Python資料分析與機器學習應用

期末報告

組別：H組

組員：陳宥誠、柯筆翔、李宇昂、高嘉彣、林晉揚

內容

一、主題與動機 4
二、資料集說明 4
三、資料前處理 4
四、分析結果 8
五、建模方式和調整方式 10
A. 分類 10
● Decision Tree 10
● Logistic Regression 11
● Random forest 11
● XGBoost 11
B. 迴歸 11
● Random forest 11
● XGBoost 12
● Ensemble Model 13
六、模型的結果/成果 13
A. 分類 14
● Decision Tree 14
● Logistic Regression 14
● 隨機森林 14
B. 迴歸 14
● 隨機森林: 14
● XGBoost: 14
● Ensemble Model: 14
C.介面 16
七、遇到的困難和解決方法 16
1. 演員資料特徵處理：Unsolved 16
2. 訓練集與測試集切點評估：Solved 17
3. 正確率不高與誤差偏高：Solved 17

4. 尋找最佳超參數過於費時：Solved..... 17
八、心得感想..... 17
宥誠：..... 17
晉揚：..... 18
宇昂：..... 18
筆翔:..... 18
嘉彣:..... 19
九、專案重要性..... 19
十、團隊分工..... 19
資料集來源：..... 20

一、主題與動機

小組此次期末報告的主題為「影視資訊與串流平台於電影分數之分析與預測」。「Netflix and chill」、「追劇」早已成為現代人不可或缺的休閒活動，也因此影視平台間的競爭也越發激烈，各自主打不同的特色，如Amazon Prime Video 就主打大量歐美作品與原創作品、Apple TV+則主打4K超高畫值以及杜比視界與杜比環繞音效，我們耳熟能詳的 Netflix 則是擁有不同地區風格迥異的影集作品，原創動畫以及紀錄片。不同的影視平台都擁有相對應的客群，因此我們希望可以去探討，同部電影在不同影視平台是否有著相同的評價、電影主打的卡司、導演，甚至影片類型對電影評分的影響，再者，能在多個平台上映的電影是否就能擁有較高的評分，以及我們是否可以透過機器學習，預測電影的評分。

二、資料集說明

我們由Kaggle取得Netflix、Amazon Prime、Disney+、Apple TV+、HBO Max與Paramount+的資料集，資料集內有兩個檔案，分別是電影資訊與演員資訊，電影資訊內容包含電影的名稱、敘述、類型、年份、時長、分級、分類、製片國家、IMDB[1]與TMDB[2]的分數；演員資訊內容有電影名稱、演員名字與飾演角色。

三、資料前處理

1. 檢查六個資料集內是否有重複項。
2. 將六個資料集的電影資訊合併，共有23358筆資料。
3. 檢查空值，並以以下方式進行處理：

欄位	處理方式
敘述	以'have no description'補上
年齡分級	以”unprovided”補上
季數	當電影是movie，season就是空值，所以以’movie’補上

其餘欄位空值數皆占約10%資料量，待進行分析或建模時才處理。

[1] Internet Movie Database，網路電影資料庫，為亞馬遜旗下公司，擁有豐富的影視資訊與專業的評分機制
[2] The Movie Database，被譽為影視資訊的維基百科

4. 年齡分級以台灣常用分級方式（普遍級、限制級…等）替代歐美分級方式（PG、NC-17…等）

歐美分級（調整前）	台灣常用分級方式（調整後）
TV-MA	成人級，未滿17歲的觀眾需有家長陪同觀看。
PG	輔導級，建議家長陪同觀看，可能包含少量不宜兒童觀看的場景。
R	限制級，未滿17歲的觀眾需有家長陪同觀看，可能包含暴力、血腥、裸露或粗口等內容。
TV-14	14歲以上觀眾觀看，可能包含暴力、性、恐怖或粗口等內容。
nan	空值，代表缺失的值。
TV-PG	家長指導級，建議家長指導觀看，可能包含一些較為成熟的主題，但不會含有過於露骨或暴力的內容。
PG-13	13歲以上觀眾觀看，可能包含部分成人內容，但未到限制級的程度。
TV-Y	所有年齡段觀眾均可觀看。
TV-Y7	七歲以上觀眾觀看，可能包含一些較為成熟的主題，但不會含有過於露骨或暴力的內容。
TV-G	所有年齡段觀眾均可觀看，通常是兒童節目。
G	普遍級，適合所有年齡層觀看。
NC-17	限制級，只適合成年觀眾觀看，可能包含過於露骨、暴力、裸露或粗口等內容。
TV-Y7-FV	七歲以上觀眾觀看，其中包含部分“暴力場景”。

5. 將"年齡分級"做以下的mapping，使該欄位以數值方式呈現

   "unprovided": 0,

   "普遍級": 1,

   "七歲以上": 2,

   "輔導級": 3,

   "13歲以上": 3,

   "14歲以上": 3,

   "家長指導級": 3,

   "限制級": 4,

   "成人級": 5

6. 去除"電影種類"、"製片國家" 中出現頻率較低者並對剩餘者以one hot encoding處理:

   a. 使欄位以數值方式呈現，方便機器學習

   b. 解決單一電影可能同時屬於多種電影種類、多個製片國家的問題

   c. 先去除出現頻率較低者，以免One hot encoding造成"Feature Explosion"

   d.

