MATH 4432 Project 2 Report

Anonymous Author(s)
Affiliation
Address
email

Abstract

This project studies the regression and prediction of the housing prices using Ames Housing dataset. The aim is to come up with a model to predict the housing price and understand major factors affecting the housing price through model interpretation. We first attempt to manipulate with the NA value, in the process of which the NA value handling methods are chosen according to the data types and thus, the complexity of the variable. Following this, we then start with the model selection by using various methods such as trees, random forests, boosting, lasso, ridge regression, etc. Finally we are able to come up with our prediction model and the ranking on Kaggle will be included at last.

1 Introduction

1.1 Background

Our report focuses on the exploration of the best model that is able to predict the housing price of the test set in the Ames Housing dataset, based on the given training set. Usual prediction on the housing prices utilize only a few number of variables, which is much less complete compared to the data set given in this Ames housing price dataset. The dramatic increase in the data dimension and volume brings about both more information and further issues in handling NA (not available) values and model selection, which are the main focuses that we are going to discuss in details in the further parts of this report.

1.2 Data Description

In the dataset given in the .csv files given on the website:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

we have 81 variables in the train.csv file and 80 variables in the test.csv (because we are lacking the reponse variable in the test set). There are catagorical and continuous variables among them and thus we may adopt different strategies in handling them. Also, there are NA values for several types that we may need to handle, the details will be discussed later.

1.3 Methodologies

Since our response variable is the housing price, which can only be reasonably treated as a continuous variable, an usual and intuitive learning model on which the response is inferred would be regression.

Before we can dive into the model selection procedure, a preliminary and essential analysis would be the handling of NA values. Our general idea is that if the cause of the NA value can be identified, we should treat the NA value according to its practical meaning in real life. If we are dealing with

Submitted to 30th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

more complicated variables where no explanation about the NA value is available, we will attempt to estimate the NA value based on the available data.

After handling the NA values, we will proceed with the model selection by using various methods including, Lasso, Ridge regression, trees, random forests, boosting, etc. The results from these models will be further interpreted to understand the major factors in determining the housing price.

Finally, a model with highest performance in prediction will be chosen and the ranking will be posted.

2 Handling NA values

2.1 Preliminary Checks

The very first check we are going to conduct is the number of NA values in each variable and the conditions for training set and test set are shown the following figures, respectively:

