1. 数据清洗及进行预处理

1.数据读取：

训练集

train\_raw=pd.read\_csv('/Users/blackhole6/Downloads/2-House+Prices-Advanced+Regression+Techniques/train.csv')

测试集

test\_raw=pd.read\_csv('/Users/blackhole6/Downloads/2-House+Prices-Advanced+Regression+Techniques/test.csv')

2.消除异常值

train = train[~((train['GrLivArea'] > 4000) & (train['SalePrice'] < 300000))]

train = train[~((train['MasVnrArea'] > 1400) & (train['SalePrice'] < 300000))]

train = train[~((train['LotFrontage'] > 300) & (train['SalePrice'] < 300000 ))]

train = train[~((train['LotArea'] > 200000) & (train['SalePrice'] < 500000 ))]

train = train[~((train['OpenPorchSF'] > 500) & (train['SalePrice'] < 100000))]

1. 消除售价和对应的ID

response\_variable = np.log1p(train.SalePrice)

id\_ = test.Idtrain = train.loc[:,'MSSubClass':'SaleCondition']

test = test.loc[:,'MSSubClass':'SaleCondition']

...

得到一个数据经过处理过的数据集

1. 模型训练
2. 读取处理好的数据

train=pd.read\_pickle('/Users/blackhole6/Downloads/2-House+Prices-Advanced+Regression+Techniques/train.pkl')

test=pd.read\_pickle('/Users/blackhole6/Downloads/2-House+Prices-Advanced+Regression+Techniques/test.pkl')

response\_variable=pd.read\_pickle('/Users/blackhole6/Downloads/2-House+Prices-Advanced+Regression+Techniques/y\_train.pkl')

2.赋值给训练集和测试集

X\_train = train

X\_test = test

y\_train = response\_variable

1. 进行5000次Lasso训练 并利用GridSearchCV进行调参

from sklearn.model\_selection import GridSearchCV

from sklearn.linear\_model import Lasso

lasso = Lasso(max\_iter=5000)

alphas = np.logspace(-4, -2, 100) # [.0004]

grid = GridSearchCV(estimator=lasso,

param\_grid=dict(alpha=alphas), cv=10, scoring='r2')

grid.fit(X\_train, y\_train)

1. 得到最好的一次预测值

grid.best\_estimator\_.score(X\_train, y\_train)

0.9360285957936789

1. 得到误差的均方差

mean\_squared\_error(y\_train, y\_hats)

0.010093502736179582

1. 得到预测值与真实值的比较图

<https://pasteme.cn/116658>