14	drama
15	comedy
16	thriller
17	action
18	documentation
19	romance
20	crime
21	family
22	animation
23	scifi
24	fantasy
25	horror
26	european
27	music
28	history
29	reality
30	sport
31	war
32	western

33	US
34	GB
35	IN
36	CA
37	FR
38	JP
39	DE
40	ES
41	KR
42	IT
43	CN
44	AU
45	MX
46	BR
47	NG
48	AR
49	BE
50	ZA
51	HK
52	PH
53	TR
54	PL

7. 新創建欄位"num_seasons"欄位，其中，0代表電影，其他值則代表影集季數

8. 刪除較難轉換成數字的文字敘述欄位(電影名稱、電影欄位、導演、演員)

9. 因為欲呈現的是「監督式學習」，所以將資料集中imdb_score、tmdb_score欄位是空值者整列刪除=>資料集大小23358->14016

10. 將imdb_score進行四分位距處理，以利將任務視為分類問題處理

四分位距處理:

imdb_score 排名在資料集前25% →0

imdb_score 排名在資料集25%~50% →1

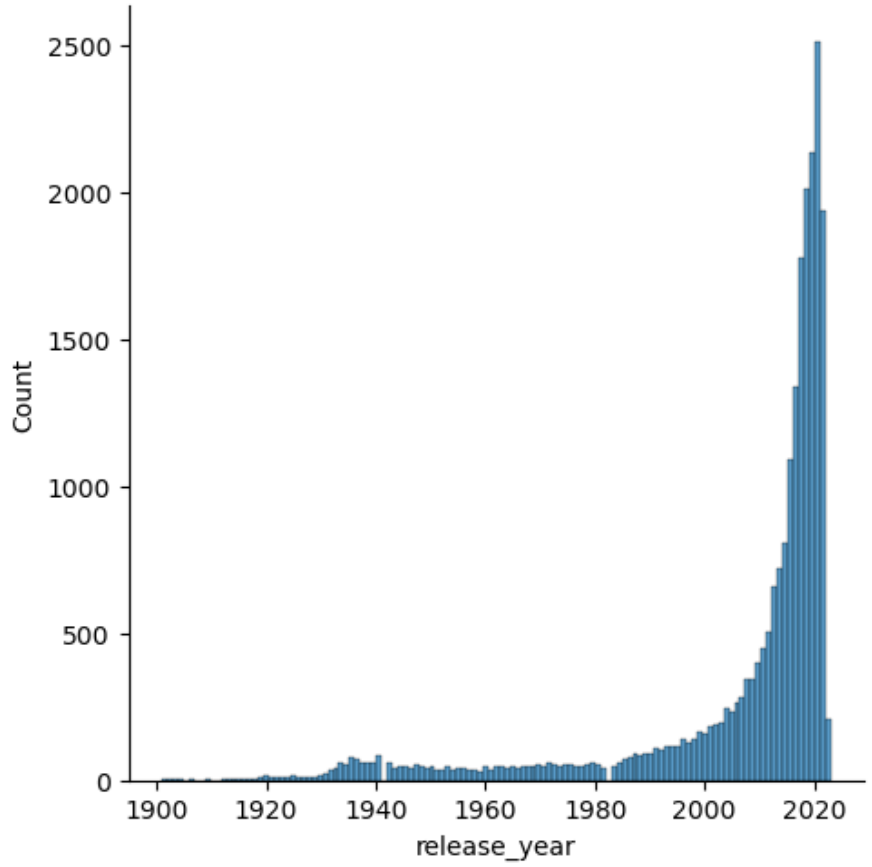
imdb_score 排名在資料集前50%~75% →2

imdb_score 排名在資料集前75%~100% →3

11. 資料進行前處理後，資料集大小: 14016*57

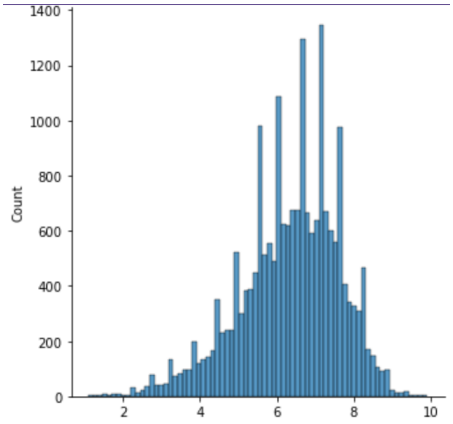
12. 切分訓練集與測試集:

