House Price Prediction Using Lasso Algorithm

Final Project for Introduction to Data Science

Student: Lina Mi

Instructor: Ricky Sethi

Contents

Introduction

Because of dynamic development of domestic and international economics, the change of the future economy is hard to predict. In this rapid developing economic environment, real estate properties are more attractive option for safe investment especially for a common person, compared with other investment alternates such as stock and fund. Real estate has property of keeping its value from decline quickly in the worse economic situation and increasing its value during prospering economic environment, so the investors will not lose their investment rapidly in bad economy and have good rewards in good economy.

When investing on real estate, what the investors concern most is the value of real estate. The price of real estate, however, is influenced by many factors and hard to estimate. Predicting the value of real estate precisely and identifying the factors that affect the value of real estate correctly could provide investors with deep insight on prospering house and help investors avoid overpaying for the candidate properties. So it is important to develop practical data mining models to give valuable estimation on the real estate market and identify the influential factors on the value of each property based on the available data.

In this project, a lasso model is chosen to analyze the dataset which contains sale prices and features of houses sold between May 2014 and May 2015 for King County, including Seattle, and then give prediction on the prices of houses and identify the important factors that influence the values of properties. The test result shows that the model is effective on identifying the influential factors and providing good estimation on the value of properties.

Research goals and Data preprocessing

Research Goals

In order to provide real estate investors who have little knowledge of real estate with profound insight on properties in the area of Kings County, one of goals of this research is to answer the question: what are most important features that determine the value of properties in Kings County? The other research goal concerns the question of what are the fare prices of houses in Kings County in the period from May 2014 to May 2015 under the given market situation? Answering this question successfully could provide a valuable reference to investors, help investors avoid overpaying or identify the properties with potential increase of value.

Data Preprocessing

Data prescription

The dataset used in this project contains 21613 observations of houses sold between May 2014 and May 2015 in area of King County, including Seattle. In each observation, 18 features of house, house ID, sale price and sale date are included. The variables in the dataset are described in the following table.

Feature name	Type of data	description
ID	Numeric	The house identification number
Date	Factor	The date house was sold
Price	Numeric	The price at which the house was sold
Bedrooms	Numeric	Number of bedrooms in the house
Bathrooms	Numeric	Number of bathrooms in the house
Sqft_living	Numeric	Square footage of the living room in the house
Sqft_lot	Numeric	Square footage of the lot
Floors	Numeric	Total floors(levels) in the house
Waterfront	Numeric	House which has a view to a waterfront
View	Numeric	Has been viewed
Condition	Numeric	How good the condition is (overall)
Grade	Numeric	The level of how the house is graded
Sqft_above	Numeric	Square footage of house apart from basement
Sqft_basement	Numeric	Square footage of the basement
Yr_built	Numeric	Year the house was built
Yr_renovation	Numeric	Year when the house was renovated

Zipcode	Numeric	The zipcode of house address
Lat	Numeric	The latitude coordination of the house address
Long	Numeric	Longitude coordination of the house address
Sqft_living15	Numeric	Living room area in 2015(implies some renovation), this might or might not have affected the lot size of the house
Sqft_lot15	Numeric	Lot size of the house in 2015(implies some renovations)

Data preprocessing

Among these variables, the types of some variables are not appropriate for analyzing directly, it is necessary to transfer these types into the ones suitable for data analysis.

1) 'Date' variable: the type of this variable is factor which is not suitable for data analysis, so I convert it as following: first it is converted into date type in R, and then the date type is transferred to numeric type, since the sale date in the dataset is between May 2014 and May 2015, the variable of sale date is expressed as number of days to the earliest date in the dataset:

```
dat<-read.csv("kc_house_data.csv")
dates<-gsub("T[0-9]+$","", dat$date)
dates<-as.Date(dates, format="%Y%m%d")
daten<-as.numeric(dates-min(dates))
dat<-data.frame(dat, daten=daten)
```

