The George Washington University

Data Mining Report

2014-15 House Price
in King County, WA Data Analysis

Submitted by

Weike Zhou, Yixuan Yang

Machine Learning and Data Ming DNSC 6279

Table of Contents

Summary

This report is mainly focused on regression analysis and classification on the data about housing price in King County, WA from May 2014 to May 2015. We want to predict future prices based on the different factors like size of living space, number of bathrooms, view, etc. And we want to build a classifier to tell the range of prices. In the preparation phase, we cleaned data and did exploration. Then we tried different models and evaluated them to come up with business perspectives, weakness and possible improvements.

Weike Zhou has contributed 1.2.2-1.2.3,Part2,3.2,4.1.2,4.2, 4.3.1 parts and Yixuan Yang has contributed 1.1,1.2.1,Part2, 3.1, 4.1.1,4.2, 4.3.2 parts and we were in charge of corresponding presentation parts of the report.

Part 1: Introduction

1.1 Background and Business Understanding

Housing price is a very important index which can reflect the economic and social development level and situation of a certain region or city. It is of great theoretical value and practical meaning to study important factors contributing to price. In addition, housing price prediction and identification can help sellers and buyers provide better guidance on the housing market.

1.2 Data Sets Preparation

1.2.1 Data Set

The data is 2014-15 Home Sales in King County, WA (https://geodacenter.github.io/data-and-lab//KingCounty-HouseSales2015/) from GeoDa Data and Lab. It contains Home sales prices and characteristics for King County, WA (May 2014 - 2015) with 21613 observations, 21 variables. Our response is “price”. And the rest of variables are the numeric variables bedrooms, bathrooms, sqft_liv, sqft_lot, floors, sqft_above, sqft_basement, yr_built, yr_renovated,squft_liv15, squft_lot15 and the categorical variables waterfront, view, grade, condition. Here are partial descriptions of variables:

Waterfront	‘1’ if the property has a waterfront, ‘0’ if not.
View	How good the view of the property was (from 0 to 4)

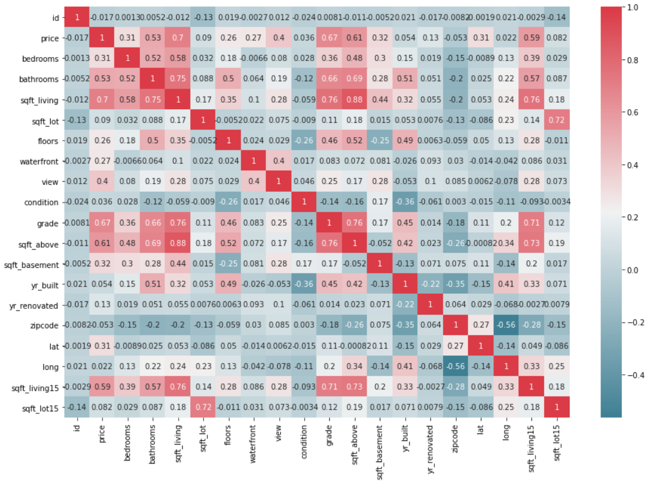
Grade	Construction quality which refers to the types of materials used and the quality of workmanship. Higher grade, higher quality. Score from 4 to 12
Condition	Condition of the house, ranked from 1 to 5
squft_liv15	Average size of interior housing living space for the closest 15 houses, in square feet
yr_renovated	Year renovated. ‘0’ if never renovated

1.2.2 Data Cleaning

For our data cleaning process, we delete some useless variables, which as ID, date. And check for missing value NA. And next, we decided to filter our dataset to delete the top 10% of most expensive homes and then limited the remaining data to just those homes within 2 to 5 bedrooms because I decided to cut it down to help me focus on midrange family homes. And next, we check for the corrected data type of each variable to transfer grade, View, condition, waterfront from numerical to category variables.

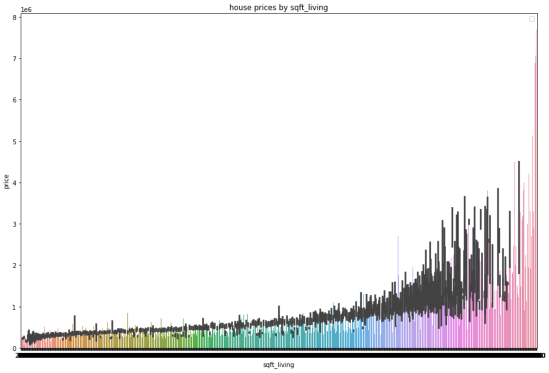
1.2.3 Correlation Testing

A correlation HeatMap is an excellent way to start to figure out what is the correlation between each variable. For example, the red means positive, green means negative. The stronger the color, the larger the correlation magnitude. Therefore, the price has a higher relationship with bathrooms, sqft_living, grade, sqft_above, which is more than 53 percent.