在本專案中，我們嘗試針對不同的機器學習演算法選擇最適切點，然而考慮到本專案使用多種算法進行訓練比較，我們決定在各算法中使用相同方式切分訓練集與測試集。考慮到在真實場景中，我們訓練完成的模型將用於預測未來上映的電影，觀察各年代發行電影數量，我們選擇2019以前的資料作為訓練集，2020以後的資料做為測試集，各自佔比75%、25%（如下圖所示）。

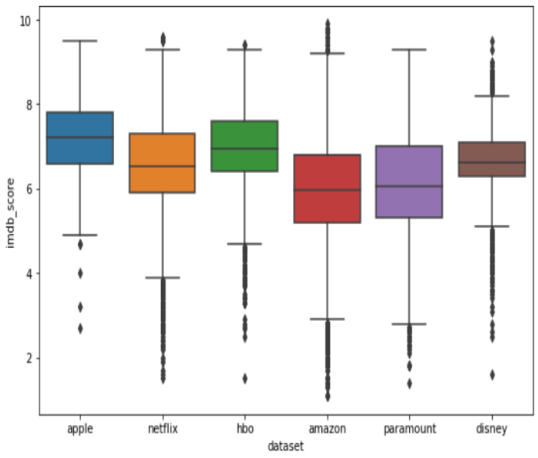


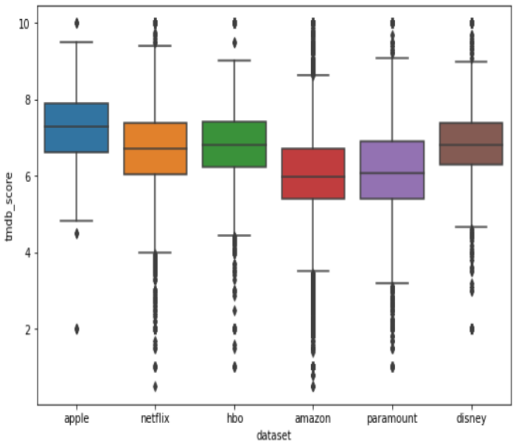
四、分析結果

1. IMDB分數平均值為6.3分，高於9分的電影極少，高於7.3分的電影即為前25%高分。

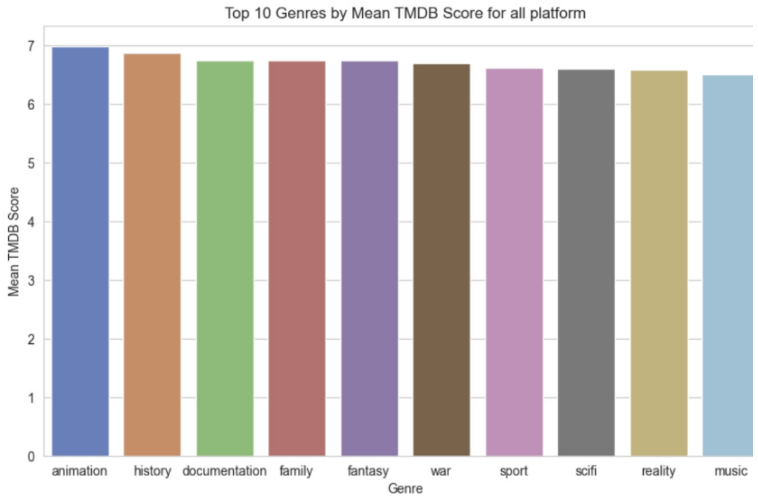


2. Amazon上架之電影/影集分數分布最廣。
3. Amazon和Paramount相較其他平台，更願意上架分數較低之電影。

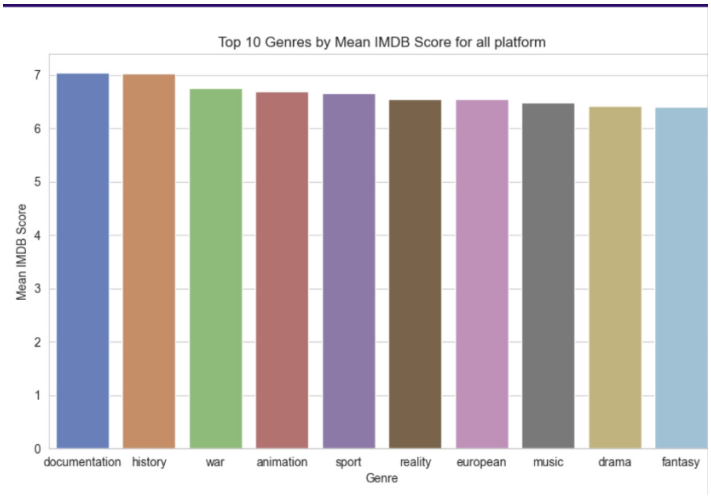




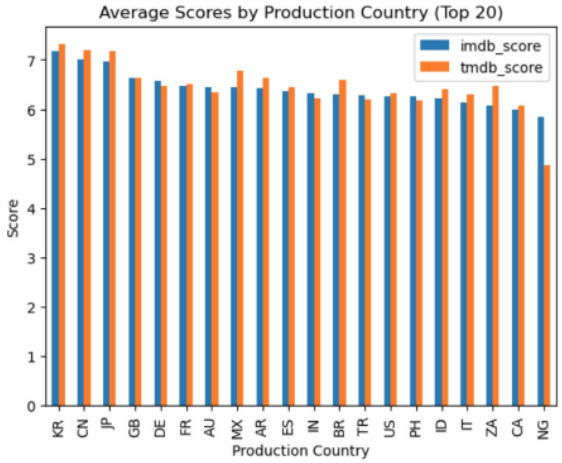
4. TMDB分數前三高類型為: Animation, History, Documentation



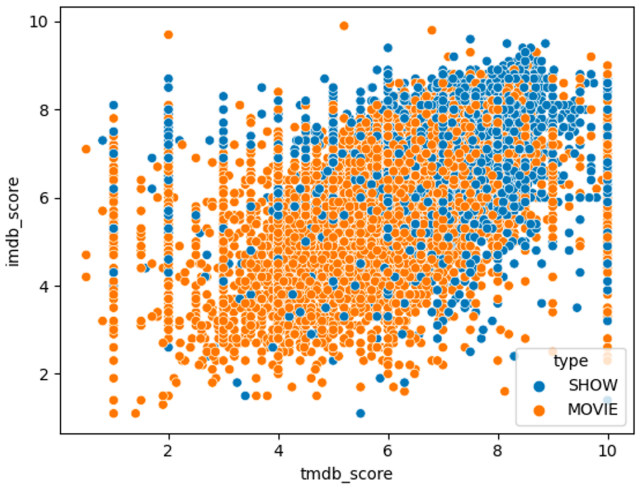
5. IMDB分數前三高類型為: Documentation, History, War



6. 電影產出數前20名當中，平均分數最高前三名為: 韓國、中國、日本。



7. show無論在 imdb score 還是 tmdb score 都比 movie 高。此外，imdb score 和 tmdb score 相關，相關係數為0.6。



8. Apple tv+獨特性最高

	HBO	Amazon	Disney+	Netflix	Paramount	Apple TV
movie	2408	9322	1314	3831	2518	62
