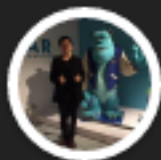


机器学习技术在 **Python** 语言的商业应用

YU-WEI (David) CHIU

丘祐玮



· 大數軟體有限公司創辦人

· 前趨勢科技工程師

· ywchiu.com

· 大數學堂

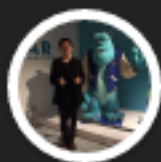
<http://www.largitdata.com/>

· 粉絲頁

<https://www.facebook.com/largitdata>

· **R For Data Science Cookbook**

· **Machine Learning With R Cookbook**



Machine Learning With R Cookbook

(机器学习与R语言实战) &

R for Data Science Cookbook



R for Data Science Cookbook

Over 100 hands-on recipes to effectively solve real-world data problems using the most popular R packages and techniques

Yu-Wei, Chiu (David Chiu)

 **Packt** open source



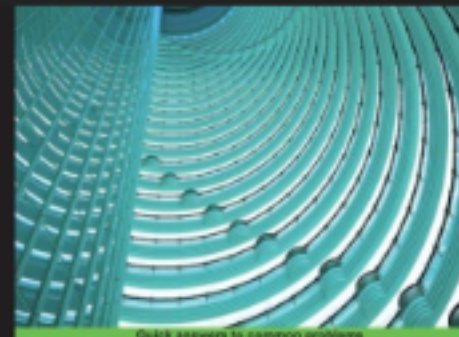
机器学习与R语言实战

第1版

第1次印刷

ISBN 7-309-05888-8

人民邮电出版社



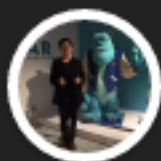
Machine Learning with R Cookbook

Explore over 110 recipes to analyze data and build predictive models with the simple and easy-to-use R code

Yu-Wei, Chiu (David Chiu)

 **Packt** open source

Author: Yu-Wei, Chiu (David Chiu)



Big Data
+ Data Analysis

= Smart Data



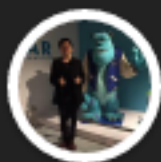
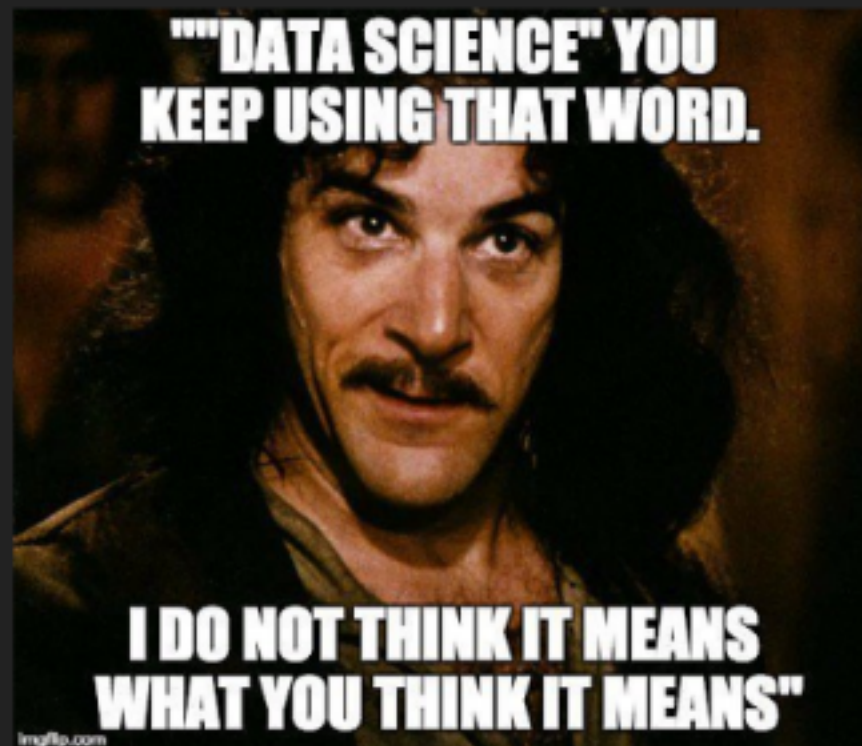
数据科学

