Housing Price Forecast

Master of Applied Economics, UCLA

Author: Zhiting Chen, Xiuqi Li, Xinwei Hu, Xinxian Li

Advisor: Patrick Convery, Rojas Randall

Content

Abstract

In this project, we would like to evaluate the performance and predictive ability of different machine learning models on predicting the house prices. We will rely on Ames Housing dataset which was compiled by Dean De Cock from kaggle.

With the use of multiple regression model as econometric approach, we used the methods of Mallows CP and StepAIC to select the influential variables on the housing price. Seven variables, which includes the lot area, overall quality, garage area and basement area play important role on deciding house price. Bootstrapping and cross validation are used to evaluate model performance.

In terms of predictive model, lasso, ridge, elastic net, Principal Component Analysis, Random Forest, Support Vector Machine are used to fit training data and forecast house price. The results show that random forest model is superior to other five models. The training score and testing score of random forest model are 0.9533 and 0.8205 respectively.

House price has strong correlations with total room above grade, number of fireplaces, size of garage in car capacity, overall material and finish quality and other five variables and Random Forest model can be set as the best performance model on predicting house price.

1. Introduction

In this project, we plan to predict the price of residential homes in Ames, Iowa using data of different aspects of home. The result could help the real estate companies to set a more reasonable price for new built home. Also, it could help customers to judge whether the aspects they care most play a significant role on house price. This paper would like to investigate the different contribution of each variable on house price and evaluate predictive ability of different machine learning models on predicting the house prices. A model trained on this data that is seen as a good fit can be used to analyze the reasons for rising house prices, make predictions and help real states business companies to make decisions whether to develop the real states in the corresponding area.

In order to realize model interpretation and predictive ability, the paper use multiple regression model as economic approach and several forecasting model via machine learning methods.

2. Workflow

Figure 1 shows detailed workflow. From economics approach, we would like multiple regression model and five steps. We also use six different machine learning model, including lasso or ridge regression model, PCA analysis, Support vector regression model random forest to fit training data. Cross validation and accuracy score will be used to evaluate model performance.

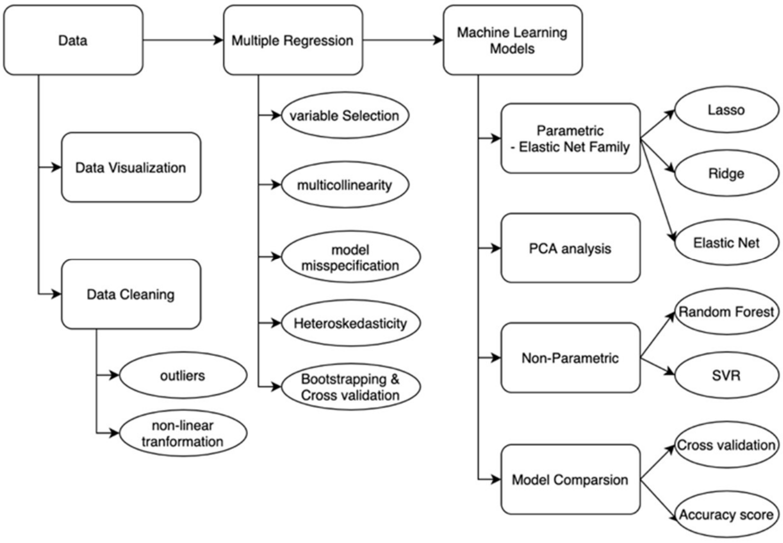


Figure 1: Workflow

3. Data

Data Source

This is a data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. It includes 17 explanatory variables and 1 independent variable (listed below). 1453 observations are collected.

Variables	**Definition**
MSSubClass	The building class
LotArea	Lot size in square feet
OverallQual	Overall material and finish quality
OverallCond	Overall condition rating
Houseage	The age of house
MasVnrArea	Masonry veneer area in square feet
BsmtFinSF1	Type 1 finished square feet
BsmtUnfSF	Unfinished square feet of basement area
TotalBsmtSF	Total square feet of basement area
1stFlrSF	First Floor square feet
GrLivArea	Above ground living area square feet
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
Fireplaces	Number of fireplaces
GarageCars	Size of garage in car capacity
GarageArea	Size of garage in square feet
WoodDeckSF	Wood deck area in square feet
OpenPorchSF	Open porch area in square feet
SalePrice	the property's sale price in dollars

Table1: Variables Definition

Descriptive Statistics

Figure 2 shows the distribution of the variables. From Figure 2 we can see that not much variables of the dataset have a normal distribution. Therefore, the future data transformation might be necessary in order to having a more stable dataset.

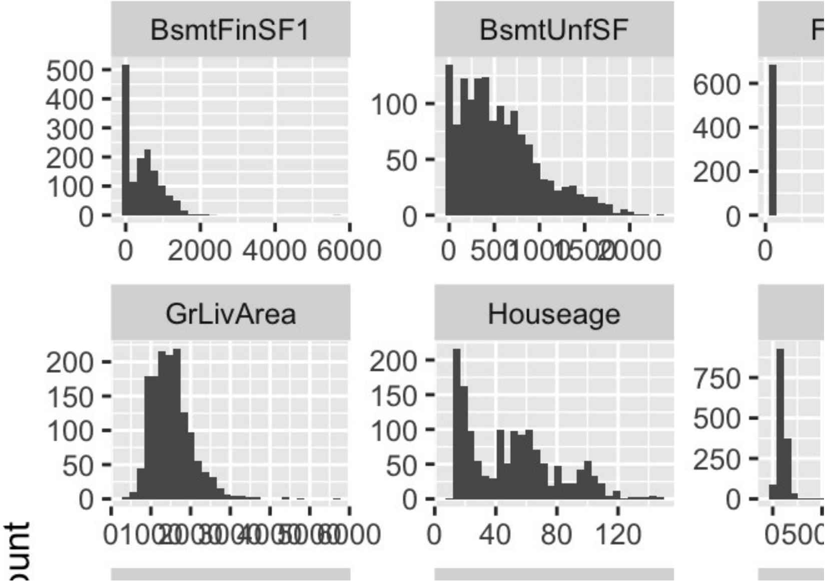


Figure 2: The Distribution of Variables

We also draw a plot to visualize the correlation between Sale price and other variables. From Figure 3, we can see that the variables we select have vary degrees of correlation, from which we can infer that the variables we pick are effective variables.

