科技部補助

大專學生研究計畫研究成果報告

＊ ********* ＊＊ ＊
＊ 計 畫 ＊
＊ ： 前進奧斯卡的秘訣—電影評價的模型建立、探討及預測 ＊
＊ 名 稱 ＊
＊ ********* ＊＊ ＊

執行計畫學生： 陳育婷
學生計畫編號： MOST 106-2813-C-006-001-M
研 究 期 間 ： 106 年 07 月 01 日至 107 年 02 月 28 日止，計 8 個月
指 導 教 授 ： 張欣民

處理方式： 本計畫可公開查詢

執 行 單 位： 國立成功大學統計學系（所）

中華民國 107 年 03 月 28 日

目錄

摘要

電影是現代人生活的必需品，在科技快速汰弱換強的今天，不但沒有被淘汰，甚至越來越受歡迎。但，就算是票房保證的電影，也不代表能在國際影壇發光。好電影能讓人回味無窮、影響世界；壞電影則是船過水無痕、浪費金錢與時間成本。藝術本身就是主觀的，電影沒有絕對客觀的評量。然而，市場中仍有一些指標被公認有代表性。大型影評網如 Internet Movie Database (IMDB)或 Rotten Tomatoes，收集來自世界各地用戶及影評評價，歷史悠久，有一定的公信力。

因此，本研究以 1916 年到 2016 年 4567 部 IMDB 網站上的電影為研究對象，透過資料視覺化初步判斷哪些特徵對 IMDB 分數有較大的影響力；在模型建立與預測方面，選取在電影上映前可以控制的因子做預測。預測分成兩種標的-直接預測分數及預測分數會大於或小於中位數。兩類的預測都嘗試不同的統計分析及機器學習方法。此份研究展現出電影確實可以依據一些特徵，在上映前做到一定程度的評價預測。

Abstract

Nowadays, movies have become everyday life commodities. With the lighting speed of technological evolution, instead of being out of step with society, movies have even turned more and more popular. However; the box-office success does not ensure being highly rated by international film critics. Great movies make people recall again and again and influence the world. On the contrary, bad movies are like fast-disappear trails leaving behind boats, wasting everyone’s time and money. Rating of a piece of art is subjective, there are no objective criteria. Nevertheless; there are still some indexes being regarded as representative criteria. For example, review aggregation websites such as IMDB and Rotten Tomatoes, collecting reviews from worldwide users and film critics both have great credibility and have already provided their services for a long time.

The dataset of this study contains 4,567 movies released from 1916 to 2016 on IMDB. Data visualization first gives us a brief look of factors having stronger impacts on movies’ IMDB rating. In terms of model building and prediction, we only select features which can be controlled before release. Our prediction can be classified into two targets: directly predict IMDB ratings, and predict whether a movie’s IMDB rating falls below or above the median, 6.5. Statistical Analysis and Machine Learning tools are both used in our analyzation. The study shows that, to some extent, we can predict movie’s rating based on several attributes.

關鍵字：電影產業(movie industry)、IMDB 分數(IMDB score)、預測分析(predictive analytics)

1. 簡介

電影可以說是最貼近人類生活的一項產業。我們身邊可能有人沒有聽過古典音樂，但幾乎不可能有人沒有看過電影。電影是最可能影響人的思想行為的產物，因為它無所不在。也因此，打造出感染人心、引人共鳴的好電影至關重要。

電影產業不管在國內國外都是產值及預算金額龐大的產業。好萊塢一年會產出 1000 部以上的電影，產值高達數十億美元以上；據統計，美國與加拿大發行之電影，平均預算為 6500 萬美元。因此，對電影相關工作者而言，一部電影要投入的心血與金錢十分龐大。如果高昂的預算換得的是一陣惡評與嫌棄，從金錢、名聲與時間的角度來看都損失鉅大。

《加勒比海盜 3：世界的盡頭》的預算為 3000 萬美元，評價卻不如預算只需約 116 萬美元的經典電影-《教父》，後者 IMDB 分數高達 9.2 分。1989 年上映之《蝙蝠俠》電影評價及獲利分別為 7.6 分及 2 億 1600 萬美元；去年上映的《蝙蝠俠大戰超人》預算比前者多接近 2 億美元，評價及獲利卻都不如前者，IMDB 分數為 6.9 分，獲利只有 8000 萬美元出頭。由上述可知，高預算不保證高評價，也因此在電影上映前，不一定總是要砸下重本，才能得到好評。

若是在電影上映之前，可以有一套預測電影評價高低的系統，告訴你到底要如何打造出一部好電影。哪些因子與評價息息相關，哪些因素又出乎意料地與評價沒有關係。這樣的模型機制，不僅可以避免把精力與金錢放在錯誤的地方，也可以為世人打造出更多動人心弦的好電影，對製作及閱聽方都是雙贏。

因此，本研究藉由建立電影評價模型，試圖找出關鍵因子，並用模型來預測即將上映的電影之評價高低。我們比較了許多不同的統計分析及機器學習方法，諸如迴歸分析、決策樹、隨機森林、支撐向量機、羅吉斯迴歸、LDA、QDA 等，用預測能力好壞為指標，選出最好的模型，並針對此模型作探討。

需注意的是：雖然資料提供的因素很多，但是我們只選用可以在上映前控制的變數。例如：雖然探討票房與評價的關係有一定的價值，可以讓我們知道高票房是否代表評價相對較高；但票房資訊不能在上映前得知，因此不能被納入預測評價之模型中。

本文章架構如下：第二節介紹如何整理及清理原始資料，第三節透過資料視覺化及敘述統計對資料有初步認識，一窺哪些因素與評價關係較強，第四節則進入資料分析及預測的階段，預測標的分別為直接預測 IMDB 分數及預測會大於或小於分數之中位數-6.5 分，最後則是結論及未來展望。

2. 資料介紹與整理

2.1 資料來源

取自兩個網站，第一是資料抓取者 Chuan Sun 在紐約大學資料科學學術網站上的部落格；第二為 Kaggle-一個數據建模和分析競賽的平台網站，發布的相似資料。原本 Kaggle 提供的是 Chuan Sun 抓取的資料，但因為美國數字千年版權法案(Digital Millennium Copyright Act)，Kaggle 改為提供 The Movie Database(TMDB)上的資料。兩邊的資料，不但電影稍許不同，包含的變數也有相當大的差異。

