機器學習 hw1　　H24031354 統計 107 江冠駒

- 資料介紹

這是一筆關於 2011 全球前 1000 大網站的資料，選取依據是由 UniqueVisitors 去選出前 1000 多的網站，第一名是 facebook，高達 8.8 億個 UniqueVisitors，第二名是 youtube，接下來的幾名也大家耳熟能詳的網站，如資料分析中的第一張圖所示。

- 欄位簡介

我們只關心 5 個欄位，Rank, PageViews, UniqueVisitors, HasAdvertising and IsEnglish.

欄位名稱	說明
Rank	依據 UniqueVisitors 由大到小的排名
PageViews	網站總瀏覽次數，若小明一年造訪 A 網站 5 次，則 A 網站 PageViews 增加 5 此欄位是本次分析的 response
UniqueVisitors	網站造訪人數，同一人多次造訪不重覆計算，若小明一年造訪 A 網站 5 次，則 A 網站 UniqueVisitors 增加 1
HasAdvertising	某網站是否有廣告
IsEnglish	某網站主要語言是否是英文，只有前 100 名有資料，後面 900 名都是 missing value

- 分析目標

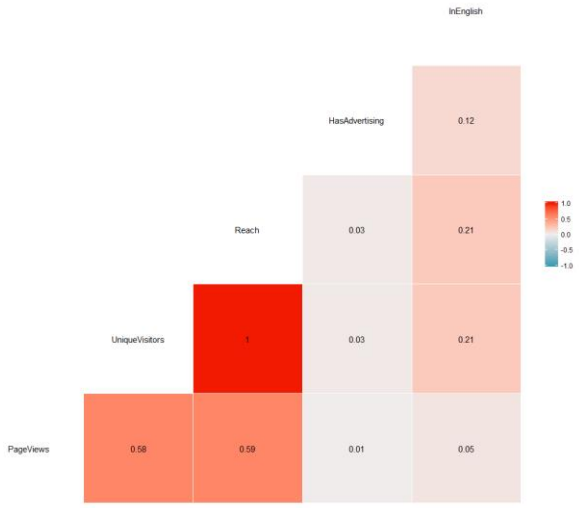
將 UniqueVisitors, HasAdvertising, IsEnglish 這三個當作變數利用迴歸分析方法預測 PageViews

- 資料分析

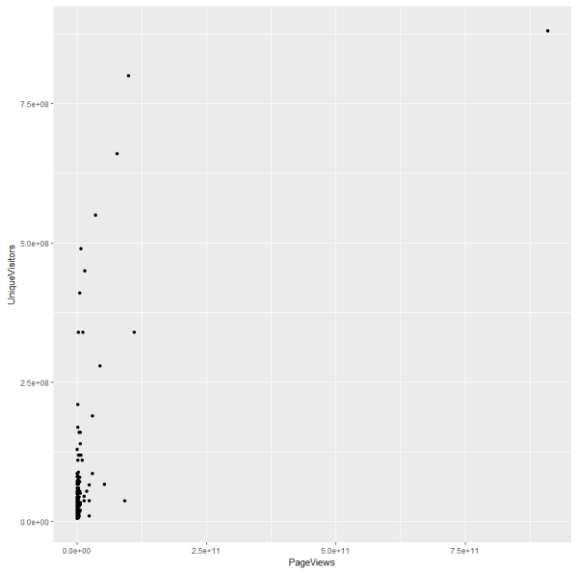
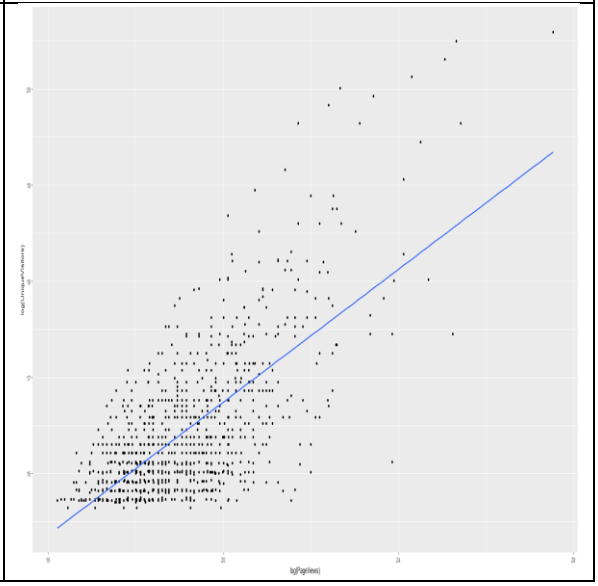
首先，先來看看我們的資料，因為有多達 1000 筆，所以我只放前 35 筆資料

