MATH4432

Result Report of Final Project, Topic 4

LAU, Wing Shing – 20342662

1 Introduction

This report is written to summarize the result of the findings in the competition named *House Prices: Advanced Regression Techniques* under the organization by Kaggle. The competition is aimed to predict the house prices for a list of properties with reference to other information about the property.

2 Method

2.1 Observation of the Data

The number of variables in training set and test set are 81 and 80, and the number of entries in training set and test set are 1459 and 1460, respectively.

The first 80 variables have identical names and meanings in both data sets, which are consisted of categorical variables and quantitative variables. The last data column in the training set is called the ***SalePrice***, which is the main target that we want to predict for the test set.

Both sets are then combined into a full set of data with 2919 entries. After the examination of all the data in the full set, the variables with missing data and the number of missing data is shown as the table below.

Variable	Number of missing data	Variable	Number of missing data
MSZoning	4	Electrical	1
LotFrontage	486	BsmtFullBath	2
Alley	2721	BsmtHalfBath	2
Utilities	2	KitchenQual	1
Exterior1st	1	Functional	2
Exterior2nd	1	FireplaceQu	1420
MasVnrType	24	GarageType	157
MasVnrArea	23	GarageYrBlt	159
BsmtQual	81	GarageFinish	159
BsmtCond	82	GarageCars	1
BsmtExposure	82	GarageArea	1
BsmtFinType1	79	GarageQual	159
BsmtFinSF1	1	GarageCond	159
BsmtFinType2	80	PoolQC	2909
BsmtFinSF2	1	Fence	2348
BsmtUnfSF2	1	MiscFeature	2814
TotalBsmtSF	1	SaleType	1

Table 1: Missing data in both datasets

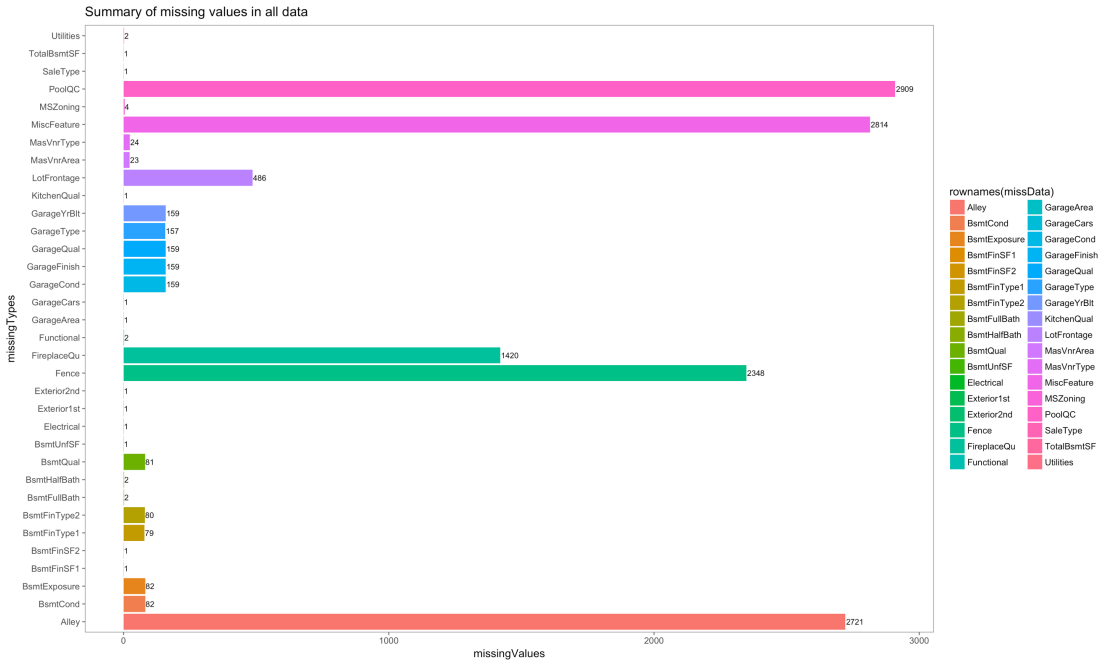

Figure 1: Visualization of the missing data

Among the categories above, the following variables will not be observed and considered in the finding as they have a critical number of missing data.