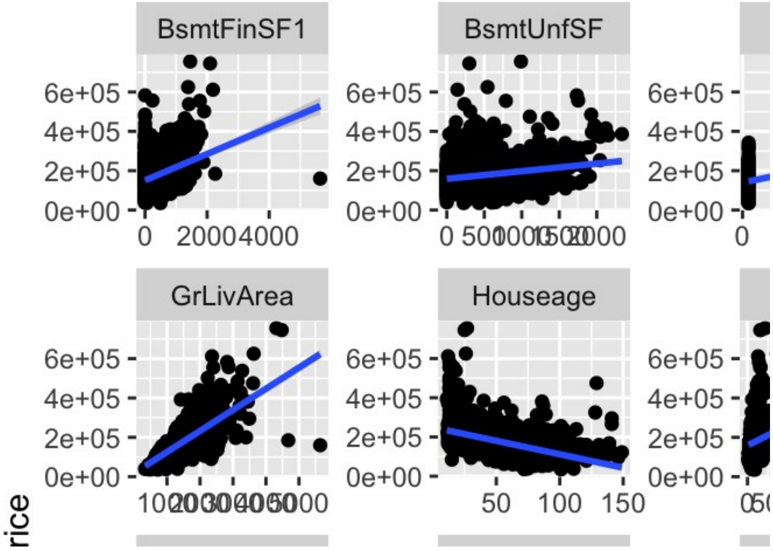


Figure 3: The correlation plot of Variables and Sale price

Figure 4 shows the correlation between the variables. From this picture we knew that the variables between houseage &overallquality, houseage & garagecars, houseage &garagearea and BsmtFinSF1 & BsmtUnfSF have strongly negative correlation. And the variables between TotalBsmtSF & **GrLivArea**, X1stFlrSF & TotalBsmtSF and garagecars & garagearea have highly positive correlation. Due to these significant correlations, multilinearity might exist in our dataset. Therefore, we need to test and eliminate the multilinearity in the model constructing process.

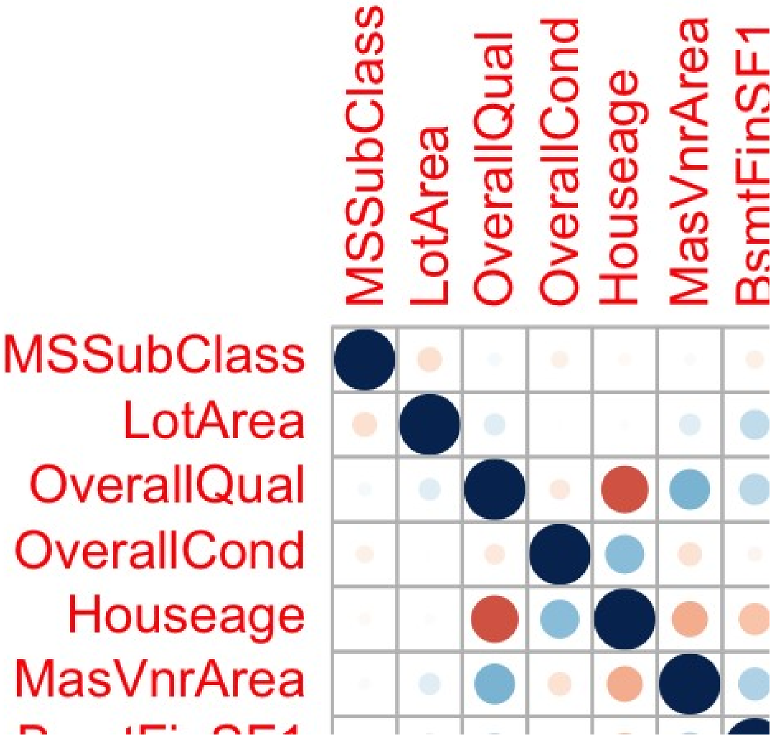


Figure 4: Correlation Plot

Data Cleaning

Figure5 and 6 are QQ Plot and Cook's Distance Plot for the dataset. Both of them are used to test whether the outliers exist in our dataset. By looking at these two pictures, we found the numbers of 523, 1177 and 1291 might be the outliers. So, we eliminated these three data respectively or entirely from the original dataset. And through comparing AIC, BIC and R square values for each elimination method, we noticed that the smallest AIC and BIC and biggest R square value were obtained when all of these 3 data were deleted.

