Homework 6

Yee Sern Tan[*]

Abstract

Regression is done to predict House Prices data on Kaggle. Linear regression and random forest are used for the task. Linear regression with interaction terms, and pruned with BIC, is found to exhibit the best result. The results are evaluated using cross validation on the training set and by Kaggle on the testing set. Linear regression diagnostics are also provided.

Keywords

data cleaning — linear regression — random forest — cross validation — Kaggle

[*] *Data Science Department,*
School of Informatics, Computing, and Engineering,
Indiana University,
Bloomington, IN, USA

Contents

1. Problem and Data Description

This data set, titled *House Prices: Advanced Regression Techniques* is from Kaggle, features 79 explanatory variables on residential homes sold in Ames, Iowa, within 2006–2010. The problem is to predict house prices based on the given explanatory variables. The data is divided into a training set and a testing set, both containing 1460 entries. Only the training set contains the sale prices of houses. The evaluation of the testing set is to be done with the root-mean-squared-error of the logarithm of sale prices.

2. Data Preprocessing & Exploratory Data Analysis

2.1 Handling Missing Values

A run of the R code shows that there are 19 variables in the training set, and 33 variables in the test set, that contain missing values.

Missing values for categorical variables are divided into two situations: in the first situation NA represents a valid category which should be added to the existing categories, in the second NA are true missing variables that are to be imputed with the mode in the train (or test) sets respectively. The code for converting NA into a new category is done differently for linear regression and random forest, because the randomForest package does not accept NA values, forcing us to use the string "NA" for a new category. There is another variable MasVnrType having "None" misrepresented as NA, which is reversed in the code.

For numeric variables, the situation is handled separately for each variable, as each has a different distribution. For variable MasVnrArea, from an exploratory plot of its histogram, the highest concentration of its value (mode) is at 0, which is imputed to missing variables. The missing values of variable LotFrontage is imputed with median as the most likely occurring value. The missing values of variable GarageYrBlt is imputed with mean to cancel out all effects that a regression analysis will have. Also, when these following variables are NA, it is assumed that they have value 0: BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath, GarageCars, and GarageArea. Upon inspection, an obvious data error is corrected: where MasVnrType is "None", MasVnrArea should be 0.

2.2 Exploratory Data Analysis

First and foremost, before doing the regression and random forest training, the sale price is converted to its logarithm. It will be exponentiated back after the final prediction.

Next, as values are to be converted to logarithms, it would be necessary to consider variables laid out in dollar terms before converting to logarithmic scales. Similarly, after the prediction is done, these values need to be added back to the sales price. The corresponding variable identified is MiscVal. It was also found that in the test set the variable MSSubClass contains a level that is not present in the train set. Its value is imputed to the mode.

Two columns are of interest: MoSold and YrSold. The month sold should be expected to exert some seasonality effects, but upon calculation of regression coefficient significance values, it is not significantly different from 0. The year sold should be tricky, because the period covered includes the subprime mortgage crisis. Therefore, the author has introduced a new variable named "time" as

$$time = (YrSold - 2006) * 12 + MoSold.$$

2.2.1 Linear Regression

Later, after fitting and pruning the results the regression coefficient is insignificant. This seems to imply that linear regression alone is not able to capture the effects of time. The author suggests using non-parametric methods like LOWESS for such variables. However, the deadline does not permit this endeavor.

The data being large (many columns and many rows) and varied, automation is required to observe the interactions between these variables. The approach taken to construct the linear regression model is described as follows:

1. Fit a linear model of sale price with all the explanatory variables.
2. Prune the linear model with AIC using step() function.
3. Separately fit new linear models with interaction of certain subsets of explanatory variables that are perceived to be related. The model is then to be pruned with AIC using step() function, and then evaluated using anova() function. Significant variables (p below 0.05) are added to the pruned linear model. The perceived related explanatory variables are:
    - MSSubClass * MSZoning
    - LotFrontage * LotConfig + LotArea * LotShape + LotFrontage * Street + Street * Alley
    - LotShape * LandContour * LandSlope
    - BldgType * HouseStyle * OverallQual * OverallCond * YearBuilt * YearRemodAdd
    - RoofStyle * RoofMatl + Exterior1st * ExterQual * ExterCond + Exterior2nd * ExterQual * ExterCond + MasVnrType * MasVnrArea * Foundation
    - Heating * HeatingQC * CentralAir
    - (BsmtQual *BsmtCond * BsmtExposure + BsmtFinType1 : BsmtFinSF1 + BsmtFinType2 : BsmtFinSF2 + BsmtUnfSF) * TotalBsmtSF
    - X1stFlrSF * X2ndFlrSF * LowQualFinSF * GrLivArea
    - BsmtFullBath * BsmtHalfBath * FullBath * HalfBath
    - BedroomAbvGr * KitchenAbvGr * TotRmsAbvGrd + KitchenAbvGr * KitchenQual
    - FireplaceQu * Fireplaces
    - (GarageType + GarageFinish + GarageQual + GarageCond + PavedDrive + GarageYrBlt + GarageCars + GarageArea)^3 (Note: this item is too large and is therefore split into two steps to speed up the computation)
    - PoolArea * PoolQC
    - SaleType * SaleCondition
4. The newly constructed model is pruned again using AIC and BIC, and comparatively evaluated with test set.
5. The linear model resultant from BIC is plotted to remove data with highest leverages, and the whole model is fitted again.

