Data Science(II) Midterm Project

Yimeng Shang

Abstract

This report discusses a study of predicting the sale price of houses. Our goal is to build a predictive model choosing from different regression methods, like ridge regression, LASSO regression, MARS method etc. We use cross validation to compare the prediction performance of these models. Our result shows that all of our models were valid and prediction ability of lasso is better among the models.

Introduction

Buying a house is an important thing for us in the whole life. Different people prefer different conditions like the material of roof and how far it is from the road. Importantly. The price of a house is influenced drametically by the different conditions of houses. In this study, we are interested in how these conditions affect the sale prices and making prediction based on our selected model.

Dataset

The data *housing.csv* have 1460 row and 81 columns. The first columns inclues covariate "ID" which lables individual housing condition and the last columns "SalePrice" in dollars is the response. Other 79 covariates are all treated as predictors of sale prices, including the type of alley access, overall material and finish quality, roof material etc. Some of the predictors are categorical data, so I factorized them. While keeping others as continious variable. To be clarify, year built is treated as continious variable and sale month can be treated as categorical variable. I treated some ordinal categorical covariates as factor with ordinal level. There are 35 numeric variables, and 21 categoric variables.

Missing value analysis

It's easy to notice that there're lots of NAs. After checking the meaning of each column, lots of NAs means no such item instead of the value is missing. To deal with this issue, I encoded these false NA into "none" indicating they're not missing. I then visualized the rest missing value to see the distribution(**Fig. 1**). The missing values lies mainly in Linear feet of street connected to property (17.74%). After removing all missing values, the data has 1120 obervations. Then we split the data into two parts, using 75% of it for training and 25% of it for testing.

Exploratory analysis

The distribution of sale price is very right skewed, so I took a log transform to make it normally distributed (**Fig. 2**). As we know, multicolinearity may cause an issue when using linear model, so I checked the correlation plot. From it (**Fig. 3**), there are some coorelation between covariates.

We made box plots and catter plots (**Fig. 4**) to see the general relationship between each variable and sale price. For example, there are obvious trends that the sale price increase with the increase of overall quality, the number of fireplaces and total rooms above grade. The sold year doesn't influence the price so much. Sale price increase with the increase of living area and Total square feet of basement area. And the variance of some variables are near zero, so we removed them.

Models

Because there are some coorelation relationship between the covariates and some trend are non linear, linear model may not be a good way to fit the data. In this report, we used Ridge Regression, LASSO Regression, Principal Components Regression and MARS. We used 10 fold cross validation to validate the models. And we calculated the MSE on test dataset to get the performance.

Ridge Regression

Ridge regression tries to minimize $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$ where $\lambda \sum_{j=1}^{p} \beta_j^2$ is a penalty term. The goal of ridge regression is to balance the two ideas: fitting a linear model and shrinking the coefficients. We added all predictors into model fitting procedure. The assumption of ridge regression is the same as the assumption of linear regression, which is the response should follow a normal distribution. The data should be scaled before fitting. There is one tuning parameter λ in ridge regression and to selecet the best λ value, here we used cross validation searching from e^{-10} to e^{15}. From the plot, the best λ we got (0.1467) minimize the RMSE. From the importance plot (**Fig. 6**), Near positive off-site feature–park, greenbelt is the most important variable in predicting the sale prices. However, ridge regression can't select variables. Instead, it keeps all covariates in the final model. Also, when the data does not satisfy the assumptions, the prediction result may not good.

LASSO Regression

Similar to ridge regression, LASSO regression tries to minimize $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$. To fit the model, we added all predictors in it. LASSO shinks the coefficients of some covariates into zero to achieve variable selection. The assumption is the same as linear model and the data should be scaled. Cross validation is used to select the best λ, here we got 0.00449 is the λ minimizing the RMSE. From **Fig. 6**, same as ridge regression, Near positive off-site feature–park, greenbelt is the most important variable for prediction. LASSO is based on linear regression, so it may not perform well for nonlinear data. Also, when the data does not satisfy the assumptions, the prediction result may not good.

Principal Components Regression

Principal components analysis can be used to define the linear combinations of the predictors. The tuning parameter in this model is how many principal components to use. Also, we used 10 fold cross validation to select the number of principal components (here is 22) minimizing the RMSE. Rates the overall material and finish of the house plays important roles in predicting the response. PCR can reduce the dimension of the data. However, it is based on linear regression, so the prediction ability is not good for nonlinear data.