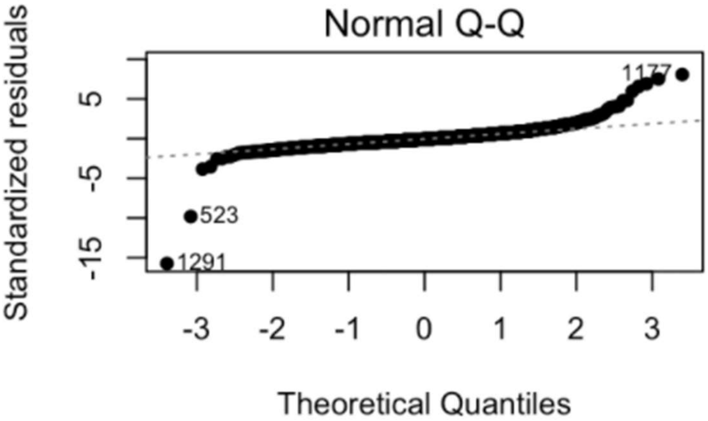


Figure 5: QQ Plot

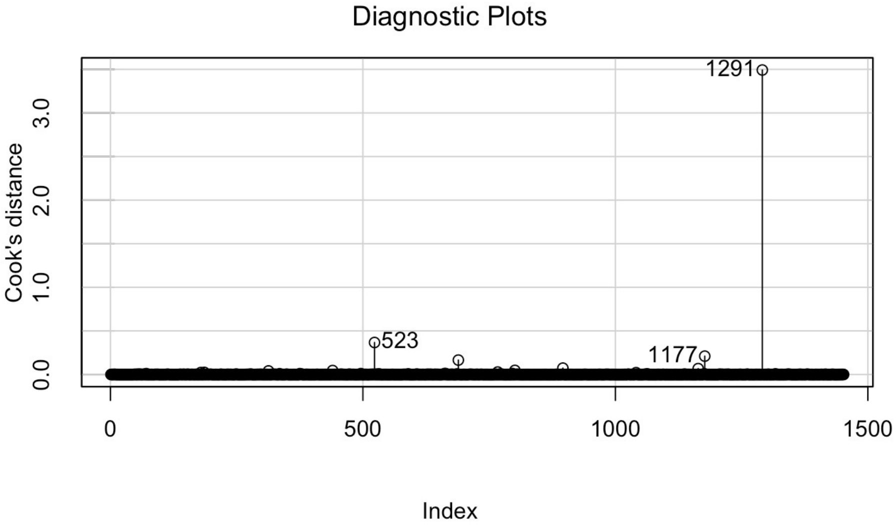


Figure 6: Cook's Distance Plot

4. Result

4.1 Multiple Regression

Transformation of Dataset

After we eliminated the outliers, we considered whether the transformation is necessary. By looking at the crPlots(), we realized that the pink curve did not fit the dashed blue line very well in the figures of “OverallQual”, so these four variables might need to be transformed. By using powerTransfrom(), the results showed that only the variable “OverallQual” needs to be transformed and the transforming power is 0.7887.

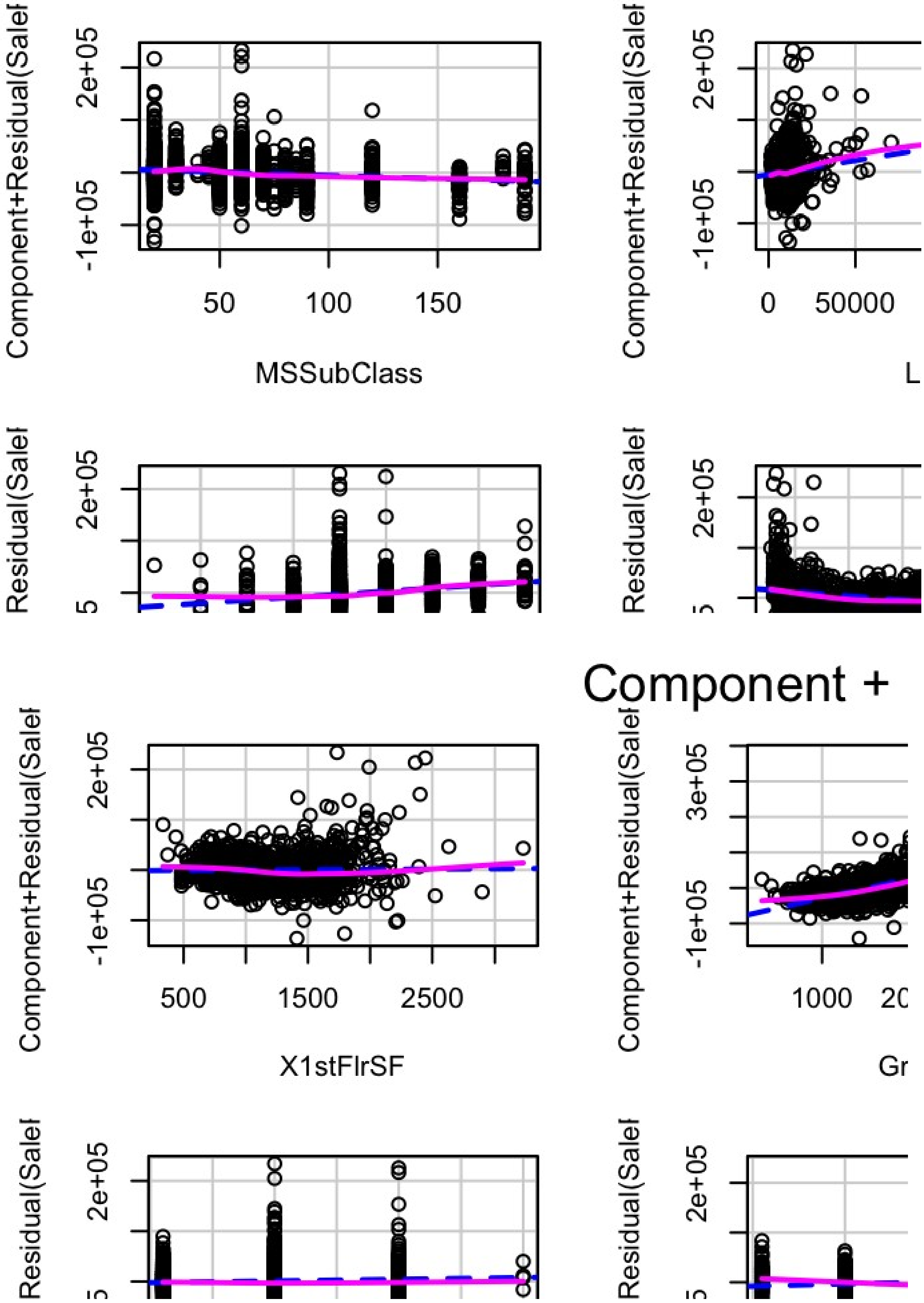
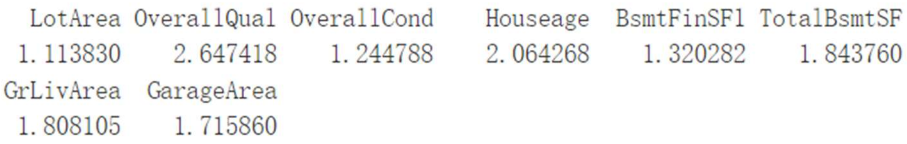