2) 'ID' variable: in the dataset, there are some observations with exactly same ID, it means that the same house was sold more than one times during the period of May 2014 and May 2015. These duplicated IDs are apparently not useful for the analysis, so the duplicated IDs are removed and only the observations with duplicated ID and largest value of variable DATEN were kept, that means only the observations of the houses which were sold at latest time are kept.

```
dat<-dat[order(dat$id, -dat$daten), ]
ids <- dat[,1]
dat<-dat[!duplicated(dat$id), ] #this will remove the duplicated record
```

3) ‘zipcode’ variable: in the dataset, the type of variable zipcode is numeric, considering the nature of zipcode feature, the ‘numeric’ type is converted into ‘factor’ type.

```
datfinal$zipcode<-as.factor(datfinal$zipcode)
```

4) ‘yr_renovated’ variable: since for newly built houses and old houses which have never been renovated since they were built, the values of this variable are same, this definitely can not reflect the real features of properties. I convert it the variable of ‘nyrs_since_last_renovate’, which reflects the number of years since last renovate to 2015, namely 2015-yr_renovated, if the value of yr_renovated is zero, then 2015-yr_built.

```
yrs<-dat$yr_renovated
yrs[yrs==0]<-dat$yr_built[dat$yr_renovated==0]
dat <-data.frame(dat, nyrs_since_last_renovation=as.numeric(2015-yrs))
```

5) Variables of ‘lat’ and ‘long’: since feature of ‘zipcode’ has indicated the location of house, variable ‘lat’ and ‘long’ which are latitude of house and longitude of house respectively are not necessary for data analysis. Therefore these two variables were not included in the final dataset

```
ycol<-"price"
xcols<-c('bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'nyrs_since_last_renovation', 'zipcode', 'sqft_living15', 'sqft_lot15','daten')
datfinal<-data.frame(price=dat$price, dat[, xcols])
```

the final dataset used for analysis is like the following:

```
 str(datfinal)
'data.frame':  21436 obs. of  18 variables:
 $ price          : num  300000 647500 400000 235000 402500 ...
 $ bedrooms       : num  6 4 3 3 4 4 5 4 3 4 ...
 $ bathrooms      : num  3 1.75 1 1 2 2.75 1.5 2.5 1 2 ...
 $ sqft_living    : num  2400 2060 1460 1430 1650 2220 1990 2540 1340 1980 ...
 $ sqft_lot       : num  9373 26036 43000 7599 3504 ...
 $ floors         : num  2 1 1 1.5 1 1 1 2 1.5 1.5 ...
 $ waterfront     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ view           : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
$ condition        : num  3 4 3 4 3 5 3 3 4 2 ...
$ grade            : num  7 8 7 6 7 7 7 9 5 6 ...
$ sqft_above       : num  2400 1160 1460 1010 760 1170 1990 2540 1340 1980 ...
$ sqft_basement    : num  0 900 0 420 890 1050 0 0 0 0 ...
$ yr_built         : num  1991 1947 1952 1930 1951 ...
$ nyrs_since_last_renovation: num  24 68 63 85 2 64 55 10 70 91 ...
$ zipcode          : Factor w/ 70 levels "98001","98002",..: 2 64 64 65 60 60
67 47 21 31 ...
$ sqft_living15    : num  2060 2590 2250 1290 1480 1540 1860 2360 1340 1360
...
$ sqft_lot15       : num  7316 21891 20023 10320 3504 ...
$ daten            : num  355 6 101 334 321 332 298 68 194 186 ...
```

