HOUSE PRICES-ADVANCED REGRESSION TECHNIQUES

Tao Wang

University of Chinese Academy of Sciences, China

Introduction

For home buyers, they generally do not buy homes with basements or near railroads, and there are other features of a home that can even much more influence the price of a home than the number of bedrooms. This project provides 79 characteristics of a house that are used to predict the price of a house.

- Project Evaluation There are 1459 data in the test set, and the output contains ID numbers and predicted house prices. Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price.
- File Description
  - train.csv - the training set
  - test.csv - the test set
  - data˙description.txt - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
  - sample˙submission.csv - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data Analysis

There are 79 variables of inconsistent types, some discrete and some continuous, and after data processing, the variables of type object can be seen as follows:

#	Column	Non-Null Count	Dtype
0	*MSZoning*	$1460 non-null$	*object*
1	*Street*	$1460 non-null$	*object*
2	*Alley*	$91 non-null$	*object*
3	*LotShape*	$1460 non-null$	*object*
...	...	...	*object*

The characteristics of variables of type object are represented by string,such as:

Mszoning:Identifies the general zoning classfication of the sale	
A	*Agriculture*
C	*Commercial*
...	...

The numeric variables are as follows:

#	Column	Non-Null Count	Dtype
0	*MSZoning*	$1460 non-null$	*int*64
1	*Street*	$1460 non-null$	*int*64
2	*Alley*	$1201 non-null$	*float*64
3	*LotShape*	$1460 non-null$	*int*64
...	...	...	...

Conclusion

The difference between the house price prediction experiment and the usual data classification is that the data classification only needs to give the data a specific prediction result and then analyze the correct rate. This time, however, the house price prediction is to give the prediction result to a more detailed point, and then to analyze the error. I think this kind of prediction is more difficult to go up to a better result, and the number of data sets itself is not large. I think the training effect can be improved by increasing the dataset through data expansion. In conclusion, I learned a lot about the new training model in this experiment. For the xgboost model, more adjustments are needed to increase the number of parameters and the depth of the tree to prevent underfitting.

Data Pre-processing

First we count the missing data in the training and test sets, and the statistics are as follows:

- Number of missing data in the training set:

type	num	persent
PoolQC	1453	0.995
MiscFecture	1406	0.963
Alley	1369	0.938
Fence	1179	0.808
FireplaceQu	690	0.472
LotFronttage	259	0.177
GarageType	81	0.055
GarageYrBlt	81	0.555
GarageFinish	81	0.555
...	...	...

- Number of missing data in the test set:

type	num	persent
PoolQC	1456	0.997
MiscFecture	1408	0.964
Alley	1352	0.926
Fence	1169	0.800
FireplaceQu	730	0.5
LotFronttage	227	0.155
GarageYrBlt	78	0.053
GarageFinish	78	0.053
...	...	...

- Remove the feature PoolQC, MiscFeature, Alley,Fence and FireplaceQu
- Use the average to fill in 8 missing numbers of masvnrarea, and use the mode to fill in the missing data of LotFrontage and GarageYrlBlt.
- Use log transformation to reduce skewness.
- Delete unreasonable data
- Use mode to fill in the missing data of string type.

Training Result

- Training Result Summary

method	result
$LinearRegression(low_corrdeleted)$	0.7313
$RandomForest(low_corrdeleted)$	0.16099
$K-neighbor(low_corrdeleted)$	0.22111
$AdaboostRegressor(low_corrdeleted)$	0.22124
$Xgboost(low_corrdeleted)$	0.0.60098
$Xgboost$	0.17429
$GradientBoostingRegressor(low_corrdeleted)$	0.15922
$GradientBoostingRegressor$	0.14173