Figure 7: CrPlots

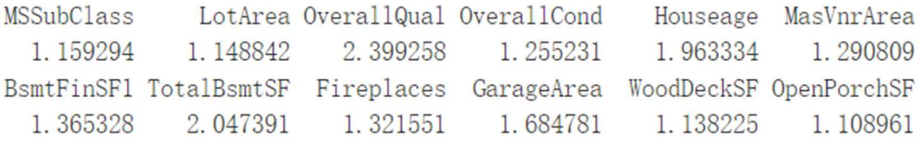
Model Selection

Firstly, we used two methods to select the relevant variables. By using Mallows CP, we found that the appropriate variables for constructing the model are “LotArea”, “OverallQual”, “OverallCond”, “Houseage”, “BsmtFinSF1”, “TotalBsmtSF”, “GrLivArea” and “GarageArea”. Because of the limit nbest we can use, we use AIC step to look for the siginificant variables. And by using StepAIC, we found the significant variables are “LotArea”, “OverallQual”, “OverallCond”, “Houseage”, “MasVnrArea”, “BsmtFinSF1”, “TotalBsmtSF”, “Fireplaces”, “GarageArea”, “WoodDeckSF” and “OpenPorchSF”, which obtained three more variables than Mallows CP. From Figure 8 and 9, we can see that the under the VIF test, both of 2 models do not show multilinearity. Then, we find that the Mallows CP model has a lower AIC and BIC value. Therefore, we should select the variables collected by Mallows CP in the further analysis



```
    LotArea OverallQual OverallCond    Houseage  BsmtFinSF1 TotalBsmtSF
   1.113830    2.647418    1.244788    2.064268    1.320282    1.843760
  GrLivArea  GarageArea
   1.808105    1.715860
```

Figure 8: VIF of Mallows CP Model



```
 MSSubClass     LotArea OverallQual OverallCond    Houseage  MasVnrArea
   1.159294    1.148842    2.399258    1.255231    1.963334    1.290809
 BsmtFinSF1 TotalBsmtSF  Fireplaces  GarageArea  WoodDeckSF OpenPorchSF
   1.365328    2.047391    1.321551    1.684781    1.138225    1.108961
```

Figure 9: VIF of AIC Model

Secondly, we checked whether the higher power terms needed to be added into our model. Because the p-value for power=2 RESET test is very small, adding square terms are necessary. After adding these higher power terms, we used StepAIC again to select the influential variables and obtain the final model (Table 1).

Parameter	Estimate	Std. Error	t	p	Sig.
LotArea	1.774e+00	1.647e-01	10.766	< 2e-16	***
OverallQual	-5.076e+04	4.408e+03	-11.516	< 2e-16	***
OverallCond	9.631e+03	6.936e+02	4.750	13.885	***
Houseage	-1.020e+03	1.090e+02	-9.360	< 2e-16	***
I(LotArea^2)	-6.777e-06	9.690e-07	-6.994	4.07e-12	***
I(OverallQual^2)	9.658e+03	5.594e+02	17.263	< 2e-16	***
I(Houseage^2)	3.126e+00	8.023e-01	3.896	0.000102	***
I(BsmtFinSF1^2)	2.084e-02	1.470e-03	14.177	< 2e-16	***
I(TotalBsmtSF^2)	1.019e-02	9.329e-04	10.923	< 2e-16	***
I(GrLivArea^2)	1.574e-02	4.924e-04	31.965	< 2e-16	***

Table 2: The Model of Multilinear Regression

From this table, we can know that the house price will increase if the overall condition is good and lot area is larger. The house price will decrease if the house is old.

The variables of LotArea and OverallQual have two terms in the regression, so the influence from these variables will be bigger.

Model Evaluation

After we obtained the model, we checked its performance by looking at the residual plot. Because the average value is almost around zero and R^2 is rather high (0.89), we think our model performs well.

