

以下內容都是還沒開始實作時，對數據該如何處理的猜想

若實作後有什麼變數的話，處理方式會再進行更動

當前未處理前的 attributes:

0	movie_name	14277	non-null	object
1	rating	14009	non-null	object
2	tomato_genre	11953	non-null	object
3	directors	13889	non-null	object
4	imdb_dirScore	6993	non-null	float64
5	tomato_dirScore	5058	non-null	float64
6	on_streaming_date	9681	non-null	object
7	title_year	4800	non-null	float64
8	runtime_in_minutes	11771	non-null	float64
9	studio_name	10280	non-null	object
10	box_office	2206	non-null	object
11	tomatometer_rating	10479	non-null	float64
12	tomatometer_count	11987	non-null	float64
13	audience_rating	10329	non-null	float64
14	audience_count	10329	non-null	float64
15	tomatometer_avg_rating	11987	non-null	float64
16	score_sentiment	10329	non-null	object
17	audience_avg_rating	10329	non-null	float64
18	num_voted_users	4903	non-null	float64
19	director_facebook_likes	4803	non-null	float64
20	ir_all_actor_1_name	4896	non-null	object
21	actor1_score_imdb_actScore	4896	non-null	float64
22	actor1_score_tomaAudience_actScore	3829	non-null	float64
23	actor1_score_tomaTmatomer_actScore	3893	non-null	float64
24	actor_1_facebook_likes	4896	non-null	float64
25	ir_all_actor_2_name	4890	non-null	object
26	actor2_score_imdb_actScore	4890	non-null	float64
27	actor2_score_tomaAudience_actScore	3384	non-null	float64
28	actor2_score_tomaTmatomer_actScore	3471	non-null	float64
29	actor_2_facebook_likes	4890	non-null	float64
30	ir_all_actor_3_name	4881	non-null	object
31	actor3_score_imdb_actScore	4881	non-null	float64
32	actor3_score_tomaAudience_actScore	3166	non-null	float64
33	actor3_score_tomaTmatomer_actScore	3275	non-null	float64
34	actor_3_facebook_likes	4881	non-null	float64
35	gross	4044	non-null	float64
36	imdb_genres	4903	non-null	object
37	cast_total_facebook_likes	4903	non-null	float64
38	imdb_num_user_for_reviews	4884	non-null	float64
39	language	4892	non-null	object
40	country	4899	non-null	object
41	budget	4423	non-null	float64
42	imdb_score	4903	non-null	float64
43	movie_facebook_likes	4903	non-null	float64

看起來可以忽略的 attribute:

Movie_name(0), 很明顯不會影響

Directors(3), 這個依據後面的 director_score 分析就好了

title_year(7), 這個要再看看・說不定老電影有加分

on_stream_date(6), 這個有點太細了(?)

studio_name(9), 這個也要再看看・主要是因為很難量化分析

box_office(10), 缺太多了・總不能補 0

num_voted_users(18), 投票人數跟分數感覺沒有絕對關係

gross(35), 主要是缺太多了而且感覺不好補 NULL

actor_name(20,25,30), 靠後面 4 個 attribute 分析就好

language(39), country(40), 主要是感覺很難量化(

runtime_in_minutes(8), 不重要。

其他不忽略的處理:

Genre(2, 36), NULL 補 0・其他不同類型分配 1~n 的分類・不過如果同時有很多個應該怎麼辦呢...

Score_sentiment(16), NULL 補 0・POSITIVE = 1, NEGATIVE = -1 (或是 123)

20~43:

感覺全部資料可以分成有 20~43 以及沒有 20~43 的來分析。

藉由觀察，IMDB 跟爛番茄的評分大概都是 2:1，所以如果其中一個 NULL 就照比例換算過去給另一邊用

至於 Tmatomer 是什麼我不知道 QQ 看起來跟 audience score 關連不大，而且還缺很多真的不知道怎麼補 <

3 個演員的分數再總合成一個，並且按照有 3 個演員的話，分數就按照 {0.4,0.3,0.3} 的比例去分配，2 個的話則是 {0.6,0.4}，一個是 {1}。

1/8 追記：

因為我不知道 null 數值要怎麼處理所以總之就先補 0 上去了

如果兩邊的評分都是 0 的話就忽略不用，一邊是 0 的話就照上面的比例補上 0 的地方。

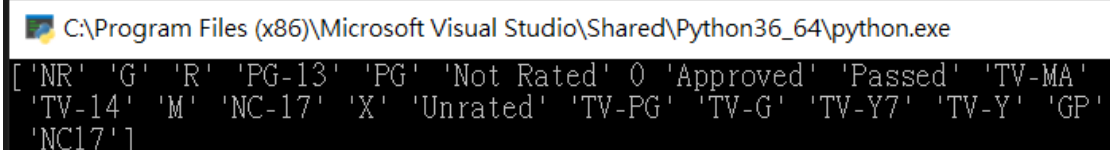
facebook_likes(19, 24, 29, 34, 37, 43) – 因為有些可能是 0 所以補 0 似乎沒甚麼問題(?)，當然如果補 0 沒辦法使用的話就先照上面說的把資料分成有/沒有 20~43 的兩筆來算。

導演在 imdb 以及爛番茄的分數(4, 5)的比例也差不多是 2:1，所以就照上面相同方式做轉換並且補齊數據。

把 0 的部分用中位數來補好了

Rating(1)

```
print(all_movie.rating.unique())
```



```
['NR' 'G' 'R' 'PG-13' 'PG' 'Not Rated' 0 'Approved' 'Passed' 'TV-MA' 'TV-14' 'M' 'NC-17' 'X' 'Unrated' 'TV-PG' 'TV-G' 'TV-Y7' 'TV-Y' 'GP' 'NC17']
```

把 Not Rated, Approved, Passed, Unrated 這幾個先當成 0

然後把其他幾種分成 1~15

tomatometer_rating(11), tomatometer_count(12), audience_rating(13),
audience_count(14):

雖然不知道是甚麼不過數據還滿齊的，總之就先用中位數來補。

/*其餘 score 的部分，就看看是不是相加取權重後再去做分群當成是 label 使用

目前個人對評分的部分是想說如果沒辦法算出太準確的評分的話，就分成 10

個群來做，並且使用 KNN 來做預測。以 audience avg rating 為準，如果沒有

值的話再用 imdb score 來分*/

1/9 追記：

把評分分成低中高 3 個區間就好。

其餘的(23, 28, 33, 38, 41)

補中位數

目前還需要處理的資料：

2, 36, 這兩個我做不來 pien

最後會用到的資料：

12~15, 44, 51~55

19, 24, 29, 34, 37, 43

目標(label): 56(audience_avg_rating)