2.2 資料介紹

IMDB 網站上抓取的原始資料有 5043 筆資料，28 個變數，電影年份橫跨 100 年、66 個國家，有 2399 位導演、超過 3 萬名演員。變數分為與電影直接相關的 14 個變數，以及與參與者直接相關的 14 個變數。

- 與電影直接相關的變數有：

電影名稱、發行年份、電影長度、國家、類別、顏色(彩色或黑白)、螢幕規格、分級、關鍵字、語言、預算、票房、電影 IMDB 鏈結網站、海報人頭數目。

- 與參與者直接相關的變數有：

導演名字、導演 FB 讚數、演員 1,2,3 名字(依讚數多寡抓取)、演員 1,2,3 FB 讚數、卡司 FB 讚數、電影 FB 讚數、IMDB 分數、投票用戶數、評論用戶數、評論影評數。

需注意：這裡之 FB 讚數是 IMDB 網站鏈結的 FB 讚數，不是個人專頁。

Kaggle 網站上的 TMDB 電影網抓取資料有 4798 筆電影資料，22 個變數，2 個資料集。第二個資料集包含完整的演職員表及工作人員表名單。22 個變數中，與 IMDB 原始資料不同的是：

- 新增下列變數：

電影首頁、id、簡介、受歡迎程度、製作公司、製作國家、發行日期、語言、狀態(發行或未發行)、宣傳標語、TMDB 分數、TMDB 投票用戶數、完整演職員及工作人員名單。

- 失去下列變數：

演員 1,2,3 FB 讚數、螢幕規格、卡司 FB 讚數、電影 FB 讚數、顏色、分級、

導演 FB 讚數、海報人頭數目、電影 IMDB 鏈結網站、評論用戶數、評論影評數。

其中有一些變數為 JSON 格式，如製作公司、類別、製作國家、語言、完整演職員及工作人員名單，需做進一步轉換。

2.3 資料整理

2.3.1 資料合併

先將資料集中，完全重複的欄位移除。IMDB 資料集從 5043 筆資料刪到 4998 筆資料；TMDB 資料集從 4802 筆資料刪到 4798 筆資料。

接著，檢查電影名稱一模一樣的資料，IMDB 資料集沒有重複的名稱；TMDB 資料集有 2 部電影名稱有重複出現。經檢查後發現，雖然名稱相同，卻是不同電影，像是有兩部不同的電影都叫做 Out of the Blue，一部在 1980 年上映，另一部則上映於 2006 年。我們在其中一部名稱尾加上 1 來突顯不同。

最後，依照共同電影名稱作為合併依據，我們將兩筆資料集合併，形成一筆有 4567 部電影、47 個變數的資料集。

2.3.2 格式轉換

資料集中，製作公司、關鍵字為 JSON 格式，利用 R 將格式做轉換後，整理成一個欄位，公司與公司之間用逗點表示，關鍵字與關鍵字之間用|表示。

雖然關鍵字在 IMDB 資料集中不是 JSON 格式，無須轉換，但也許 TMDB 資料集中這個相同變數的遺失值較少，或者可以提供 IMDB 關鍵字缺少的欄位資訊，所以我們還是對 TMDB 資料集中的關鍵字變數做轉換。

2.3.3 新增演員變數

TMDB 的第二個資料集為完整演職員及工作人員名單，紀錄方式也是 JSON 格式，目前仍未新增到合併資料集中。我決定把演職員表中，最先出現的三名演員加到資料集中。這三名演員理應是電影前三重要的角色。相對於原始資料前三名的演員是以 FB 讚數多寡做排序，導致小角色的演員可能因人氣高被排在第一；這樣的選取方式更能找到電影的主要演員。現在這筆資料共有 50 個變數。

2.3.4 預算單位轉換，新增利潤、ROI 變數

IMDB 網站上有些電影之預算會隨製作國家不同，而以該國家的貨幣單位表示；但是 Chuan Sun 在爬取資料時並沒有特意做轉換，全部以美元為單位。有人

在運用這筆資料時發現並留言在他的部落格，說這樣的情況特別發生在韓國、土耳其及日本電影中。因此我針對製作國家為上述三個國家之電影預算，做匯率轉換，以免預算過於龐大的離群值出現，但事實上僅為單位的不同。

轉換完預算單位後，我自行新增了兩個變數，:利潤及投資回報率(Return on Investment(ROI))變數，整筆資料目前有 52 個變數。計算公式分別如下：

利潤(Profit)=票房(Gross)-預算(Budget)

ROI(%)=利潤(Profit)/票房(Gross)*100

2.3.5 變數選取

兩筆資料集合併後，有些變數意義是重疊的，例如：預算、票房、電影長度、類別、關鍵字。為了得到最多的資訊，預算、票房、放映長度、類別、關鍵字都選擇 IMDB 資料集中的變數，因為它們相對於 TMDB 資料集中同樣意義的變數，遺失值較少；接著再用 TMDB 資料集中的變數，對 IMDB 資料集遺失的筆數作插補。

電影長度由原本 6 筆遺失值變為 4 筆；預算由原本 361 筆遺失值變為 269 筆；票房由原本 670 筆遺失值變為 465 筆；關鍵字由原本 121 筆遺失值變為 51 筆，類別則沒有遺失值。

導演名稱此變數有 8 筆遺失值、語言有 6 筆遺失值，這部分我自行上網搜尋相關資訊將欄位補齊。

最後，把不會用於視覺化與分析的變數刪除-電影 IMDB 網站鏈結、電影首頁、發行年份(與發行日期重疊，資訊卻更少)、TMDB 資料集之口語(spoken language)、原始語言(original language)及製作國家變數(與 IMDB 網站重疊)、狀態(幾乎所有電影都是上映狀態)、原始電影名稱(與現在之電影名稱幾乎一模一樣)。形成一筆有 4567 部電影，39 個變數之資料集。變數列表如下：

