicates that some of the missing values are actually none value. In that case, we replace NA value with factor variable named None. However, there are some numerical predictors with missing values in an unsystematical manner. In that case, we impute them with mean values of predictors.

As a next step, we apply log transformations to area related predictors such as GrLivArea and LotArea as well as target variable. The log transformation has an effect to remedy skewness of data [8] by making original distributions of predictors to more normally distributed. Consequently, it helps regression model to work better.

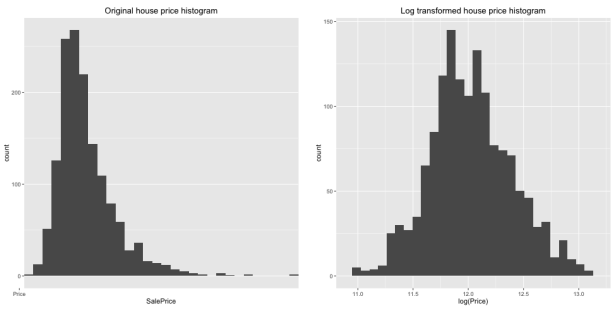

Fig. 3. Effect of Log Transformation - the distribution of original sale price and the distribution of sales price after log transformation

After the log transformation, we still notice few outliers and eliminate lowest and highest 0.1% data points. Furthermore, majority of predictors are categorical under which regression models cannot be used directly. Therefore, we apply one-hot encoding [9] to convert categorical values to numerical ones. It consequently expands features from 79 to approximately 500. The Table 1 below summarizes this procedure.

Table 1: Data Preprocess and Variable Transformation

Factorization	Log Transformation	Removing Outliers
MSSubClass	SalePrice	17 points ->
YearBuilt	LotArea	below 10.91511
YrSold	GrLivArea	and above
MoSold		SalePrice
GarageYrBlt		

C. Data preparation

Before fitting the model, we first split dataset into train and test. We could have a separate validation set. However, R library caret has a built-in cross validation as a generic interface. So it automatically takes care of cross validation. We use train data to train and tune our models using 5-kold cross validation, and later compare RMSLE using hold-out test data [13].

D. Featuring engineering

Considering the complexity of the problem as well as the number of observations and predictors, we assume that the success of this analysis is largely dependent on informative, feature engineered predictors that can reveal the subtle relationship to our target variable. Given the small size of dataset with 1460 observations, we conclude that feature learning, which is a set of techniques that learn a feature: a transformation of raw data input to a representation that can be effectively exploited in machine learning tasks [1], is not a feasible option because feature learning often involves very complicated models with multiple layers, which tends to cause an overfitting issue when dataset is small [3].

With this observation, we therefore focus more on manual feature engineering [4]. This process is a important stepping-stone in that it helps reveal significant predictors that are previously not represented well in original dataset. By explicitly designing what the input x's should be, our predictive models can solve a problem easily.

E. Model description and hyper-parameter tuning

While training each model, we need to find optimal paramters for each model. In order to effectively select hyper-parameters, we use 5-fold cross-validation. For lasso and ridge, lambda is the tuning parameter. It determines how much we will penalize models for high weights on predictors. If lambda is high, it penalizes models more and ends up generating sparse models. This kind of models is called shrinkage method because they shrinken weights or even remove predictors by penalizing models. These models are especially good options when there are many predictors. By penalziing or removing unnecessary predictors, they provide more interpretable results. Thus, they are often utilized in genomic and pharametical analysis.

F. Modeling

We utilize both shrinkage regression models and ensemble models. Practitioners often favor ensemble models [11] due to their conveniences. Ensemble models such as Random Forest [12], which uses the averaged result from the randomly grown decision trees such as CART [13], often work well with unscaled, missing data and are used for both classification and regression problems. Also, since our dataset contains a hug number of predictors, shrinkage methods such Lasso and Ridge regressions [13], which penalize predictors by shrinking their weights, can be highly effective. Furthermore, we utilize a dimension reduction technique called PCA[13] in order to compress information into lower dimension. The results of modeling will be further explained in the following section in detail.