show	622	1551	540	2306	664	108
上架兩個平台	42	1759	21	211	1668	1
上架三個平台	9	18	0	12	18	0
重複率	1.7%	16.3%	1.1%	3.6%	53.0%	0.6%

五、建模方式和調整方式

一共採用了七種方式建模，包含四種分類方式及三種回歸方式。分類採用Decision Tree、Logistic Regression、XGBoost與Random forest；回歸使用到Random forest、XGBoost與Ensemble Model。

A. 分類

- Decision Tree

在Decision Tree中有嘗試使用二元的結果（imdb 於中位數以上為1，中位數以下為0）及四分位距作為結果。二元在模型表現上雖較好，但卻無法進行更細部的解釋，因此選擇使用四分位距最為應變項。

Decision Tree 未調參測試集正確率為0.364，手動調參後測試集正確率為0.435，用GridSearchCV調參後測試集正確率為0.404。手動調參中最佳深度為6

，葉節點所需最小樣本數為1，訓練集正確率為0.460，沒有過擬合的問題，但正確率不足50%。

- Logistic Regression

  Logistic Regression 未調參測試集正確率為0.382，用GridSearchCV調參後測試集正確率為0.407，正確率不足50%。

- Random forest

  Random forest 利用迴圈方式找到超參數後，測試集的預測正確率為0.48，正確率不足 50%。

- XGBoost

  XGBoost的訓練資料集結果為0.67，測試資料結果為0.56，結果不如預期。

B. 迴歸

- Random forest

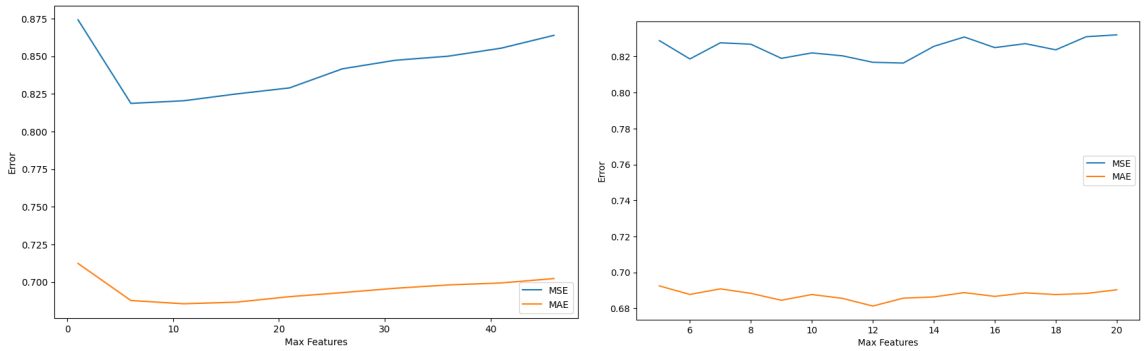
使用Random Forest、XGBoost各別訓練，並嘗試將兩者合併成Ensemble Model