變數類型	變數
屬質型變數（20 個）	電影名稱、id、分級、顏色(彩色或黑白)、語言、國家、類別、螢幕規格、發行日期、導演名字、演員 1,2,3 名字(按照讚數多寡)、演員 1,2,3 名字(按照演職員表出現順序)、製作公司、簡介、關鍵字、宣傳標語
屬量型變數（19 個）	IMDB 分數、TMDB 分數、IMDB 用戶投票數、IMDB 用戶評論數、IMDB 影評評論數、TMDB 投票數、電影長度、預算、票房、利潤、投資回報率、受歡迎程度、海報人頭個數、電影 FB 讚數、演員 1,2,3 FB 讚數(按照讚數多寡)、導演 FB 讚數、卡司 FB 讚數

表 1:變數介紹

3. 資料視覺化及敘述統計探討

3.1 屬量變數

3.1.1 IMDB 與 TMDB 分數

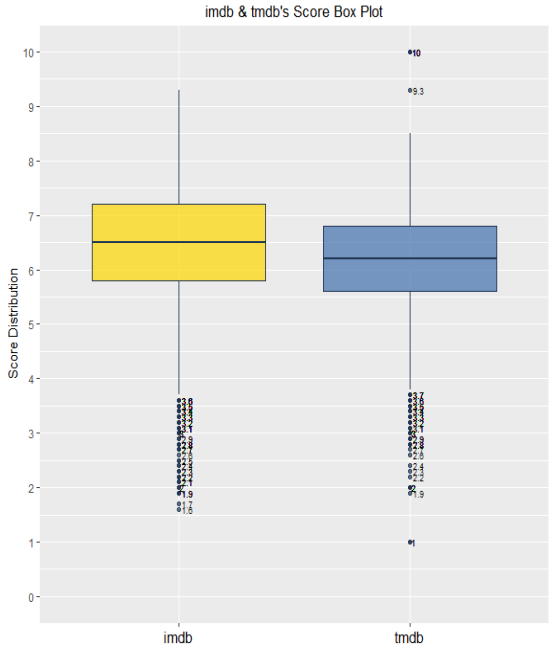


圖 1:IMDB 與 TMDB 分數盒型圖

由兩筆資料集分數盒型圖發現，分數的分布形狀在兩個網站中大致相同，都有些微的左偏分配，中位數稍高於平均。IMDB 分數整體較 TMDB 分數高，中位數為 6.5 分，平均為 6.42 分；對比於 TMDB 的 6.2 及 6.1 分。另外，低分的離群值相較於高分離群值較多，代表相較於特好的電影，似乎更常打造出特糟的電影。

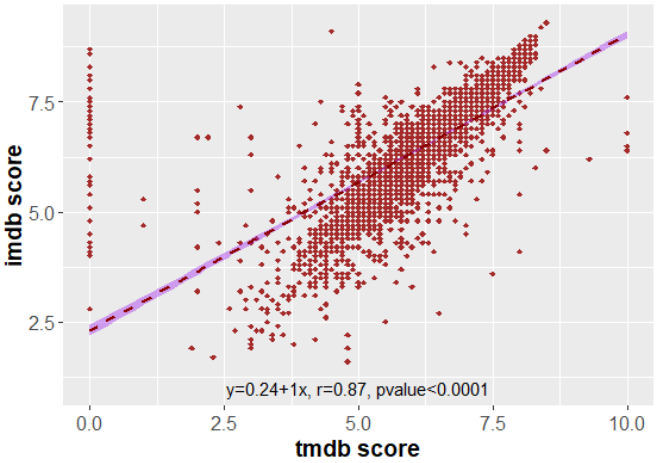


圖 2:IMDB 與 TMDB 分數散布圖

由散布圖可知，兩網站分數呈高度正相關，相關係數 r 高達 0.87，代表同一部電影在不同影評網的評價高低相當一致。因此，我們之後的分析，只選用 IMDB 網站的分數為代表，因為它沒有遺失值，相反的，TMDB 分數有 51 筆遺失值。

3.1.2 投票用戶數與 IMDB 分數

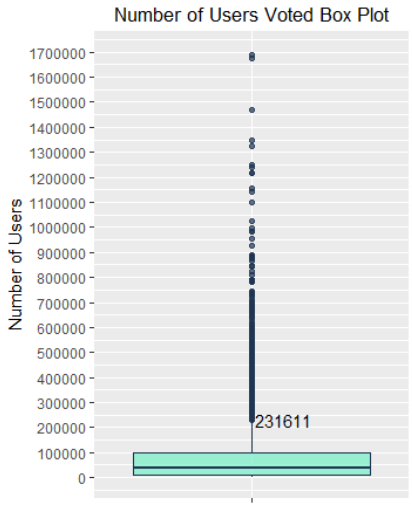


圖 3:用戶投票數盒型圖

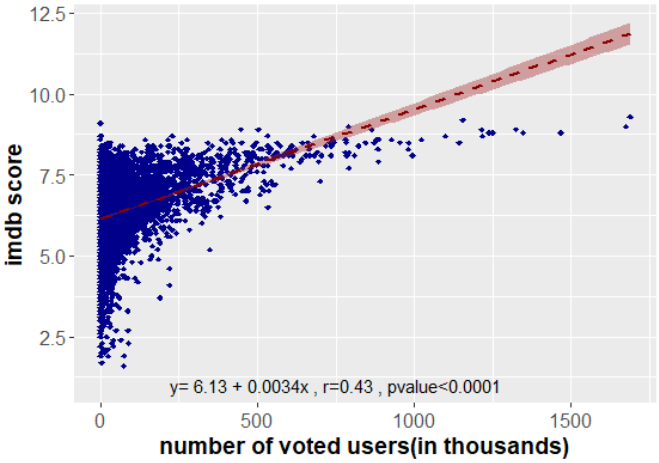


圖 4:用戶投票數與 IMDB 分數散布圖

用戶投票數呈右偏分布，中位數為 35990 票，遠小於平均投票數 85569 票，最高的投票數量為經典電影《刺激 1995》，約為 168 萬，分數 9.3 分，是資料中分數最高的電影，感覺戶投票數似乎與分數有些相關；而真正的相關性可由右邊散布圖得知，用戶投票數與 IMDB 分數相關程度(r=0.44)在所有屬量變數中是最高的，達中度正相關，平均而言，多 100000 個投票用戶，分數可增加 0.34 分。