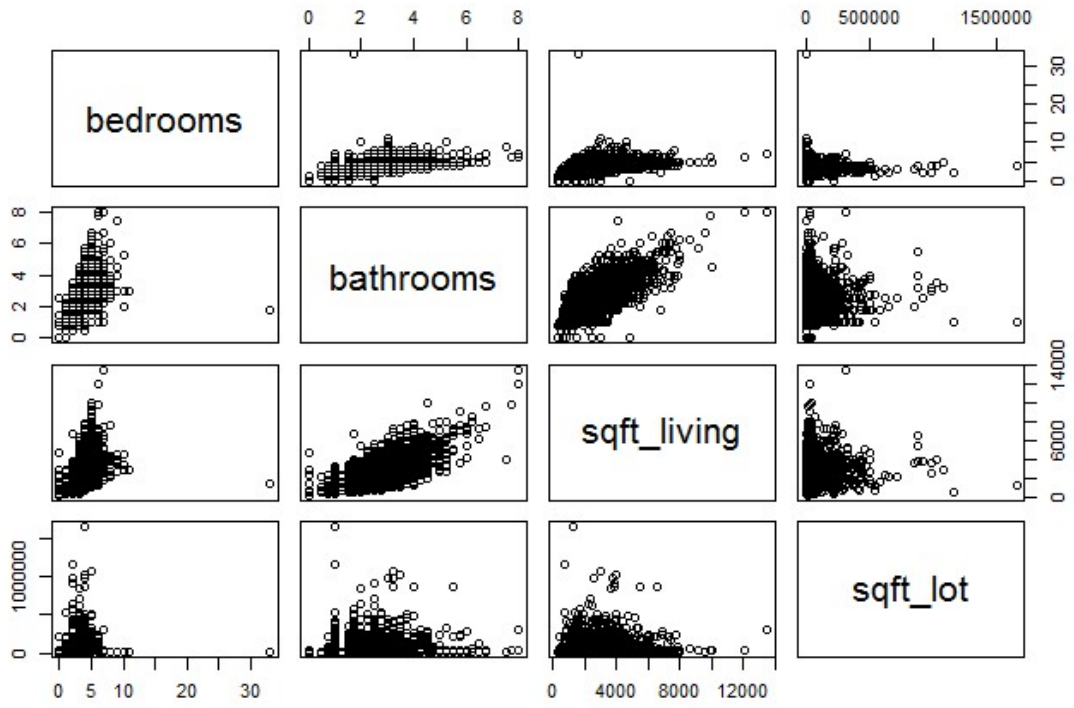
Analysis on Variables

The degree of correlation among variables has critical effect on the model we used to analyze the dataset, so before choosing the model, I first check the collinearity between each pair of variables.

```
pairs(datfinal[, xcols])
```

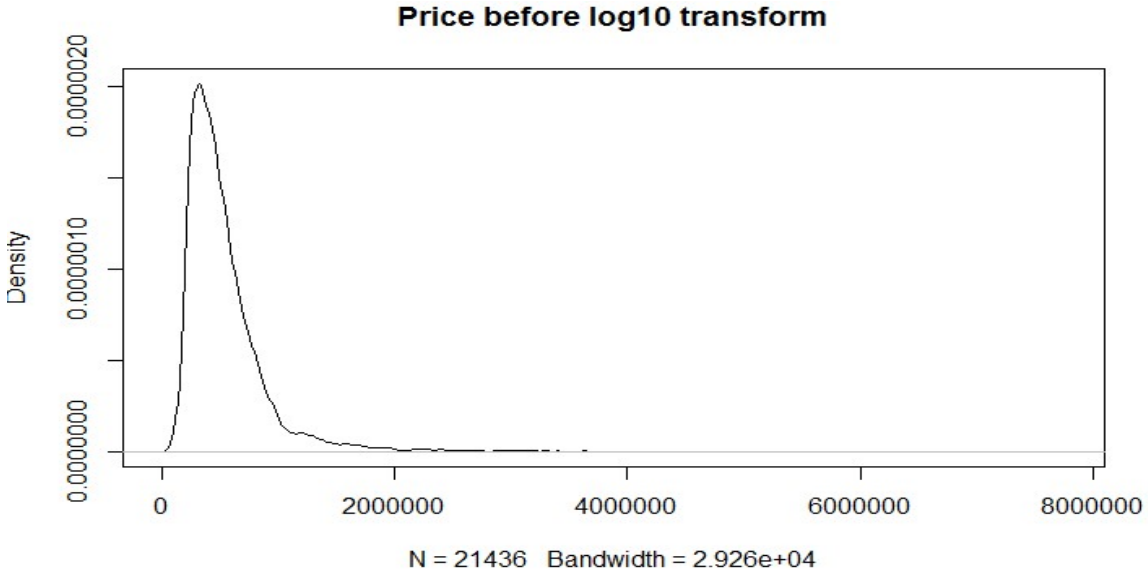
since there are 18 variables in the final dataset, the plot on correlation between each pair of variables will be too large to show here, I only show the plot on the correlations of first four variables here.

```
pairs(datfinal[, xcols[1:4]])
```