Pruning and doing model selection using the step() function is not a perfect solution, as the ordering of variables can influence the result and may lead to non-optimal results. Nevertheless, this has been the best that the author has tried with linear regression.

Due to the presence of too many interaction terms having many different levels, it would be too troublesome to do *k*-fold cross validation on the data as many factor levels present in a testing set would not be present in the training set. Instead, the author proceeds to generate the sale prices of the test set and evaluate them on Kaggle. It is found that BIC yields a better result. Therefore, it is deemed that AIC overfits. The final BIC-pruned model obtained has formula:

$$\begin{aligned}
&\log(SalePrice - MiscVal) = \\
&MSZoning + LotFrontage + LotArea + OverallQual + \\
&OverallCond + YearBuilt + YearRemodAdd + \\
&Foundation + BsmtFinSF1 + BsmtFinSF2 + \\
&CentralAir + X1stFlrSF + X2ndFlrSF + \\
&LowQualFinSF + KitchenAbvGr + Functional + \\
&Fireplaces + GarageCars + GarageArea + \\
&WoodDeckSF + EnclosedPorch + ScreenPorch + \\
&SaleCondition + BsmtQual * TotalBsmtSF + \\
&X1stFlrSF * X2ndFlrSF + X1stFlrSF * GrLivArea + \\
&X2ndFlrSF * GrLivArea + GarageCars * GarageYrBlt + \\
&GarageArea * GarageYrBlt + \\
&X1stFlrSF * X2ndFlrSF * LowQualFinSF + \\
&X1stFlrSF * X2ndFlrSF * LowQualFinSF * GrLivArea
\end{aligned} \tag{1}$$

2.2.2 Random Forests

Random forests are reputed to be a model that is easy to construct computationally and is one that is less prone to overfitting. It is tried here to compare with linear regression results.

In this regression forest, we implement with seed 1234, the default mtry=26, and ntree=1200. We have chosen ntree=1200 after doing cross validation with ntree=200, 500, and 1200, with 5 repetition of 10 folds. The result of cross validation is presented as follows:

```
== Summary of a  Cross Validation
Performance Estimation Experiment ==

Task for estimating  rmse  using
 5 x 10 - Fold Cross Validation
 Run with seed =  1234

* Predictive Tasks ::  randomForest
* Workflows  ::  randomForest.v1,
randomForest.v2, randomForest.v3

-> Task:  randomForest
  *Workflow: randomForest.v1
             rmse
avg     0.13590544
std     0.01484563
med     0.13714514
iqr     0.02200960
min     0.10870770
max     0.16833535
invalid 0.00000000

  *Workflow: randomForest.v2
             rmse
avg     0.13570398
std     0.01478097
med     0.13707702
iqr     0.02205024
min     0.10889140
max     0.16812614
invalid 0.00000000

  *Workflow: randomForest.v3
             rmse
avg     0.13547014
std     0.01473427
med     0.13691643
iqr     0.02223300
min     0.10871343
max     0.16858564
invalid 0.00000000
```

The best result, i.e. that for ntree=1200, is selected.

3. Algorithm and Methodology

The algorithm and methodology has been thoroughly explained in the previous section, as the author considers it an iterative procedure to alternate between exploratory data analysis and model selection.