1. 在測試資料集中切分部分驗證資料集出來，進行Random Forest、XGBoost的最佳超參數組合:

Random Forest—

先找尋最合適的Max Feature

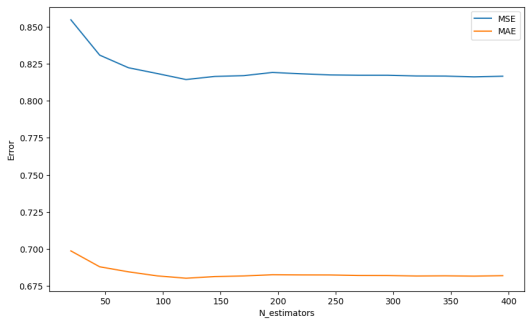
先大範圍、大間隔搜尋;再依據結果縮小範圍以小間隔搜尋

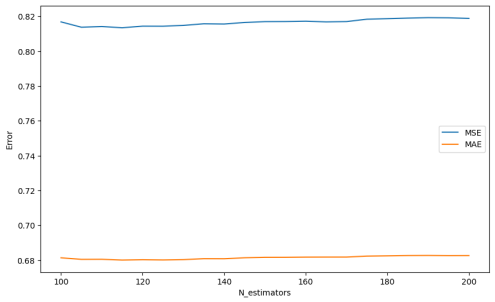


固定上步驟找到的最合適的Max Feature

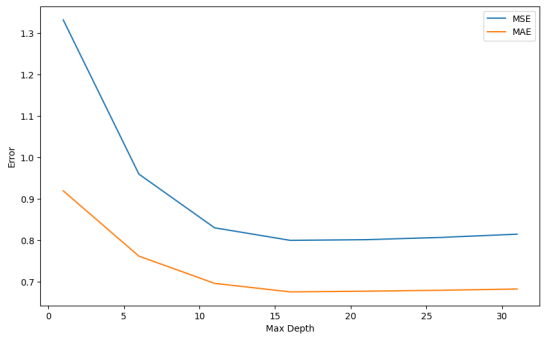
找尋最合適的N_estimators

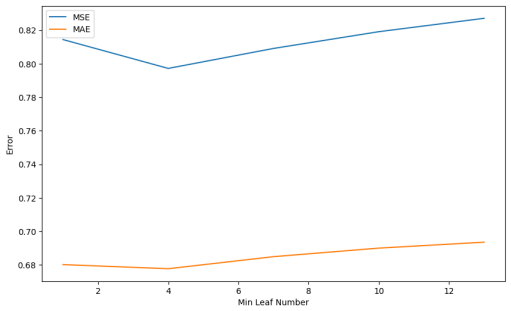
先大範圍、大間隔搜尋;再依據結果縮小範圍以小間隔搜尋





依據相同方法再找max_depth、min_leaf_number

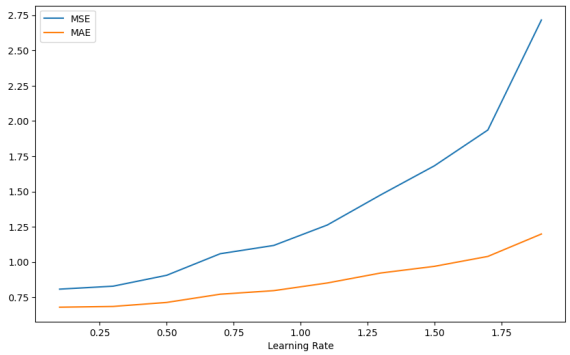




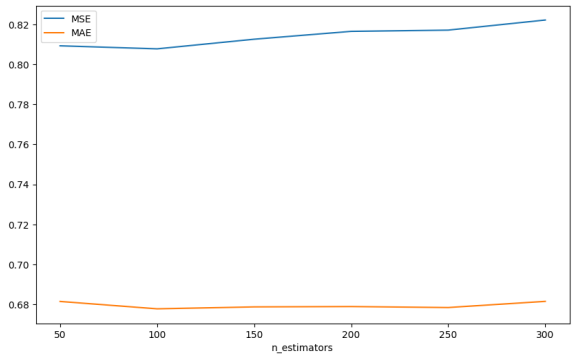
=>得到random forest最佳超參數組合: n_estimators=120, max_features=12, min_samples_leaf=4

- XGBoost

先調整Learning Rate



再調整N estiamtors



p.s. 用這個方式找到的最佳超參數也許不是Global Min但至少是個Local Min!!