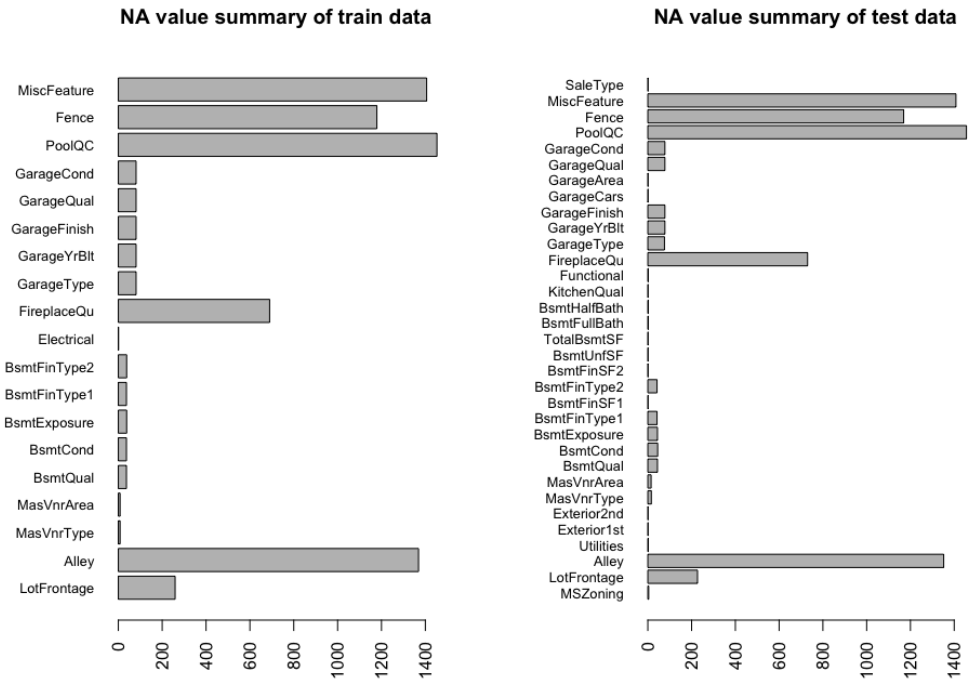

Figure 1: NA Values in Training Set & Test Set

Based on the number, cause and type of NAs in different predictors, we categorize them into three different situations: Meaningful, Meaningless&Negligible, Meaningless&Nonnegligible. The treatment of each kind of situations will be briefly introduced and the detail of the treatment can be found in the NA value handling section in the source code. More specifically:

a) Meaningful NAs:

Meaningful NAs are those NAs whose cause and meaning are clearly explained in the supporting file from Kaggle. For example, NAs in the predictor, *Fence*, means that the house does not have fence. For such kind of NAs, we simply replace the NA with its practical meaning. A take-away message would be replacing 'null' with 'no'.

b) Meaningless & Negligible NA values:

While most of the NAs in the dataset are meaningful and can be easily handled, there are still several predictors with unexplained NAs. Fortunately, most of these predictors contain fewer than 10

meaningless NAs. For example, though unexplained in the supporting file, the predictor, *Electrical*, in train data contains only one NA which is negligible when compared with the size of the whole dataset. Therefore, for this kind of NAs, we impute the NAs with median for numerical variables or mode for categorical variables of the available data.

c) Meaningless & Nonnegligible NAs:

The only predictor left after the treatment stated above is *LotFrontage*. As suggested by the supporting file, *LotFrontage* measures linear feet of street connected to property. With 259 and 227 unexplained NAs in train and test data, simply replacing NAs with median can be quite risky.

2.2 Details of Dealing With Meaniingless & Nonnegligible NAs

We treat the NAs in two steps. In the first step, we create a new predictor called *LotFrontageAva*. 'Yes' in *LotFrontageAva* means the *LotFrontage* of the house is available and 'No' in *LotFrontageAva* means the *LotFrontage* is not available. In future regression analysis, this predictor can suggest whether NAs in *LotFrontage* can affect price of the house. In the second step, we apply random forest model to impute the value of NAs in *LotFrontage* based on other predictors. In the random forest model, *LotFrontage* is regarded as the response variable and all other variables, except for *SalePrice*, the response variable we eventually want to predict, are regarded as predictors. All observations in training and test data with available *LotFrontage* values are used to train the random forest model. The NAs are then predicted by the trained model.