80%都在做加總與平均

1. 数据处理

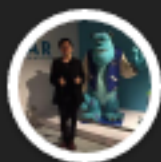
2. 数据分析

3. 诠释结果



工欲善其事
必先利其器

使用 **Python** **R** 来做数据分析



为什么使用Python

网络爬虫



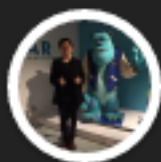
数据分析



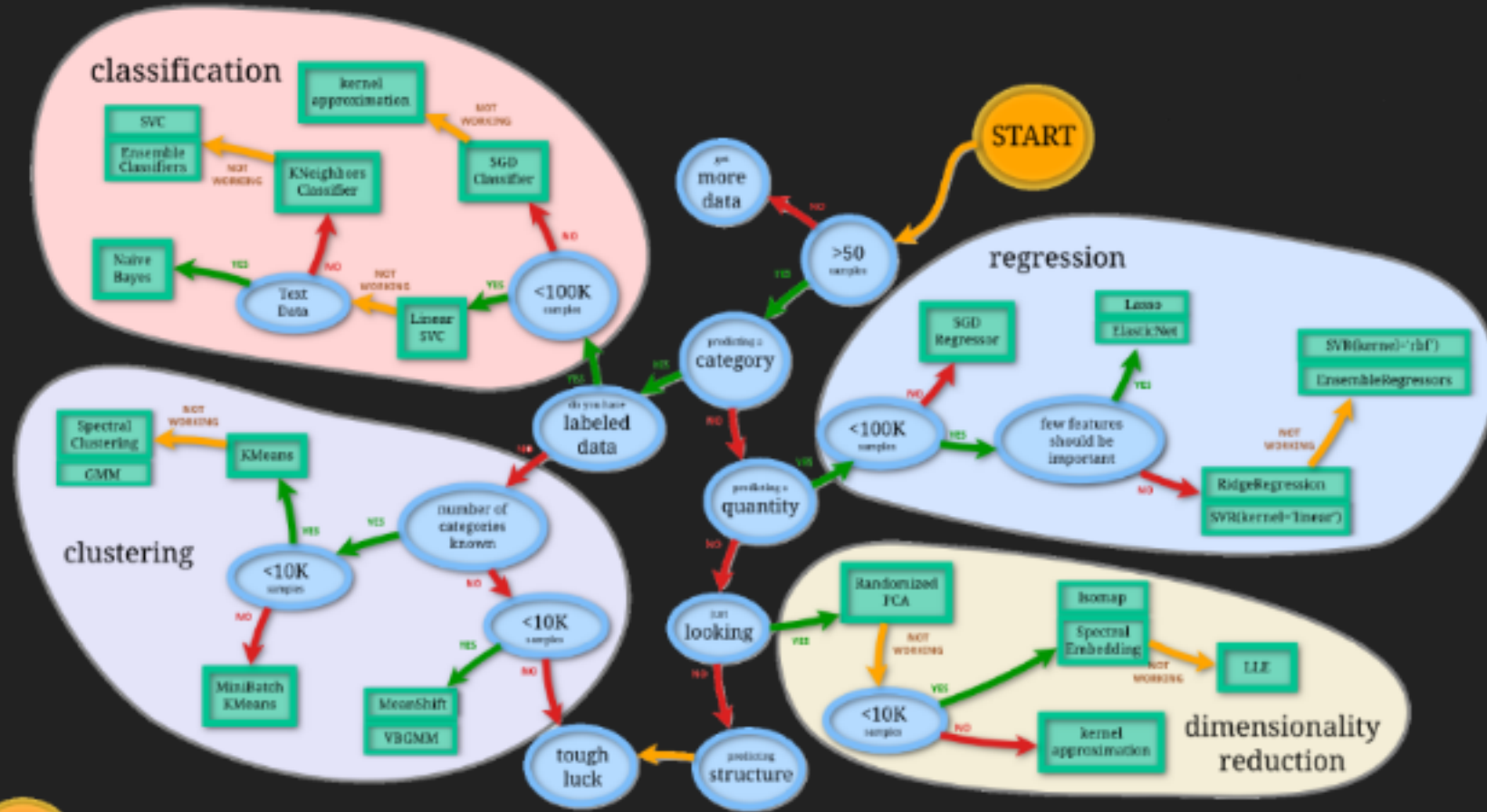
网页呈现



1. 拥有众多第三方分析套件
2. 语法简单，上手容易
3. 一条龙分析



机器学习



预测房价信息

所有物件

屋主

已成交

共找到9,181個房屋

默认排序

刊登時間

坪數

金額



大安名人巷稀有物件，全新裝潢四百萬配備

黃金屋光

整層住宅 | 4房2廳2衛 | 45.3坪 | 樓層：7/7

· 大安區-大安路一段

仲介 許先生 / 昨日更新 / 52人瀏覽

68,000 元/月



套房3分鐘捷運行天宮站榮星花園

黃金屋光

獨立套房 | 10坪 | 樓層：5/5

中山區-民權東路二段152巷5弄2號

屋主 王太太 / 14分鐘內更新 / 248人瀏覽

11,000 元/月



捷運象山精緻套房2萬2包水電網路瓦斯

黃金屋光

分租套房 | 13坪 | 樓層：7/7

信義區-信義路六段30號

代理人 余先生 / 14分鐘內更新 / 73人瀏覽

22,000 元/月

刊登出租

推薦物件

我也要出現這裡



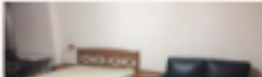
全新裝潢+採光好+空間大+近公園+環境佳

73,000 元/月

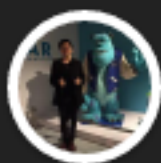


近八德路長安東路，高樓景採光大三房

60,000 元/月



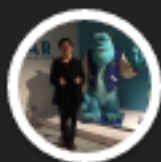
返回舊版



读取房价数据

```
import pandas  
