Advanced Regression Techniques on House Price Prediction

Weiqing Li, Dekun Geng, Yang Tao

November 30, 2017

Abstract

In the final project, we performed Linear Regression, LASSO, Random Forest and Xgboost methods on the real estate data which contains 80 variables and the price. After data cleaning, we focus on the comparison of the performance of the four method we use. It turns out the Xgboost and LASSO are better methods with RMSE = 0.1481 and 0.1440. Random Forest's RMSE is 0.1540 and Linear Regression is 0.1834.

1 Introduction

Real estate has always been an attractive topic because of its unscientific and illogical performance. People are trying to find out the inner pattern of real estate. The prediction of real estate's estimated price has its value in both theory and practice. Anyway, a proper price evaluation model can help both sellers to decide their selling price and buyers to make decision.

There are several points we should consider about. First, quality difference makes the estimation difficult. People have used an index on prices of the same property at different times to avoid it and use a regression method to build the index [1]. Second, the feature selection will be a difficulty. Since we would have a lot of features, we need to find out those who would contribute to the price model. When it comes to the inference and prediction, simple linear regression would be a basic approach, however, the model may not be linear.

We will explore the House Prices Data Set[1] posted on Kaggle.The housing data set contains 1460 rows and 81 features with the target feature Sale Price. In this report, we will build up four prediction models based on training data set and evaluate the models with RMSE. We will apply the linear regression and LASSO method, followed by random forest and boosted tree appraoch to fit the model.

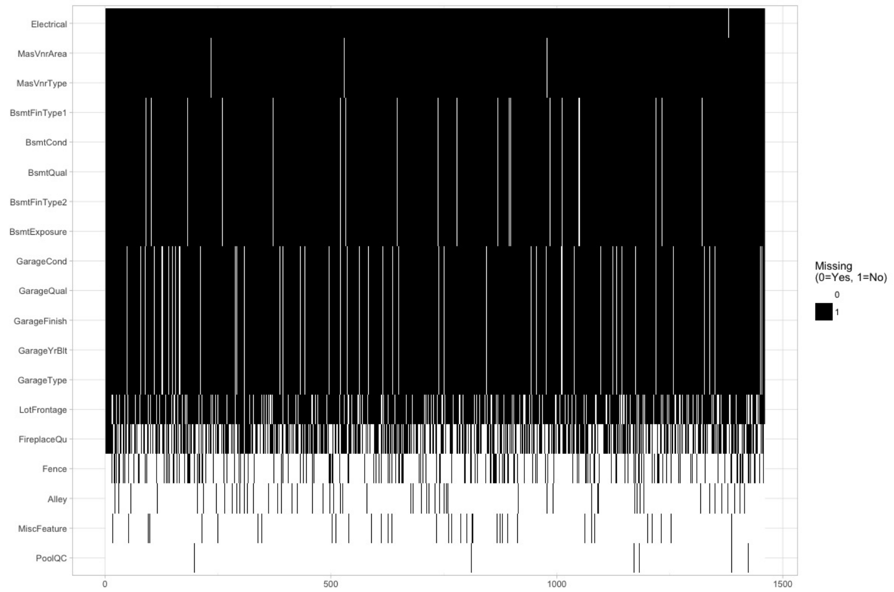
[1] It could be found at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

2 Explore The Data Set

2.1 Data Pre-processing

Since the raw data set contains lots of character variables and NA values. More precisely, there are 43 character variables and over 2000 NA values. Thus we think the best steps to start with would be reformatting some character variables that we can easily convert to numeric and substitute the NA values with 0. The street type for instance, we can just make that paved or not, which reformatting the street type with "0" and "1". We also explored the correlation between different variables, we create intersection variables when the absolute value of correlation is greater than 0.6.

In the end, we got a data set of 92 variables.



2.2 Feature Engineering

After completing the data cleaning process, we will now apply the feature engineering to the processed data set, which leads to a decision that allows us to choose the reletively significant variables from over 80 features. We mainly applied Lasso method.

2.3 Cross Validation

We randomly seperate the data set into two subset equally. One served as training set, one as testing set. We use training set to tune the model, and test set to get the result.

3 Experiment

3.1 Linear Regression Model

In statistics, linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X.

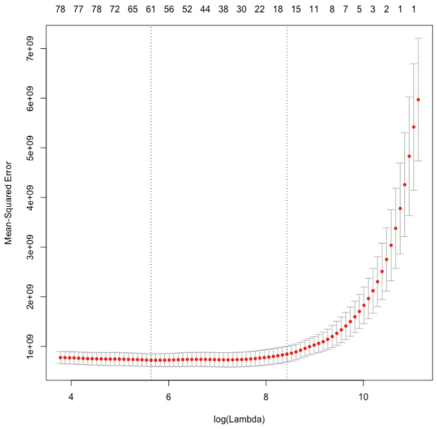
After data pre-processing, we have our data and can build some models. Since our outcome is a continuous numeric variable, we want a linear model, not a GLM First, let's just toss all variables in there. We always like to use a proper regression model as our first examination of the data, to get a feel for what's there.