Alley, FireplaceQu, PoolQC, Fence, MiscFeature

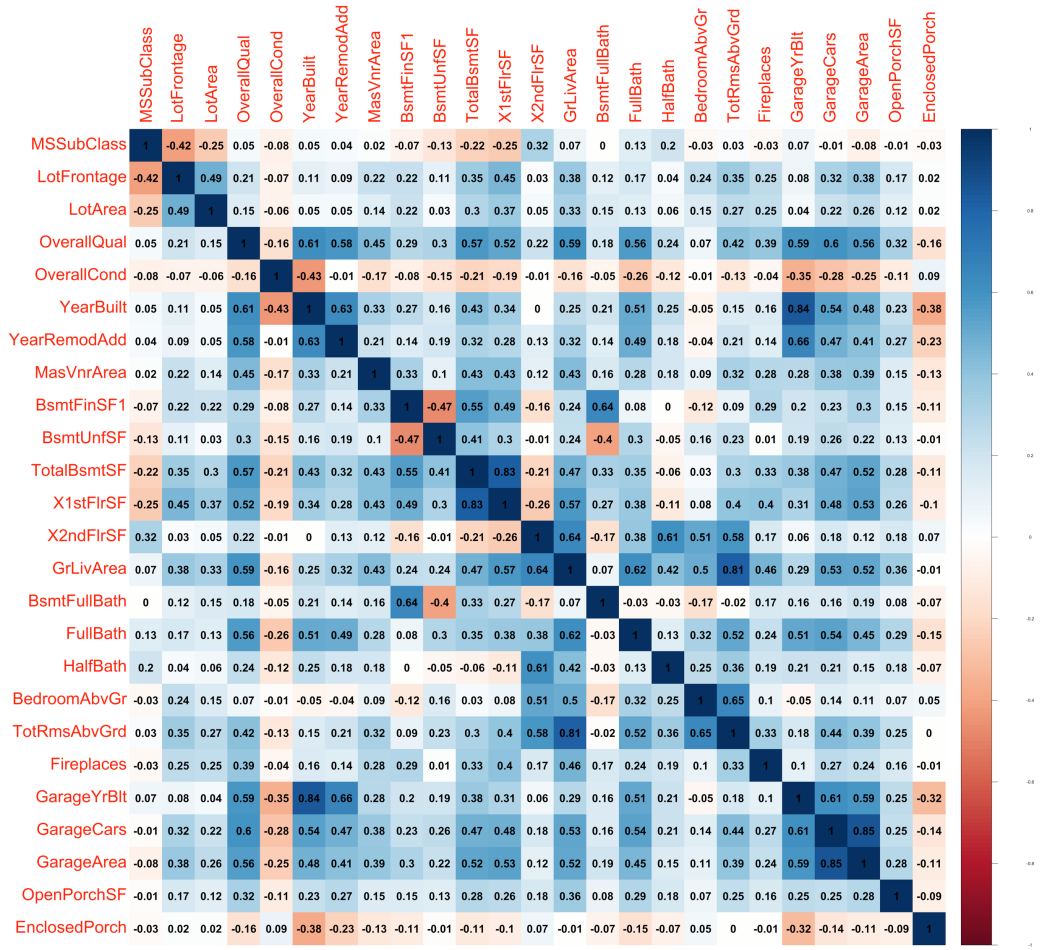

Figure 2: Correlation coefficients of the numeric variables without missing values

In order to obtain a more precise estimation of the missing data, a chart for the correlation coefficients of the numeric variables is produced for observing which kind of variable is useful or effective in predicting the missing values in some variables.

