Final Project Report

Group 56: Zhongyang Wang (zw3057) / Yuxuan Zhao (yz8370)

1. Introduction:

Housing price is a concept closely related to our real life. In China, many people spend their entire lives just to have a house of their own in a bustling big city. The establishment of housing prices is closely related to many factors. From the basic housing area and the number of rooms to the complex geographical environment and social policies, these factors all affect housing prices to a greater or lesser extent.

In this project, our group chose to discuss the relationship between housing prices and more than ten related factors, such as house area, house quality, and construction year. The basic method adopted by our group is to perform regression analysis on the collected house prices and more than ten related factors, including linear regression, which includes model selection, decision tree model, random forest and so on. After analyzing the results, our group has obtained some relevant factors such as housing quality and condition, which will greatly affect the housing price. In addition, we also compared the regression effects of each model, and finally found that linear regression works best in the data set selected for this project.

2. Related work:

In fact, there are many existing studies on the factors affecting housing prices. For example, Madhuri tried to forecast house prices in 2019[1]. Madhuri estimates speculative prices by analysing commodities, fare ranges and warning of developments. He used different regression techniques for prediction such as multilinear, ridge, LASSO, elastic net, gradient boosting and Ada Boost regression. Madhuri did predict house prices well in his experiments, and the regression analysis method he used was very complete and efficient. However, the factors that Madhuri considers that affect house prices are mostly factors that change over time, such as housing demand in different time periods. The starting point of his research is to find the time factor of house price changes. Varma also conducted research to predict house prices in 2018[2]. His main research method is the weighted average of the results of various regression models to give the most accurate house price predictions. He gave us a lot of inspiration on the algorithm, but it is worth mentioning that the influencing factors studied by Varma are slightly

[1] Madhuri, CH Raga, G. Anuradha, and M. Vani Pujitha. "House price prediction using regression techniques: a comparative study." *2019 International Conference on Smart Structures and Systems (ICSSS)*. IEEE, 2019.

[2] Varma, Ayush, et al. "House price prediction using machine learning and neural networks." *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2018

insufficient in our opinion. Lu proposed a hybrid Lasso and Gradient boosted regression model in 2017 to predict individual house prices.[3] The proposed approach has recently been deployed as the key kernel for Kaggle Challenge "House Prices: Advanced Regression Techniques". The performance is promising as the latest score was ranked top 1% out of all competition teams and individuals. We were encouraged by Lu's results, so we decided to do a good job of researching housing prices. The research conducted by our group is also quite different from that of Madhuri and Varma. Madhuri's research focuses on the impact of time factors on house price changes, and his starting point is to give some people advice on when to buy a house. The factors that affect house prices selected in this paper are the basic conditions of the house, such as housing area, house quality, kitchen quality, heating system evaluation, and the number of rooms. Our two studies have different research points, and we prefer to start with the physical conditions of the house. Compared with Varma's study, we believe that we have selected more influencing factors to obtain a more comprehensive answer.

3. Methods:

In this project, the questions we finally selected were to study the year of sale, the overall quality of the house, the overall condition of the house, the year of construction, the quality of the basement, the quality of the heating system, the area of the first and second floors, the quality of the kitchen, the occurrence of fire, and the parking lot. Size, pool size, number of rooms above ground, distance from the street, and floor space, these 15 factors relationship with the house price in Boston area.

The main research method of this project is to perform regression processing on the above 15 influencing factors and housing price information. The regression models used include linear regression, decision trees, random forests, bagging, boosting, and PCA. The linear regression model includes the choice of linear regression model in addition to the basic linear regression. We used Best Subset Selection with BIC, Forward Stepwise Selection with R^2, Backward Stepwise Selection with Cp, Ridge Regression with 10-fold CV, Lasso with 10-fold CV, KNN with k=5 these methods to achieve linear regression Model selection optimization.