- Ensemble Model

使用最佳超參數組合的random forest、XGboost模型進行Ensemble Learning，得到:

Ensemble Model = 1.2128 * (predictions of Random Forest) - 0.2128 * (predictions of XGBoost)

使用整筆training data進行訓練，得到真正拿來預測的Random Forest、XGboost、Ensemble model

六、模型的結果/成果

以一部電影的時間長度、上映年等特徵作為X，imdb 分數會落在所有電影的imdb分數的四分位距的哪一個位距為Y，進行預測。

A. 分類

- Decision Tree

  → 模型的預測準確度為 0.46

  → 決策樹中影響 IMDB 分數的重要特徵包含：電影時長、上映年份、季數

- Logistic Regression

  → 模型的預測準確度為 0.40

  → Logistic Regression 中影響 IMDB 分數的重要特徵包含：電影時長、年齡分級、季數

- 隨機森林

  → 模型的預測準確度為 0.47，平均平方誤差為 1.2；重要特徵包含季數、電影時長，及年齡分級。

B. 迴歸

- 隨機森林:

  Mean Squared Error (MSE): 1.0392

  Mean Absolute Error (MAE): 0.7715

- XGBoost:

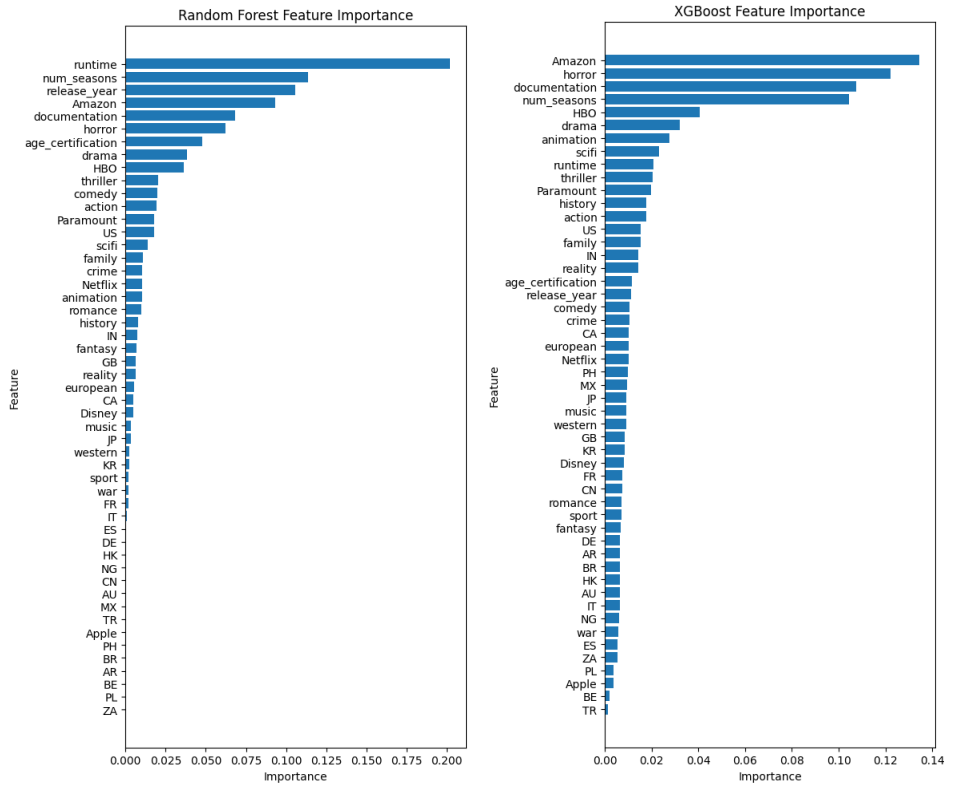
  Mean Squared Error (MSE): 1.0547

  Mean Absolute Error (MAE): 0.7747

- Ensemble Model:

  MSE: 1.0505

  MAE: 0.7769



1.「年齡分級、電影時長、影集季數」在各個模型中皆是前幾重要的Features

2. 在各個模型中,「電影種類」重要性皆大於「製片國家」

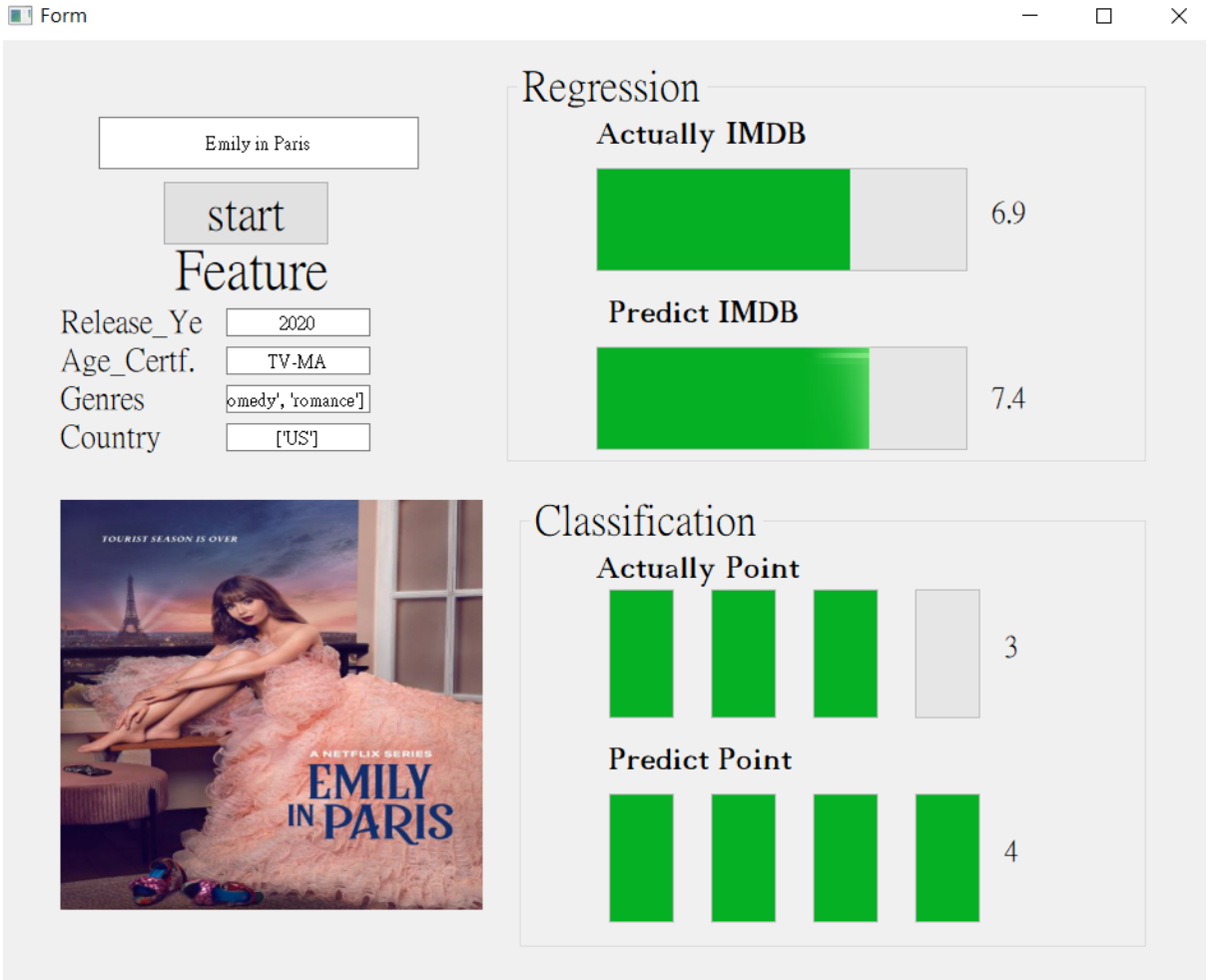
C.介面

1.在使用者介面上分為兩部分:輸入的電影資訊與機器學習結果

2.輸入電影資訊的部分由輸入電影名稱找到相對應的電影資訊包含:上映年分、分類、電影種類、國家...

3.回歸模型的的部分以實際的imdb分數與預測的imdb分數以能量條的方式呈現，並且顯示實際分數與預測分數。

4.分類結果以亮燈的方式顯示實際分類與預測的分類結果，亮幾個燈表示為第幾類。



七、遇到的困難和解決方法

1. 演員資料特徵處理：Unsolved

- 困難點：一部電影有多位演員與導演，難以評估各演員與導演所代表的意義

- 嘗試解法：將導演和演員依照過去其所參與過之電影的 IMDB 平均分數「分類」，轉換成新的欄位
- 結論：因時間有限，優先處理其他 Feature

2. 訓練集與測試集切點評估：Solved

- 困難點：不同的模型有不同的最佳切點，然而本專案最終將比較各種模型對於結果的表現，因此需統一切點
- 結論：討論後考量到真實情境中的時序性，決定以2019年以前的資料為訓練集（75%），2020年以後的資料為測試集（25%）

3. 正確率不高與誤差偏高：Solved

- 困難點：分類器的演算法預測出來的正確率不高與誤差偏高
- 嘗試解法：將特徵重要性前三高的資料再進行標準化，成效不佳
- 結論：問題不適合分類，先利用迴歸分析先針對 IMDB 分數進行預測，再以其結果用來進行分類的預測