3.1.3 用戶評論數與 IMDB 分數

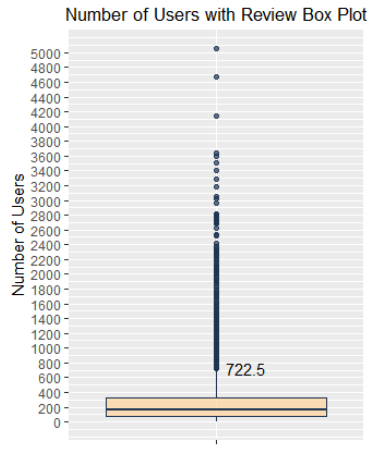


圖 5:用戶評論數盒型圖

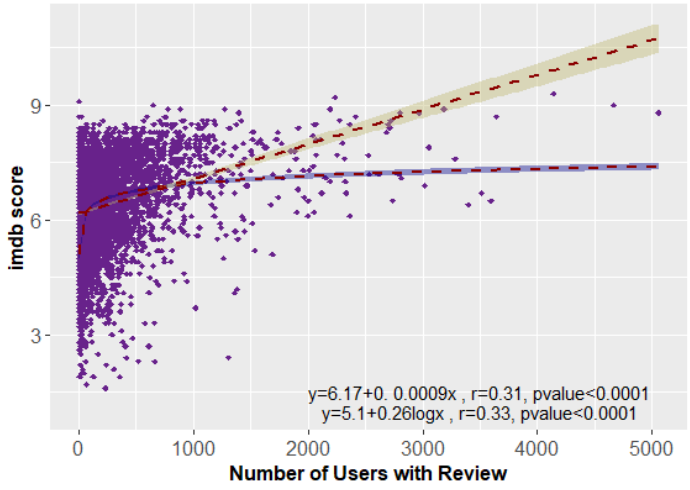


圖 6:用戶評論數與 IMDB 分數散布圖

IMDB 網站上平均用戶評論數為 277.3 篇，中位數為 161.5 篇，有 11 筆遺失值。評論數最高的電影為《魔戒首部曲》，高達 5060 篇！用戶評論數與 IMDB 分數相關性，沒有用戶投票數與 IMDB 分數相關性來得高，相關係數 r 為 0.31。若對用戶評論數取 log 再配適迴歸線，相關係數稍微提升(r=0.33)，但提升不多。

3.1.4 影評評論數與 IMDB 分數

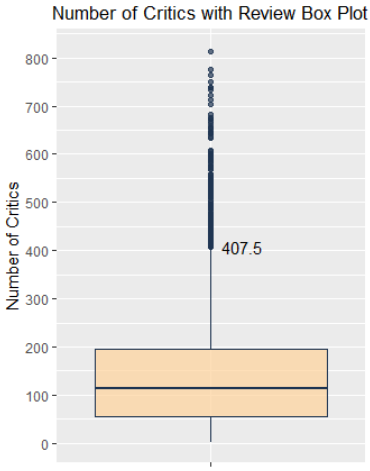


圖 7:影評評論數盒型圖

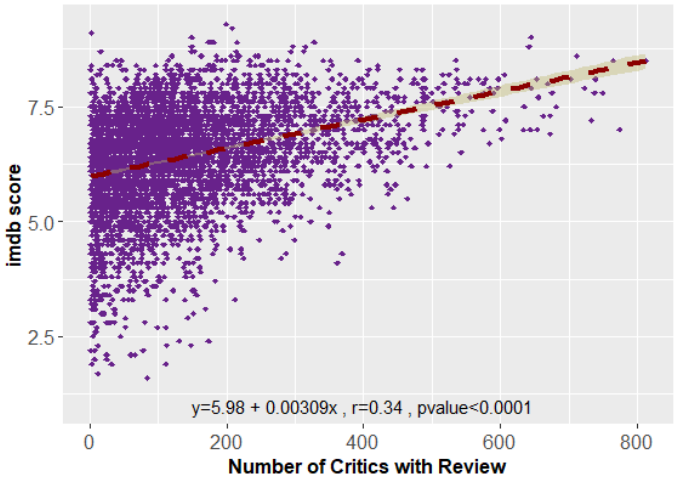


圖 8:影評評論數與 IMDB 分數散布圖

影評評論數與 IMDB 分數的相關程度，相較於用戶評論數與 IMDB 分數之相關性來得高一點，相關係數 r 達 0.34，平均每增加 100 篇，分數約提高 0.31 分。影評平均評論數為 143 篇，中位數 113 篇，有 51 筆遺失值。192 部電影評論數大於 407.5 篇，屬於離群值；評論數最高之電影為《黑暗騎士：黎明崛起》，IMDB 分數 8.5 分，相當得高。

3.1.5 預算與 IMDB 分數

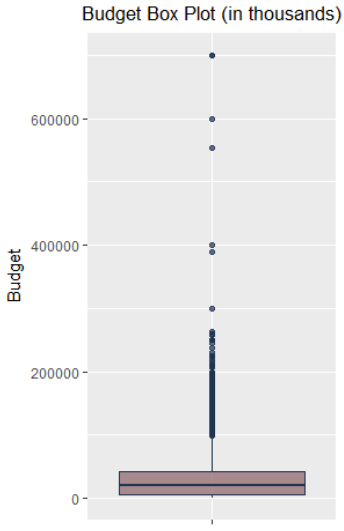


圖 9:預算盒型圖

有 6 部電影的預算明顯的超出其他電影許多，全部都大於 3 億美金，也就是約 90 億台幣，實在很不尋常。所以我把這 6 部電影叫出電影名稱來檢視，看看是否有問題。六部電影列表如下：

電影名稱	國家
《The Messenger: The Story of Joan of Arc 聖女貞德》	美國
《Red Cliff 赤壁》	中國大陸
《Kabhi Alvida Naa Kehna 永不說再見》	印度
《The Legend of Suriyothai 暹羅女王》	泰國
《Kites 寶萊塢之我倆沒有明天》	印度
《Tango》	西班牙