G. Model comparison

As mentioned on in Evaluation section, we use RMSE to compare models and select the best model. Although RMSE does not provide an absolute means of model accuracy, it provides a relative measure to compare models. Thus, we finalize our model with the lowest RMSE. In summary, the following is the the procedure for each model.

1) Split the data into train and test, 80% and 20% repectively.
2) Train a model using train data with 5-fold cross-valdation.
3) Pick the optimal hyper-parameters.
4) Predict balance using the model with the optimal parameters.
5) Calculate Mean Square Error.
6) Record both Mean Square Error and coefficients.

V. ANALYSIS

Due to the high dimensionality of dataset, we first experiment to decide whether to either project predictors into lower dimensional bases or select predictors using shrinkage methods. First, we try Principal component analysis which is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [15]. Since our one-hot encoded dataset has a lot of predictors that have near zero variances, we first need to remove near zero variances. This is because PCA considers variability of data and compress the predictors with high variances into first principal components.

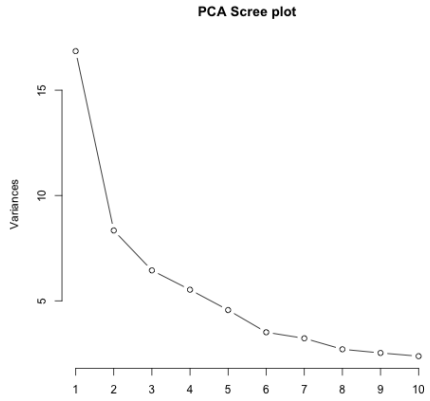

Fig. 4. PCA scree plot for each principal component

After removing near zero variance 519 to 128. This indicates that a great number of predictors have apporximately identical across observations. Figure 4 shows a PCA scree plot for each principal component. It indicates that first few principal components effectively explain the variances of data, and overall graph exponentially decays. Cumulative Proportion in $Tablemodel_pca$ shows that first 10 pcs and 61 pcs explain approximately 44% and 90% of variance in dataset respectively. It means that if we decide to reduce dimensionality at the expense

of a bit little of prediction accuracy, we could use 61 principal components which are approximately 1/8 of total predictors.

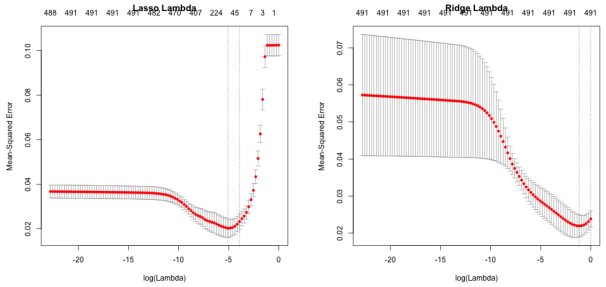

Fig. 5. MSE values for Lambda Values - Lasso Regression (left) and Ridge Regression (right)

Furthermore, we try to fit both lasso and regression and tune lambda values using 5-fold cross validation and then examine coefficients. Figure 5 shows respectively the Mean Square Error for corresponding lambda values for both models. The Figure suggests that MSE decreases as log(lambda) increase up to log lambda equals -5 and -1.15 respectively. Based on these results, we decide optimal lambda for both model.

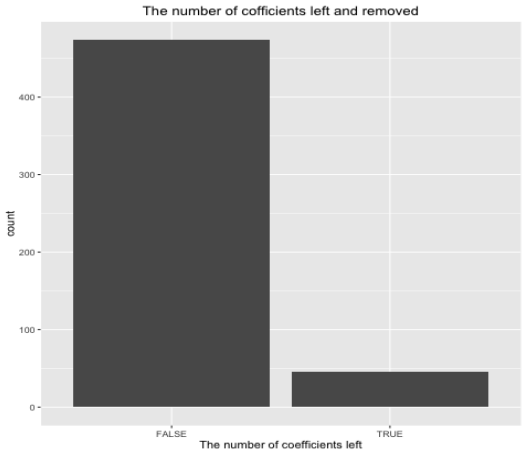

Fig. 6. Number of Coefficients - Lasso Regression

We also examine coefficients of lasso regression. Figure 6 indicates that most of weights for corresponding predictors become zeros, which effectively eliminates a great portion of predictors. After elimination, only 78 predictors are left. We plot top 10 coefficients for both lasso and ridge to investigate how much each coefficient contribution to the model.

