HOUSE PRICES-ADVANCED REGRESSION TECHNIQUES

Tao Wang

University of Chinese Academy of Sciences

(None)

Overview

Project Description & Evaluation

Project Description & Evaluation

Description

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.

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.

Dataset Description

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

Outlying Aspects Mining vs Outlier Detection

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	$MSZoning$	$1460non-null$	$object$
1	$Street$	$1460non-null$	$object$
2	$Alley$	$91non-null$	$object$
3	$LotShape$	$1460non-null$	$object$
...	...	...	$object$

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	$Commercial$
...	...

Outlying Aspects Mining vs Outlier Detection

The numeric variables are as follows:

#	Column	Non-Null Count	Dtype
0	$MSZoning$	$1460 non-null$	$int64$
1	$Street$	$1460 non-null$	$int64$
2	$Alley$	$1201 non-null$	$float64$
3	$LotShape$	$1460 non-null$	$int64$
...	...	...	...

Data Pre-processing

Date 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
...	...	...

Date Pre-Processing

- 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
...	...	...

Data cleaning of training set

Step1

We choose to remove the feature PoolQC, MiscFeature, Alley,Fence and FireplaceQu because we do not have a suitable method to replenish a large amount of data.

Step2

Since the propeties of numeric types are conveniently complemented by medians or averages, etc. We first examine the properties of numeric types.

We first examine the properties of numeric types. After studying it, we can find that only LotFrontage, MasVnrArea and GarageYrlBlt have missing features for the number types in the training set.

For the masvnrarea with only 8 missing numbers, we can use the average to fill in, while for the other two with more missing numbers, we choose to use the plural to fill in.

Data cleaning of training set

Step3

For some positively biased data, we try to use log transformation to reduce their skewness. By using function displot, we can find that the feature LotArea have positive skew like figure1.