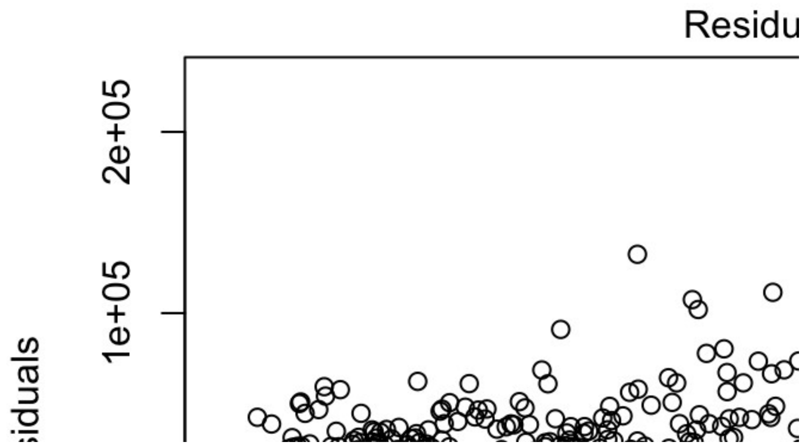
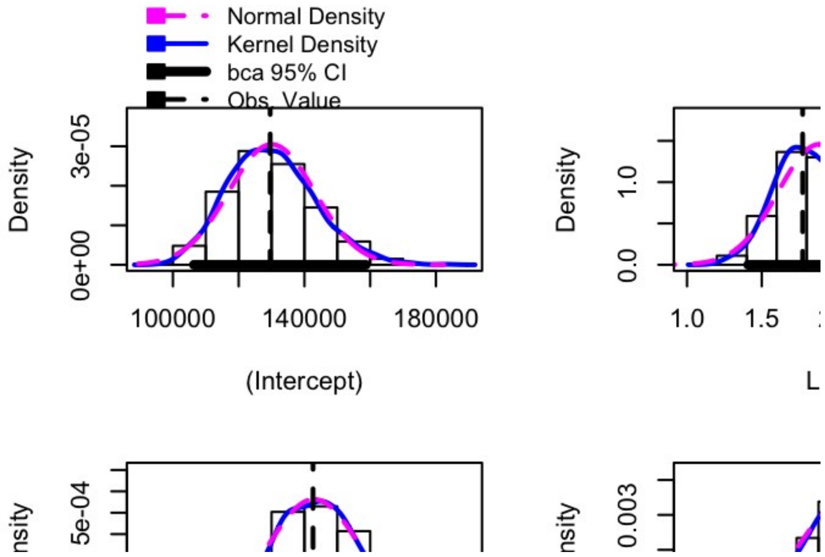


Figure 10: Residuals Plot of Our Constructed Model

Moreover, we also used the method of bootstrapping and cross validation to test the robustness of our model. Figure 11 shows the result of bootstrapping. Here we noticed that the blue curves almost overlapped with the pick curve, which meant that the distribution of our results almost followed the normal distribution and our model was robust.



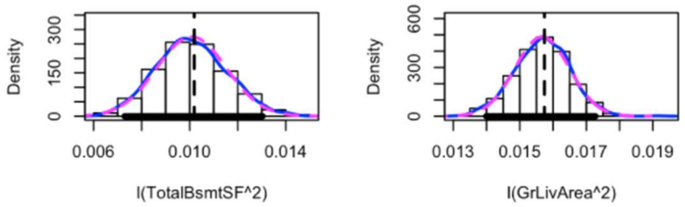


Figure 11: Bootstrapping of the Constructed Model

Figure 12 was the result of 5-fold cross validation. Here we found that the five lines almost had the same slop and looked very similar. In addition, the MSE for each fold was 6.8e+08, 6.33e+08, 6.82e+08, 6.97e+08 and 7e+08, which did not change too much. Therefore, this method also indicated our model had good performance.

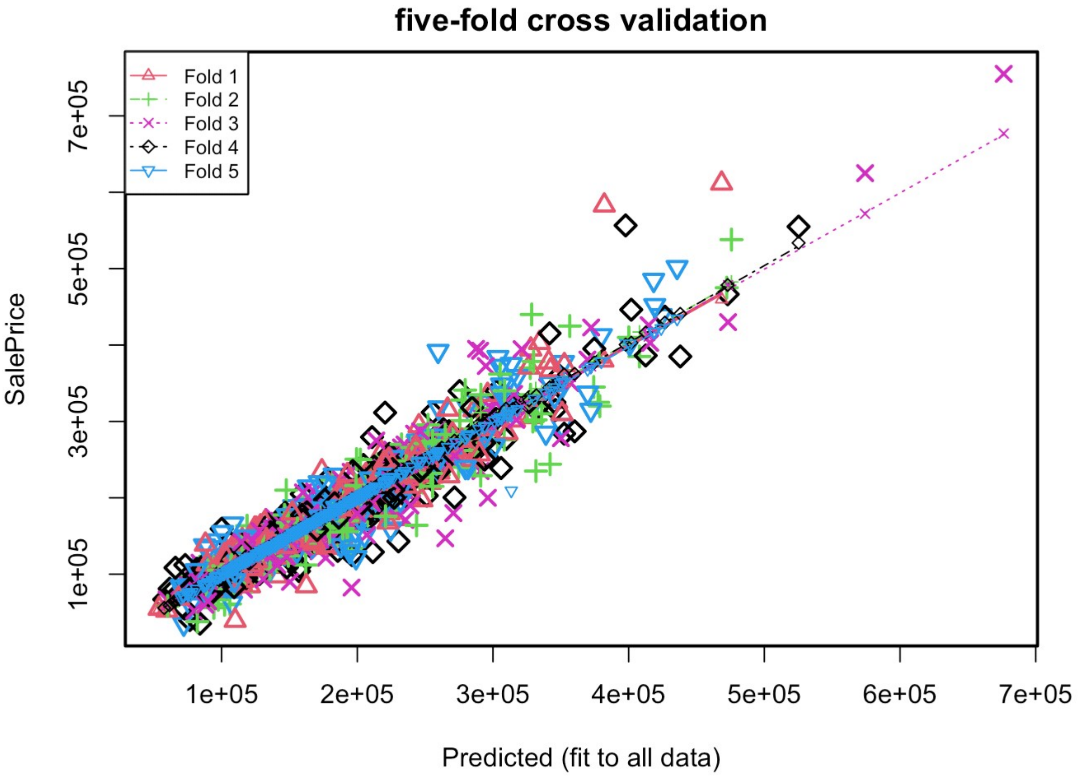


Figure 12: 5-Fold Validation of the Constructed Model

4.2 Lasso

Algorithm Introduction

Lasso is the abbreviation of Least Absolute Shrinkage and Selection Operator, which is a linear regression method combined with L1 regularization in absolute value.

$$L(\beta) = \|Y - X\beta^T\|_2^2 + \alpha\|\omega\|_1$$

The L1 penalty norm could set the partially learned feature weight to 0, thus achieving the purpose of sparsity and feature selection.