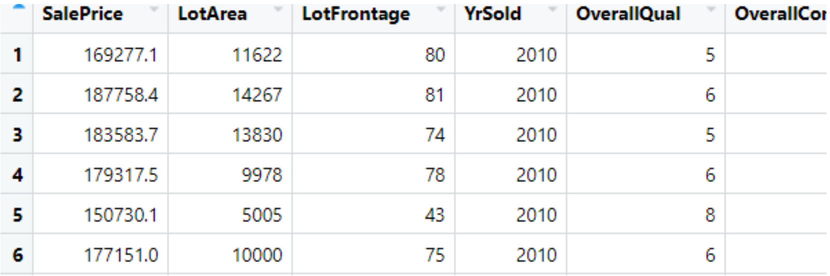
Our assumption is that the 15 influencing factors mentioned above will have a certain degree of impact on housing prices and most of them are positive. For example, the better the overall quality of the house, the higher the housing price. In fact, in the follow-up link, we will introduce in detail the results of our regression prediction of 15 influencing factors and housing prices. The results show that most of the 15 influencing factors we selected show a very strong relationship with housing prices. This proves that our assumptions and variables are quite reasonable. And all the modeling finally achieved good results, which also shows that the modeling method we chose is more reasonable.

4. Data and Experiment setup:

The data set we use is come from Kaggle. It is a set of housing price information in Boston, including 2919 sets of records for 80 variables, including sales price, sales year and month, house area and so

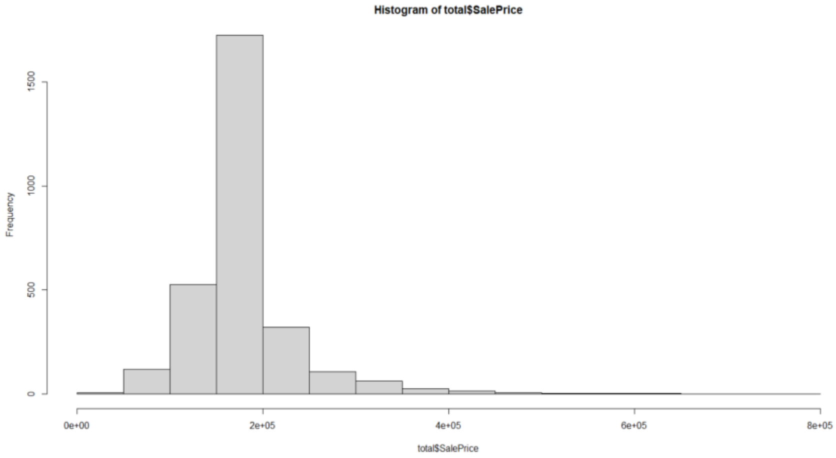
[3] Lu, Sifei, et al. "A hybrid regression technique for house prices prediction." *2017 IEEE international conference on industrial engineering and engineering management (IEEM)*. IEEE, 2017

on. On the basis of this set of data sets, we randomly divided 1460 sets of information as training data sets and 1459 sets as test data sets. Because there are too many variables, we finally choose 15 variables as our predictors, and the outcome is the sale price variable. In the preprocessing, we delete all the missing values in the dataset and finally we got 1170 observations in our train set and 1193 observations in our test set. The following picture shows what our data set look like:



