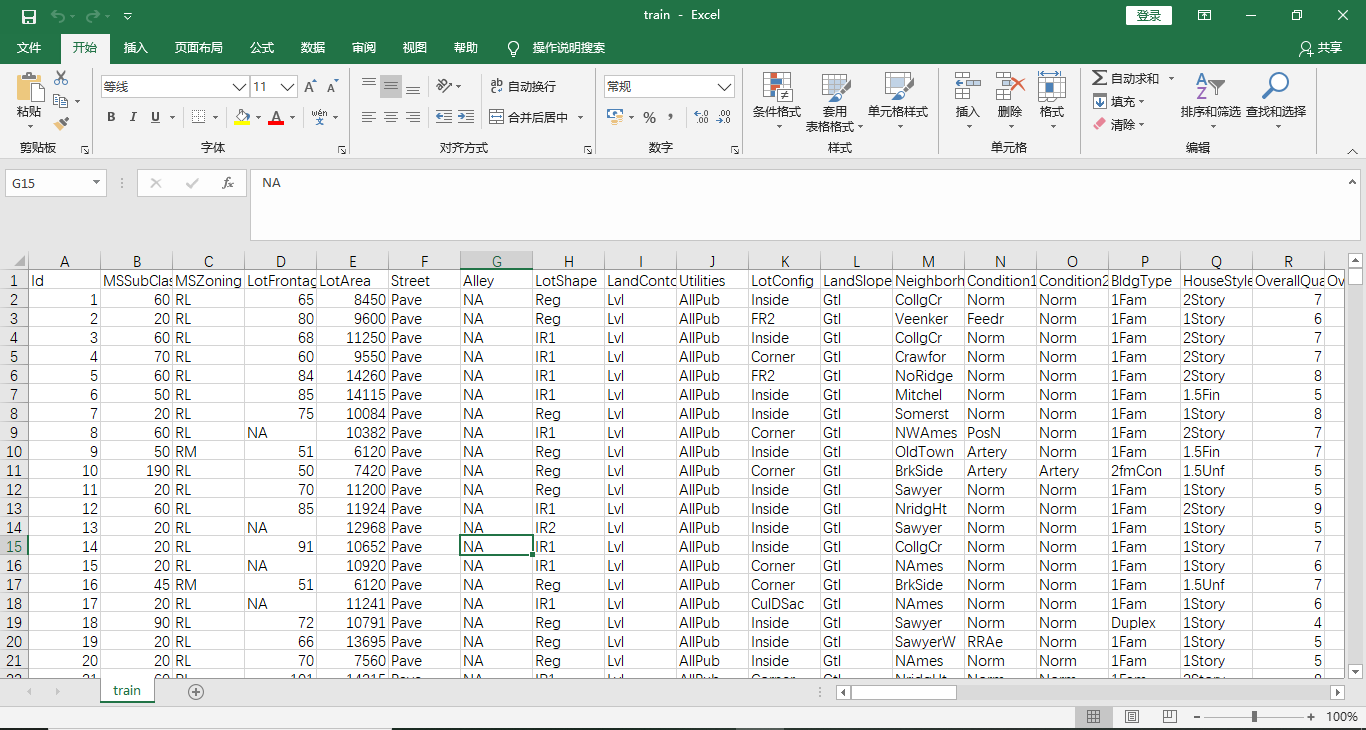
1. **数据来源**

数据来源于<http://jse.amstat.org/v19n3/decock.pdf>，论文名称为Ames, Iowa: Alternative to the Boston Housing Data as anEnd of Semester Regression Project，发表于Journal of Statistics Education，数据集名称为Ames Housing dataset。