Model Selection

In order to choose the optimal scale of the regularization, we draw the plot of the relationship of coefficients weights and penalty parameter alpha in Figure 13. It could be found that when the penalty alpha is larger than 1, the coefficients would lose the efficacy.

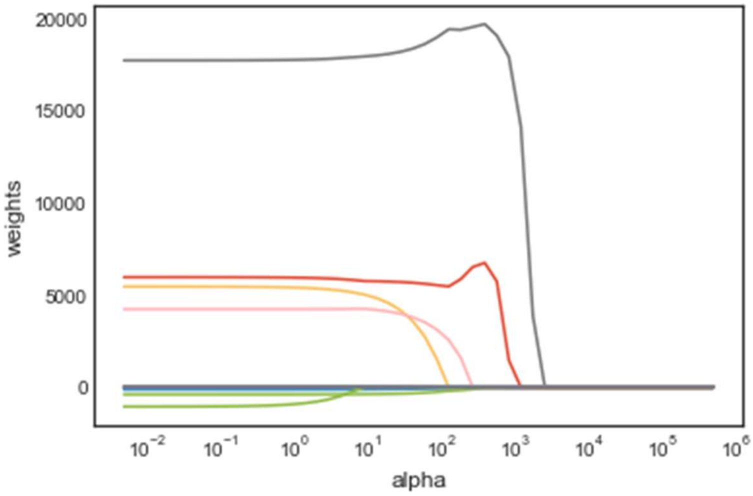


Figure 13: Weights of coefficients in Lasso

Model Evaluation

The general Lasso method with penalty equals 1 has the outcomes that the training score is about 0.8487 and testing score is about 0.8620, and mean square error is about 884161841.

In order to evaluate the performance of Lasso model, we try the cross validation and the optimal penalty is about 14.72. Then training score is about 0.8482 and testing score is about 0.8608, which the difference is eliminated by CV method. The MSE also increase to 884182428. It's worth mentioning that the model includes 15 independent variables, while the general lasso method uses all 18 variables. Because the scores are similar, there’s no overfitting problem.

4.3 Ridge

Algorithm Introduction

Ridge is dedicated to a total of linear biased estimation of regression data analysis method. The main difference of lasso and ridge is that they use different penalty norm. Lasso uses L1 norm which is the absolute value and ridge uses L2 norm which is in quadratic form.

$$L(\beta) = \|Y - X\beta^T\|_2^2 + \alpha\|\omega\|_2^2$$

Model Selection

Also plot the relationship of coefficients weights and penalty parameter alpha in Figure 11. It could be found that when the penalty alpha is larger than about 1, the coefficients would totally lose the efficacy.

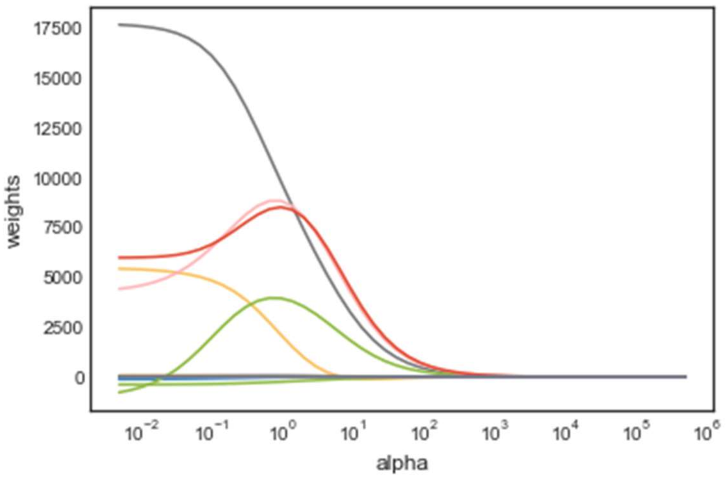


Figure 14: Weights of coefficients in Ridge

Model Evaluation

In ridge we use cross validation method to determine the optimal penalty parameter, alpha, which is about 0.03, and the training score is about 0.8481, the testing score is about 0.8602, also similar, so no overfitting problem, either. The mean squared error of the testing sample is also about 885561319. And ridge model doesn’t drop any features. There’s very little difference between the outcomes of two models, while lasso does a little bit better.

4.4 Elastic Net

Algorithm Introduction

The algorithm of elastic net model is the combination of Lasso and Ridge, it contains two types of the regularizations. In the loss function, the sum of the weights of L1 norm and L2 norm equals 1.

$$L(\beta) = \|Y - X\beta^T\|_2^2 + \alpha\|\omega\|_2^2 \quad + \alpha\|\omega\|_1$$

Model Selection

We also use CV to find the optimal value of alpha and the penalties ratio. The ratio is 0.5, and optimal alpha is 0.005. The plot on the right is the contract of normal elastic net model and the model with CV. In plot 15, the scale of coefficients of CV method, which the orange triangles showed in the plot, is larger than the normal model. And, the accuracy result is better.