2.2 Handling of the missing data

After reading the description of the data columns and observing the pattern of the missing data, there could be three types of variables with missing data.

2.2.1 Quantitative variables with missing values

Variables belongs to this type: LotFrontage

LotFrontage is the only variable in this category which needs the estimation from other variables. Correlation coefficients will be considered for variable selection in the estimation of LotFrontage.

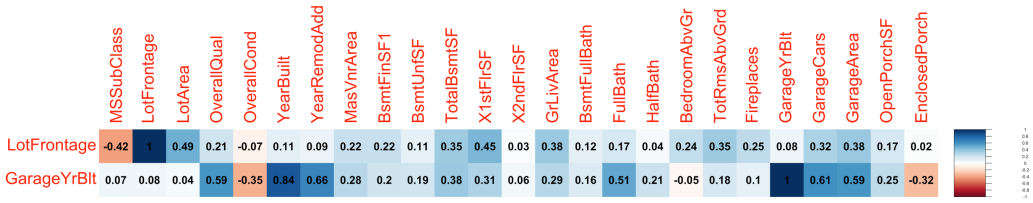

Figure 3: Correlation between numeric variables and LotFrontage

After the observation of correlations and the description of data columns, LotShape, LotArea, LotConfig and MSSubClass are chosen to estimate the value of LotFrontage, because the first three variables are named with the prefix Lot which refer to some characteristics of the property(Lot). The last variable MSSubClass has a correlation coefficient -0.42 with LotFrontage, which is relatively higher than other variables, is chosen as one of the predictors. The method used for the prediction of LotFrontage is **random forest model**.

2.2.2 Categorical variables with missing values

Variables belongs to this type: MSZoning, Utilities, Exterior1st, Exterior2nd, Electrical, KitchenQual, Functional, SaleType

The handling of each missing data is described as below.

- MSZoning: It identifies the general zoning classification of the sale. Filled with a new level in the factor, “Unknown”.
- Utilities: Type of utilities available. Filled with a new level in the factor, “Unknown”.
- Exterior1st: Exterior covering on house. Filled with existing level in the factor, “Other”.
- Exterior2nd: Exterior covering on house (if more than one material is used). Filled with a new level in the factor, “Other”.
- Electrical: Electrical system. Filled with a new level in the factor, “Unknown”.
- KitchenQual: Kitchen quality. Filled with the level with the highest frequency in the factor, which is “TA” – Typical/Average.
- Functional: Home functionality. Filled with a new level in the factor, “Unknown”.
- SaleType: Type of sale. Filled with existing level in the factor, “Oth”.

Most of the variables here have not been estimated through using other variables, because most of these variables have number of missing values less than 5, which is just a little proportion of the full number of data. The values in these variables will not affect too much even if they are unknown.

2.2.3 Variables with missing values representing non-existence of prerequisite

Variables belongs to this type: MasVnr~, Bsmt~, TotalBsmtSF, FireplaceQu, Garage~ (~ represents some characters ended with a prefix, for example, BsmtQual is one of the Bsmt~.)

The missing variables in this category have a characteristic in common. If the prerequisite of the variables does not exist or is not in the property, there will be missing values in those variables.

The prerequisites of the variables are described as below:

- MasVnrType~: The masonry.
- Bsmt~, TotalBsmtSF: The basement.
- FireplaceQu: The fireplace.
- Garage~: The garage.

In other words, if the facilities related to these variables does not exist, we have to fill in "None" as a new level to replace the NA values for these factors if the variable is categorical, 0 if the variable is numerical.

However, some of the data could be in special case that although the facilities do exist, there are missing values in other correlated variables. For example, some of the data has existing GarageType, but their GarageYrBlt, which is the built year of the garage, is missing in the records.