1. **数据描述**

数据集中包含训练集、测试集，训练集中比测试集多出SalePrice属性，我们要做的就是预测出测试集的SalePrice。



关于数据集中各个属性的描述、取值、意义如下：

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

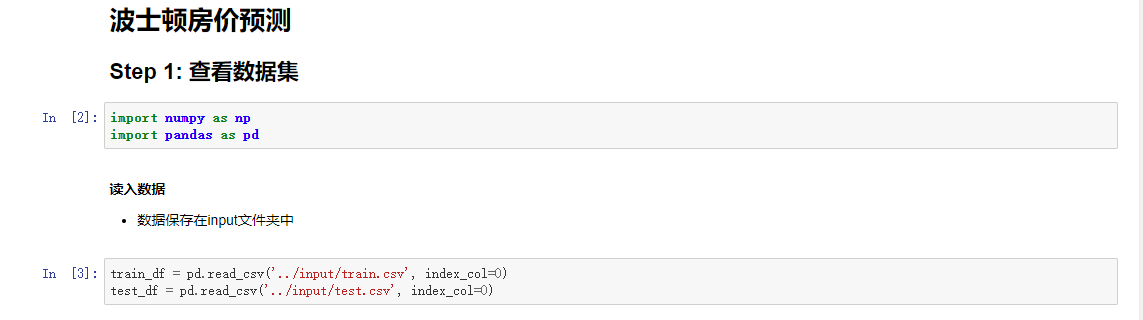
Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

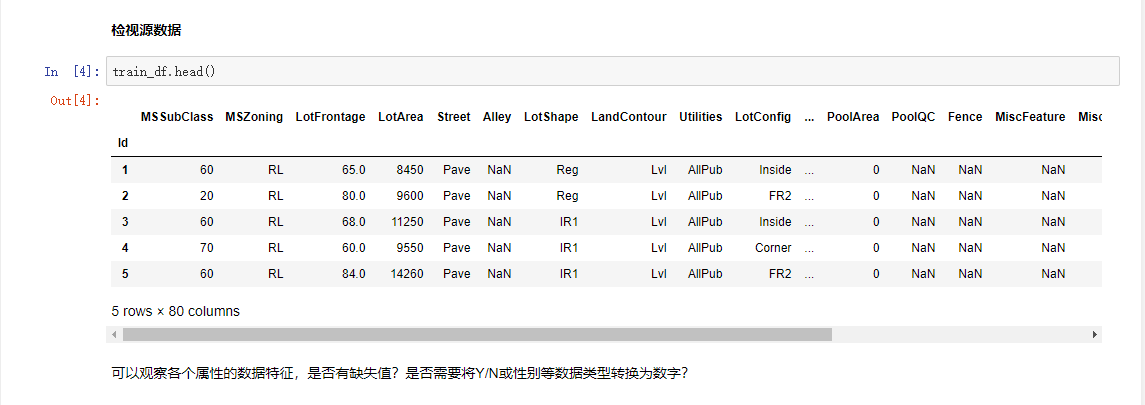
Family Sale between family members

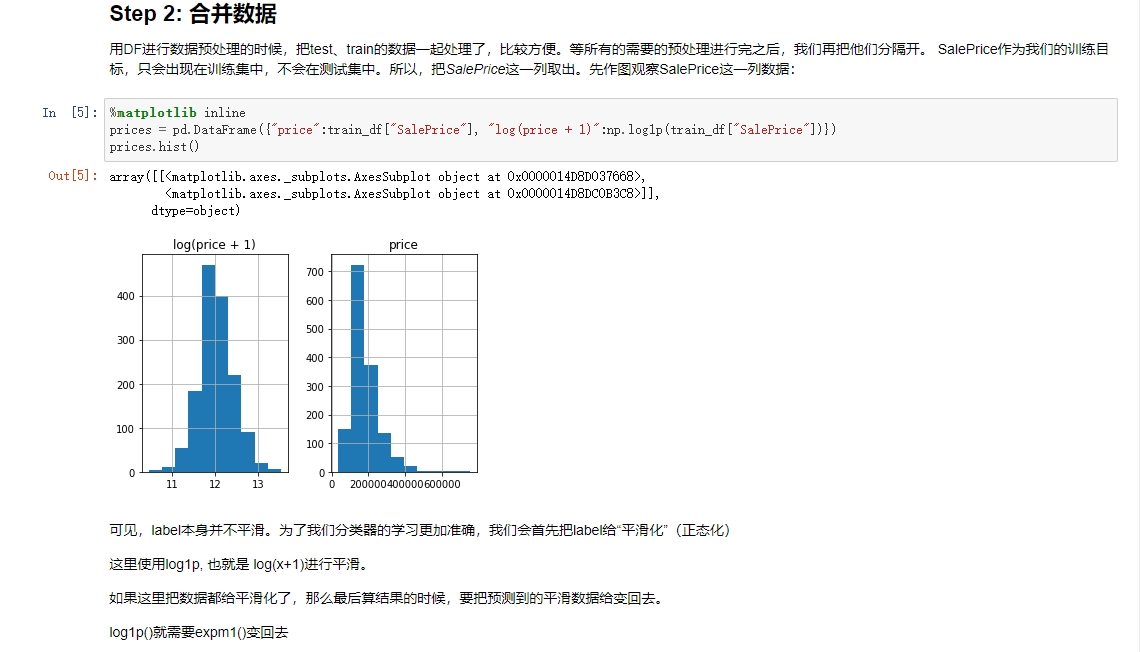
Partial Home was not completed when last assessed (associated with New Homes)