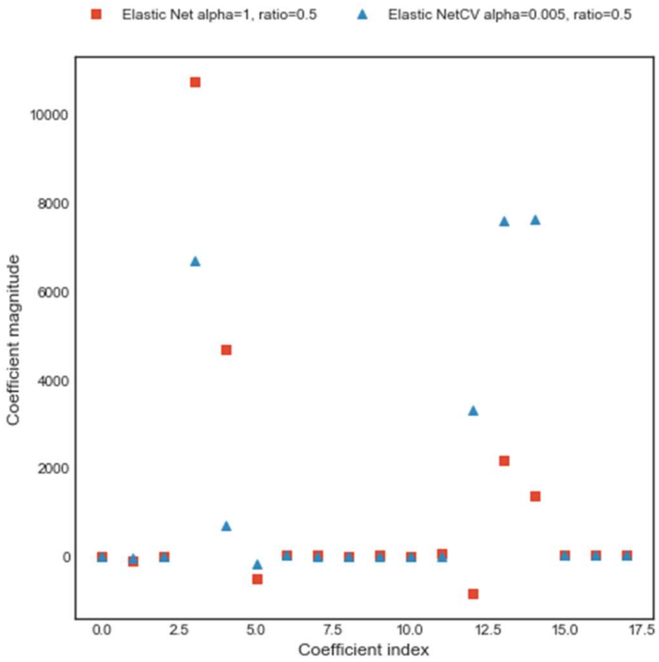


Figure 15: Coefficients of normal ElasticNet and CV Elastic Net

Model Evaluation

We select the elastic net model after crossing validation. Then the training score is about 0.7293 and testing score is about 0.7098. the mean squared error is about 1860025888. There’s no overfitting problem in elastic net model, either.

4.5 Principal Component Analysis

Algorithm Introduction

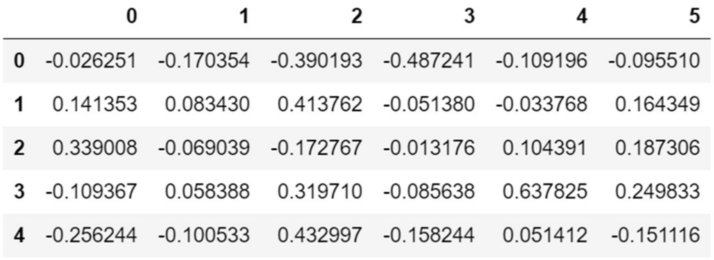
When number and dimensions of right-hand side variables are too large, estimation with ordinary least square is not robust anymore. Instead, there are other novel methods to reduce dimensions of predictors before estimation. One of the commonly used approaches is Principal Component Analysis (PCA).

The main idea of PCA is to utilize mathematical algorithms, such as orthogonal transformation, to combine original variables into a new set of independent synthetic variables where information is cohered. And then according to different circumstance, to choose specific number of reconstructed variables to do estimation and analysis.

There are several functions of PCA which are reducing dimension of data set, solving multicollinearity, visualizing high-dimensional data and on forth. Specifically, in this paper, because predictors may be highly correlated with each other, the main purpose of PCA is to solve multicollinearity problem.

Due to various factors may change housing price, PCA is an extraordinary method to integrate information. Next, we will train PCA model and test the performance.

Model Evaluation



	0	1	2	3	4	5
0	-0.026251	-0.170354	-0.390193	-0.487241	-0.109196	-0.095510
1	0.141353	0.083430	0.413762	-0.051380	-0.033768	0.164349
2	0.339008	-0.069039	-0.172767	-0.013176	0.104391	0.187306
3	-0.109367	0.058388	0.319710	-0.085638	0.637825	0.249833
4	-0.256244	-0.100533	0.432997	-0.158244	0.051412	-0.151116

Table 3: Normalized data

In PCA analysis, we divided it into three steps.

The first step is to normalize data into normal distribution with mean 0 and variance 1. The new data set is shown in Table 3. And then to reduce dimension with pca.fit function in python, where we set n_components parameter as "MLE". MLE algorithm is to select principal component features according to the variance distribution of features. In this project, variables stay the same with 17.

In the second step, we plot MSE and accuracy score results in training data with 10-folds Cross Validation, shown in Figure 16 and 17 respectively. In the two plots, we can notice that after 2, MSE and accuracy become stable and model performs quite good. And in order to avoid over- fitting problem, we choose number of principal components as 2.

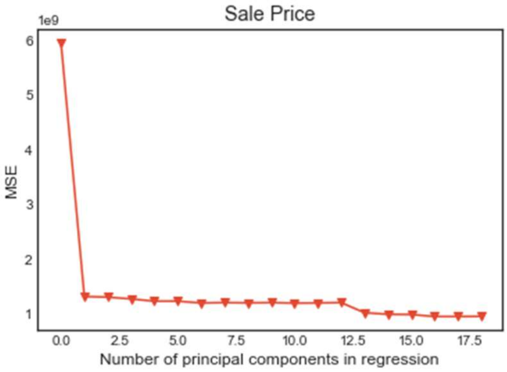


Figure 16: MSE of PCA 10-folds CV

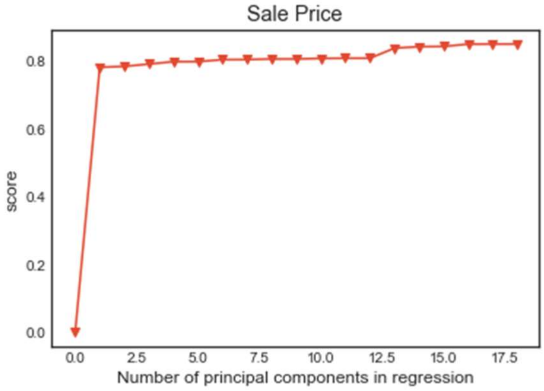


Figure 17: Accuracy of PCA 10-folds CV

Step 3 is to train PCA regression model and calculate accuracy scores with previous selected parameter. The results are 0.800809 and 0.836609 respectively, which can be regarded as an indicator to compare with other models

4.6 Random Forest Algorithm Introduction

The basic algorithm of Random Forest is Decision Trees which is a prediction Machine Learning method with several branches of decision splits. In each step m, we need to decide which group to split, based on which variable ($X!$), using what threshold (t). Imagine we have a set of data and want to get the label of them. We will start from the bottom of tree, i.e. all data information, until we meet the first branch.