表 2：預算高於 3000 萬美金之電影列表

經檢查後發現，除了第一部電影(金額本身也有誤)，其他五部電影都非美國片，IMDB 上預算都用當地貨幣計算，轉換成美元後金額就會大幅降低。

我們不把所有外國片的預算都轉換，是因為有些外國片的預算仍然是以美元為單位；我曾做過轉換，錯誤的情況比一開始沒有轉換還要嚴重。所以才採取上述方式，對預算過高的片做檢查，確認單位不對後，才做轉換。

轉換後的預算與 IMDB 分數資料視覺化如下：

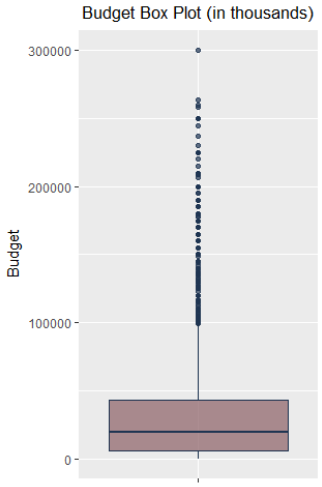


圖 10:調整後預算盒型圖

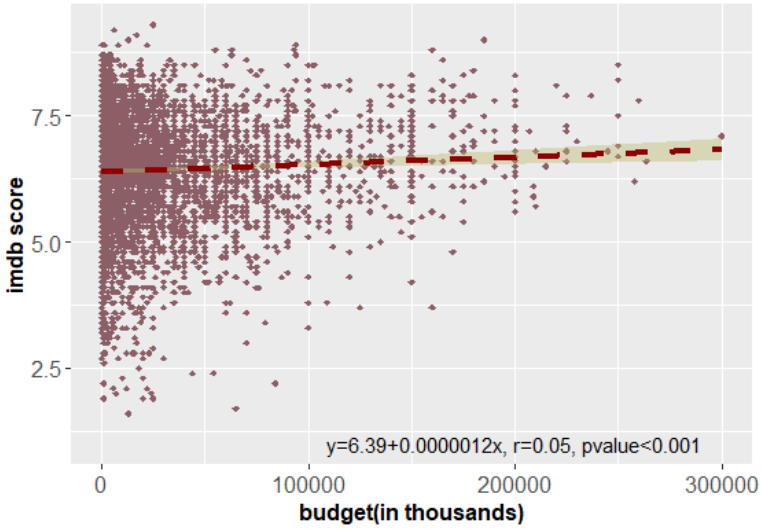


圖 11:調整後預算與 IMDB 分數散布圖

資料中，預算之中位數約為 1980 萬美金，平均約為 3300 萬美金，遺失值有 270 筆。預算最高的電影為《加勒比海盜 3：世界的盡頭》，高達 3 億美金！但其票房卻只比預算多 942 萬美金；IMDB 分數為 7.1 分，屬中間偏上。預算特高的許多電影讓分布傾向右偏，預算特高(大於 9850 萬美金)致屬於離群值的電影有 324 部。從右邊散布圖可發現，預算與 IMDB 分數的相關性很低，相關係數 r=0.05。顯然地，高預算不代表高評價。

3.1.6 票房與 IMDB 分數、預算與票房

票房小於 1000 美金的電影有 6 部，其中有 4 部記錄有誤，分別為《A Farewell to Arms 戰地春夢》、《Out of the Blue 走出憂鬱》、《F.I.S.T.》、《Skin Trade 浴血拳霸》，前兩部沒有記載票房，後兩部票房大於 1000 美金；修正以後，再去作圖形繪製。

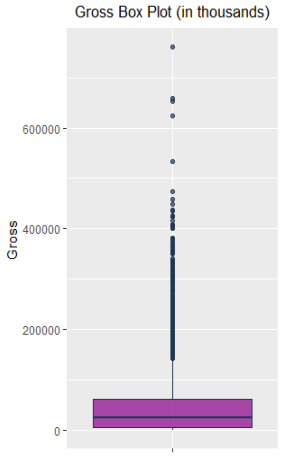


圖 12:票房盒型圖

資料集中，票房中位數約為 2440 萬美金，平均約為 4680 萬美金，有 467 筆遺失值。票房最高的電影為《阿凡達》，光票房收入就有 7 億 6000 多萬美金，IMDB 分數也有 7.9 分。票房最低的電影為 The Jimmy Show，收入只有 703 美金。

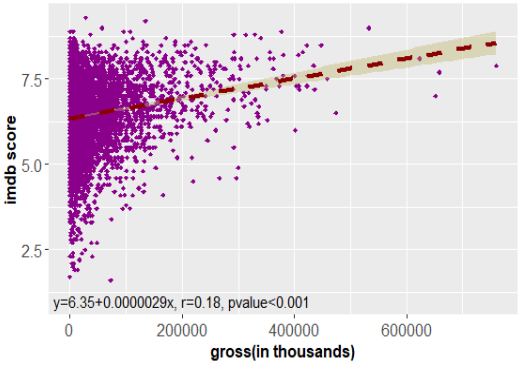


圖 13:票房與 IMDB 分數散布圖

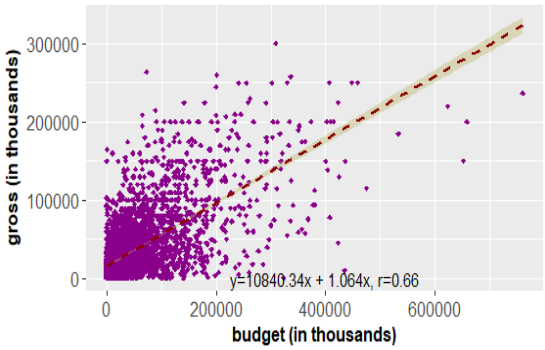


圖 14:預算與票房散布圖

票房與 IMDB 分數的相關程度，較預算與 IMDB 分數之相關程度(r=0.05)高，相關係數 r 為 0.18，但絕對性來看，仍然為低度相關；相對的，預算與票房之相關係數高達 0.66，達到中度相關！