	Rank	Site	Category	UniqueVisitors	Reach	PageViews	HasAdvertising	InEnglish	TLD
1	1	facebook.com	Social Networks	8.8e+08	47.2	9.1e+11	Yes	Yes	com
2	2	youtube.com	Online Video	8.0e+08	42.7	1.0e+11	Yes	Yes	com
3	3	yahoo.com	Web Portals	6.6e+08	35.3	7.7e+10	Yes	Yes	com
4	4	live.com	Search Engines	5.5e+08	29.3	3.6e+10	Yes	Yes	com
5	5	wikipedia.org	Dictionaries & Encyclopedias	4.9e+08	26.2	7.0e+09	No	Yes	org
6	6	msn.com	Web Portals	4.5e+08	24.0	1.5e+10	Yes	Yes	com
7	7	blogspot.com	Blogging Resources & Services	4.1e+08	21.9	5.4e+09	Yes	Yes	com
8	8	baidu.com	Search Engines	3.4e+08	18.0	1.1e+11	Yes	No	com
9	9	bing.com	Search Engines	3.4e+08	18.3	1.1e+10	Yes	Yes	com
10	10	microsoft.com	Software	3.4e+08	18.3	2.7e+09	Yes	Yes	com
11	11	qq.com	Web Portals	2.8e+08	15.0	4.4e+10	Yes	No	com
12	12	ask.com	Search Engines	2.1e+08	11.2	2.0e+09	Yes	Yes	com
13	13	taobao.com	Classifieds	1.9e+08	10.3	3.0e+10	Yes	No	com
14	14	adobe.com	Multimedia Software	1.7e+08	9.2	1.0e+09	No	Yes	com
15	15	wordpress.com	Blogging Resources & Services	1.7e+08	9.2	1.0e+09	Yes	Yes	com
16	16	twitter.com	Email & Messaging	1.6e+08	8.4	6.0e+09	Yes	Yes	com
17	17	youku.com	Online Video	1.6e+08	8.4	3.6e+09	Yes	No	com
18	18	soso.com	Search Engines	1.6e+08	8.4	3.6e+09	No	No	com
19	19	sohu.com	Web Portals	1.4e+08	7.6	5.9e+09	Yes	No	com
20	20	163.com	Web Portals	1.4e+08	7.7	6.5e+09	Yes	No	com
21	21	windows.com	Windows OS	1.3e+08	7.0	5.4e+08	Yes	Yes	com
22	22	hao123.com	Web Portals	1.2e+08	6.3	7.2e+09	Yes	No	com
23	23	tudou.com	Online Video	1.2e+08	6.3	2.7e+09	Yes	No	com
24	24	amazon.com	Shopping	1.2e+08	6.3	4.4e+09	Yes	Yes	com
25	25	apple.com	Mac OS	1.1e+08	5.8	1.1e+09	Yes	Yes	com
26	26	ebay.com	Auctions	1.1e+08	5.8	1.0e+10	Yes	Yes	com
27	27	sogou.com	Search Engines	8.9e+07	4.8	2.3e+09	Yes	No	com
28	28	mozilla.com	Internet Clients & Browsers	8.7e+07	4.7	5.9e+08	No	Yes	com
29	29	yahoo.co.jp	Web Portals	8.7e+07	4.7	3.0e+10	Yes	No	co.jp
30	30	paypal.com	Merchant Services & Payment Systems	8.2e+07	4.4	1.7e+09	Yes	Yes	com
31	31	tmall.com	Apparel	8.1e+07	4.3	2.1e+09	Yes	No	com
32	32	go.com	Web Portals	8.1e+07	4.4	3.3e+09	Yes	Yes	com
33	33	about.com	How-To‚ DIY & Expert Content	8.1e+07	4.3	6.0e+08	Yes	Yes	com
34	34	flickr.com	Photo & Image Sharing	8.1e+07	4.3	1.7e+09	Yes	Yes	com
35	35	56.com	Online Media	8.0e+07	4.3	1.1e+09	Yes	No	com

再來看看我們要拿來做迴歸的欄位間的 correlation plot，在此我特別也把 Reach 加進來看，Reach 和 UniqueVisitors 呈現完全正相關，因此待會只會將 UniqueVisitors 加入迴歸模型，防止發生共線性問題；我們可以看到 PageViews 只和 UniqueVisitors 及 Reach 這兩個變數有較高相關，HasAdvertising 及 IsEnglish 這兩個變數和 PageViews 沒什麼關係，UniqueVisitors,HasAdvertising, IsEnglish 三者間相關性也不大。



就直覺來說 PageViews 和 UniqueVisitors 的相關性應該會很高，所以我們先將這兩欄的資料用散布圖畫出來，如(左圖)所示，這個圖非常難看，大部分的點都集中在左下角，原因是數值太大了且資料不服從常態，導致軸上的刻度間距很大，只有少數幾個值不會擠成一團；因此我同時對 PageViews 和 UniqueVisitors 取 log，再去畫散布圖，如(右圖)所示，藍線為迴歸線，果然可以看出兩者有高度的正相關，因此待會 PageViews 和 UniqueVisitors 都會取 log 再丟下去建模。