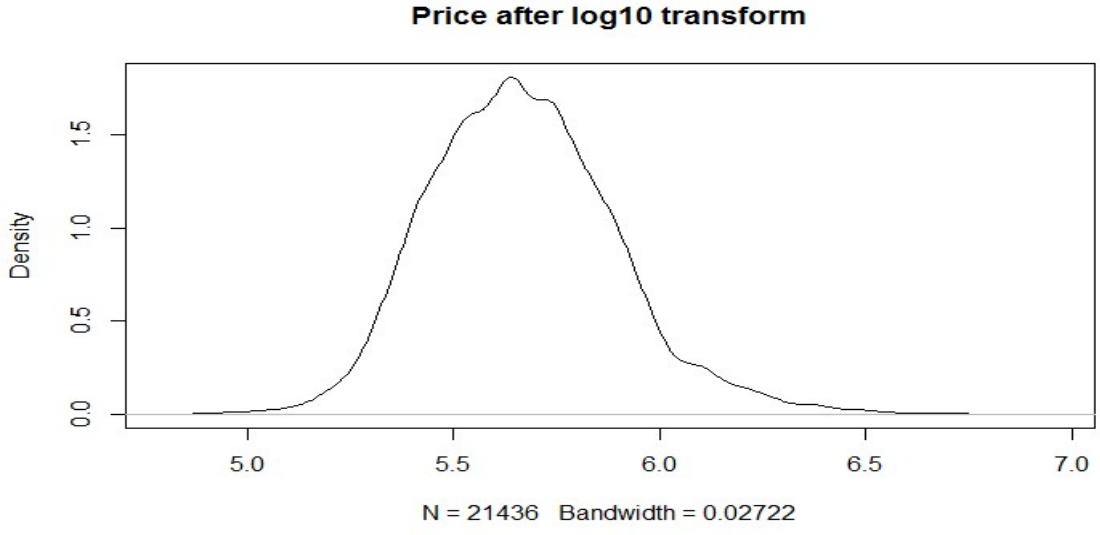


According the figures above, it is easy to tell that the pair of variable ‘bedrooms’ and ‘bathrooms’ and the pair of variable ‘bathrooms’ and ‘sqft_living’ are highly correlated (linear relationship), that means the model of multiple linear regression will not work for this dataset analysis.

In the dataset, ‘price’ is dependent variable and chosen as goal variable, it is necessary to examine the distribution of ‘price’ first.