1. **实验过程**

实验环境为jupyternotebook，使用python3.7以及sklearn、matplotlib等工具包。

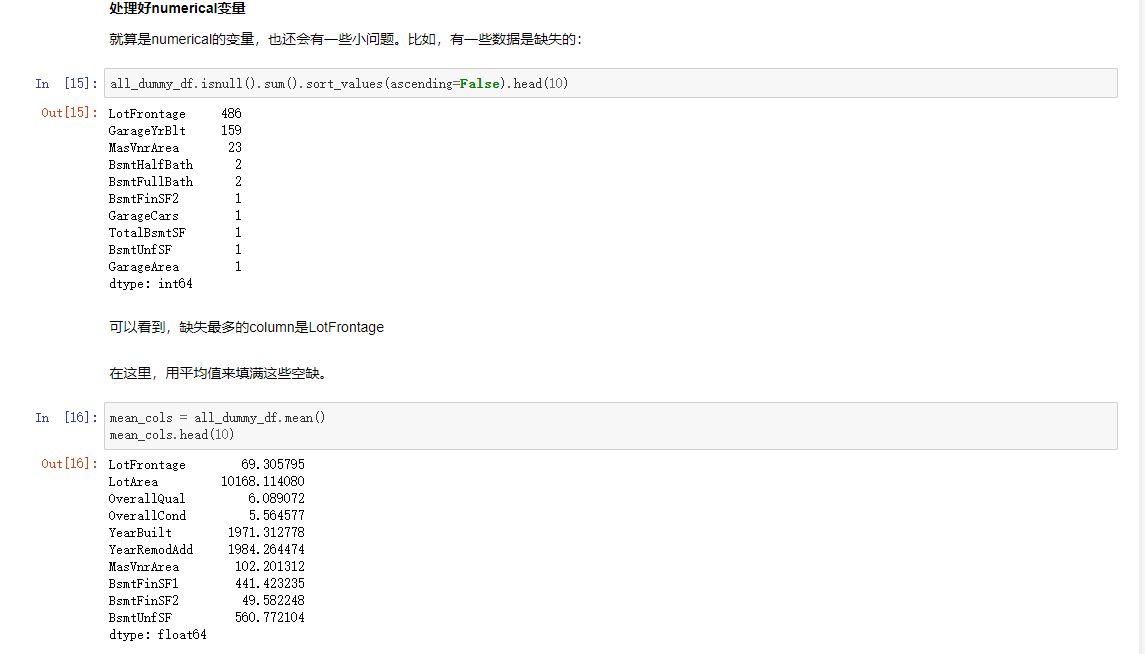






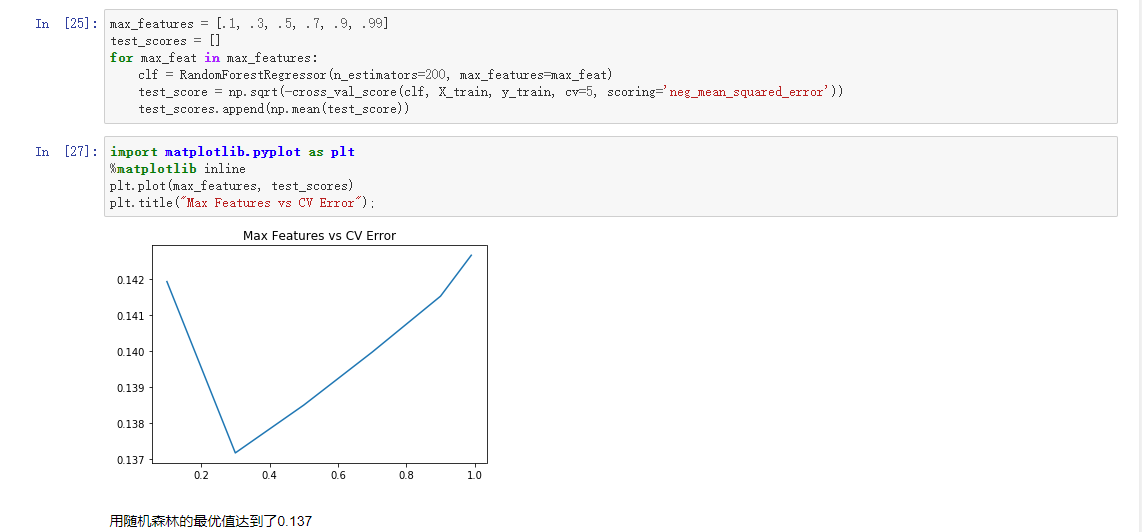
 













1. **实验代码**

#!/usr/bin/env python

# coding: utf-8

# # 波士顿房价预测

#

# ## Step 1: 查看数据集

# In[2]:

import numpy as np

import pandas as pd

# #### 读入数据

#

# \* 数据保存在input文件夹中

# In[3]:

train\_df = pd.read\_csv('../input/train.csv', index\_col=0)

test\_df = pd.read\_csv('../input/test.csv', index\_col=0)

# #### 检视源数据

# In[4]:

train\_df.head()

# 可以观察各个属性的数据特征，是否有缺失值？是否需要将Y/N或性别等数据类型转换为数字？

# ## Step 2: 合并数据

# 用DF进行数据预处理的时候，把test、train的数据一起处理了，比较方便。等所有的需要的预处理进行完之后，我们再把他们分隔开。

# SalePrice作为我们的训练目标，只会出现在训练集中，不会在测试集中。所以，把\*SalePrice\*这一列取出。先作图观察SalePrice这一列数据：

# In[5]:

get\_ipython().run\_line\_magic('matplotlib', 'inline')

