数据集 <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

一个来自葡萄牙银行机构的某次营销活动的访谈记录，目的是预测客户接下来是否会办理定期存款。

数据集共21列，其中特征20列，标签1列。

特征：

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

# social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

标签：

21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

数据集预处理：

1. 年龄（由上至下影响因素由高至低，下同）

40~60

30~40

30以下

60以上

1. 职业

Admin.,entrepreneur,management,technician

Blue-collar,housemaid,retired,self-employed,services,student,unemployed

1. 婚姻状况

Married

Divorced

Single

1. 文化程度

professional.course

university.degree

basic.4y,basic.6y,basic.9y,high.school,illiterate

1. Pdays 客户上一次办理服务距采访日的天数（999代表客户没有办理过该服务）

分为两类：非999一类，999一类

其余字符型的特征重新按（0，1，2，3，4，...）赋值

最后调用 sklearn.preprocessing.scale函数将特征集归一化

将数据集按4：1划分样本集和测试集

样本集：32950

测试集：8238

创建lg模型 对prediction做出评价

precision： 0.65565

Accuracy： 0.90823

Recall： 0.40321

F1\_score: 0.49933