MATH4432 Mini-Project 2 :
Kaggle contest regression: Predict house sales price
Chow Wing Ho(20279607)

This project is going to explore the predict value of house sales price by multiple regression and lasso regression .

Methods

- Missing values
  - PoolQC, MiscFeature, Alley, Fence, FireplaceQu had deleted in the dataset as these variables had many missing values
  - All train and test contain same Utilities, therefore, Utilities dropped in the dataset
  - LotFrontage and GarageYrBlt are used median of whole data to cover
  - Other factor type variables are used the highest observations in whole data with corresponding variables to cover
  - Other value type variables are used the highest observations in whole data with corresponding variables to cover
- Dataset
  - There are two dataset: train data set with the sales prices and the test set with the missing sales prices

Test the multiple regression

- Step1: choose some important dependent variables for regression and do the one dependent variable regression for each one

  - HouseStyle, LotArea, Neighborhood, Condition1, Condition2, BldgType, YearBuiltm YearRemodAdd, OverallQual, Mosold, OverallCond, YrSold

  - lm1 = lm(SalePrice ~ LotArea + Neighborhood + Condition1+ Condition2+ BldgType + HouseStyle + YearBuilt + YearRemodAdd + OverallQual + OverallCond +MoSold +YrSold,Tr)

  - Drop MoSold and YrSold for the dependant variables as these two variables are not significant in both lm 1 and the lm 11 (lm(SalePrice~MoSold,Tr)) and 12 (lm(SalePrice~YrSold,Tr)

- Step 2 : one more time for regression

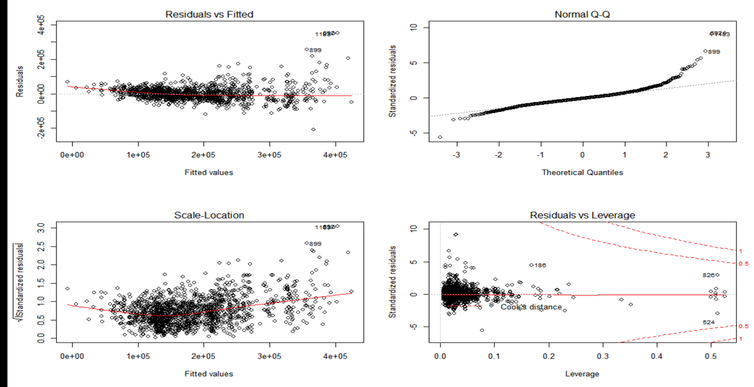
  - lm14 = lm(SalePrice ~ LotArea + Neighborhood + Condition1+ Condition2+ BldgType + HouseStyle + YearBuilt + YearRemodAdd + OverallQual + OverallCond, Tr)

  - Drop the OverallCond for dependent variables as OverallCond are insignificant in lm 1 and lm 14

- step 3: lm15 = lm(SalePrice ~ LotArea + Neighborhood + Condition1 + Condition2 + BldgType + HouseStyle + YearBuilt + YearRemodAdd + OverallQual,Tr)

  - All dependent vaiables are significant with small p value with 0, 0.001, 0.01,0.5

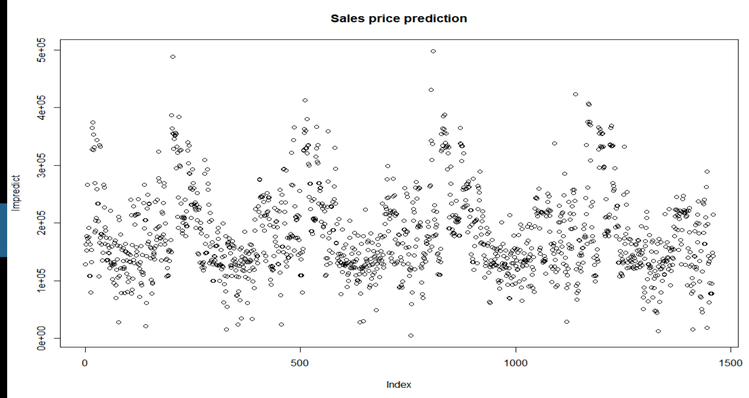
Diagram of m15



- Step 4 : Prediction with the dependent variable in lm15

  - Impredict = predict(lm(SalePrice ~ LotArea + Neighborhood + Condition1 + Condition2 + BldgType + HouseStyle + YearBuilt + YearRemodAdd + OverallQual,Te))

Diagram of sales prices by multiple regression

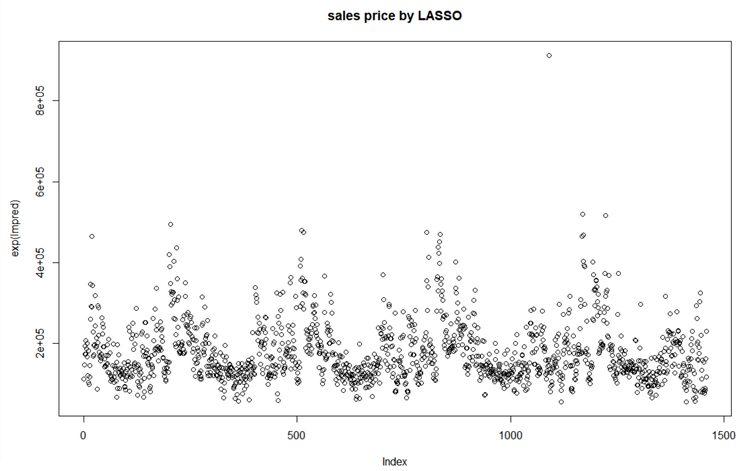


. The first submission with the prediction of house sales prices by using the multiple regression method to the Kaggle. The score is 0.55. It means the prediction is not good and need to modify.
Then, try to use the LASSO Regression to predict the house sales price

Lasso Regression

- Since lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable.
- The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model. Variables with non-zero regression coefficients variables are most strongly associated with the response variable.
- Try to use the Lasso Regression for the prediction of sales prices which can reduce the prediction error or not
- use the package “glmnet” for lasso regression

Diagram of sales prices by lasso regression



Conclusion

- Use the lasso regression to predict the house sales price is better than use the multiple regression means that the prediction error of house sales prices is improved and being more accurate.
- The score (RSME) is improved from 0.55 to 0.13187
- The score is 0.13187 and rank 1573 in the leaderboard.