4. 尋找最佳超參數過於費時：Solved

- 困難點：Python套件GridSearch因交叉驗證導致過於費時
- 嘗試解法：依序尋找不同超參數的最佳位值，找到以後就固定不動，變動下一個超參數
- 結論：雖然這樣有可能落入Local Minima之中，但效率高，且MSE與MAE皆在可接受範圍

八、心得感想

宥誠：

經過這次實作發現可以用機器學習完成的事非常多，只要有足夠的時間與創意即可，不過如何在短時間內完成目標也是我在本次報告學習到的。而藉由與組員的合作也發現大家的意見與思考的邏輯雖有時會不同，但透過交流都可以讓團隊的目標離完成更進一步。

晉揚:

在不管是資料前處理還是後續的建模，最大的體悟是在機器學習裡面寫程式不見得是最為困難的，對資料處理的想法反而更為困難。資料前處理中要對缺失值、文字與各項特徵進行處理，到訓練模型時要採用的 feature，可能的 output，適合的演算法，都是在寫程式之外需要去想像、思考與判斷的，這樣才能知道每一步的成敗原因，在調參或是解決問題時才能針對根本原因去改進，畢竟從資料前處理的方式，到 feature 的選擇，以及演算法的使用，都是環環相扣，每個環節的改變都將影響到結果。

宇昂:

資料前處理技術十分重要，如One Hot Encoding、feature mapping、缺失值處理；
在實作project的過程有感受到集成學習的強大，但因時間心力有限，尚未嘗試集成學習的各種方法，製作的子模型也還不夠多；
除了對資料的X(feature)做前處理、轉換以外，使用domain knowledge重新定義機器學習的問題也是很重要的(如: 將回歸問題換為分類問題、先做回歸預測再轉成分類問題評估效能...)；
報告時，使用視覺化界面demo會比起講解複雜的訓練背景更有效；
在面對有文字描述的資料，自然語言處理等技術是我所欠缺的；

筆翔:

實際完成project的過程中深刻感受到，在實作機器學習的路上，有70%的時間其實是在做資料清洗的部分，因為有好的資料集在機器學習的過程中才能避免很多操作上面的困難，例如有缺失資料的部分、又過者有些字串的資料要做one-hot encoding才能夠順利進行分析。

另外30%的時間實作機器學習，最難的工作其實是找到妥善的參數與找到重要feature進行模型訓練，我覺得在實作上真的是一大挑戰，只能不斷的try and error找到相對較好的答案，但不太能夠確定是否已經找到最好的參數與預測結果真的是一大難題，真的只能透過實作project逐漸掌握這項技能。

嘉彣:

在學習機器學習時，由於是一部分一部分的學習，透過專案的進行才能將每個步驟合併、更清楚過去練習的部分。這樣統整的方式很能夠幫助我了解每個步驟的處理及需要注意的細節。能夠如此完成一項專案，很有成就感。

原本以為進行機器學習專案時，只要按照步驟做就不太會遇到問題。然而在實際進行專案時會遇到許多之前沒有想到的部分，如缺失值該如何處理、正確率太低、對於模型的解釋怎麼樣才算合理。遇到這些問題時，除了先獨立思考怎麼做之外，透過和大家討論可以發現自己原先沒有想到的解決方法，最後補足缺失，共同合力完成一項專案，因此如此無論是在機器學習或是團隊合作上都學習到許多。

九、專案重要性

本專案可以讓影音串流平台使用者可以更容易找到品質比較好的電影跟影集，同時可以讓使用者輕易的了解評分背後的機制，在挑選影片時，可以更客觀的去選擇自己喜愛的電影，避免單單被分數影響。對於電影、影集工作者來說，也可以將模型進行調整，透過對模型的優化，找出時代的趨勢與喜好，進而得出一套高分電影方程式。找出電影種類的變化趨勢，與影響評分的關鍵因素，未來在製作影集與電影時可以將模型結果作為參考，預測電影的評分高低，甚至有望作為預估票房的依據。

十、團隊分工

姓名	團隊分工
陳宥誠	Data Analysis, ModelTraining, Tuning Parameter, PPT presentation
柯筆翔	Data Analysis, ModelTraining, PPT presentation, UI interface production, Code organization
李宇昂	Data Pre-processing, Data Analysis,

	ModelTraining, Tuning Parameter, PPT production
高嘉彣	Data Cleaning, Data Preprocessing, Web Scraping, ModelTraining, Tuning Parameter, PPT production
林晉揚	Data Cleaning, Data Preprocessing ,Web Scraping, PPT production, Report Organization

資料集來源:

Netflix、Amazon Prime、Disney+、Apple TV+、HBO Max與Paramount+