由上面兩張圖之對比可以得到如下推論：預算對票房有一定的影響力，但是預算、票房跟評價的關係不大。

3.1.7 利潤與 IMDB 分數、預算與利潤

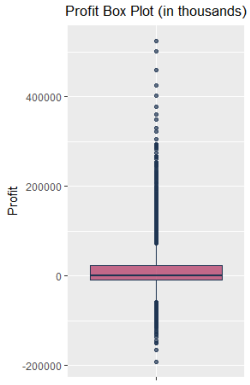


圖 15:利潤盒型圖

4567 部電影中，有 1792 部電影賠錢，5 部不賺不賠，2123 部賺錢，剩下的為 647 個遺失值。賺最多錢的電影與最賣座的電影相同—都是《阿凡達》，傳 5 億多美元；最賠錢的電影為《異星爭霸戰:尊卡特傳奇》，賠 1 億 9000 多萬美元。

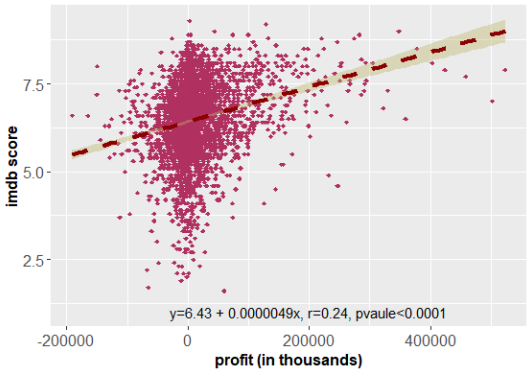


圖 16:利潤與 IMDB 分數散布圖

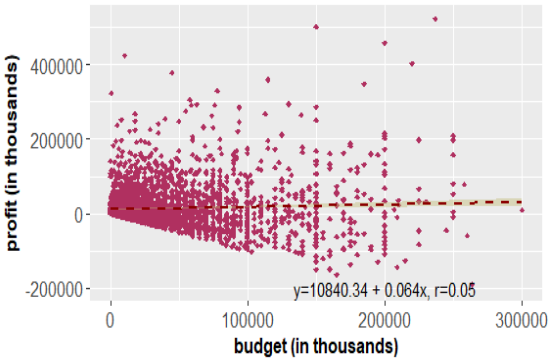


圖 17:預算與利潤散布圖

利潤與 IMDB 分數關係，比票房或預算與 IMDB 分數之關係還要高！相關係數達 r=0.24。預算與利潤的關係，相反地，不比預算跟票房的關係(r=0.66)強，且低很多，相關係數只有 r=0.064。顯示高預算雖然對應到高票房，但跟賺不賺錢幾乎沒甚麼關係，而賺錢程度與評價的關係則相對地高。

3.1.8 ROI 與 IMDB 分數

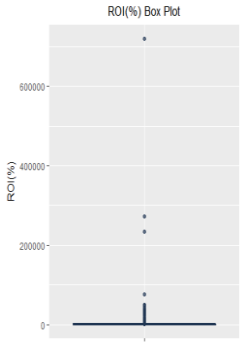


圖 18:投資回報率盒型圖

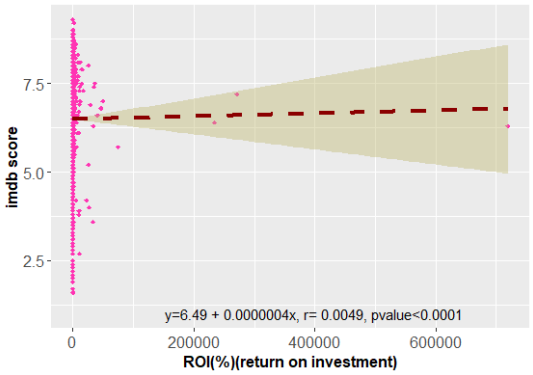
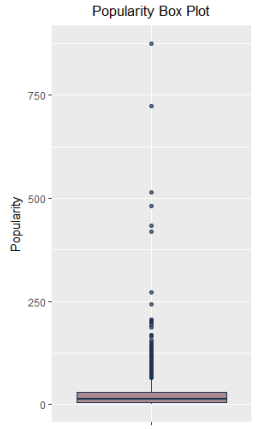


圖 19: 投資回報率與 IMDB 分數散布圖

投資回報率的分配中，極端離群值十分龐大，反映出電影有時會有料想不到的結果。例如，投資回報率最高的電影-《靈動：鬼影實錄》成本只有 15000 美元，卻有 1 億 9300 多萬的票房，投資回報率接近 720000%！恐怖片的低成本與高票房似乎也是這幾年越來越多恐怖片出品的原因。

投資回報率中位數為 12.8%，平均為 630.5%，有 404 筆回報率極高之離群值，為右偏分配，代表多數電影的投資回報率不是太高。它與 IMDB 分數的關係非常低，相關係數 r 只有約 0.005。高報酬率的電影，不代表評價也是高昂的。

3.1.9 受歡迎程度與 IMDB 分數



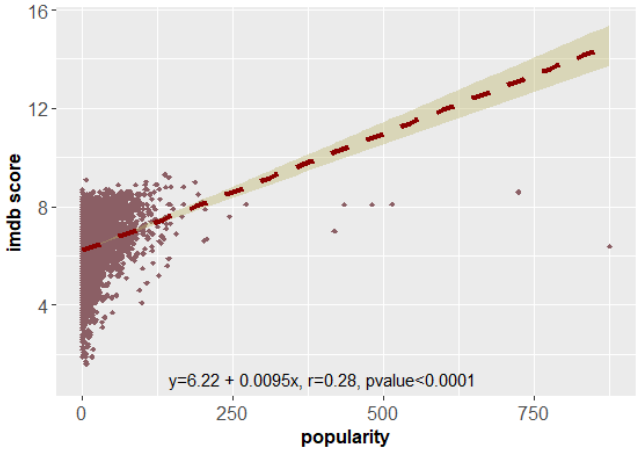
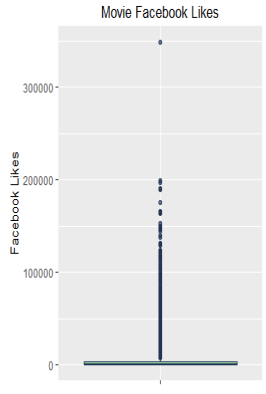