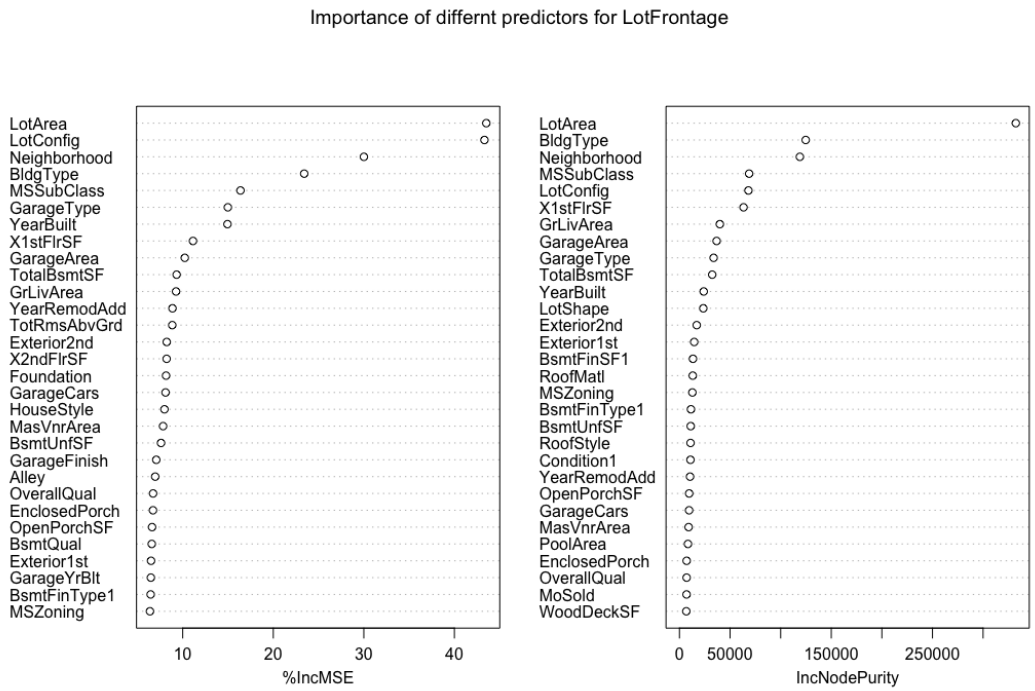

Figure 2: Importance of Different Predictors for *LotFrontage*

The incentive of imputing NAs in *LotFrontage* with this method is that similar house may have similar *LotFrontage*. For instance, houses in the same neighbourhood are more likely to have similar *LotFrontage* due to homogenous design patterns. Another example would be the lot area, as large buildings are typically built with wider streets. Results from random forest agrees with our intuition as high importance is indeed given to predictors such as *LotArea* and *Neighborhood* (Figure 2). To further demonstrate the validity of this method for imputation, a validation set approach is applied to estimate the imputation error.

By comparing random forest model with the estimation based on median, we can obtain the following results:

```
> rf.imputation.error
[1] 12.83242
> median.error
[1] 21.96269
```

Figure 3: MSEs of Random Forest Model & Median-Based Model

The MSE from the random forest model is 12.9 which is a significant improvement when compared with the MSE of 22.0 from the estimation based on median. Therefore, we claim that imputation of NAs in *LotFrontage* with random forest model is a more suitable approach.

2.3 Last Check

We examine at last if every NA has been eliminated in both training set and test set. And we indeed obtain two zero-matriices as indicator matrices, meaning that all NAs have been cleared.

3 Model Selection and Interpretation

3.1 Log Transformation of Response Variable, *SalePrice*

Instead of using the original response variable, we choose to log transform the response variable and fit model to the log transformed data. The reason is that the change in log-transformed response variable is a good approximation of percent change when the change is small. When compared with absolute change in price, we are more interested in percent change of house price caused by predictors because the significance of absolute change highly depends on the price itself. For example, an absolute change of 10000 can be substantial for a house with price of 34900 (the lowest price in the train data), but it is subtle for a house with price of 755000 (the highest price in the train data). For the similar reason, minimizing percent error rather than absolute error is more desirable during model fitting and this is another advantage of log transformation of response variable.

Still, we split the dataset into training and validation dataset. The validation set is mainly for us to estimate the performance of each model and decide which model to submit for fianl evaluation.

3.2 Lasso

We choose to strat our model selection from Lasso. The incentive for starting from Lasso is that even though the dataset includes 80 predictors, not all of them may be closely related to housing price and thus a sparse model generated from Lasso may help us identify important predictors. The estimation reults are given below:

(Intercept)	MSSubClass	MSZoningFV	MSZoningRL
6.273780e+00	-2.063209e-04	6.159001e-04	2.132119e-02
MSZoningRM	LotArea	StreetPave	LotShapeIR3
-4.985122e-02	1.129055e-06	3.191015e-02	-3.318612e-02
LotConfigCulDSac	NeighborhoodBrDale	NeighborhoodClearCr	NeighborhoodCrawfor
9.754193e-03	-6.680611e-03	6.364557e-02	7.406905e-02
NeighborhoodEdwards	NeighborhoodIDOTRR	NeighborhoodMeadowV	NeighborhoodNoRidge
-3.587405e-02	-6.801192e-02	-3.113599e-02	4.157181e-02
NeighborhoodNridgHt	NeighborhoodSomerst	NeighborhoodStoneBr	NeighborhoodVeenker
9.738273e-02	2.933325e-02	9.836686e-02	5.732645e-03
Condition1Feedr	Condition1Norm	Condition2PosN	BldgTypeDuplex
-4.876402e-03	2.753755e-02	-4.273247e-01	-1.235003e-03
BldgTypeTwnhs	OverallQual	OverallCond	YearBuilt
-5.370896e-02	7.921446e-02	2.792544e-02	1.134985e-03
YearRemodAdd	RoofStyleGable		
1.079536e-03	-2.333821e-03		