4. Experiments and Results

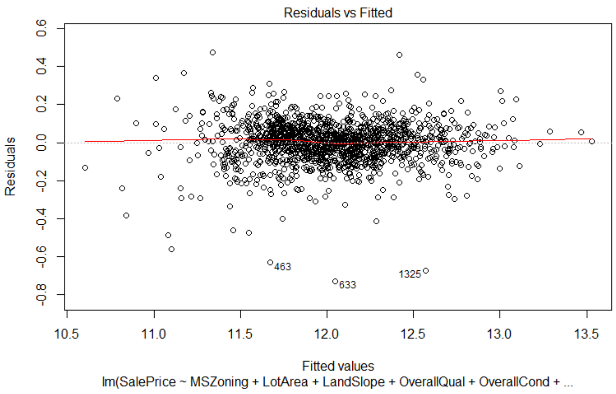
Having removed the data points identified with high levelrage (584, 667, 1004, 1231, 826, 524, and 1299), the linear model is fitted again and pruned with BIC. The results for linear regression are uploaded using the test set to Kaggle for evaluation, with the best score achieved using BIC pruning of the linear model with interactions being 0.12654 (with submission name ystan87). To generate results for random forests, the following error occurred:

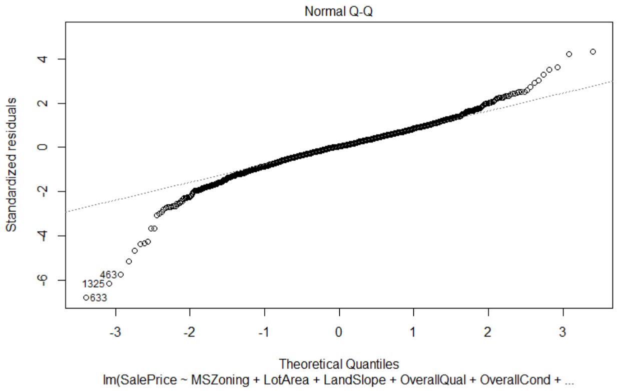
```
> preds <- exp(predict(trained.model,
  HPTest)) + HPTest\$MiscVal
Error in predict.randomForest(
  trained.model, HPTest) :
    New factor levels not present
in the training data
```

Random forest package does not indicate which factor in which variable is not present in the training data, and given that the average results from our cross validation is not as good as the result in the linear regression model, effort is not put in to debug and resolve this error.

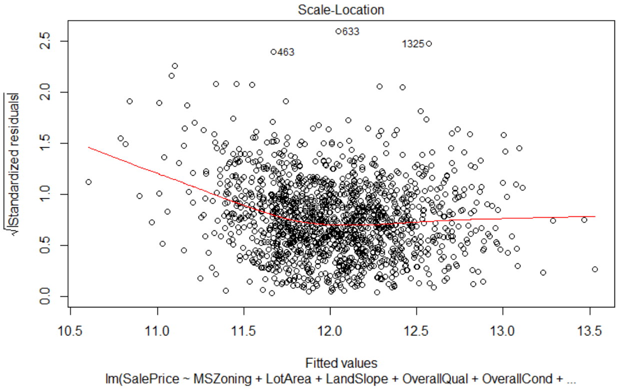
4.1 Linear Regression Diagnostics

Without diagnostics, any regression analysis would not be complete. The residuals and standardized residuals are plotted against the fitted values. Apart from few outliers, there is no obvious trend on their distributions.

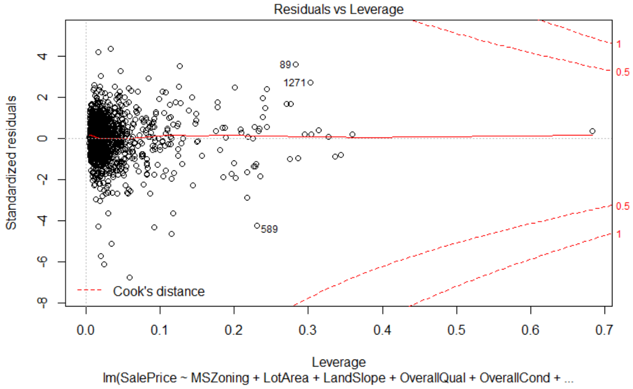




The QQ-plot below shows that the model used deviates from normal at the upper and lower end of residuals. Thus, the assumptions of least squares regression are not entirely valid. Having only few samples that deviate extremely may suggest that there are other factors influencing the sale price present in these samples.



The following is a plot of residuals of samples over their leverages, with level curves of Cook statistic at 0.5 and 1 plotted. It should be reported that there are samples with leverage 1 and are not plotted. They are rows 121, 272, 1270, and 1293 .



Other than a few outliers, the majority of data have good fit to the training data, and do not exert disproportionately large influences on the regression model.

5. Summary and Conclusions

In this work, the author attempted to predict house prices using linear regression and random forest. Linear regression with interaction terms, pruned with BIC, is found to be the best performing model. Several improvements can be suggested: non-parametric regression (such as LOWESS) for application on year and month sold, fitting models by discarding outliers, detailed testing for linear model improvements with cross validation, or ensemble methods such as AdaBoost for regression trees. As a homework, the work will stop here.

Acknowledgments

The author is indebted to open source programming language R and troubleshooting advice given on StackExchange.

References