We now get the adjusted R-Squared : 0.9231, p-value <2.2e - 16 and then test the model with RMSE : 0.1834.

3.2 LASSO method

In statistics and machine learning, lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates need not be unique if covariates are collinear.

We run the training data set with LASSO method and evaluate the model with RMSE when actually do the predicting, the RMSE is 0.1440, much better than linear regression model.



Lasso shrink 90 variables to 65 variables. See Table.1

Feture Name	Coef.	Feture Name	Coef.
(Intercept)	-5.829838e+05	MSZoning	-6.282371e+03
LotFrontage	4.896984e+00	LotArea	5.450050e-01
Street	3.362577e+04	LotShape	-1.529167e+03
LandContour	-4.924609e+02	LotConfig	-2.352634e+00
LandSlope	- 1.322996e+02	Neighborhood	-1.939768e+02
Condition2	-4.739957e+03	BldgType	-8.751606e+02
OverallQual	-1.889713e+01	OverallCond	8.070050e+03
YearBuilt	1.733180e+02	RoofStyle	-2.963215e+02
RoofMatl	-2.517580e+03	Exterior2nd	-1.121893e+01
MasVnrType	4.304220e+03	MasVnrArea	1.095773e+01
ExterQual	-5.732028e+03	ExterCond	9.772344e+02
Foundation	-2.072490e+03	BsmtQual	-8.163616e+03
BsmtFinType1	8.062700e+02	BsmtFinSF1	1.335974e+00
BsmtFinSF2	1.478744e+00	BsmtUnfSF	-4.920934e+00
Heating	-2.537431e+03	HeatingQC	-1.797482e+03
Electrical	1.454621e+03	X2ndFlrSF	-3.398747e+01
LowQualFinSF	-6.721986e+01	GrLivArea	-7.708308e-01
BsmtFullBath	-3.811554e+03	BsmtHalfBath	-3.790915e+02
FullBath	-3.199190e+03	BedroomAbvGr	-6.842280e+03
KitchenAbvGr	-1.500204e+04	KitchenQual	-7.343990e+03
TotRmsAbvGrd	3.275203e+03	Functional	2.644124e+03
Fireplaces	1.421483e+03	GarageType	2.109724e+02
GarageYrBlt	1.045589e+02	GarageFinish	1.539300e+03
GarageArea	3.671830e+00	GarageCond	-3.346074e+02
PavedDrive	3.793755e+03	WoodDeckSF	2.218694e+00
OpenPorchSF	8.330252e+00	EnclosedPorch	-4.068948e-01
X3SsnPorch	4.831392e+00	ScreenPorch	5.253292e+01
PoolQC	1.663007e+04	Fence	1.988279e+03
MiscFeature	-5.752141e+03	MiscVal	9.896720e+00
YrSold	3.655679e+01	SaleType	-8.050931e+02
SaleCondition	6.211596e+03	livarea_qual	6.597914e+00
bsmtFin_Bsmtfb	1.596383e+01	X2ndFlrSF_livarea	1.546996e-02
livarea_TotRmsAbvGr	1.331264e-02		

Table 1: The Features Selected by Lasso

3.3 Random Forest

The results seemed not to be so bad considering that it's just linear model. Now just apply Random Forest method to fit the model. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

We will use all the variables since randomforest does its own feature selection and will evaluate the model by testing the test data set with RMSE: 0.1540.

3.4 XGBoost

XGBoost is an open-source software library which provides the gradient boosting framework for R. It works on Linux, Windows, and macOS. From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library". Other than running on a single machine, it also supports the distributed processing frameworks Apache Hadoop, Apache Spark, and Apache Flink. It has gained much popularity and attention recently as it was the algorithm of choice for many winning teams of a number of machine learning competitions.

We run the training data set on XGBoost and evaluate the model with RMSE when actually do the predicting, the RMSE is 0.1481.

4 Results

We have performed four regression method on the data.

Method	RMSE	Interpretability
Linear Regression	0.1834382	good
Lasso	**0.1440249**	good
Random Forest	0.1540485	bad
XGBoost	0.148113	bad

Table 2: Comparison Among Four Regression Method

5 Discussion

From table.2, we could see that Lasso performance best on the data. XGBoost also performs very well. However, Linear Regression and Lasso have better interpretability than the other two. We can easily know which factor is important to the estimation of property price.

When taking interpretability into account, Lasso is the best method among these four.

Beacause we have a large number of features and both factor variables and numeric variables, it is complicated to perform other feature selection on to this data. In this report, we only perform Lasso as the feature selection. In future, we could try PCA or PLS to select some features to do the regression.

References

[1] Bailey, Martin J., et al. “A Regression Method for Real Estate Price Index Construction.” *Journal of the American Statistical Association*, vol. 58, no. 304, 1963, pp. 933–942. JSTOR, JSTOR, www.jstor.org/stable/2283324.