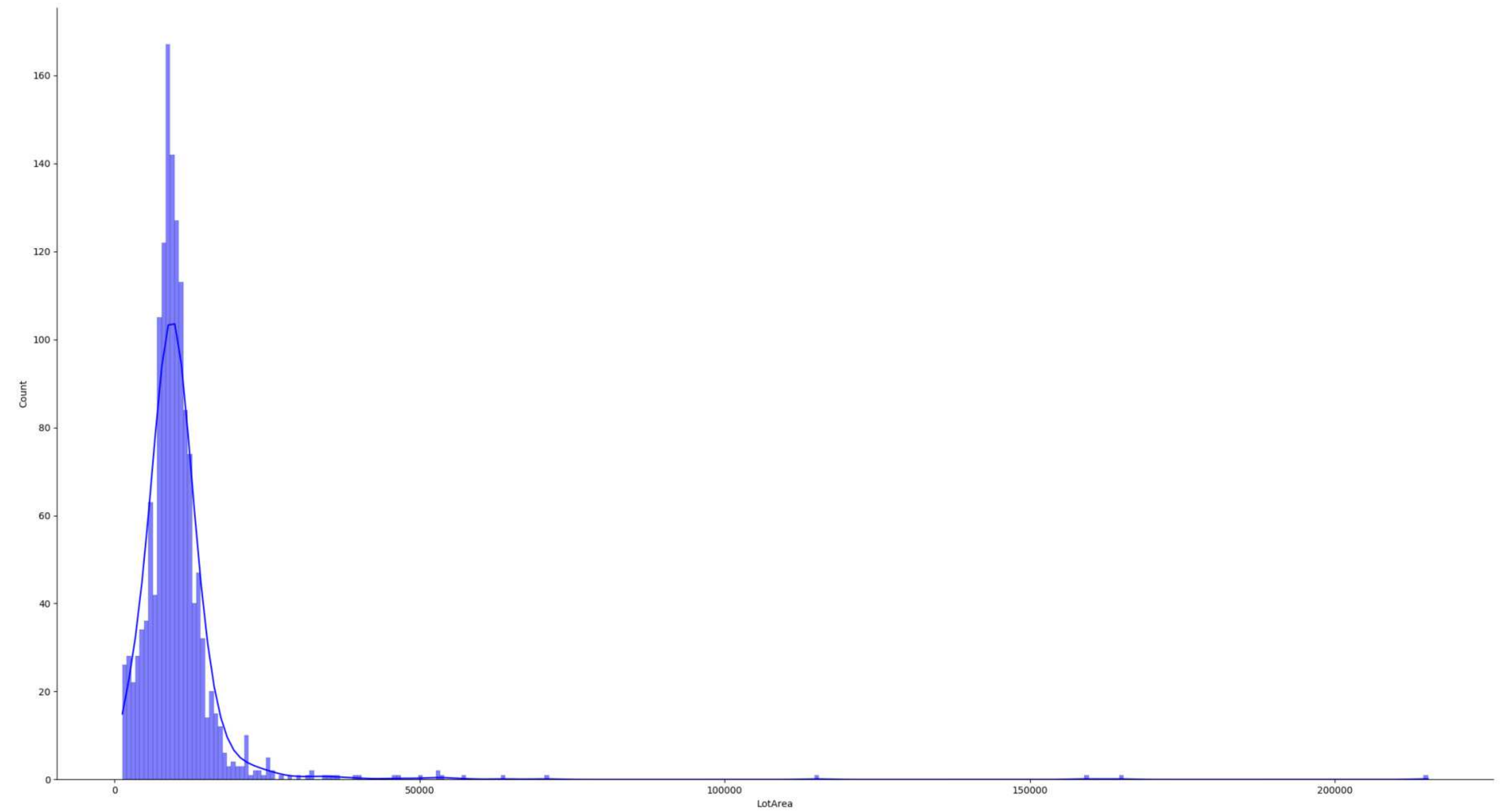

Figure 1: An example of positive skew

Data cleaning of training set

Step4

Some of the data will be unreasonable, may be too large or too small. We need to remove this data before we put it into the training model. We first use the data visualization to look at it and decide on the removal method in figure2.

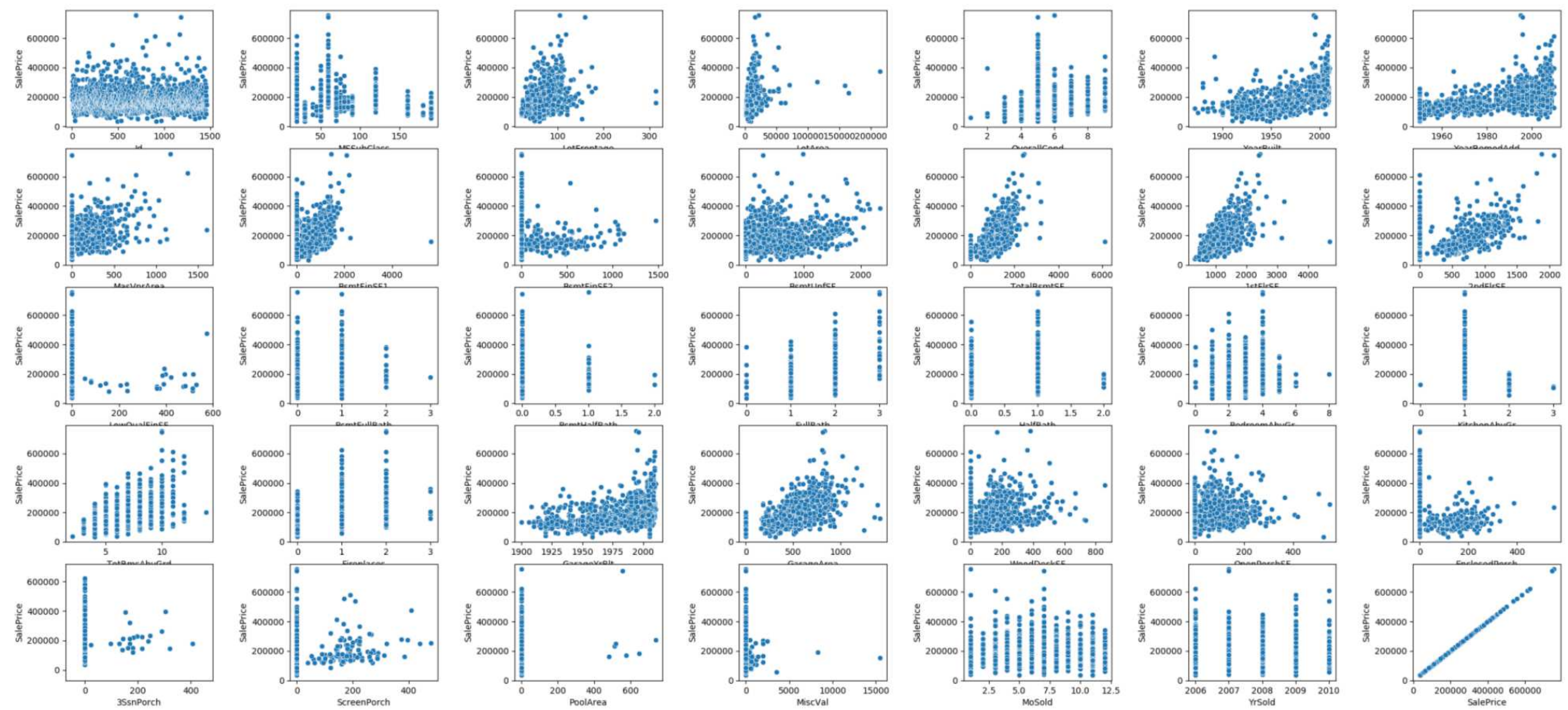

Figure 2: Visualization of Training data

Data cleaning of training set

Step5

Now, we analyze the categorical data. For this string type of data, we choose to use the plural to fill in the missing data.

- Data cleaning of testing set All the previous steps are the same, but with the previous steps, we can find that the missing data in the test set is not the same as the training set, so we just need to simultaneously use the same analysis idea for these new missing data.
- Date normalization and removal of weakly correlated data After we clean the training and test set data, we need to consider if all the characteristics are related to the sales price. So we calculate the correlation coefficient of each characteristic and the sales price, we consider the correlation coefficient below 0.3 as almost irrelevant and remove these characteristics

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

Conclusion