Figure 4: Lasso Estimation

As can be seen from the reuslts (Figure 4), Lasso generates a model with 31 variables and a MSE of 0.1424. Although a detailed interpretation of the results may require expertise in real estate business, we can still draw some preliminary conclusions: 1) Some neighborhoods have positive coefficeints

while others have negative coefficients. This means type of neighborhoods can play significant role in determining the housing price. For example, neighborhoods with positive coefficients are probably high-end neighborhoods. 2) *LotArea*, *OverallQual* and *OverallCond*, not surprisingly, have positive coefficients because larger houses with higher quality are usually more expensive. Furthermore, the coefficient of *OverallQual* suggests that generally, the housing price increases by approximately 8 percent if the rates of material and finish of the house increases by one. Therefore, we expect *OverallQual* to be a very influential factor in determining the housing price.

3.3 Ridge Regression

We next apply Ridge regression to the dataset to see whether Ridge regression can give higher performance in prediction. However, the error from Ridge regression is 0.1420, which is not a very significant improvement. Since Lasson and Ridge regression are both linear models, while relationship between housing price and predictors may be non-linear, we decide to apply tree-based methods (tree, random forest and boosting) to see whether further improvement can be achieved.

3.4 Tree

Although a single tree usually cannot have high performance in prediction, its high interpretability may give extra information for understanding the relationship between housing price and predictors.The preliminary tree can be further pruned to the tree shown in Figure 6 based on cross validation (Figure 5). As can be seen from the pruned tree, three predictors, *OverallQual*, *Neighborhood* and *GrLivArea*, are involved and this agrees with the conclusion we draw from the Lasso model: 1) Neighborhoods influence housing price. 2) Larger house with higher quality generally has higher price. Additionally, from the point of view that tree can more closely reflect human diecision making process, the fact that the first split is made based on *OverallQual* seems to suggest that *OverallQual* may be the first factor to be considered during the decision making process of housing price.

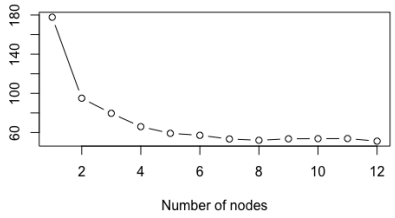

Figure 5: Cross validation error versus number of terminal nodes

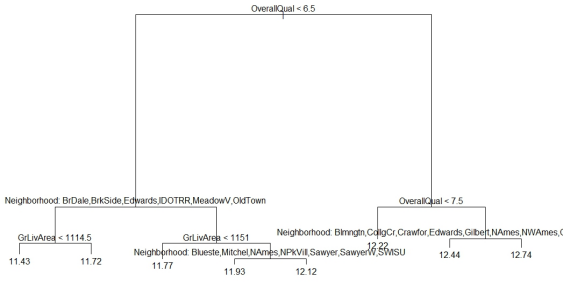

Figure 6: Tree (Pruned)

3.5 Random Forest

Next we apply random forest to achieve higher prediction performance becasue when compared with a single tree, random forest can have significantly lower variance. The MSE estimated from validation set indeed improves from 0.2322 (single tree) to 0.1436, but the improvement is still not significant when compared with linear models. The importance calculated from random forest suggests *OverallQual*, *Neighborhood* and *GrLivArea* are the top three preditors and this qualitatively agrees with our previous analysis.