df = pandas.read_csv('house_rental.csv')  
df
```

	Unnamed: 0	Sqft	Floor	TotalFloor	Bedroom	Living.Room	Bathroom	Price
0	1	1177.698	2	7	2	2	2	62000
1	2	2134.800	5	7	4	2	2	78000
2	3	1138.560	5	7	2	2	1	58000
3	4	1458.780	2	7	3	2	2	45000
4	5	967.776	11	14	3	2	2	45000

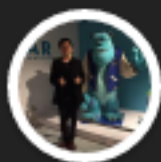


进行回归分析

```
import numpy as np
from sklearn import datasets, linear_model

# Create linear regression object
regr = linear_model.LinearRegression()

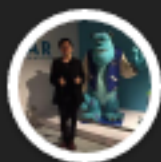
# Train the model using the training sets
regr.fit(df['Sqft'].to_frame(), df['Price'].to_frame())
```



验证回归分析结果

```
# The coefficients
print('Coefficients: \n', regr.coef_)

# The mean squared error
print("Mean squared error: %.2f"
      % np.mean((regr.predict(df['Sqft'].to_frame()) - df['Price'].to_frame()) ** 2))
```

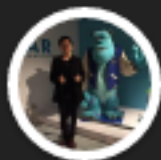
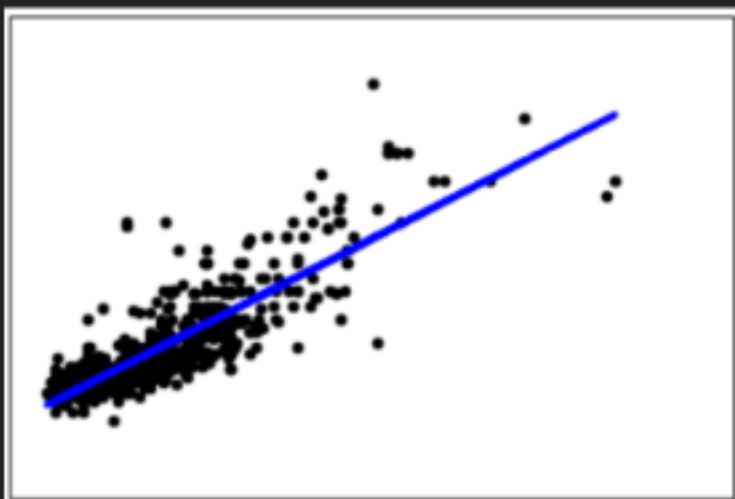