After the investigation on data, Garage~ is found to have such case. GarageType is chosen as the indicator variable in this family of variables to indicate the existence of garage. If GarageType is not "None", that means the garage exists, we build up a **random forest model** to estimate the missing values for other variables in the family, which the predictors include GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea, GarageQual, GarageCond, YearBuilt, YearRemodAdd and OverallQual, with the existing records. The last three variables are not in the Garage family but still added into the list of predictors as they have a relatively correlation coefficient with GarageYrBlt, GarageCars and GarageArea as shown in Figure 2.

Finally, all missing values are confirmed and filled with some proper values. The prediction of SalePrice is based on the columns except

Id, Alley, FireplaceQu, PoolQC, Fence, MiscFeature

as specified in earlier section. We also need to exclude Id as it is just an identification of a piece of data.

2.3 Prediction on Test Data

Three models are chosen for the prediction of the SalePrice in the test dataset, which includes random forest model, LASSO model and the gradient boosting model. For each prediction model, there are 74 predictor variables, which exclude the ones mentioned in last page of the report.

2.3.1 Random Forest

Random Forest is applied on the test dataset with number of trees = 500 and number of random variables selection in each split = 74/3 = 24. After the completion of the regression, a chart of the importance of the 74 variables (in term of Logarithmic Mean Decrease Gini) is plotted as below.

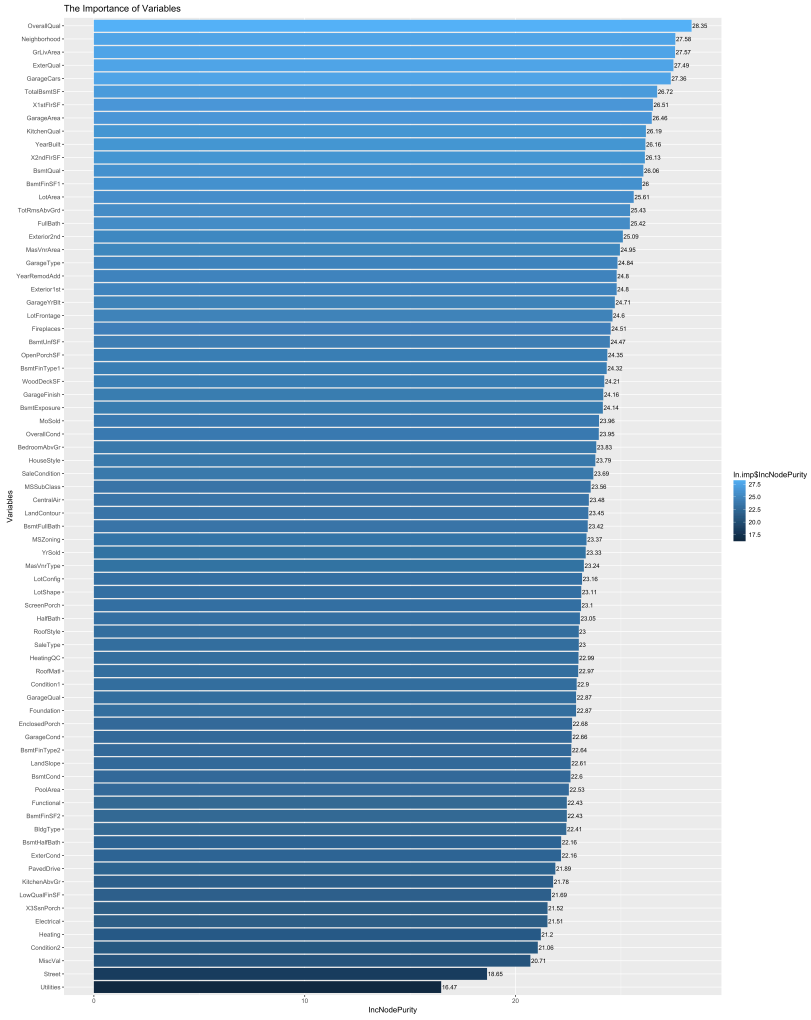

Figure 4: Logarithmic Mean Decrease Gini of the 74 variables

There is not a significant difference between the importance of the variables, which is ranged from 16.47 to 28.35.