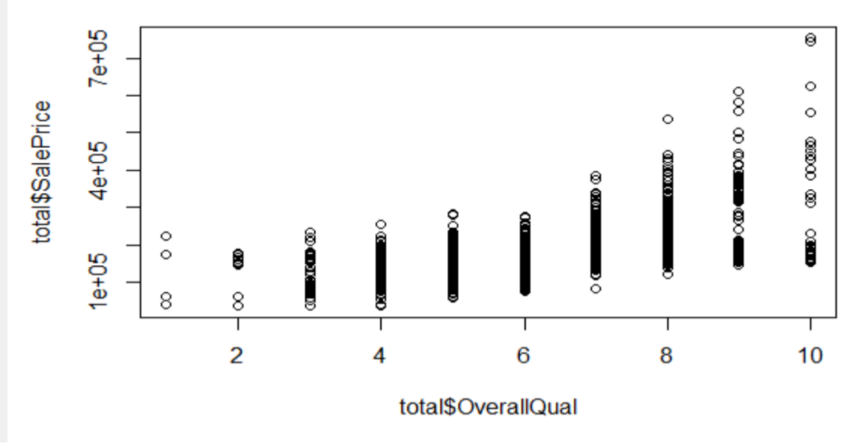
	SalePrice	LotArea	LotFrontage	YrSold	OverallQual	OverallCor
1	169277.1	11622	80	2010	5	
2	187758.4	14267	81	2010	6	
3	183583.7	13830	74	2010	5	
4	179317.5	9978	78	2010	6	
5	150730.1	5005	43	2010	8	
6	177151.0	10000	75	2010	6	

After the data preprocessing, we did some univariate analysis and bivariate analysis on our data, here is what the distribution of the sale price look like:



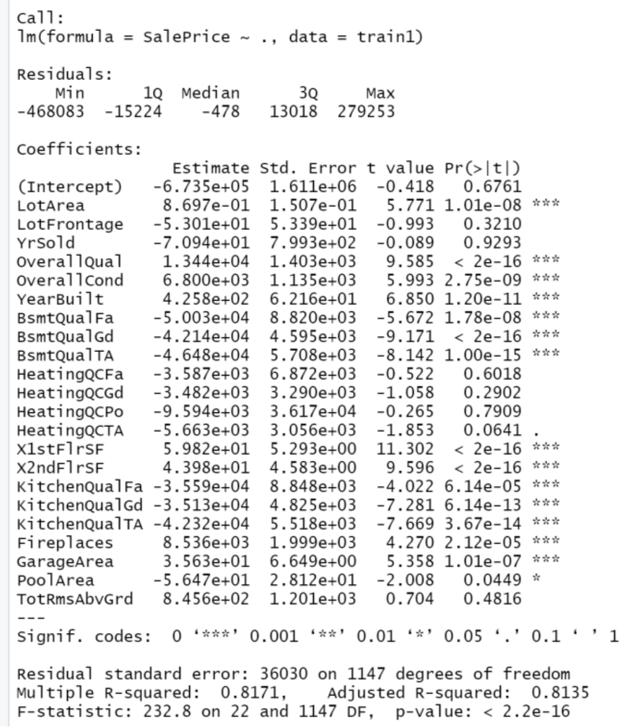
We can see that the distribution of the sale price is not normal. And after we did all the univariate analysis, we find that nearly all of the distribution of the variables are not normal.

And here is an example of bivariate analysis:



From the picture we can see that there may be a positive relationship between the OverallQuality and the sale price.

After doing all these descriptive statistics, we find that our predictors did have some relationship with the sale price. So, it is meaningful to use our machine learning method to do the house price prediction. Our benchmark is the linear regression for the linear regression is the most common used machine learning method.