MARS

Multivariates adaptive regression splines creates a piecewise linear model. Given a cut point c for a predictor, two new features are hinge functions of the original. This is a good way to fit nonlinear data. The tuning parameter in this model is the product degree and the number of terms. To select the best tuning parameter, we used cross validation. From the **Fig. 6**, when product degree is 2 and the number of terms if 13, we got the smallest RMSE. From the importance plot, Rates the overall material and finish of the house plays important roles in the prediction.

Model selection

After resample several times from the data, we can draw box plots of RMSE of each model. **Fig. 7** shows, the overall RMSE of LASSO is the smallest, which means that the prediction ability of LASSO is better among these models. So we choosed LASSO as the final model. Among the MSE of test data, mars got the smallest 0.02 wheras LASSO got the largest 0.039, which shows that although the prediction ability of LASSO is good, the performance of it on test data is not good.

Appendix A

Figure 1: missing value distribution

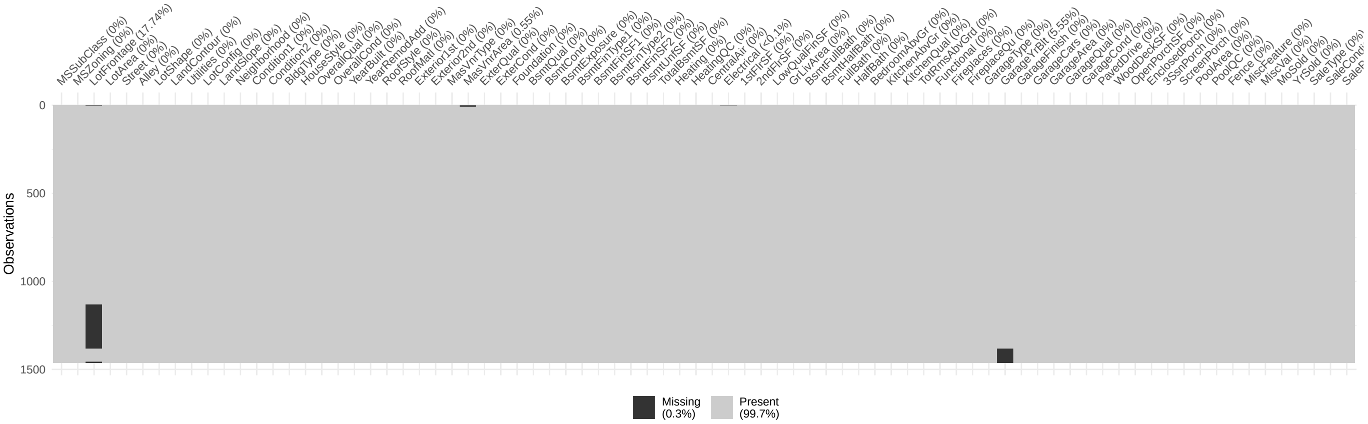

Figure 2: transformation of response

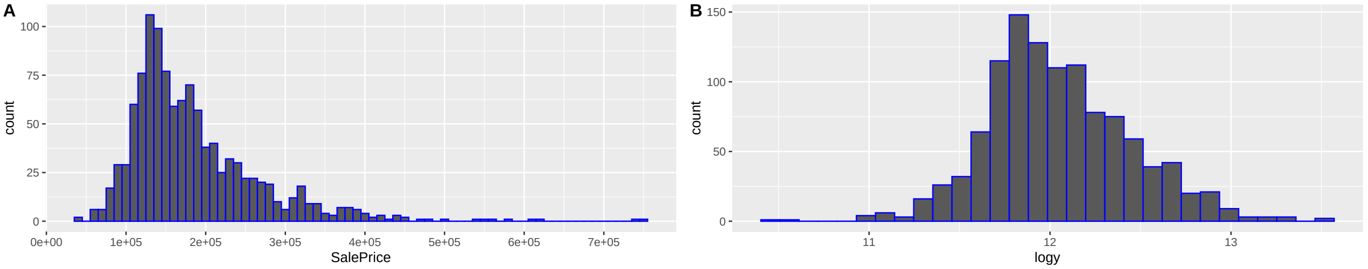

Figure 3: Coorelation plot

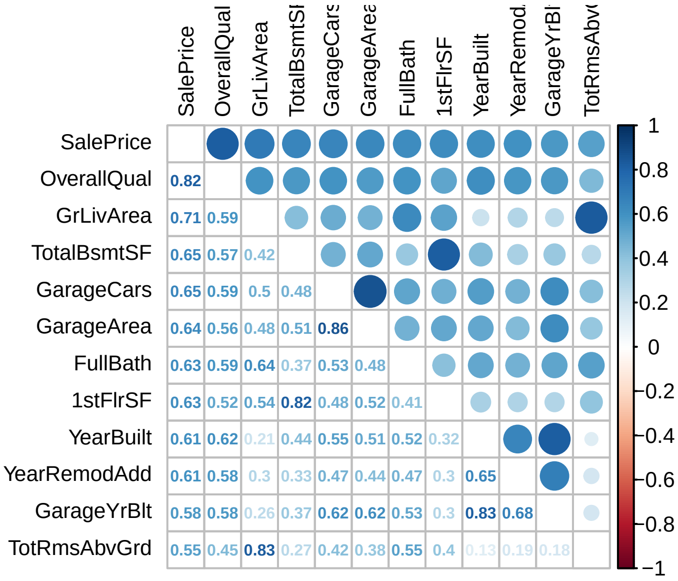