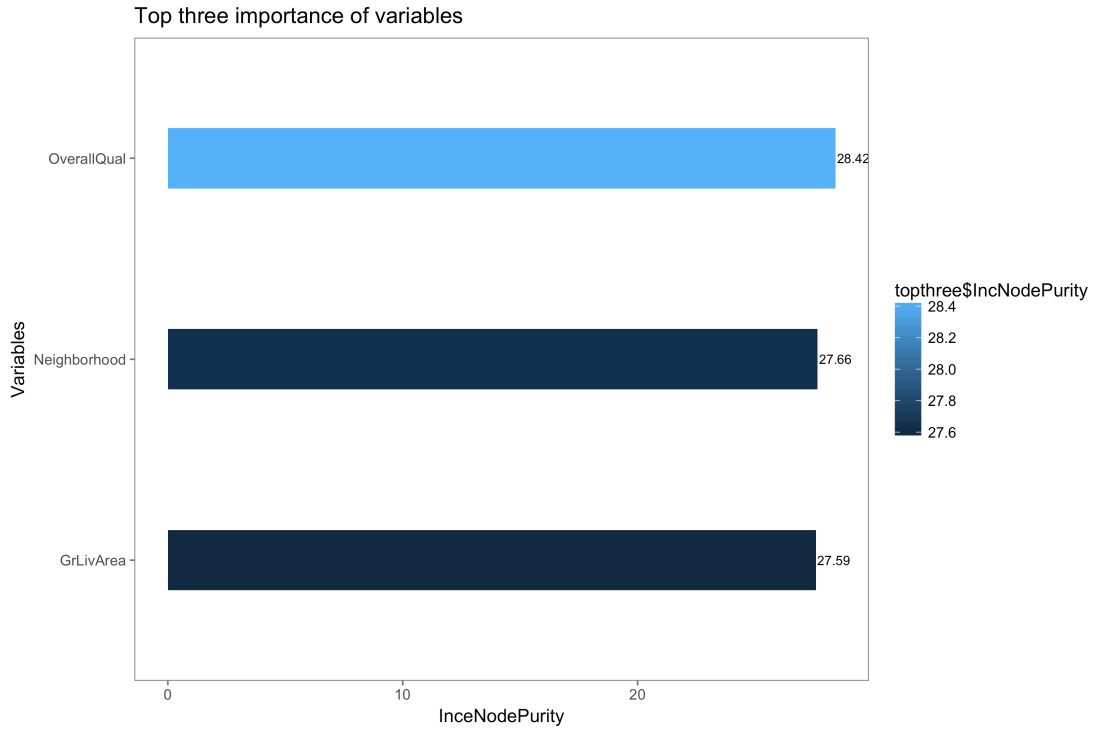

Figure 5: Top three of the most important variables

As we can observe, top three of the most important variables are OverallQual, Neighborhood and GrLivArea, with log(Mean Decrease Gini) values 28.42, 27.66 and 27.59 respectively. They are most significantly important in predicting the value of SalePrice.

2.3.2 LASSO

LASSO is applied on the test dataset with minimum penalty on the absolute value of the coefficient of the predictors.