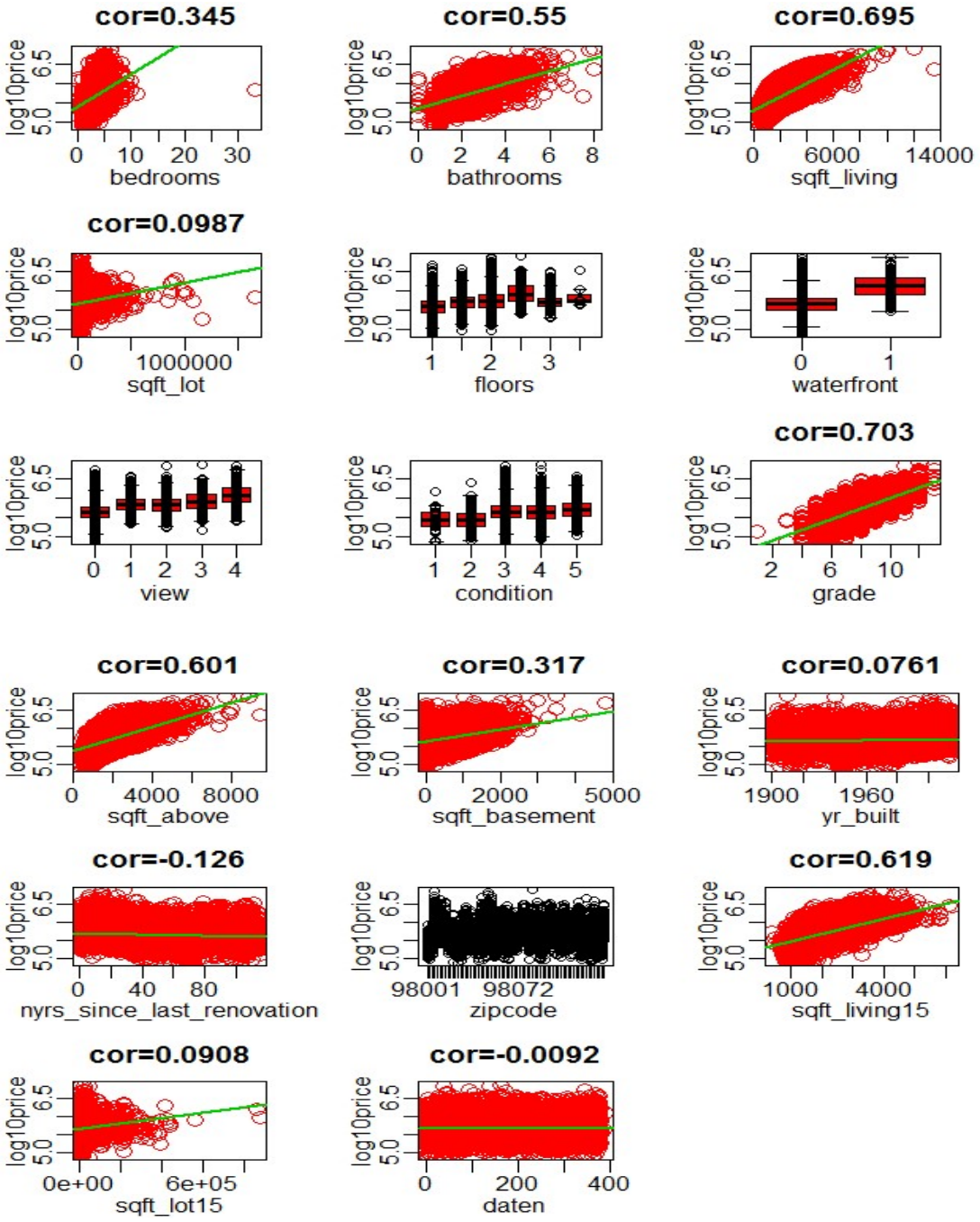
The above figure shows that the distribution of ‘price’ variable is not normal distribution, so the value of variable ‘price’ should be manipulated to get normal distribution. Here, the ‘price’ value are taken logarithm with 10 as base, the distribution of ‘price’ variable after taken logarithm is shown in following:



From the above figure, it can be seen that the distribution of 'price' variable after taken logarithm become normal and the manipulation is appropriate.

Marginal analysis

Before using the model to analyze the dataset, it needs to check the correlations between goal variable 'Y and each of the independent X variables, the results are shown in the following figures:



It can be seen from above figures that some X variables are highly correlated with dependable variable Y (such as variable ‘grade’ and ‘sqft_living’), some X variables are much less correlated with Y. That means that different X variable has different degree of influence on variable Y.

Data Analysis Using Lasso Model

Lasso Model

Considering that the independent variables in the dataset are highly correlated and the mothod of traditional multiple linear regression is not appropriate for this situation. Hence the Lasso Model is employed to analyze the dataset and give the prediction on the prices of houses. Lasso is acronym of least absolute shrinkage and selection operator, Lasso method is type of regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Compared with multiple linear regression, Lasso Model has advantage of allowing to input variables as many as you want and allowing the high collinearity among independent variables.