绘制回归结果

```
# Plot outputs
plt.scatter(df['Sqft'], df['Price'], color='black')
plt.plot(df['Sqft'].to_frame(), regr.predict(df['Sqft'].to_frame()), color='blue',
         linewidth=3)

plt.xticks(())
plt.yticks(())

plt.show()
```



电商网站

**買東西**
udn shopping

全品項
最高回饋

購物金

18%

登入 加入會員 my東西 帳戶 購物車 訂單 新聞網  下載

快速到貨 聯合超市 台灣物產 元氣樂活 旅

美妆保健 流行服飾 精品名錶 鞋包配飾 NB/PC 消費電子 大小家電 居家生活 傢俱寢具 食尚良品 圖書文創 運動交通 品牌專

商品搜尋: 全類別 熱門: 抽獎皇+西堤雙人券 10

商品分類

訂閱報紙

my東西推薦

美妆保健
今日最靚 專櫃保養 彩妝
塑身保健 醫美美妆 香水

流行服飾
結帳9折 流行女裝 男裝
內衣 專櫃內衣 bossini

精品名錶
COACH挑戰全台最低價 LV
SEIKO CASIO CITIZEN

鞋包配飾

Brand 5/13-5/15

品牌典藏

CARVEN

全品項
回饋

18%

購物金

lenovo

全品項
回饋

13%

購物金

快速到貨

生活好物

3C家電

流行時尚





14吋
靜音電扇





限時搶購



5/13 2倍送 母親節活動•贈
JSmax 白色戀人7吋雙核平板(WiFi)
限時價\$3390

優惠結束倒數 01:01:52

公益競標

1/2 <

CHARITY SALE

公益競標

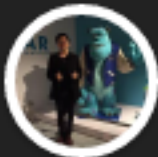
買馬可多一點給孩子溫暖
也給孩子更好



商业问题

```
203.145.207.188 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0005158462;cat=J,J_007,J_007_001,
1.171.24.21 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0022007845;cat=I,I_001,I_001_003,I_00
1.172.0.185 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0006604986;cat=G_001_006_002;erUId=6c
101.13.126.139 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=cart;uid=;plist=0022356972,3,139,0009755432,1,19
61.58.145.131 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0024134891;cat=L,L_006,L_006_002,L_
114.41.4.218 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=order;uid=U312622727;plist=0006944501,1,1069;erUId
114.43.89.52 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=view;uid=;pid=0022226912;cat=D,D_004,D_004_028,D_0
```

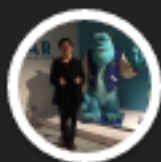
拿第一个月的电商网站日志设计数值模型，第二个月网站日志做为测试数据，进行第二个月的**销售排行榜预测**。



原始资料Log

```
203.145.207.188 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0005158462;cat=J,J_007,J_007_001,
1.171.24.21 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0022007845;cat=I,I_001,I_001_003,I_00
1.172.0.185 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0006604986;cat=G_001_006_002;erUId=6c
101.13.126.139 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=cart;uid=;plist=0022356972,3,139,0009755432,1,19
61.58.145.131 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0024134891;cat=L,L_006,L_006_002,L_
114.41.4.218 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=order;uid=U312622727;plist=0006944501,1,1069;erUId
114.43.89.52 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=view;uid=;pid=0022226912;cat=D,D_004,D_004_028,D_0
```

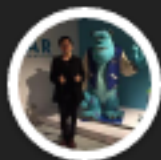
参数	意义
act=view	检视商品内容
act=search	搜寻商品
act=cart	将商品放入购物车
act=order	正式订购商品