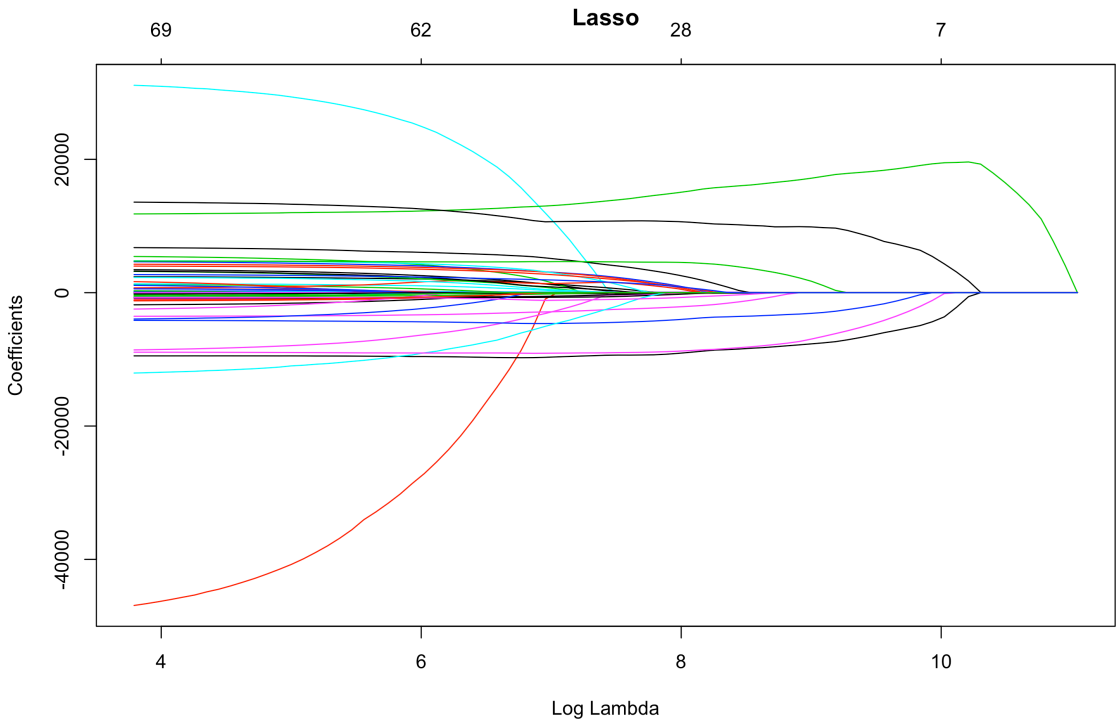

Figure 6: Shrinkage of coefficients with respect to the increment of penalty

All coefficients are shrunk to 0 as the log-version of penalty increased, this is expected in a LASSO model.