Figure 7 shows the top 10 coefficients from both models. It is noteworthy that GrivArea has the highest coefficient in Lasso and YearBuilt2010 has the highest coefficient in Ridge. Both make sense because the area and built year are important aspects when it comes to making a decision about purchasing house. Both models indicate that GrLiArea, NeighborhoodCrawfor and NeighborhoodStoneBr belong to top 10 highest coefficients

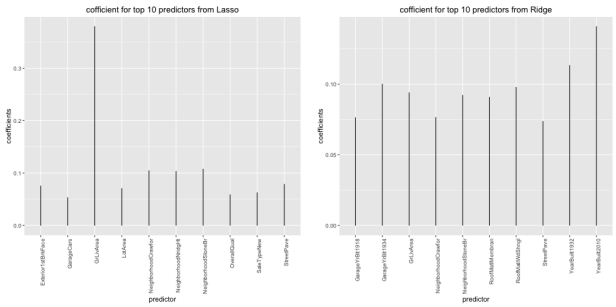

Fig. 7. Top 10 Coefficients - Lasso Regression (left) and Ridge Regression (right)

and are highly associated with house price. Furthermore, we try to figure out feature importance using Gradient boosting machine. Feature importance is automatically measured as GBM grows its tree while minimizing entropy. Figure $model_gbm_predictorimportance.png$ shows that overall quality has the highest feature importance

VI. RESULTS

As a result, Figure 8 shows model comparison between true price and predicted price.

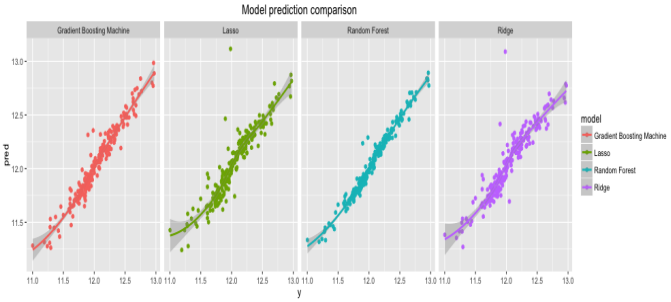

Fig. 8. Model Comparison between True Price and Predicted Price

Four models are fairly accurate and most of data points are located on diagonal line, which indicates a strong linear relationship between true price and predicted. Especially, Random Forest is especially well fitted among these four.

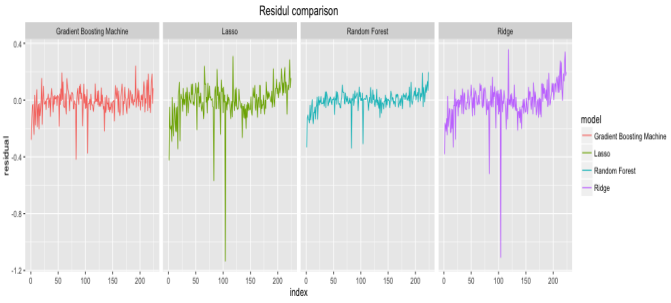

Fig. 9. Model Comparison - Residual Plots of Four Models

Figure 9 shows the residual plots for four models. The residual is the difference between real house price and

predicted price from models. We can see from Figure 9 that all models overestimate the price for cheaper house and underestimate the price for expensive house. This phenomenon is not surprising because we have few data points in cheapest and most expensive houses. Thus, the models incur a high bias in these end zones. This can be alleviated if we gather more data if possible. Compared to Lasso and Ridge, Gradient boosting machine and Random Forest are more stable. Both Lasso and Ridge have huge errors in the middle of graphs.