prices = pd.DataFrame({"price":train\_df["SalePrice"], "log(price + 1)":np.log1p(train\_df["SalePrice"])})

prices.hist()

# 可见，label本身并不平滑。为了我们分类器的学习更加准确，我们会首先把label给“平滑化”（正态化）

#

# 这里使用log1p, 也就是 log(x+1)进行平滑。

#

# 如果这里把数据都给平滑化了，那么最后算结果的时候，要把预测到的平滑数据给变回去。

#

# log1p()就需要expm1()变回去

# In[6]:

y\_train = np.log1p(train\_df.pop('SalePrice'))

# 然后把剩下的部分合并起来

# In[7]:

all\_df = pd.concat((train\_df, test\_df), axis=0)

# 此刻，可以看到all\_df就是合在一起的DF

# In[8]:

all\_df.shape

# 而\*y\_train\*则是\*SalePrice\*那一列

# In[9]:

y\_train.head()

# ## Step 3: 预处理

#

# 类似特征工程，把不方便处理的数据给统一。

#

# #### 正确化变量属性

#

# ①\*MSSubClass\* 的值其实是一个category，但使用DF的时候，这类数字符号会被默认记成数字。这就很有误导性，需要把它变回成\*string\*

# In[10]:

all\_df['MSSubClass'].dtypes

# In[11]:

all\_df['MSSubClass'] = all\_df['MSSubClass'].astype(str)

# 变成\*str\*以后，可以统计一下每个类别出现的次数

# In[12]:

all\_df['MSSubClass'].value\_counts()