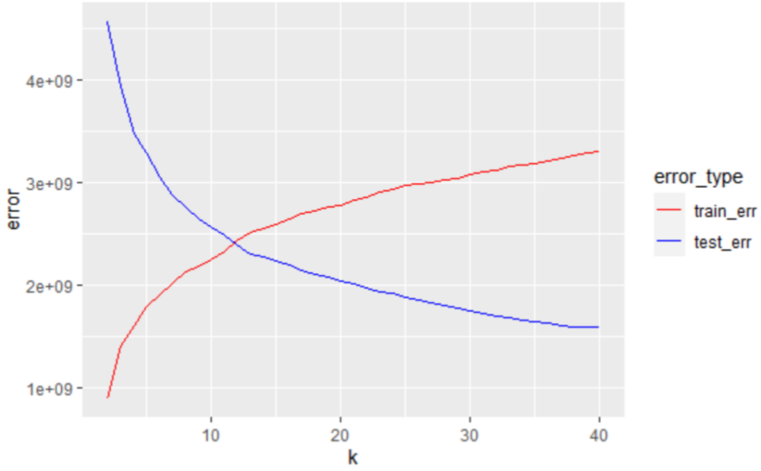


```
Call:
lm(formula = SalePrice ~ ., data = train1)

Residuals:
    Min      1Q  Median      3Q     Max 
-468083  -15224    -478   13018  279253 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -6.735e+05  1.611e+06  -0.418   0.6761    
LotArea        8.697e-01  1.507e-01   5.771 1.01e-08 ***
LotFrontage   -5.301e+01  5.339e+01  -0.993   0.3210    
YrSold        -7.094e+01  7.993e+02  -0.089   0.9293    
OverallQual    1.344e+04  1.403e+03   9.585  < 2e-16 ***
OverallCond    6.800e+03  1.135e+03   5.993 2.75e-09 ***
YearBuilt      4.258e+02  6.216e+01   6.850 1.20e-11 ***
BsmtQualFa    -5.003e+04  8.820e+03  -5.672 1.78e-08 ***
BsmtQualGd    -4.214e+04  4.595e+03  -9.171  < 2e-16 ***
BsmtQualTA    -4.648e+04  5.708e+03  -8.142 1.00e-15 ***
HeatingQCFa   -3.587e+03  6.872e+03  -0.522   0.6018    
HeatingQCGd   -3.482e+03  3.290e+03  -1.058   0.2902    
HeatingQCPo   -9.594e+03  3.617e+04  -0.265   0.7909    
HeatingQCTA   -5.663e+03  3.056e+03  -1.853   0.0641 .  
X1stFlrSF      5.982e+01  5.293e+00  11.302  < 2e-16 ***
X2ndFlrSF      4.398e+01  4.583e+00   9.596  < 2e-16 ***
KitchenQualFa -3.559e+04  8.848e+03  -4.022 6.14e-05 ***
KitchenQualGd -3.513e+04  4.825e+03  -7.281 6.14e-13 ***
KitchenQualTA -4.232e+04  5.518e+03  -7.669 3.67e-14 ***
Fireplaces     8.536e+03  1.999e+03   4.270 2.12e-05 ***
GarageArea     3.563e+01  6.649e+00   5.358 1.01e-07 ***
PoolArea      -5.647e+01  2.812e+01  -2.008   0.0449 *  
TotRmsAbvGrd   8.456e+02  1.201e+03   0.704   0.4816    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36030 on 1147 degrees of freedom
Multiple R-squared:  0.8171,	Adjusted R-squared:  0.8135 
F-statistic: 232.8 on 22 and 1147 DF,  p-value: < 2.2e-16
```

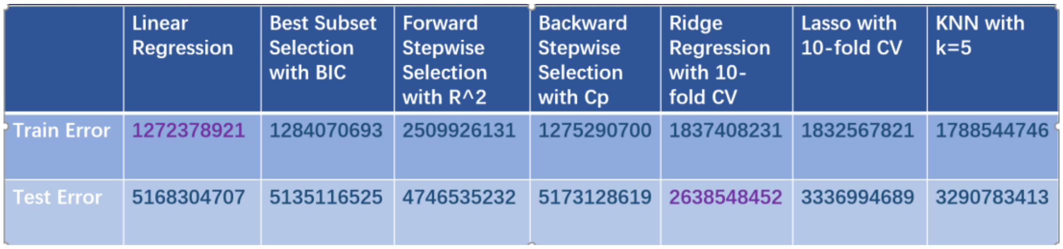
This is our model of the linear regression, we can see that not all the variables are statistically significant. And the adjusted R-squared for our original model is 0.8135 which is quite good.

5. Results:

Firstly, we did more things on the linear regression model. First is the linear regression with KNN, below is the plot of the train error and the test error of KNN when k changes. From the plot, we can find that when the k increases, the train error keeps increasing and the test error keeps decreasing, so it is hard to say which k is the best k.