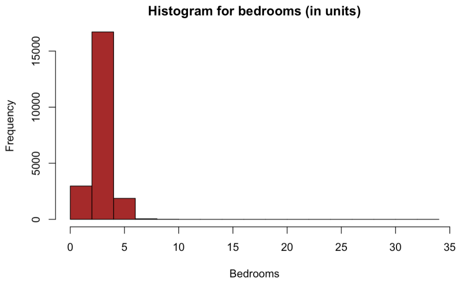


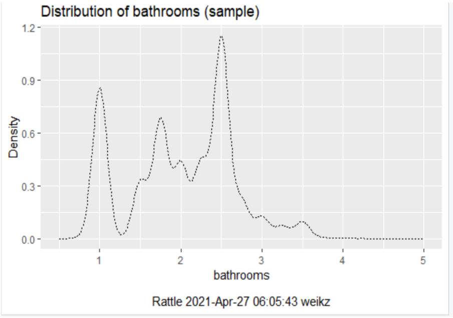
Part 2: Data Exploration

The next picture shows our first graph how the price will influence the sqft_living area; as you can see from the chart, the living area also increased when the price increased.

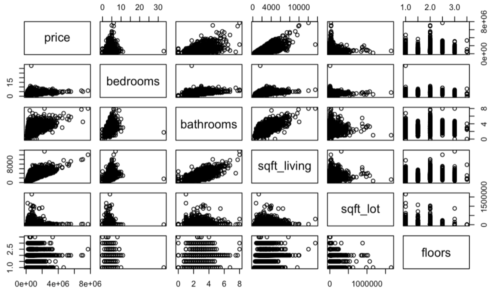


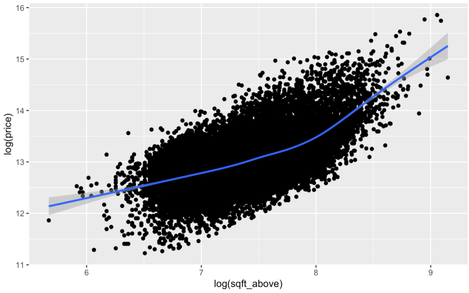
Furthermore, the next two graphs show how many bathrooms and bedrooms people most have in their house. That bedrooms and bathrooms people almost have around 2~3 and 1~3.



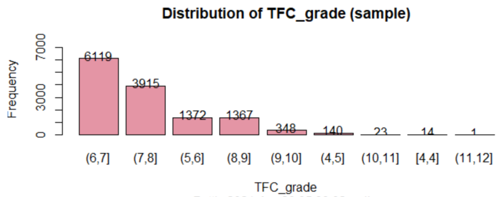


We drew the scatterplot about partial variables and to explore if there exists any linear relationship, especially the relationship between price and other variables. We also chose specific variables sqft_above here to explore, we used a log-log model to avoid the scale problem, as we can see, there does exist a positive linear relationship between them.





The following graph shows the level of building construction and design; as we can see, most of the building design around 6~8, which refer to an average level of construction and design. And there is not a very poor level of building construction from 1~3. And also a few of higher quality of building design from 11~13.



Rattle 2021-Apr-29 05:33:02 weikz

Part 3: Modeling

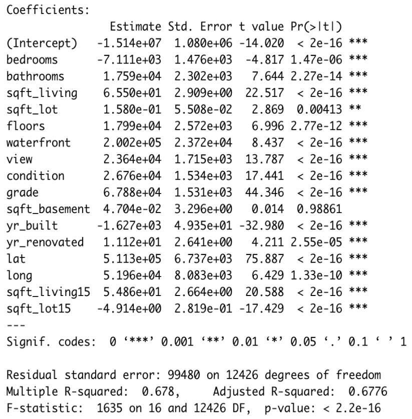
3.1 Regression

3.1.1 Random Forest

In order to check the importance of each variable, it’s a good choice to fit a random forest, and then we put all the 17 independent variables into the model, we got the result below and the test error (MSE) is 4,450,194,793. The rank of importance is lat, sqft_living, long, grade. Then we put 6 variables into the model to decrease variance but unexpectedly, the test error increases to 4,669,091,136 and the rank of importance changes into lat, long, sqft_living, yr_built. Therefore, it can be concluded that the sqft_living is an important factor contributing to the house price without doubt.