Lastly, Table shows RMSLE for each model, which is measured from hold-out dataset. It indicates that Random Forest predicts target variable the best.

VII. CONCLUSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word ÃŠacknowledgmentÃŞ in America is without an ÃŠeÃŞ after the ÃŠgÃŞ. Avoid the stilted expression, ÃŠOne of us (R. B. G.) thanks . . .ÃŞ Instead, try ÃŠR. B. G. thanksÃŞ. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

References

[1] G. O. Young, ÃŠSynthetic structure of industrial plastics (Book style with paper title and editor),ÃŞ in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15ÃŘ64.

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123ÃŘ135.

[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, ÃŠAn approach to graphs of linear forms (Unpublished work style),ÃŞ unpublished.

[5] E. H. Miller, ÃŠA note on reflector arrays (Periodical styleÃŚAccepted for publication),ÃŞ IEEE Trans. Antennas Propagat., to be publised.

[6] J. Wang, ÃŠFundamentals of erbium-doped fiber amplifiers arrays (Periodical styleÃŚSubmitted for publication),ÃŞ IEEE J. Quantum Electron., submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, ÃŠElectron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style),ÃŞ IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740ÃŘ741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, ÃŠInfrared navigationÃŚPart I: An assessment of feasibility (Periodical style),ÃŞ IEEE Trans. Electron Devices, vol. ED-11, pp. 34ÃŘ39, Jan. 1959.

[11] S. Chen, B. Mulgrew, and P. M. Grant, ÃŠA clustering technique for digital communications channel equalization using radial basis function networks,ÃŞ IEEE Trans. Neural Networks, vol. 4, pp. 570ÃŘ578, July 1993.

[12] R. W. Lucky, ÃŠAutomatic equalization for digital communication,ÃŞ Bell Syst. Tech. J., vol. 44, no. 4, pp. 547ÃŘ588, Apr. 1965.

[13] S. P. Bingulac, ÃŠOn the compatibility of adaptive controllers (Published Conference Proceedings style),ÃŞ in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8ÃŘ16.

[14] G. R. Faulhaber, ÃŠDesign of service systems with priority reservation,ÃŞ in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 3ÃŘ8.

[15] W. D. Doyle, ÃŠMagnetization reversal in films with biaxial anisotropy,ÃŞ in 1987 Proc. INTERMAG Conf., pp. 2.2-1ÃŘ2.2-6.

[16] G. W. Juette and L. E. Zeffanella, ÃŠRadio noise currents n short sections on bundle conductors (Presented Conference Paper style),ÃŞ presented at the IEEE Summer power Meeting, Dallas, TX, June 22ÃŘ27, 1990, Paper 90 SM 690-0 PWRS.

[17] J. G. Kreifeldt, ÃŠAn analysis of surface-detected EMG as an amplitude-modulated noise,ÃŞ presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.

[18] J. Williams, ÃŠNarrow-band analyzer (Thesis or Dissertation style),ÃŞ Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[19] N. Kawasaki, ÃŠParametric study of thermal and chemical nonequilibrium nozzle flow,ÃŞ M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[20] J. P. Wilkinson, ÃŠNonlinear resonant circuit devices (Patent style),ÃŞ U.S. Patent 3 624 12, July 16, 1990.

VIII. Below should be removed

A. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation ÃŠFig. 1ÃŞ, even at the beginning of a sentence.

TABLE I

An Example of a Table

One	Two
Three	Four

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

Fig. 10. Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity ÃŠMagnetizationÃŞ, or ÃŠMagnetization, MÃŞ, not

just ÃŠMÃŞ. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write ÃŠMagnetization (A/m)ÃŞ or ÃŠMagnetization A[m(1)]ÃŞ, not just ÃŠA/mÃŞ. Do not label axes with a ratio of quantities and units. For example, write ÃŠTemperature (K)ÃŞ, not ÃŠTemperature/K.ÃŞ