In each branch, there is a condition to decide which way to continue. Repeat this process until reach the top of this tree where prediction can be given.

Based on Decision Trees, Random Forest is developed. Different with Decision Tress, when building these trees, Random Forest randomly chooses m predictors from the full dataset for splitting. In this paper, we will also use random forest decision tree to regress house prices and exam model performance.

Model Evaluation

First, we adjust the maximum number of depth. In the accuracy plot (Figure 18), we can notice that when max depth = 15, there is a small peak in testing sample and the gap between training and testing score is smallest. Therefore, we select 15 as depth number. Under this circumstance, training score and testing score are 0.9768 and 0.8655 respectively.

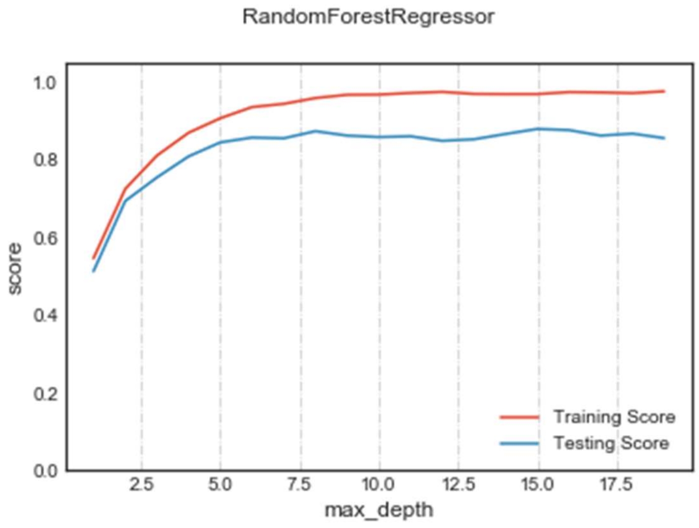


Figure 18: Adjust maximum number of depth

4.7 Support Vector Machine

Algorithm Introduction

For regression purpose, there is a branch of Support Vector Machine called Support Vector Regression (SVR), which shares the same margin maximum algorithm with classification method. SVR tends to estimate true value of data instead of classifying into groups.

The algorithm of SVR estimation is shown below:

Minimize $\frac{1}{2}\|\omega\|^2$

Subject to $|y - f(x, w) - b| \leq \varepsilon$

where x is vector with predictors; y is independent values; ω is weight vector; b is constant; ε is threshold.

The most important inner algorithm is transiting input data into target form. In this process, it can use different kinds of mathematical functions that are named as the kernel function. There are various types of kernel function, for instance linear, polynomial, radial basis function (RBF), and sigmoid, etc. Table 2 shows the difference.

SVM	Kernel Function	Parameter
Linear	$k(x_i, x_j) = {x_i}^T x_j$	
Polynomial	$k(x_i, x_j) = ({x_i}^T x_j)^d$	$d \geq 1$, Degree of Polynomial
RBF	$k(x_i, x_j) = \exp\,(-\frac{\|x_i - x_j\|}{2\sigma^2})$	$\sigma > 0$, Width
Sigmoid	$k(x_i, x_j) = \tanh\,(\beta {x_i}^T x_j + \theta)$	$\beta > 0, \theta < 0$

Table 4: Three Kernel functions

Model Evaluation

During the SVR process, first step is to process initial data and divided into training sample and testing sample. And then estimate regression model with training data and calculate accuracy score or error ratios. The last step is to predict testing data with trained model and also calculate performance indicators.

SVM	Training Accuracy	Testing Accuracy
Linear	0.781	0.789
Polynomial	0.773	0.784
RBF	-0.046	-0.057
Sigmoid	-0.046	-0.057

Table 5: SVM Results

In Table 2, it is obvious that SVM with Linear Kernel out-performs than the other models. The accuracy score is 0.781 and 0.789 for training and testing sample respectively.

5. Conclusion and Future Work

According to estimation results of different models, it is safe to conclude that both of multi-linear regression model and novel Machine Learning models can be regarded as a powerful tool to predict house prices. Moreover, it is noteworthy that Random Forest has the most outstanding performance with highest training score and testing score.

In future work, this paper can be improved in several ways.

First, this data set could be too old to reflect recent real estate status because the last house sold in this dataset is in 2010, and the house are mainly in America, New York city. The performance results could be biased and not universal. In other words, model performance may vary with different regions and different time period. Therefore, more recent and multi-region house price data can be added in the training process.

Moreover, model optimization is essential for Machine Learning. In this paper, the hyperparameters considered in this paper only account for a small part. Hence, researchers could devote more effort to adjust parameters and test the performance with difference parameter combinations. Except for individual models, hybrid models can also be trained in stock price prediction.

Last but not least, interpretability is a big problem for Machine Learning methods even though they show great result to predict data set. As a consequence, computer-based model can be improved to explain the predicting result.

Reference

[1] Cock, Dean De. “House Prices - Advanced Regression Techniques.” Kaggle, www.kaggle.com/c/house-prices-advanced-regression-techniques/overview.