3.1.2 Multiple Linear Regression

Based on the previous data exploration, there exists a linear relationship so we fit linear regression. After removing a insignificant variable sqft_above, we ran the model again and then we got the coefficient results below. As the result shows the bedrooms and yr_built has a negative relationship with the house price and the coefficients of bedroom and grade are high so they have a large impact on the price.



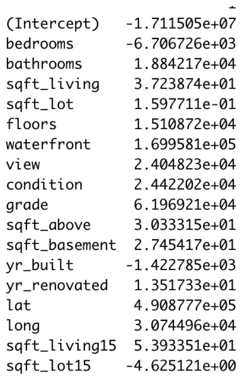
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.514e+07  1.080e+06 -14.020  < 2e-16 ***
bedrooms      -7.111e+03  1.476e+03  -4.817 1.47e-06 ***
bathrooms      1.759e+04  2.302e+03   7.644 2.27e-14 ***
sqft_living    6.550e+01  2.909e+00  22.517  < 2e-16 ***
sqft_lot       1.580e-01  5.508e-02   2.869  0.00413 ** 
floors         1.799e+04  2.572e+03   6.996 2.77e-12 ***
waterfront     2.002e+05  2.372e+04   8.437  < 2e-16 ***
view           2.364e+04  1.715e+03  13.787  < 2e-16 ***
condition      2.676e+04  1.534e+03  17.441  < 2e-16 ***
grade          6.788e+04  1.531e+03  44.346  < 2e-16 ***
sqft_basement  4.704e-02  3.296e+00   0.014  0.98861    
yr_built      -1.627e+03  4.935e+01 -32.980  < 2e-16 ***
yr_renovated   1.112e+01  2.641e+00   4.211 2.55e-05 ***
lat            5.113e+05  6.737e+03  75.887  < 2e-16 ***
long           5.196e+04  8.083e+03   6.429 1.33e-10 ***
sqft_living15  5.486e+01  2.664e+00  20.588  < 2e-16 ***
sqft_lot15    -4.914e+00  2.819e-01 -17.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99480 on 12426 degrees of freedom
Multiple R-squared:  0.678,	Adjusted R-squared:  0.6776 
F-statistic:  1635 on 16 and 12426 DF,  p-value: < 2.2e-16
```

3.1.3 Best Subset Selection

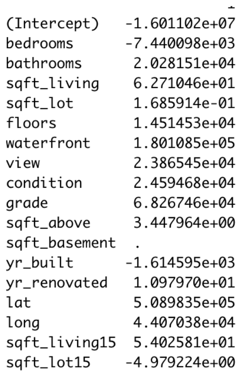
For selecting useful features among all the 17 independent variables, we used best subset selection to choose variables and lastly the model has 15 variables left based on the minimum BIC. The test error is 17,167,066,295. As we can see from the coefficient results, the variable sqft_living is positively correlated with the price and its coefficient is the highest, so we can conclude that the size of living space is important to predict the price.

3.1.4 Ridge Regression & Lasso Regression

In order to decrease the dimension and improve the prediction, we fit two models here, ridge regression and lasso regression. After finding the optimal λ (10091.72) of ridge regression, the model shows that the test error is 9,964,802,522 but the variables are still 17 variables. The optimal λ of lasso regression is 217.4194 and the test error is 9,927,462,627. The two shrunken models didn't shrink any coefficients into 0. But we got reliable results, the test error both shows decrease a lot compared with the previous models. The left picture shows the coefficients of ridge regression and the right shows the coefficients of lasso. The number of bedrooms is negatively related to the price and grade is important for the price prediction.