原始	取 log 後
	

接下來就是建立迴歸模型
先建立 full model: log(PageViews)~log(UniqueVisitors)+HasAdvertising+InEnglish

```
Call:
lm(formula = log(PageViews) ~ log(UniqueVisitors) + HasAdvertising +
    InEnglish, data = rawdata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4283 -0.7685 -0.0632  0.6298  5.4133

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.94502    1.14777  -1.695  0.09046 .
log(UniqueVisitors)  1.26507    0.07053  17.936  < 2e-16 ***
HasAdvertisingYes    0.30595    0.09170   3.336  0.00088 ***
InEnglishNo          0.83468    0.20860   4.001 6.77e-05 ***
InEnglishYes        -0.16913    0.20424  -0.828  0.40780
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.067 on 995 degrees of freedom
Multiple R-squared:  0.4798,    Adjusted R-squared:  0.4777
F-statistic: 229.4 on 4 and 995 DF,  p-value: < 2.2e-16
```

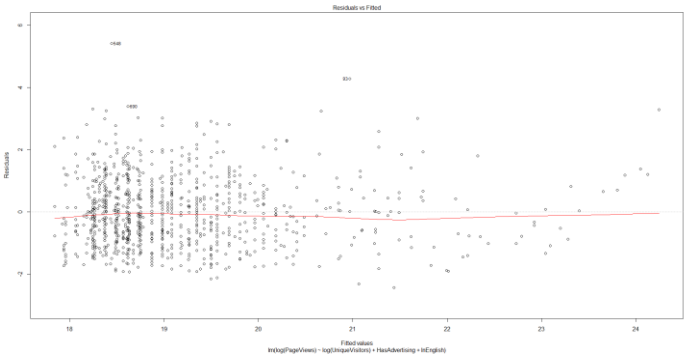
這邊會出現 InEnglishYes 和 InEnglishNo 是因為我將遺失值"NA"也當成此變數的一個 level 且是 baseline，若不這麼做的話，R 只會拿前 100 筆資料去做迴歸。
我們可以看到 log(UniqueVisitors),HasAdvertising,InEnglish 三個變數對 log(PageViews)都是有影響的

接下來分別對 log(UniqueVisitors),HasAdvertising,InEnglish 三個變數做簡單迴歸

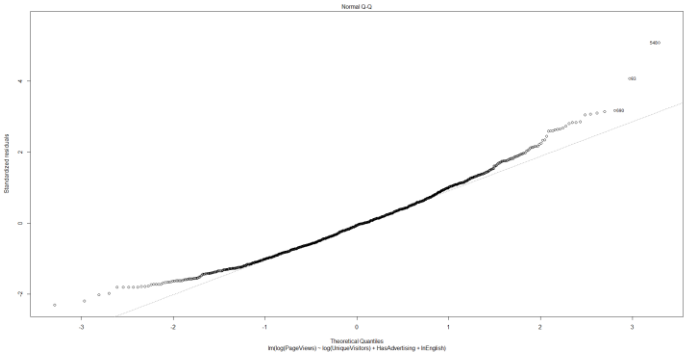
模型	判定係數 R^2	RMSE
log(PageViews)~log(UniqueVisitors)	0.4616	1.0826
log(PageViews)~ HasAdvertising	0.0107	1.4674
log(PageViews)~ InEnglish	0.3043	1.2306
full model	0.4798	1.0641

由上表可知，log(UniqueVisitors)的 R^2 高達 0.4616，和 full model 的 R^2 已經差不多了，代表其他兩個變數對 log(PageViews)的解釋力不高；雖然看起來 InEnglish 的 R^2 也不小有 0.3043，但此變數 9 成的資料都是 missing value，可信度不高，我便再拿前 100 筆 InEnglish 欄位有值的資料和 log(PageViews)做簡單迴歸，發現 R^2 僅有 0.0312，代表 InEnglish 只能解釋前 100 筆 log(PageViews)總變異的 3%左右，因此推斷 InEnglish 對 log(PageViews)的解釋力不大。

最好的模型是 full model，因此我要來看它是否符合模型假設，首先是殘差圖，變異沒有隨預測值增加而波動，大致都在正負 2 倍標準差內，符合同質變異數假設。



接下來看殘差是否符合常態分配，從 qqplot 看來殘差並沒有很符合常態



做 Shapiro-Wilk 看看殘差是不是真的不符合常態，虛無假設 H0:殘差服從常態分配，因為 p-value < 0.05，拒絕 H0，果然不服從常態分配

```
        Shapiro-Wilk normality test

data:  fullmodel$residuals
W = 0.97627, p-value = 1.037e-11
```

最後看看殘差有無符合獨立性假設，虛無假設 H0:殘差間相互獨立，因為 p-value > 0.05，代表不會拒絕 H0，符合獨立性假設

```
 lag Autocorrelation D-W Statistic p-value
   1     -0.03899812      2.068281     0.3
 Alternative hypothesis: rho != 0
```

- 結論:雖然常態假設不符合，但不影響預測力。UniqueVisitors 也可用 Reach 代替去建模。另外 InEnglish 該欄的 missing value 比例太高，可能影響預測準確率。

- Reference: https://rpubs.com/skydome20/R-Note5-First_Practice

Drew Conway and John Myles White (2012) Machine Learning for Hackers
https://molecular-service-science.com/2013/11/27/r-ggplot-tutorial-1/
http://www.sthda.com/english/wiki/ggally-r-package-extension-to-ggplot2-for-correlation-matrix-and-survival-plots-r-software-and-data-visualization