1) Lambda parameter

   Before employing the Lasso Model, it is necessary to get the best value of the parameter of 'lambda'.

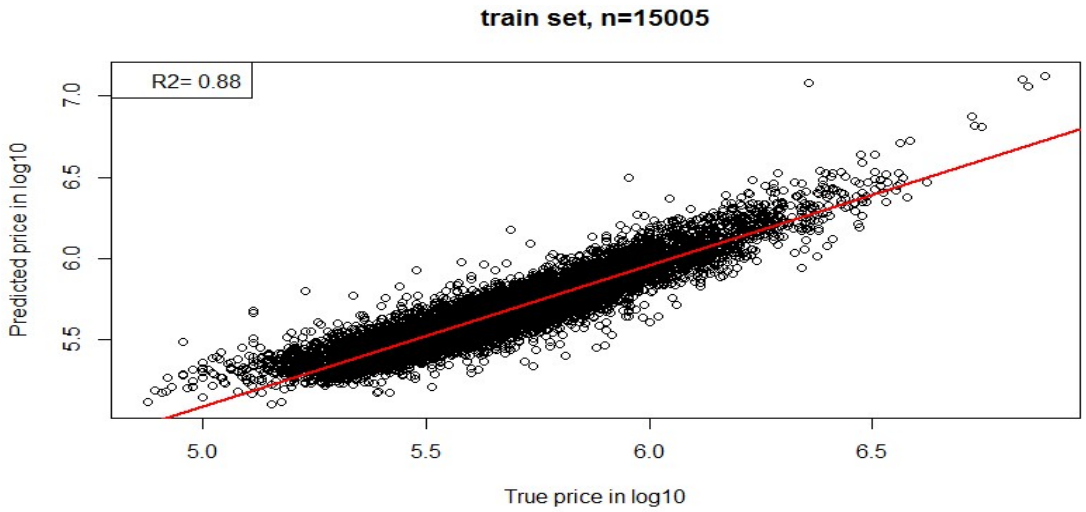
```
cvs  =  cv.glmnet(x=modx[train_ids, ],y=y[train_ids],  alpha=1)
lam<- cvs$lambda.1se
lam 0.0002873867
```

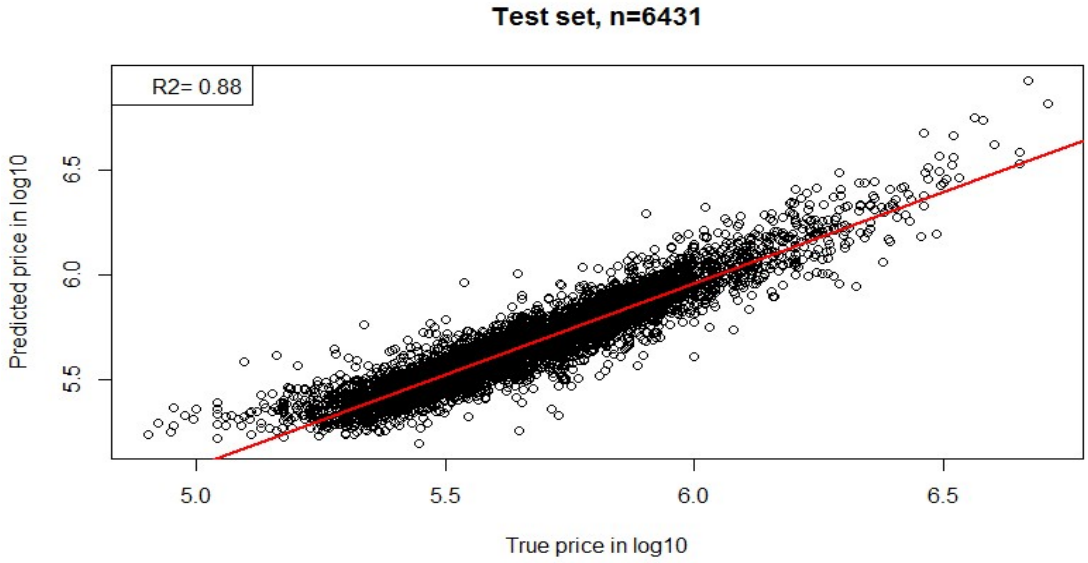
2) Applying Lasso Method and Result Interpretation