原始资料Log (2)

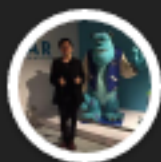
```
203.145.207.188 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0005158462;cat=J,J_007,J_007_001,1.171.24.21 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0022007845;cat=I,I_001,I_001_003,I_001.172.0.185 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0006604986;cat=G_001_006_002;erUid=6c101.13.126.139 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=cart;uid=;plist=0022356972,3,139,0009755432,1,1961.58.145.131 - - [01/Feb/2015:00:00:00 +0800] "GET /action?;act=view;uid=;pid=0024134891;cat=L,L_006,L_006_002,L_114.41.4.218 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=order;uid=U312622727;plist=0006944501,1,1069;erUid114.43.89.52 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=view;uid=;pid=0022226912;cat=D,D_004,D_004_028,D_0
```

参数	意义
uid=XXXX	已登入的用户标识符 (一个账号对应一个 uid)
erUid=XXXX	未登入的用户标识符 (清除 Cookie 会取得另一组)
pid=XXXX	产品标识符
cat=XXXX	商品于购物网站内部定义的群组标识符
plist=AAA,BBB,CCC,DDD,....	放入购物车的产品信息



使用正规表达式剖析数据

```
import re, sys
f = open(n, 'r')
for line in f.readlines():
    ele = line.strip()
    m0 = re.match(".*- - \[(.*?)\].*", ele)
    dt = m0.group(1) if m0 else None
    m1 = re.match('.*act=(.*?);.*', ele)
    act = m1.group(1) if m1 else None
    m2 = re.match('.*uid=(.*?);.*', ele)
    uid = m2.group(1) if m2 else None
    m3 = re.match('.*erUid=(.*?);.*', ele)
    erUid = m3.group(1) if m3 else None
    m4 = re.match('.*pid=(.*?);.*', ele)
    pid = m4.group(1) if m4 else None
    m5 = re.match('.*cat=(.*?);.*', ele)
    cat = m5.group(1) if m5 else None
    m6 = re.match('.*plist=(.*?);.*', ele)
    plist = m6.group(1) if m6 else None
```

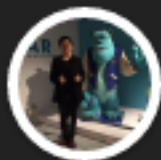


资料清理

01/Feb/2015:00:00:01 +0800	order	U312622727	252b97f1-25bd-39ea-6006-3f3ebf52c80	None
01/Feb/2015:00:00:01 +0800	view	U129297265	e88c3e9a-7e55-76a9-3f89-70c6b334cba	002346
01/Feb/2015:00:00:03 +0800	view	U321001337	cbcaa5a5-4587-abac-b2f3-e2f2db2683eb	001892
01/Feb/2015:00:00:03 +0800	order	U239012343	92e720da-17be-2b67-3383-9e5ccbd9499f	None
01/Feb/2015:00:00:04 +0800	search	U238612347	27e6f25e-3acf-5951-3dd-dd7350e3aacb	None
01/Feb/2015:00:00:05 +0800	view	U10070718237	fa3f1184-d8db-17f9-b09c-7d927c1a429f	000006
01/Feb/2015:00:00:05 +0800	view	U393805241	f2805eb7-a7f7-d805-645a-45e51678e524	000554
01/Feb/2015:00:00:10 +0800	view	U10004621134	b2f454fb-9e52-7ca3-969c-e6056624b5f3	002213

取得使用者的

- act 动作
- uid 标识符
- eruid Cookie
- pid 产品
- cat 群组
- plist 购物车的产品信息



转换成分类问题

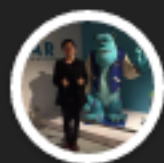
“ 如果知道使用者会根据浏览几次商品以后才下购买决定，是否可以将问题变成一个分类问题？”