```
(Intercept)   -1.711505e+07
bedrooms      -6.706726e+03
bathrooms      1.884217e+04
sqft_living    3.723874e+01
sqft_lot       1.597711e-01
floors         1.510872e+04
waterfront     1.699581e+05
view           2.404823e+04
condition      2.442202e+04
grade          6.196921e+04
sqft_above     3.033315e+01
sqft_basement  2.745417e+01
yr_built      -1.422785e+03
yr_renovated   1.351733e+01
lat            4.908777e+05
long           3.074496e+04
sqft_living15  5.393351e+01
sqft_lot15    -4.625121e+00
```



```
(Intercept)   -1.601102e+07
bedrooms      -7.440098e+03
bathrooms      2.028151e+04
sqft_living    6.271046e+01
sqft_lot       1.685914e-01
floors         1.451453e+04
waterfront     1.801085e+05
view           2.386545e+04
condition      2.459468e+04
grade          6.826746e+04
sqft_above     3.447964e+00
sqft_basement  .
yr_built      -1.614595e+03
yr_renovated   1.097970e+01
lat            5.089835e+05
long           4.407038e+04
sqft_living15  5.402581e+01
sqft_lot15    -4.979224e+00
```

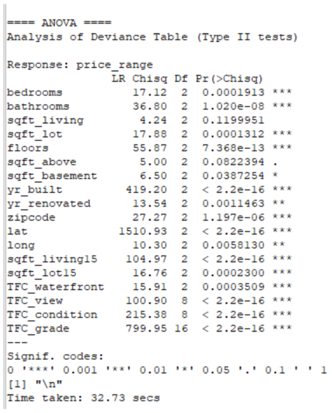
3.1.5 Principal Components Regression & Partial Least Squares

We fit PCR and PLS to maximize the amount of variance explained in the predictors. We firstly put 17 components into the PCR model to capture 100% of the variance or information. The test error is 9,917,015,840. For PLS, because the lowest CV error occurs when M = **9** partial least squares directions are used, we chose 9 components into the model. Its test error is 10,299,700,761. As a result, PLS explains 67.97 % variance in the price which is close to PCR fit that explains 68% in the price when M=17 because PLS searches for directions that explain variance in both the predictors and the response.

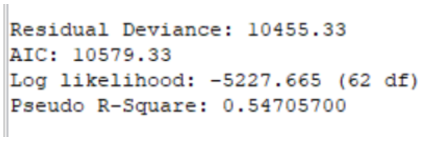
3.2 Classification

First, we decide to divide our price into 3 ranges, the lower, 0~ 200,000, median: 200,000 ~ 600,000, high: 600,000~ 886,000. And then we created a new column called: price range. We load it into a rattle, using partition 70/15/15 as our train/validation/test set, and set seed to 42.

3.2.1 linear regression



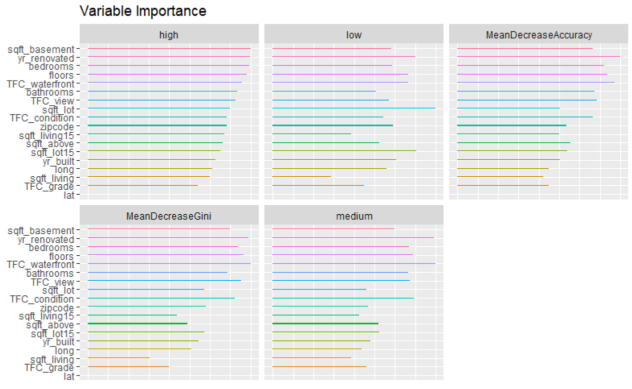
```
==== ANOVA ====
Analysis of Deviance Table (Type II tests)

Response: price_range
               LR Chisq Df Pr(>Chisq)
bedrooms          17.12  2  0.0001913 ***
bathrooms         36.80  2  1.020e-08 ***
sqft_living        4.24  2  0.1199951
sqft_lot          17.88  2  0.0001312 ***
floors            55.87  2  7.368e-13 ***
sqft_above         5.00  2  0.0822394 .
sqft_basement      6.50  2  0.0387254 *
yr_built         419.20  2  < 2.2e-16 ***
yr_renovated      13.54  2  0.0011463 **
zipcode           27.27  2  1.197e-06 ***
lat             1510.93  2  < 2.2e-16 ***
long              10.30  2  0.0058130 **
sqft_living15    104.97  2  < 2.2e-16 ***
sqft_lot15        16.76  2  0.0002300 ***
TFC_waterfront    15.91  2  0.0003509 ***
TFC_view         100.90  8  < 2.2e-16 ***
TFC_condition    215.38  8  < 2.2e-16 ***
TFC_grade        799.95 16  < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
Time taken: 32.73 secs
```