圖 20:受歡迎程度盒型圖　　圖 21: 受歡迎程度與 IMDB 分數散布圖

受歡迎程度(popularity)是 TMDB 的一項變數，計算方式是考慮用戶與 TMDB 網站的互動，包含用戶投票次數、加至「觀看」或「最愛」次數、上映日期等。受歡迎程度在所有屬量變數中，與 IMDB 分數相關程度為第 5 高，相關係數 r 為 0.28。最受歡迎的電影是動畫《小小兵》。

3.1.10 電影 fb 讚數與 IMDB 分數



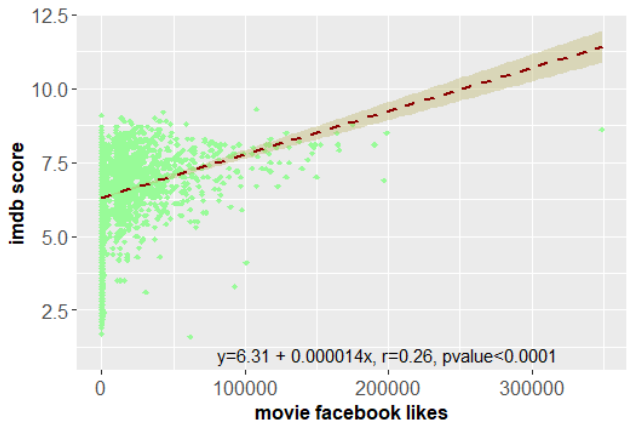


圖 22:電影 FB 讚數盒型圖　　圖 23: 電影 FB 讚數與 IMDB 分數散布圖

電影 FB 讚數有許多極端離群值，為右偏分配，讚數少的電影偏多；平均讚數為 7603，遠高於中位數 168。電影 FB 讚數與 IMDB 分數相關係數 r=0.26，在所有屬量變數中偏高。讚數最高的電影為 2014 年上映的《星際效應》，IMDB 達 8.6 分。在這邊需注意，所有提及之 FB 讚數指的是 IMDB 網站鏈結上的按讚次數。

3.1.11 卡司 FB 讚數與 IMDB 分數

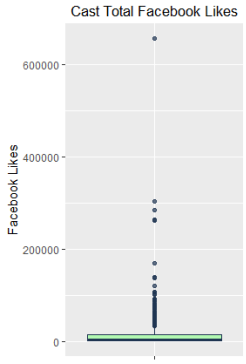


圖 24:卡司 FB 讚數盒型圖

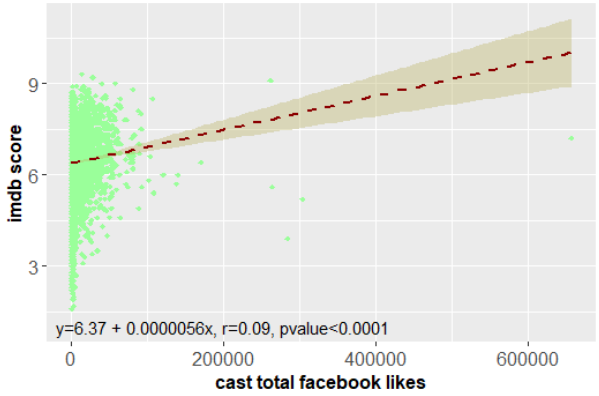


圖 25: 卡司 FB 讚數

卡司 FB 讚數平均為 9962 個讚，中位數為 3163 讚，讚數最多的電影是一部喜劇片-威爾法洛主演的《銀幕大角頭》，有 65 萬多讚。卡司 FB 讚數與 IMDB 分數相關性非常低，相關係數 r 只有 0.09。

3.1.12 演員 1,2,3 與 IMDB 分數

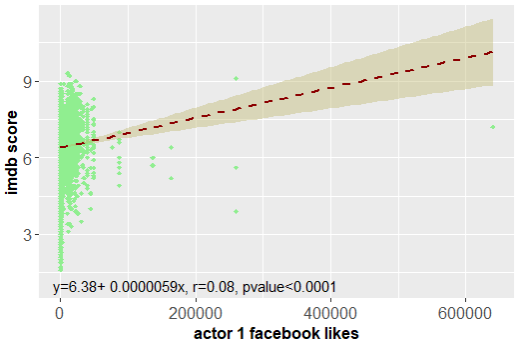


圖 26:演員 1 與 IMDB 分數散布圖

圖 27: 演員 2 與 IMDB 分數散布圖

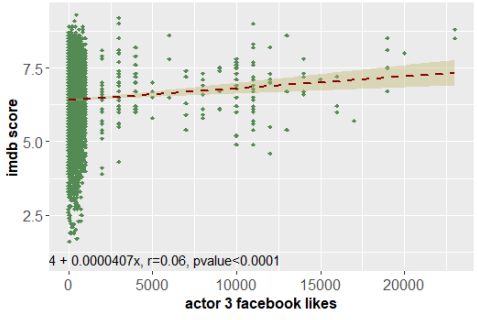


圖 28: 演員 3 與 IMDB 分數散布圖

演員 1,2,3 的讚數與 IMDB 分數的相關性都很低，相關係數 r 分別為 0.08、0.09 與 0.06，遺失值分別為 3、6、12 筆。演員 1 中，讚數最多的演員為 Darcy Donavan，不是特別知名的演員。她有參與《銀幕大角頭》的演出，也解釋了為何這部片卡司FB讚數如此多的原因。FB讚數最多的演員 2 為 Andrew Fiscella，也是位名不見經傳之演員。演員 3 FB 讚數最高的演員為喬瑟夫戈登李維 Joseph Leonard Gordon-Levitt，相對於前兩名較不知名的演員，他算是揚名國際的演員，獲得許多獎項的肯定。