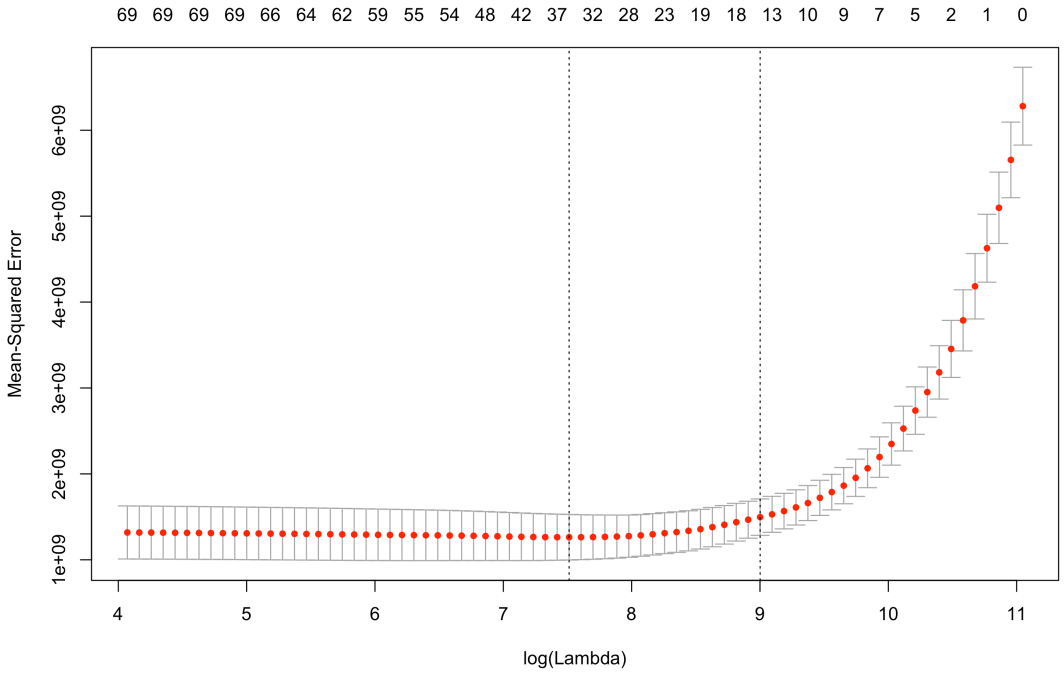

Figure 7: Increment of MSE with respect to the increment of penalty

The value of mean squared error has increased when the log-version of penalty increases. Therefore, it is better for us to choose the least value of the penalties.

```
cv.lasso$lambda.min
```

```
## [1] 1520.229
```

```
log(cv.lasso$lambda.min)
```

```
## [1] 7.326616
```

Figure 8: Minimum Penalty and its log-version

After the model fitting, the minimum penalty is computed as 1831.121 with logged version as 7.51. Therefore, we can choose this value as the penalty for all coefficients in the prediction for the sale price in test dataset.

2.3.3 Gradient Boosting

GBM is applied on the test dataset with number of trees = 20000, shrinkage parameter to each tree = 0.01 and the split is stopped when there are 20 observations in each terminal node as to lower the complexity of extending the tree.