```
Residual Deviance: 10455.33
AIC: 10579.33
Log likelihood: -5227.665 (62 df)
Pseudo R-Square: 0.54705700
```

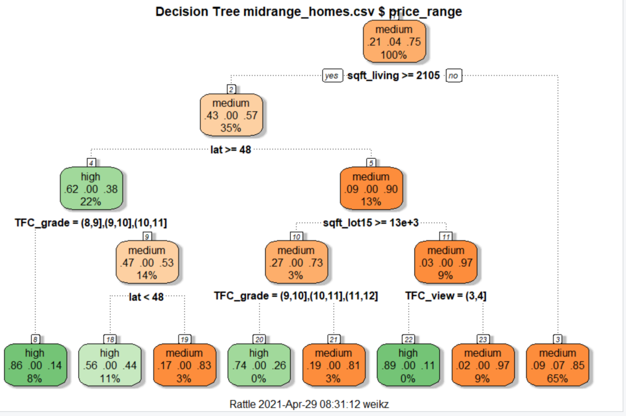
Our first model is a linear regression with category response. We choose this model because it can show the coefficient between each variable from a different category. The reference group is a high price category variable. And we can see the p-value from the chi-square test, the yr_bulit, sqft_ling 15, lat, View, condition grade, those variables p-value is significant so we can conclude that the price range will be influenced by those variables changed. And for R-square is around 0.54, which means that value is considered a Moderate effect on price_range.

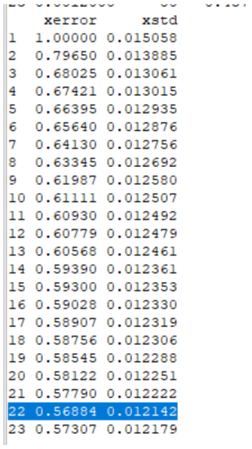
3.2.2 Random Forest



The following model is random frost; we choose this model because Random forest adds additional randomness to the model while growing the trees. A graph showing the mean decrease in the Gini coefficient is a way to measure how each variable contributes to the homogeneity of the nodes and leaves.The result shows from the graph; we can see from the high price range the most important feature is sqft_living, yr-renovated, bedrooms, floor. And for medium-range is :sqft_lot,yr_renovated, sqft_lot15, yr_bulit. For low price range: yr_renovated, waterfront, condition, floors.

3.2.3 Decision tree

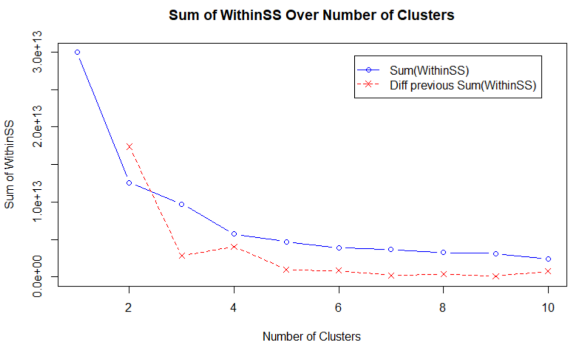


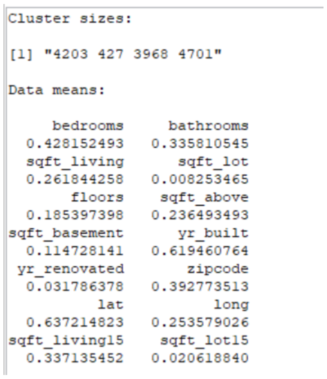