   With obtained value of parameter 'lambda', the lasso model is applied on the training dataset.

```
fits<-glmnet(x=modx[train_ids, ],y=y[train_ids],  alpha=1, lambda=lam)
predtrain<- predict(fits, newx=modx[train_ids, ], type="response")
predtest<-predict(fits, newx=modx[test_ids, ], type="response")
```

   the prediction results for training dataset and testing dataset are show as following:





From the above figures, it can be seen that the model fit not only training data but also testing data very well. In the above figures of analysis results, R square is a statistical measurement of how close the data are to the fitted regression line. It is also known as the coefficient of multiple determination for multiple regression. It is defined as the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%, In general, the higher the R-squared, the better the model fits your data. In above two figures, the R square of both datasets are same (0.88), that means model generated by Lasso algorithm fit not only training dataset but also testing dataset very well, and the model could provide relatively accurate prediction on the house price in King County.

Part of the coefficient for independent variables generated by the model are listed to show the importance of each feature of house to the price:

zipcode98002	-0.106
zipcode98032	-0.098
zipcode98023	-0.097
zipcode98003	-0.082
zipcode98038	-0.018
floors	-0.008
yr_built	-0.001
nyrs_since_last_renovation	0.000
daten	0.000
bedrooms	0.002
bathrooms	0.017
zipcode98146	0.020
zipcode98010	0.022
condition	0.024
view	0.025
grade	0.042
zipcode98008	0.182
zipcode98007	0.184
zipcode98144	0.185
zipcode98136	0.191
waterfront	0.206
zipcode98116	0.223
zipcode98199	0.270
zipcode98105	0.297
zipcode98102	0.305
zipcode98119	0.315
zipcode98109	0.323
zipcode98112	0.335
zipcode98004	0.380
zipcode98039	0.424

The larger the absolute value of coefficient is, the more important the feature is. It can be seen that the most important feature affecting price of the house is some zipcode, it is reasonable because zipcode reflected the district where the property is located, so some zipcode have almost deterministic effect on the value of the house. The importance of some features(such as nyrs_since_last_renovated), however, were not correctly reflected in the results. This defect of algorithm need to be further examined.

Conclusion

The lasso model works very well on predicting the price of houses in King County using the available dataset, this fact can be demonstrated by the results figures which showed that the predicted data fit the regression line very well. It also successfully identified the some property features that have great influence on the price of the properties. It also failed, however, to identify some house characters which have apparent influence on the value of the properties. The reason for this defect needs further explored, application of Lasso algorithm may need to be deeply examined.

The constrains on the dataset could also contribute the inaccuracy of the prediction results, one of this limitation is the time span that the data were collected. Short coverage of time span (one year) can not provide deep insight on the influential features and may lead some features appear more important than they should be. The other constrain on the dataset is locality of the dataset. Since the data were collected only within King County, the irrationality of results may reflect the specialty of properties of King County.

References

- Kaggle.com
- Data Mining With Rattle and R- Graham William
- Data Mining –concepts, Models & Techniques.
- R Data Mining -Yanchang Zhao
- https://en.wikipedia.org/wiki/Lasso_(statistics)
- https://drsimonj.svbtle.com/ridge-regression-with-glmnet
- http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- http://ricardoscr.github.io/how-to-use-ridge-and-lasso-in-r.html

Data Analysis reference

Excel dataset

kc_house_data.csv

R Code

project_house_price_prediction_Lasso.R