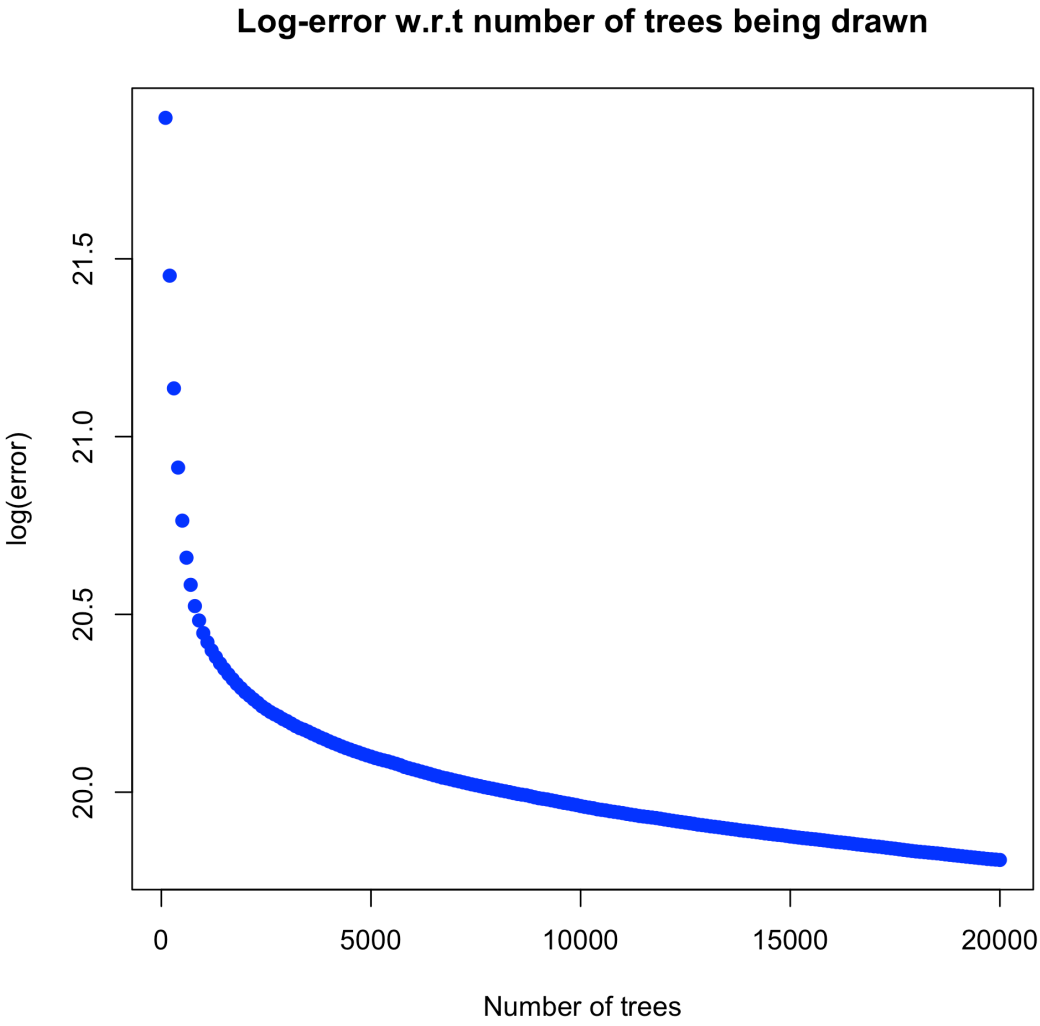

Figure 9: Logarithmic error versus number of trees being drawn

Error is then calculated by the prediction of model on the train dataset subtracted by the real value of the SalePrice, and take log function on the MSE. As we can observe, as the number of trees being drawn is increased, the logarithmic error will gradually decrease. Theoretically, the error will decrease as more as the number of trees being drawn, for the sake of preventing the over-fit problem, a suitable number of tree as 20000 is being chosen.

3 Result of the prediction

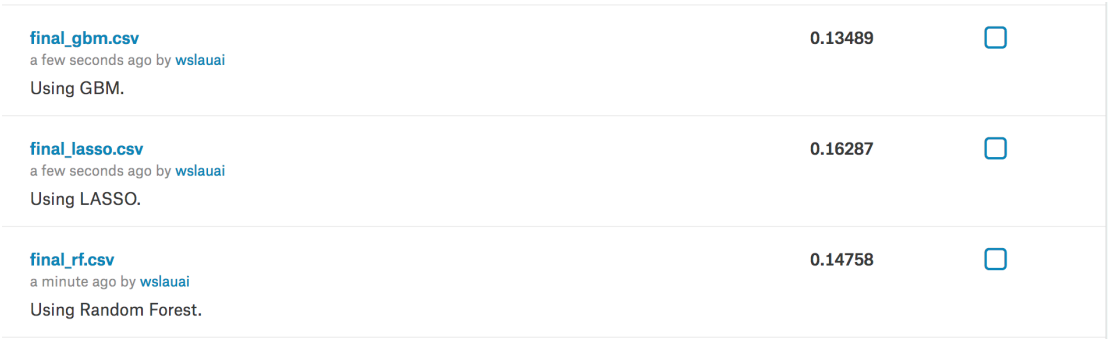

Figure 10: Score achieved after submission of the solution

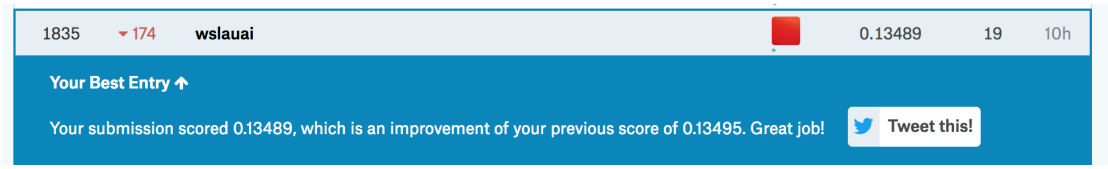

Figure 11: Ranking on the Kaggle Competition

On the Kaggle Leaderboard, scores are presented as the Root Squared Mean Logarithmic Error(RSMLE).

4 Conclusion

The RMSLE between the prediction of the estimation and the true values of the sale prices for the three models are shown below.

Model	RMSLE
Random Forest	0.13489
LASSO	0.16287
GBM	0.14758

Table 2: RMSLE of the models

In which Gradient Boosting performs the best.

5 Reference

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017) *An Introduction to Statistical Learning with Applications in R*. Springer Science+Business Media New York.