```
   xerror     xstd
1  1.00000 0.015058
2  0.79650 0.013885
3  0.68025 0.013061
4  0.67421 0.013015
5  0.66395 0.012935
6  0.65640 0.012876
7  0.64130 0.012756
8  0.63345 0.012692
9  0.61987 0.012580
10 0.61111 0.012507
11 0.60930 0.012492
12 0.60779 0.012479
13 0.60568 0.012461
14 0.59390 0.012361
15 0.59300 0.012353
16 0.59028 0.012330
17 0.58907 0.012319
18 0.58756 0.012306
19 0.58545 0.012288
20 0.58122 0.012251
21 0.57790 0.012222
22 0.56884 0.012142
23 0.57307 0.012179
```

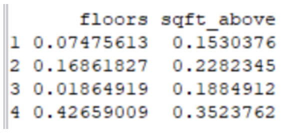
Furthermore, in the following model, we choose to use decision trees which is another good way for helping you to choose important variables associated with price. They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options. So, we need to find a size of the decision tree that could have a small standard error, so we see from cp:0.0012 is optimal tree size, if the cost of adding value to the decision tree is above this value, then the tree building is not continuing.

3.2.4 KNN

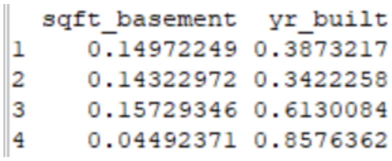




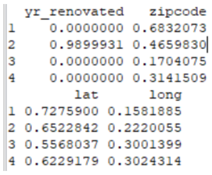
```
Cluster sizes:

[1] "4203 427 3968 4701"

Data means:

     bedrooms    bathrooms
  0.428152493  0.335810545
  sqft_living     sqft_lot
  0.261844258  0.008253465
       floors   sqft_above
  0.185397398  0.236493493
sqft_basement     yr_built
  0.114728141  0.619460764
 yr_renovated      zipcode
  0.031786378  0.392773513
          lat         long
  0.637214823  0.253579026
sqft_living15   sqft_lot15
  0.337135452  0.020618840
```



```
      floors sqft_above
1 0.07475613  0.1530376
2 0.16861827  0.2282345
3 0.01864919  0.1884912
4 0.42659009  0.3523762
```



```
  sqft_basement  yr_built
1    0.14972249 0.3873217
2    0.14322972 0.3422258
3    0.15729346 0.6130084
4    0.04492371 0.8576362
```



```
  yr_renovated   zipcode
1    0.0000000 0.6832073
2    0.9899931 0.4659830
3    0.0000000 0.1704075
4    0.0000000 0.3141509
        lat      long
1 0.7275900 0.1581885
2 0.6522842 0.2220055
3 0.5568037 0.3001399
4 0.6229179 0.3024314
```

In addition, the reason why we choose KNN is that *KNN* is an excellent method to classification that estimates how likely a data point is to be a member of one group or the other. And the result shows that we need to choose 4 cluster sizes because, after cluster 4, there is not much difference. And for the first cluster, floor and lat is an essential element to influence price_range. For the 2 and 4 sets, the ye-renovated and yr_bulit is a crucial factor influencing price because the correlation is quickly high above the mean.

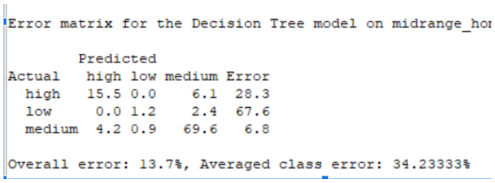
Part 4: Evaluation 4.1 Evaluation

4.1.1 Regression Models

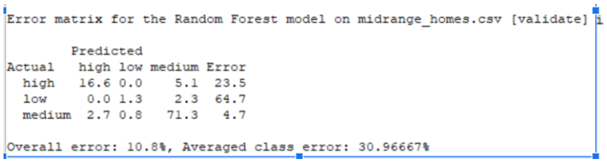
	Random Forest	Multiple Linear Regression	Best Subset Selection	Ridge Regression	Lasso Regression	PCR PLS
Test Error (MSE)	4,669,091,136	52,188,694,666	17,167,066,295	9,964,802,522	9,927,462,627	9,917,015,840 10,299,700,761
Advantages	-Importance rank; -low error rate(good prediction); -Simple model-fitting procedure	-Coefficients results; - +/- relationship; -Simple model-fitting procedure	-Coefficients results; -decrease variables(reduce complexity); -Simple model-fitting procedure	-Coefficients results; -standardize	-Coefficients results; -standardize; -low error rate	-Standardize; -Dimension reduction(PLS); -low error rate; -Simple model-fitting procedure
Disadvantages	-No coefficients; -no standardize;	-Interaction; -no standardize; -bad prediction	-So-so prediction	-Complex model-fitting procedure	-Complex model-fitting procedure	-No coefficients; -PCR not sparse

We compared all the models based on the prediction, simple fitting procedure, sparsity, whether it produced coefficient results or not and whether it does rescale. Finally we think random forest and lasso both would be suitable models for this dataset. Because random forest shows clear importance rank, good prediction and simple model fitting procedure, and lasso shows specific coefficients and avoids the scale problem even though the fitting procedure is a little bit complicated.

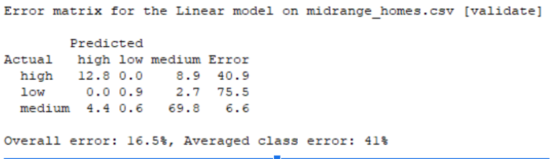
4.1.2 Classification models