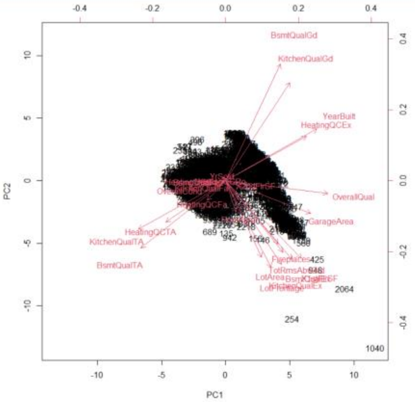
And then we did some linear model selection using different methods such as forward stepwise selection with R-squared, best subset selection with BIC, lasso with 10-fold CV and so on. The total result is shown in the following table:



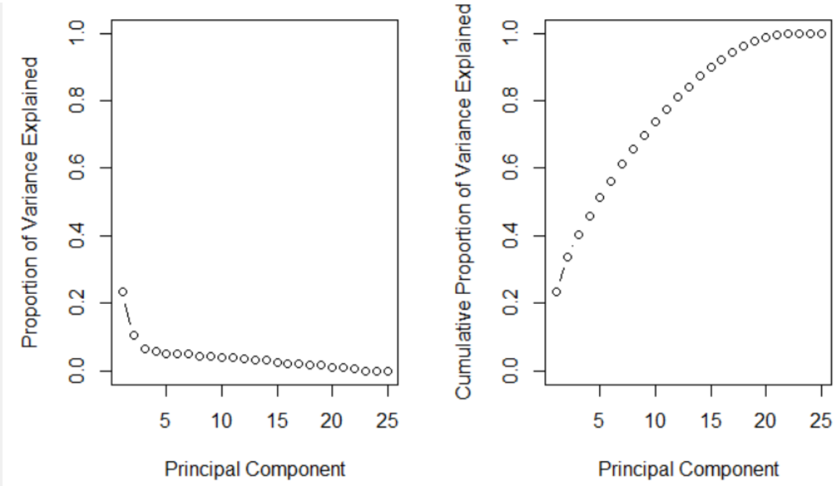
	Linear Regression	Best Subset Selection with BIC	Forward Stepwise Selection with R^2	Backward Stepwise Selection with Cp	Ridge Regression with 10-fold CV	Lasso with 10-fold CV	KNN with k=5
Train Error	1272378921	1284070693	2509926131	1275290700	1837408231	1832567821	1788544746
Test Error	5168304707	5135116525	4746535232	5173128619	2638548452	3336994689	3290783413

We use the different color to show the best train error and test error. From the table, we can see that, the original linear regression has the smallest train error but its test error is big, the ridge regression with 10-fold CV has the smallest test error and its train error is also small. Therefore, compared all these different linear regression models, we think the ridge regression with 10-fold CV is the best linear model.

And then we use PCA to see if it can help us better predict the house price. This is the biplot of our PCA method, we can see that because our data is quite big, it is difficult to see much information from this plot.

