Prediction on House Sale Prices

Tong Chun Ho, Department of Mathematics, HKUST
Lai Cheuk Man, Department of Mathematics, HKUST
Wong Ngo Cheung, Department of Mathematics, HKUST

Introduction

We predicted the house price by analyzing the dataset provided in
https://www.kaggle.com/c/house-prices-advanced-regression-techniques
We utilized some non linear methods, like regression tree, bagging and random forest, and regularized regression - ridge regression and LASSO regression to do the prediction.

Data Preprocessing

Missing Value Imputation

According to “data_description.txt”, if NA is stated as none, we refilled the missing data with “none” for categorical data and 0 for quantitative factors. Otherwise, we refilled the missing data with median for quantitative data or mode for categorical data.

Revaluation

We revalued the categorical data with ascending order if the data have preference difference. Moreover, the “None” factor was revalued as 0. For other factors without preference difference, each of them was assigned a number ranged from 1 to the number of types randomly. Finally, we transformed the sale price by log as it is very large in number.

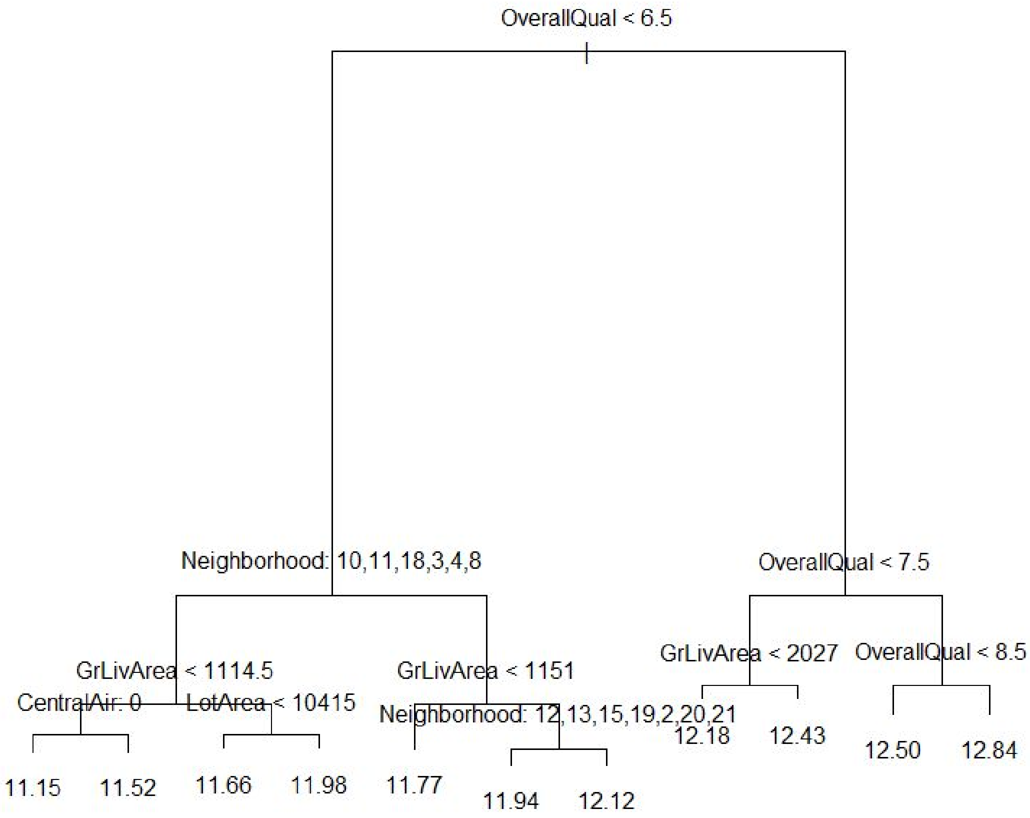
Spltting training data and test data

We separated the data of “train.csv” into 70% of training data and remaining 30% of test data.