# #### 把category的变量转变成numerical表达形式

#

# 当使用numerical来表达categorical的时候，数字本身有大小的含义，可能会给之后的模型学习带来麻烦。于是可以用One-Hot的方法来表达category。

#

# pandas自带的get\_dummies方法，可以做到One-Hot。

# In[13]:

pd.get\_dummies(all\_df['MSSubClass'], prefix='MSSubClass').head()

# 此刻\*MSSubClass\*被我们分成了12个column，每一个代表一个category。是就是1，不是就是0。

# 同理，把所有的category数据，都进行One-Hot

# In[14]:

all\_dummy\_df = pd.get\_dummies(all\_df)

all\_dummy\_df.head()

# #### 处理好numerical变量

#

# 就算是numerical的变量，也还会有一些小问题。比如，有一些数据是缺失的：

# In[15]:

all\_dummy\_df.isnull().sum().sort\_values(ascending=False).head(10)

# 可以看到，缺失最多的column是LotFrontage

# 在这里，用平均值来填满这些空缺。

# In[16]:

mean\_cols = all\_dummy\_df.mean()

mean\_cols.head(10)

# In[17]:

all\_dummy\_df = all\_dummy\_df.fillna(mean\_cols)

# 检查是否所有的空缺都被填上平均值

# In[18]:

all\_dummy\_df.isnull().sum().sum()

# #### 标准化numerical数据

#

# 把源数据给放在一个标准分布内。不让数据间的差距太大。

#

# 先显示是numerical的属性：

# In[19]:

numeric\_cols = all\_df.columns[all\_df.dtypes != 'object']

numeric\_cols

# 计算标准分布：(X-X')/s

#

# 让我们的数据点更平滑，更便于计算。

# In[20]:

numeric\_col\_means = all\_dummy\_df.loc[:, numeric\_cols].mean()

numeric\_col\_std = all\_dummy\_df.loc[:, numeric\_cols].std()

all\_dummy\_df.loc[:, numeric\_cols] = (all\_dummy\_df.loc[:, numeric\_cols] - numeric\_col\_means) / numeric\_col\_std

# ## Step 4: 建立模型

#

# #### 把数据集分回 训练/测试集

# In[21]:

dummy\_train\_df = all\_dummy\_df.loc[train\_df.index]

dummy\_test\_df = all\_dummy\_df.loc[test\_df.index]

# In[22]:

dummy\_train\_df.shape, dummy\_test\_df.shape

# 把DF转化成Numpy Array，符合Sklearn

# In[23]:

X\_train = dummy\_train\_df.values

X\_test = dummy\_test\_df.values

# 用Sklearn自带的cross validation方法来测试模型

# #### Random Forest 随机森林

# In[24]:

from sklearn.ensemble import RandomForestRegressor

from sklearn.model\_selection import cross\_val\_score

# In[25]:

max\_features = [.1, .3, .5, .7, .9, .99]

test\_scores = []

for max\_feat in max\_features:

clf = RandomForestRegressor(n\_estimators=200, max\_features=max\_feat)

test\_score = np.sqrt(-cross\_val\_score(clf, X\_train, y\_train, cv=5, scoring='neg\_mean\_squared\_error'))

test\_scores.append(np.mean(test\_score))

# In[27]:

import matplotlib.pyplot as plt

get\_ipython().run\_line\_magic('matplotlib', 'inline')

plt.plot(max\_features, test\_scores)

plt.title("Max Features vs CV Error");

# 用随机森林的最优值达到了0.137

# #### Decision Tree 决策树

# In[28]:

from sklearn.tree import DecisionTreeRegressor

# In[29]:

max\_depths = [2,4,6,8,10,12,14,16,18,20]

test\_scores = []

for max\_depth in max\_depths:

regressor = DecisionTreeRegressor(random\_state=0,max\_depth=max\_depth)

test\_score = np.sqrt(-cross\_val\_score(regressor, X\_train, y\_train, cv=5, scoring='neg\_mean\_squared\_error'))

test\_scores.append(np.mean(test\_score))

# In[30]:

import matplotlib.pyplot as plt

get\_ipython().run\_line\_magic('matplotlib', 'inline')

plt.plot(max\_depths, test\_scores)

plt.title("Max Depth vs CV Error");

# 可以看出，随机森林的结果目前好于决策树的结果