```
Error matrix for the Decision Tree model on midrange_hor

        Predicted
Actual   high low medium Error
  high   15.5 0.0    6.1  28.3
  low     0.0 1.2    2.4  67.6
  medium  4.2 0.9   69.6   6.8

Overall error: 13.7%, Averaged class error: 34.23333%
```



```
Error matrix for the Random Forest model on midrange_homes.csv [validate] i

        Predicted
Actual   high low medium Error
  high   16.6 0.0    5.1  23.5
  low     0.0 1.3    2.3  64.7
  medium  2.7 0.8   71.3   4.7

Overall error: 10.8%, Averaged class error: 30.96667%
```



```
Error matrix for the Linear model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high   12.8 0.0    8.9  40.9
  low     0.0 0.9    2.7  75.5
  medium  4.4 0.6   69.8   6.6

Overall error: 16.5%, Averaged class error: 41%
```

Overall, the error for the random forest is our best model than the decision tree and linear model because the overall is lower to 10.8%.

4.2 Business Perspective

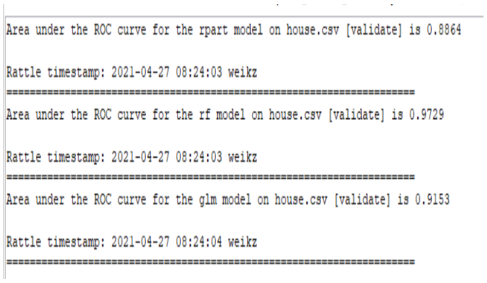
Based on the results, we can conclude from regression models we fit before that higher construction quality, size of living space, and average size of living space for the nearest 15 neighbors will result in higher price. Therefore we think for the sellers, they could highlight those 3 important factors in the ads to increase their own house market value and to gain competitive advantages further. Buyers can forecast market trends and appraise the price based on them.

Based on the result, we can conclude that from category models the yr_renovated is a most significant variable for price_range. People who have high income may care more about the sqft_basement,and people for medium-income may care more about if this building has a waterfront or not. And for low income, people care more about sqft_lot.

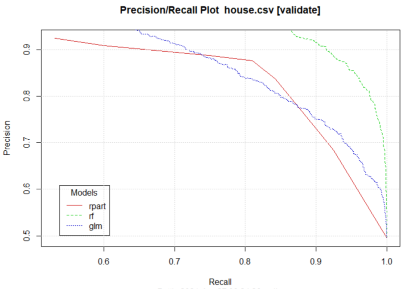
4.3 Weakness and Possible Improvement

4.3.1Classfication models

From the category model, It is hard to divide the price into a specific range because some value may be closed to that range; you cannot simply define theirs into a different category. (so more range, more error occurs) But when the price only has two sorts, high or low, the accuracy of AOC up to 0.9 that’s good. And the precision shown on the graph is good as well.



```
Area under the ROC curve for the rpart model on house.csv [validate] is 0.8864

Rattle timestamp: 2021-04-27 08:24:03 weikz
======================================================================
Area under the ROC curve for the rf model on house.csv [validate] is 0.9729

Rattle timestamp: 2021-04-27 08:24:03 weikz
======================================================================
Area under the ROC curve for the glm model on house.csv [validate] is 0.9153

Rattle timestamp: 2021-04-27 08:24:04 weikz
======================================================================
```



4.3.2 Regression Models

In the regression part, we realize some risks existing the prediction, firstly, the variable called yr_renovated which means the year renovated. In the data it shows specific numbers like 1980 but if the house is brand-new, it would show 0. So there is a large difference between them which may bring some influence on predicting the price when inputting yr_renovated into regression models. Therefore, for improvement, we can transform it into categorical variables. The house has been renovated can be labeled as 1, but it is not renovated, it can be labeled as 0. This method may help decrease prediction deviation.

Part 5: Reference

Data source: https://geodacenter.github.io/data-and-lab//KingCounty-HouseSales2015/