Modelling

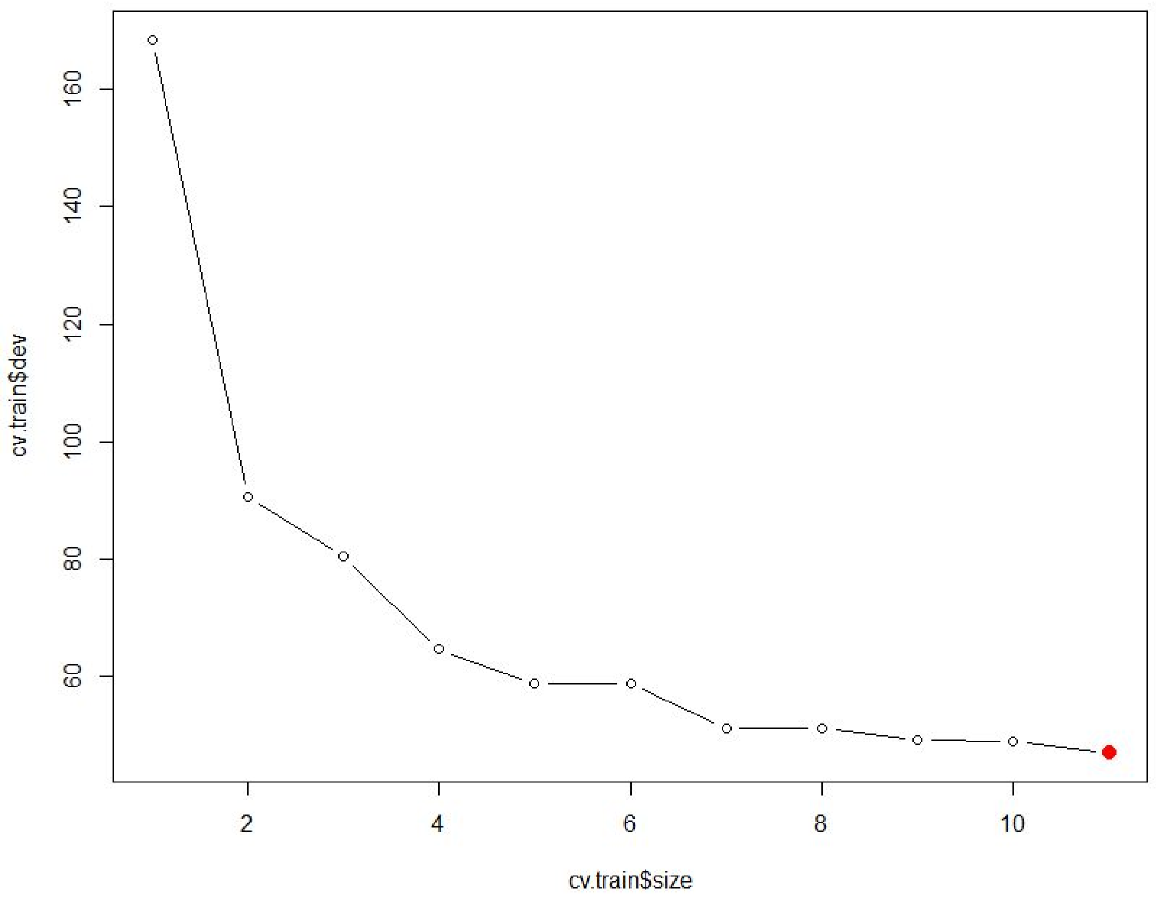
Regression tree

We performed a regression tree and found out that it stopped splitting at the node of 12. It is determined by “OverallQual”, "Neighborhood", "GrLivArea", "CentralAir" and "LotArea".



Prune tree

we used Prune tree in order to find the subtree that has the best test error, so we used cross-validation to examine the test errors for sequence of subtrees during the growing or pruning, instead of all possible subtrees which is too large a model space. From the graph, we can see that the optimal node is 12, which is the same as the regression tree. It means that the pruning of tree cannot minimize the test error. So, we decided to give up this method.



Bagging

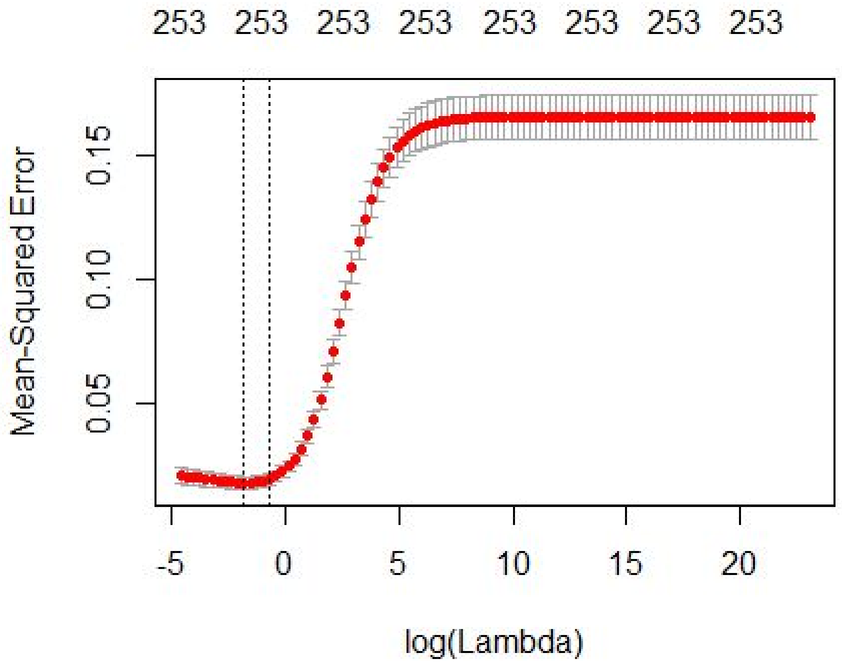
We used all the remaining 79 predictors for each split of the tree and grew 1000 trees.