Figure 4: Categorical data distribtuion

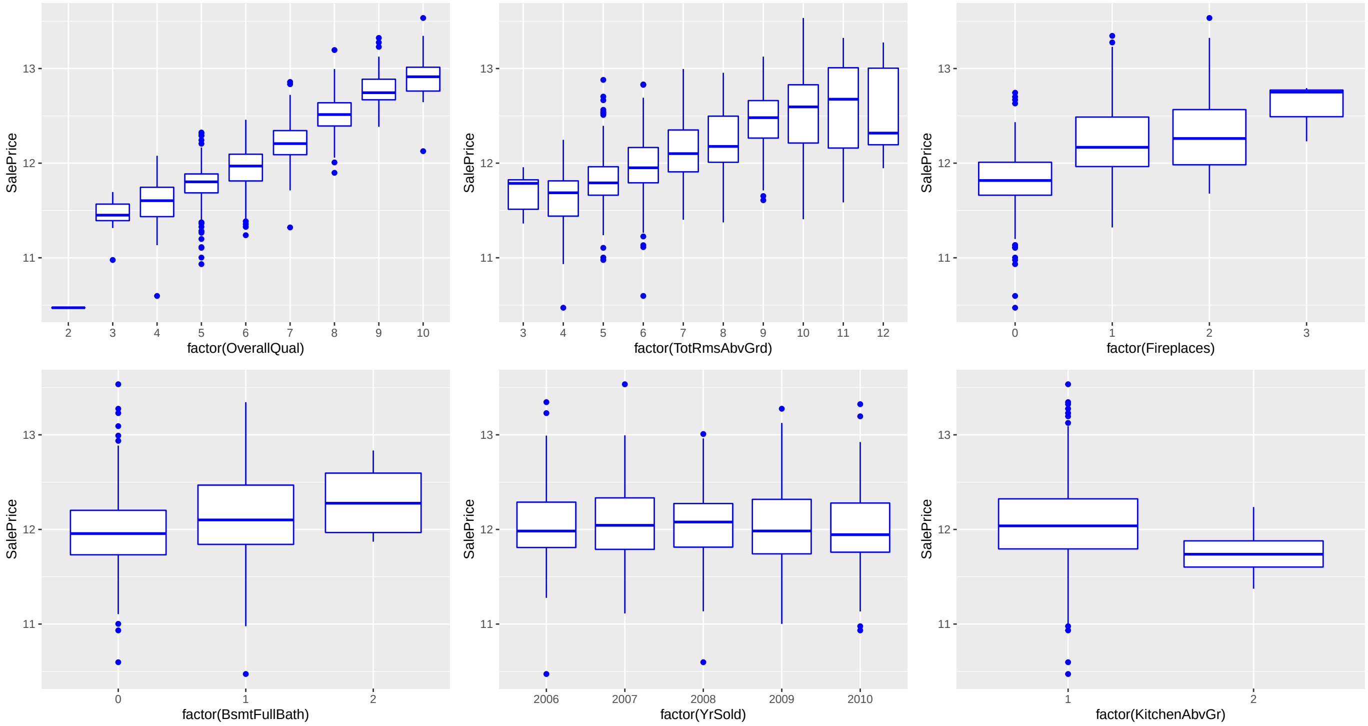

Figure 5: Feature plot for continious data

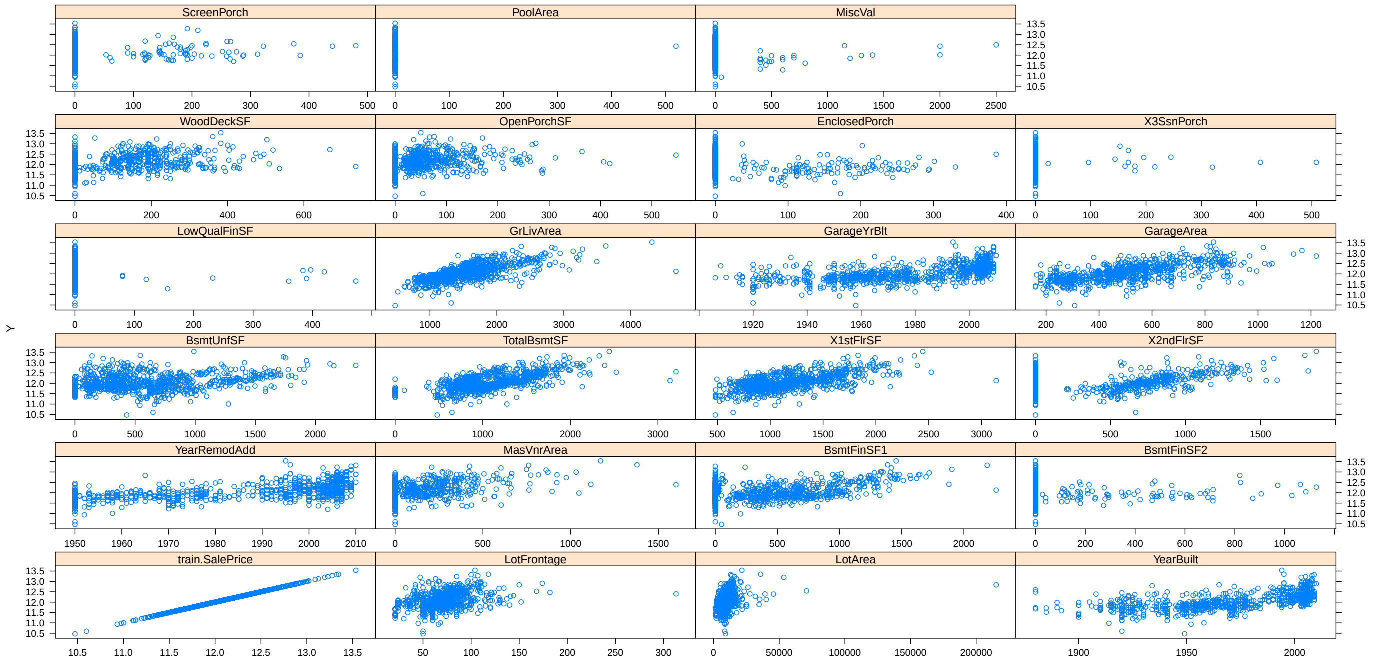

Figure 6: Tuning parameter and VIP plot

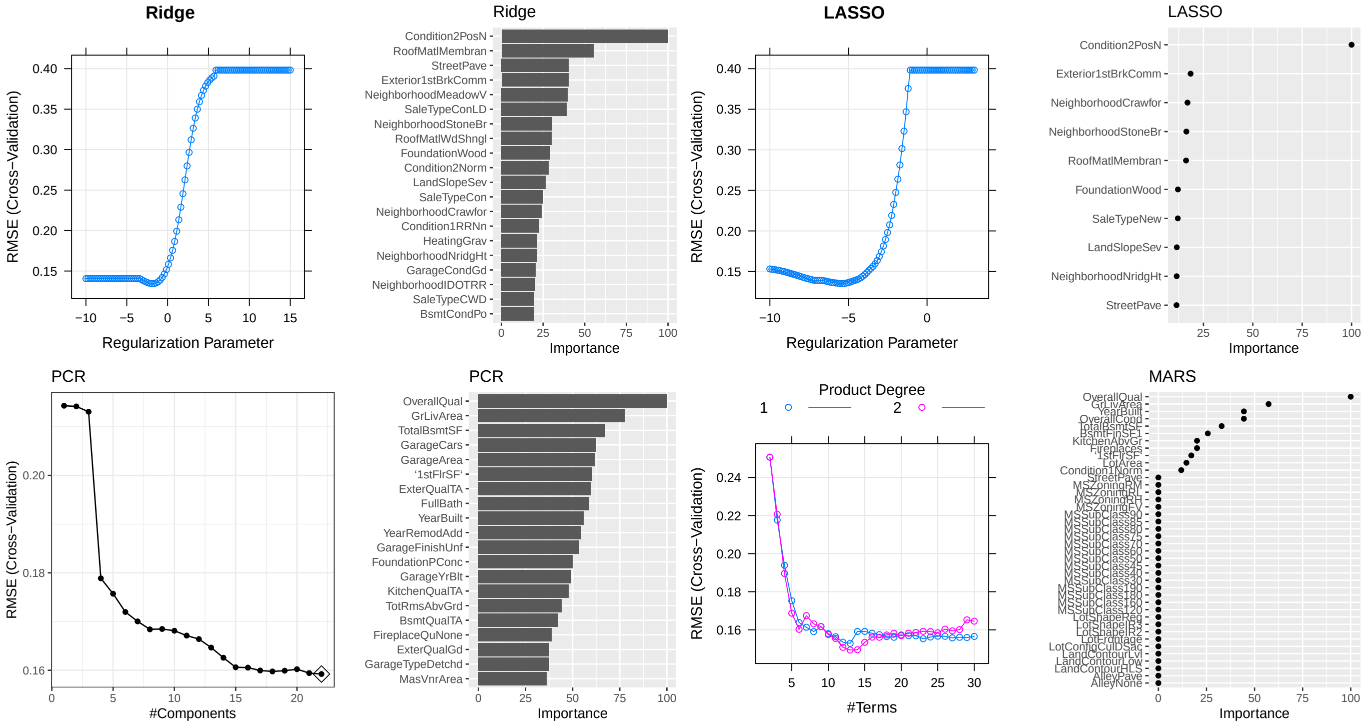

Figure 7: RMSE distribution

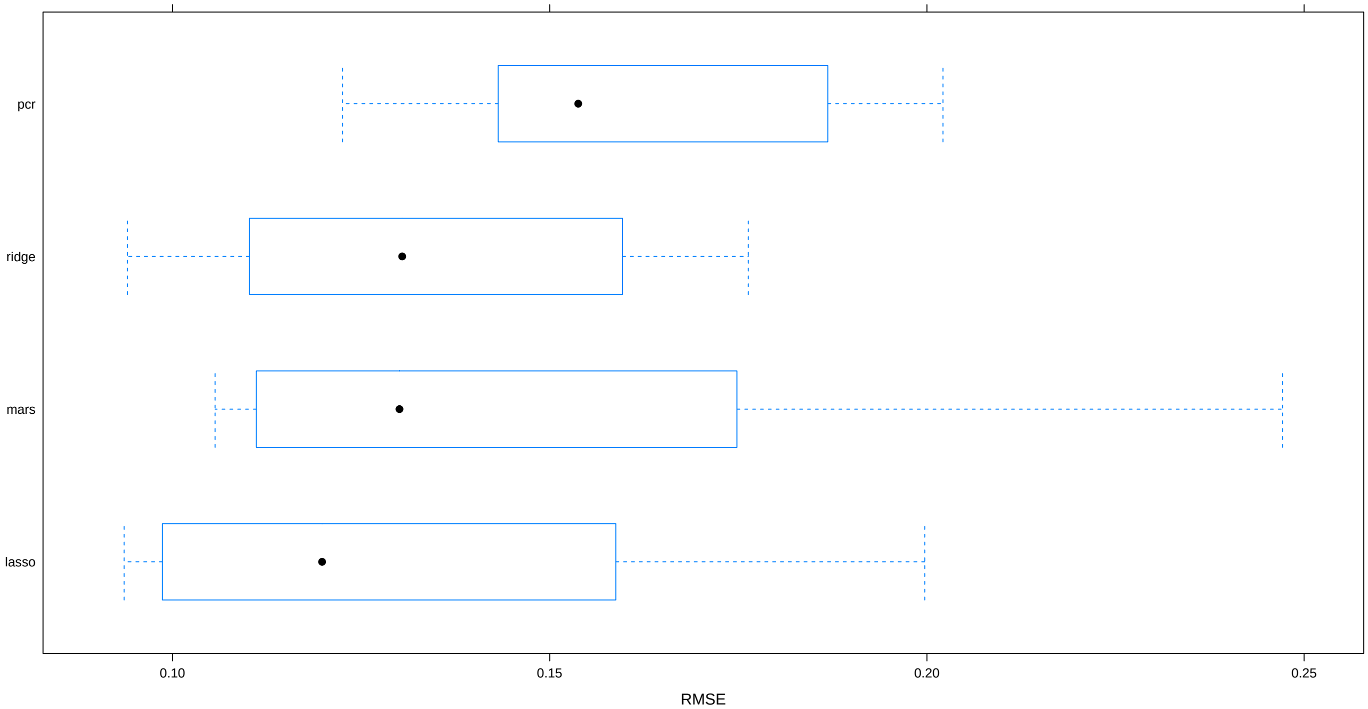

Table1: MSE for test data

Table 1: MSE of four models

mars_mse	ridge_mse	pcr_mse	lasso_mse
0.0201592	0.0323984	0.0345776	0.0392343