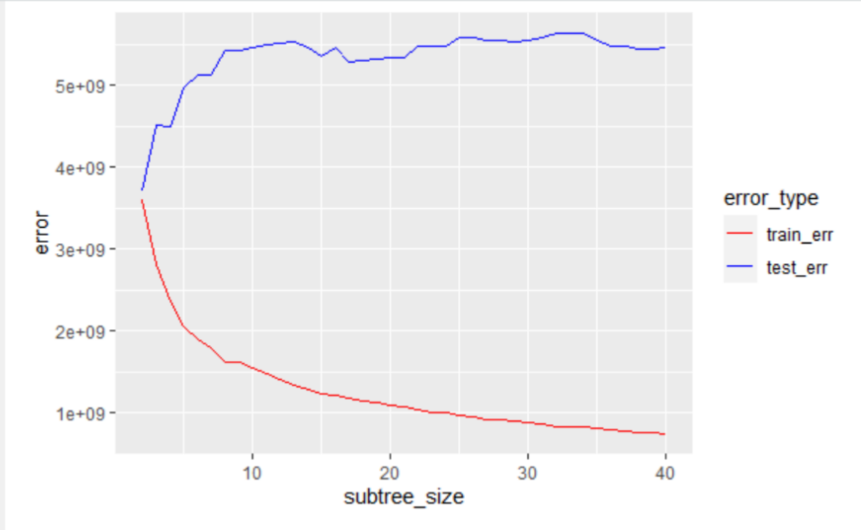


And then we make a plot about the proportion of variance explained by each of the principal component. From the plot, we can see that the variance explained by most of the principal components are quite close and not big. This shows that the PCA method may not help us much on the predicting.



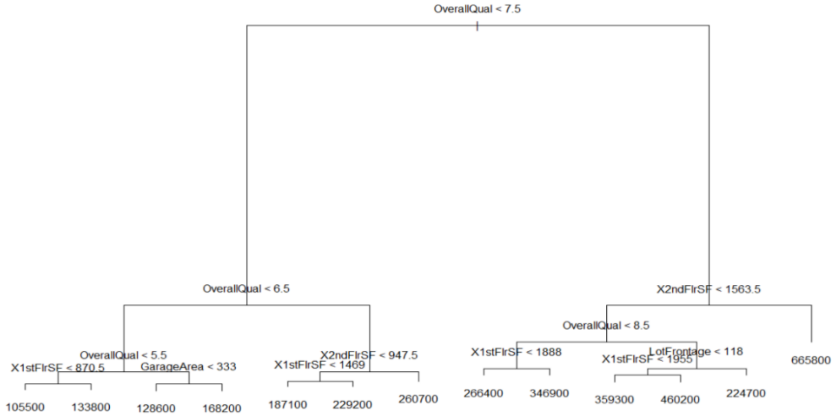
Finally, we choose the first 3 principal components to do linear regression on the data set. Compared this to the original linear regression, we find that there is nearly no difference between the PCA linear model and the original linear model. So, we think the PCA is not very useful in our house price prediction project.

Then we did some other methods. Firstly, we use the regression tree to do the prediction. We tried to prune the tree to different subtrees with the number of terminals nodes ranging from 2 to 40, for each subtree, we compute the training error and prediction error. The following is the plot:



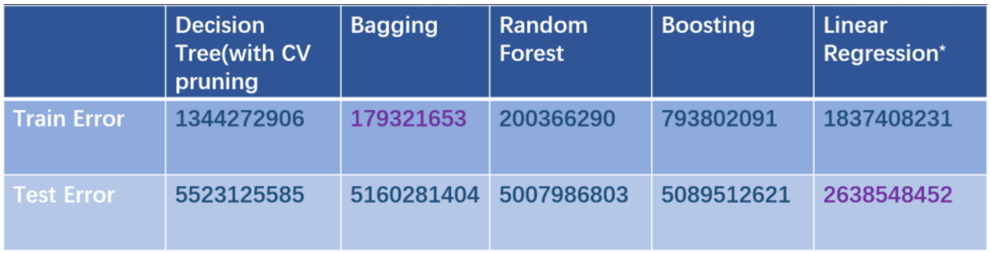
From the plot, we can see that when the subtree size increases, the training error keeps decreasing, however, the prediction error keeps increasing. This shows that the regression tree may be not suitable for our house prices prediction.

And then we use cross validation method to help us choose the best sub tree size and the result is 13. The following picture is what our final tree looks like.



From the tree, we can see that the overall quality of the house is an important factor that influence the house price. And the space of first floor and second floor are also important factors in this tree. Though there are 13 nodes in the tree, we can see that the result of prediction value is quite different from each other. This may cause the test error to be big. We think this is why the regression tree method is not very suitable for the house price predictions.