Random Forest

Random Forest can improve the variance reduction of bagging by reducing the correlation between trees, without increasing the variance too much. So we used 9 predictors, which is approximately equal to square root of the number of predictors for each split of the tree and set the number of trees as 1000.

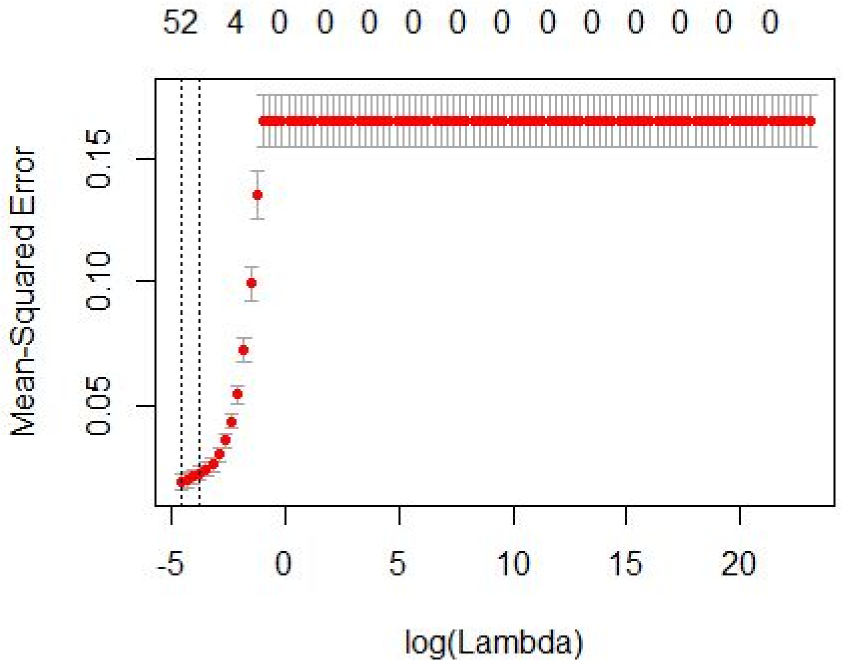
Ridge Regression

We set a grid of values for 100 λ ranging from 10^{-10} to 10^{-2} and standilized the variables. We then used cross validation to find the minimizer λ and used this λ to perform a ridge regression.



LASSO Regression

The perocedure was similar to that of the ridge regression.



Data Validation

Random Forest performed the best in the aspect of R-square.

	R-square
Regression Tree	0.6719312
Bagging	0.8326187
Random Forest	0.8525111
Ridge Regression	0.7662264
LASSO Regression	0.7520053

Analysis

The LASSO regression and ridge regression have a high test MSE because the house price may have a non-linear relationship with the predictors. They are inflexible models which always try to capture the linear relationship with shrinkage penalty to the number of parameters. In this case, it may underfit the data due to the non-linearity.

Regression tree has a higher test MSE than a random forest and bagging do because a little change in data may cause a large change in the tree. Random forest is more stable as it is a collection of many tree models. A great change may happen in an individual tree but overall change in the collection is small.

The importance of the random forest implies that "GrLivArea", "Neighborhood" and “TotalBsmtSF” are the most important factors. The construction of the regression tree is also mainly based on these factors. By these two evidences, we can assume that these three factors are important factors affecting house price.

Conclusion

Comparing the test MSE of the above models, the one produced by random forest performed the best. We validated the models by R-square. The R-square of the one built by random forest is the largest, which shows that it has the highest accuracy. In conclusion, random forest is the most suitable method in this case.

Contribution

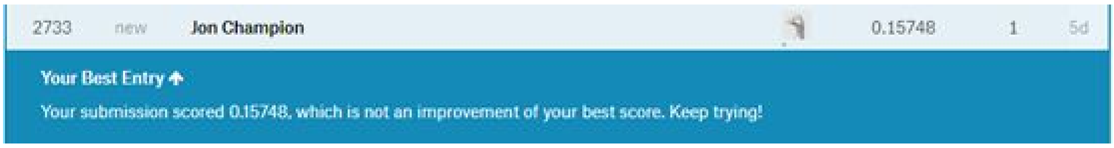
Data Preprocessing

- Wong Ngo Cheung

Model

- Tong Chun Ho
- Lai Cheuk Man

Result in Kaggle



2733 new Jon Champion 0.15748 1 5d

Your Best Entry ↑

Your submission scored 0.15748, which is not an improvement of your best score. Keep trying!