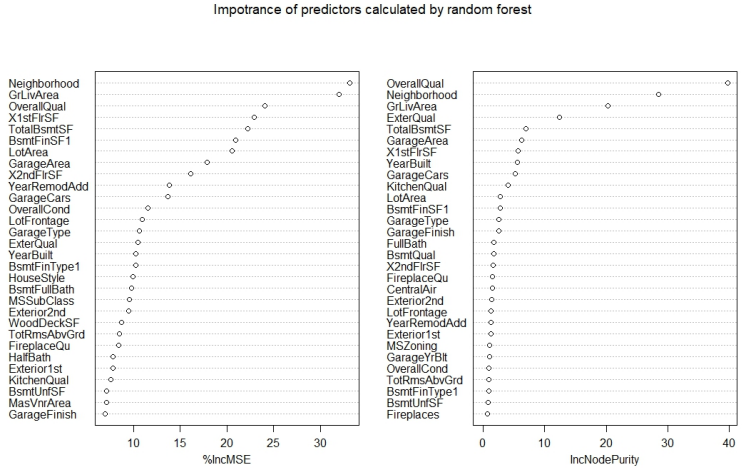

Figure 7: Importance of predictors calculated from random forest

3.6 Boosting

Finally we apply boosting to improve the performance in prediction. Unlike random forest, boosting has a risk of over fitting when interaction depth or number of trees is too large and thus we use the validation set to optimize interaction depth and number of trees. The performance under different interaction depth is plotted in the Figure 8. Plots with different number of trees have also been generated, but are not shown here.

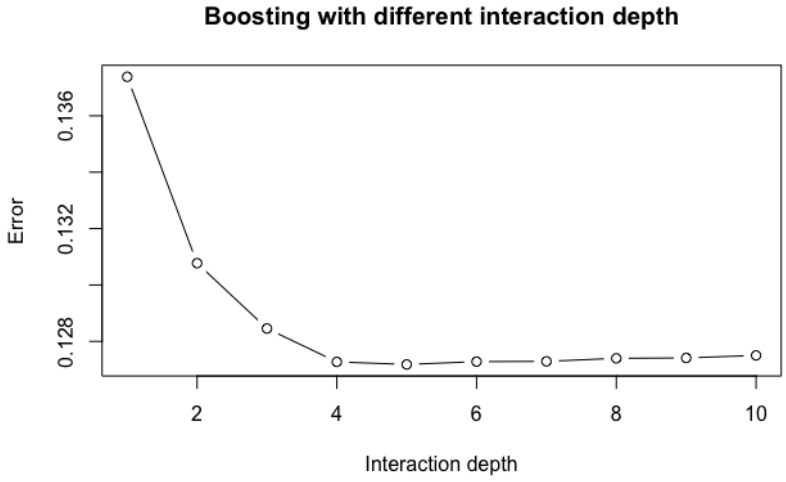

Figure 8: Interaction Depth against Error

3.7 Final Selection

The best model we choose is with the number of trees = 20000 and the interaction depth = 4.

4 Conclusions

We first try to handle the NA values, which can be divided into three types: Meaningful, Meaningless&Negligible, Meaningless&Nonnegligible. The first is dealt with by replacing 'null' with 'no'. The second is manipulated by replacing with sample median (or mode) of train and test. The last type of NA value is replaced by values estimated by a trained random forest model.

Next we try to select the best model by trying out Lasso, Ridge Regression, Tree, Random Forest and Boosting. The final model we adopt is random forest with n.trees=20000 and boosting with interaction depth=4. Also, by interpreting models, especially Lasso and Tree, we manage to identify several important factors in determining housing price:*OverallQual*, *Neighborhood* and *GrLivArea*.

As a matter of fact, the precition results coming from our model ranked 1194 on

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/leaderboard

under the name of Xingbo SHANG.

5 Contributions

Coding in R: Xingbo SHANG;

Code Commenting: Xingbo SHANG, Kao ZHANG;

Project Report: Kao ZHANG.