And then we use other methods such as bagging, random forest and boosting. The train error and test error of each method are shown in the table.



	Decision Tree(with CV pruning	Bagging	Random Forest	Boosting	Linear Regression*
Train Error	1344272906	179321653	200366290	793802091	1837408231
Test Error	5523125585	5160281404	5007986803	5089512621	2638548452

From the table, we can see that the train error of bagging and random forest is quite small, however, the test errors of these methods are quite big. We think that this is because these methods may cause overfitting on the train set, so the test error will be big. Compared all those methods, we can see that the linear regression may be the best method for its test error is the smallest and the train error is not very bigger than other methods. So, we choose the ridge regression with 10-fold CV as our final model. Compared this method to the benchmark, we can see that though the train error is a little bigger, the test error is a lot smaller. So, we think this method is better than the benchmark.

6. Discussion:

Through the description of our group's experimental results by the above result module, we can see that housing prices are indeed closely related to many factors. By observing the result graph of linear regression, we can see that only five of the more than ten elements we selected showed no correlation with house prices in linear regression, while the rest of the elements showed even at the level of 0.001. correlation with house prices. This shows that our initial assumptions are still quite correct. The R-SQUARED value obtained by linear regression is 0.8135, which also shows that the established linear regression model has high prediction accuracy for the data set we set. In other words, the linear regression model performs well in predicting housing prices.

Interestingly, we initially thought that the quality of the heating system configured in the house would have a greater impact on the house price, but in fact, the data analysis results showed that the quality of the heating system has nothing to do with the house price. We believe this is due to the fact that the area where house price information is collected is Boston. In areas where the temperature is low all year round, the heating system is a necessities of the house configuration, so its quality difference has less influence on the house price. Another interesting finding was the effect of the number of rooms on house prices. Although the results show that the number of surface rooms does not have a very high impact on the housing price, we can find that for houses with a lower number of rooms, the increase in the number of rooms does will cause house prices to rise. When the number of rooms exceeds a certain amount, the change in the number of rooms will no longer affect the price. We think this is also an interesting finding that conforms to the laws of reality. As we all know, buyers have completely different attitudes towards one-bedroom, one-bedroom and two-bedroom, two-bedroom apartment types. For this category, the number of rooms is a very significant determinant. But for a mansion with a large number of rooms, the number of rooms has become meaningless. Those who can afford to buy a luxury home consider other factors more than the number of rooms.

In our view, this study has two shortcomings. The first problem is the choice of analysis method. In this study, although our group used six different regression models for predictive analysis, we did not combine these six models well, and we mainly compared the results of these six models. As mentioned in the related work module, many studies innovatively average the prediction results of multiple models to obtain a model with higher prediction accuracy. Therefore, in future work, we can try to combine multiple predictive models for joint analysis, or use popular machine learning methods to create better and more accurate models. Another area where we can improve is that we can expand the area of house price information in future work to obtain more general conclusions. The dataset we used for this research project is information on housing prices in the Boston area. Therefore, our conclusions may be affected by geographical factors. Or we can directly use the established model to predict housing prices in other regions. By observing the prediction accuracy, we can better judge whether the model needs to be modified.

7. Team members' contributions:

Code: Zhongyang Wang, Yuxuan Zhao

Presentation & PPT； Zhongyang Wang, Yuxuan Zhao

Report: Zhongyang Wang, Yuxuan Zhao

8. Reference:

1. Madhuri, CH Raga, G. Anuradha, and M. Vani Pujitha. "House price prediction using regression techniques: a comparative study." *2019 International Conference on Smart Structures and Systems (ICSSS)*. IEEE, 2019.
2. Varma, Ayush, et al. "House price prediction using machine learning and neural networks." *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2018.
3. Lu, Sifei, et al. "A hybrid regression technique for house prices prediction." *2017 IEEE international conference on industrial engineering and engineering management (IEEM)*. IEEE, 2017.