属性

- 浏览次数
- 总浏览次数
- 浏览秒数
- 购买/浏览比例

目标属性

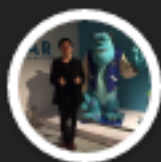
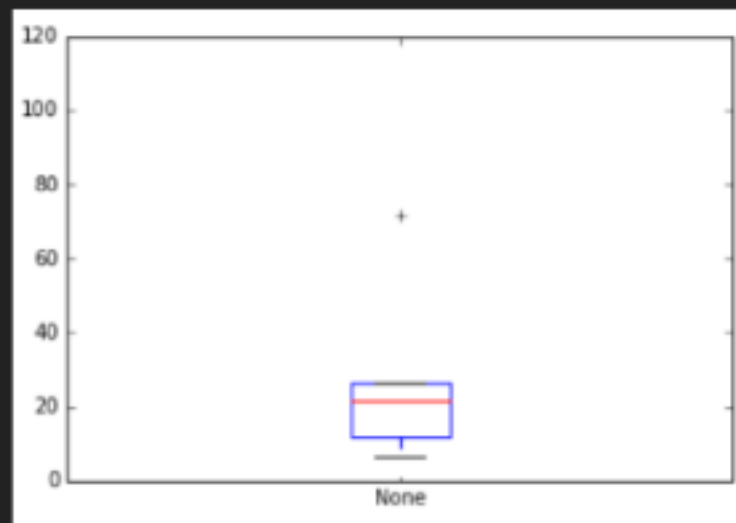
- 购买/不购买



如何将行为归为同一个Session的动作?

01/Feb/2015:00:09:08 +0800	view	U409491353	4fa5c44c-fd08-e0c5-530a-741312ad292b	06
01/Feb/2015:11:21:02 +0800	cart	U409491353	4fa5c44c-fd08-e0c5-530a-741312ad292b	Nc

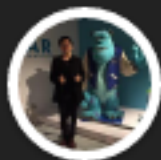
使用Boxplot 推断最多以
每20分钟为一个Session



将数据整理成可分类数据

U339736346	0000445944	2	22	1	1
U339736346	0000178080	2	22	1	0
U46498056	0002689481	1	0	1	1
U234579365	0024026973	2	257	2	1
U438707274	0023733533	1	0	1	1
U328995856	0000515060	5	143	1	0

取得UID,PID, 总浏览次数, 浏览
秒数, 浏览次数, 是否购买

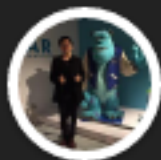
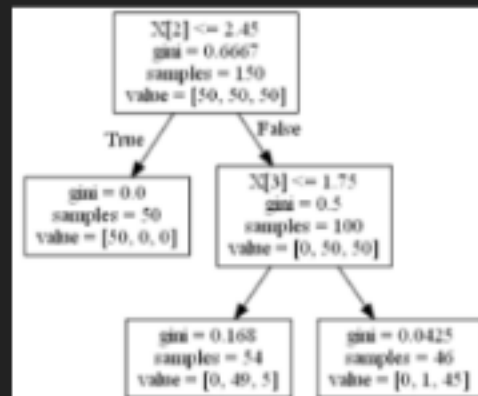


建立判定树

```
from sklearn import tree
treemodel = tree.DecisionTreeClassifier(max_depth = 5)
clf = treemodel.fit(data, target)

tree.export_graphviz(clf, out_file='tree.dot')
```

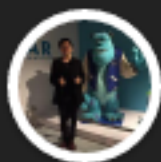
使用GraphViz



产生预测结果

```
predicted = clf.predict(data)
```

1. 预测是否购买 X 观看商品价格
= 预期是否成功销售
2. SUM(所有预期销售)



验证预测结果

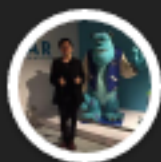
1. 20 个商品有13 个被预测正确
2. 可使用Average Precision at K 验证分数

```
import ml_metrics as metrics

## Answer
actual=['0000143500','0004880654','0024239865','0005721

metrics.apk(actual,answer)
```

$$ap@n = \sum_{k=1}^n P(k) / \min(m, n)$$



Thank You

大數軟體有限公司
丘祐瑋 (**David Chiu**)
david@largitdata.com