3.1.13 電影長度與 IMDB 分數

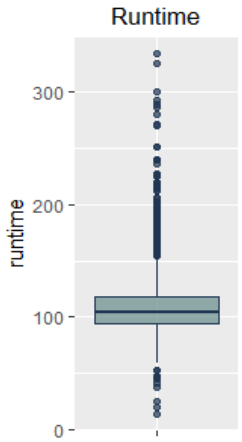


圖 29:電影長度盒型圖

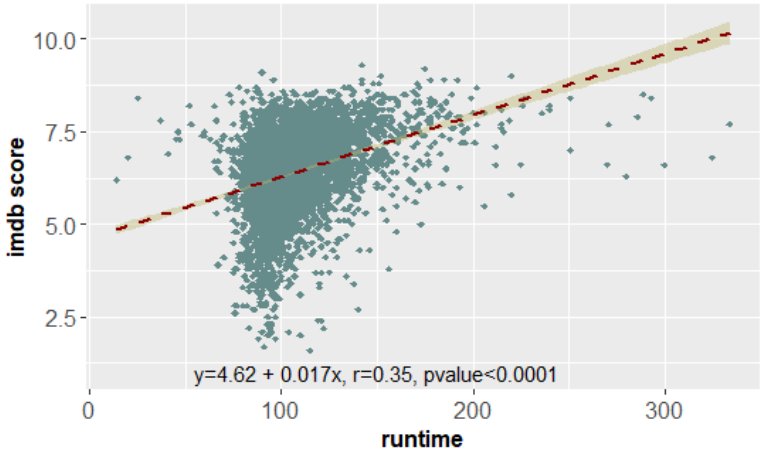
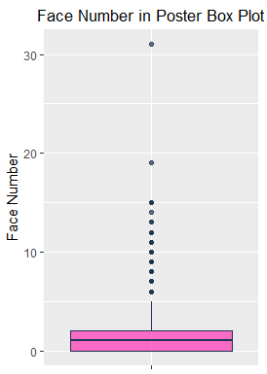


圖 30: 電影長度與 IMDB 分數散布圖

電影長度平均約為 108 分鐘，多 10 分鐘評價平均會上升 0.17 分。我們意外的發現，電影長度跟 IMDB 分數相關程度是所有屬量變數中第 2 高，相關係數 r 達 0.35，僅次於用戶評論數。但是後者並不能在上映前得知，所以後續預測分析不會採用，代表電影長度應是關鍵預測因子。

3.1.14 海報人頭個數與 IMDB 分數



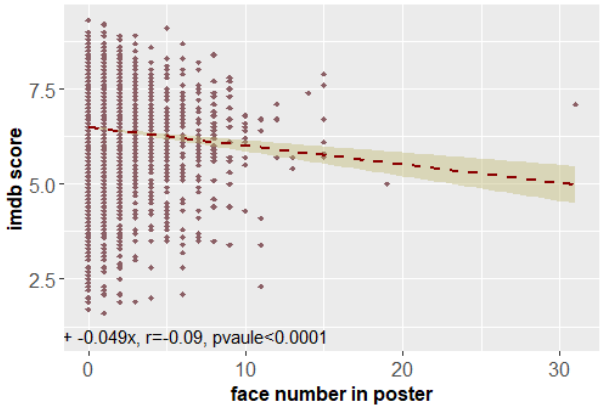
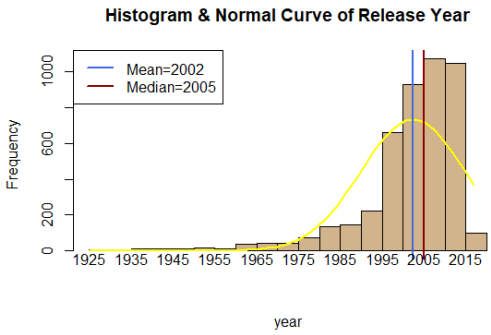


圖 31:海報人頭個數盒形圖　　圖 32: 海報人頭個數與 IMDB 分數散布圖

海報人頭個數與 IMDB 分數關係不大，相關係數 r 為-0.09，有 11 筆遺失值。值得注意的是-相較於其他屬量變數與 IMDB 的正向關係，海報人頭數目越多，評價有下降的趨勢，雖然關係不大，仍然值得注意。

3.1.15 上映年份與 IMDB 分數



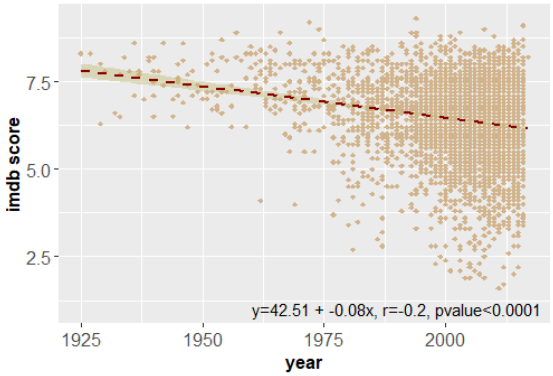
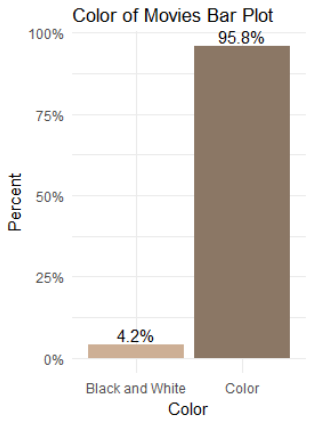


圖 33: 上映年份直方圖　　圖 34: 上映年份與 IMDB 分數散布圖

多數電影上映年份在 2000 年後，2000 年前上映之電影有 1259 部，2000 年後(包含 2000 年)上映之電影則有 3308 部，占了近 4 分之 3。上映年份與評價的相關性是第二個呈負相關之屬量變數；年份增加一年，IMDB 分數平均下降 0.08 分。可能與現今電影數量越來越多，品質參差不齊，從而平均水準下降有關。

3.2 屬質變數

3.2.1 電影顏色與 IMDB 分數



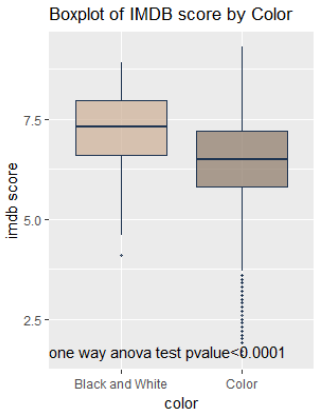


圖 35:電影顏色柱狀圖　　圖 36: 電影顏色與 IMDB 分數盒型圖

多數資料集中電影是彩色的，只有 191 部黑白電影，有 12 筆遺失值。黑白電影分數分布明顯高於彩色電影，不過因為數量差異太大，不一定有代表性。

3.2.2 電影語言與 IMDB 分數

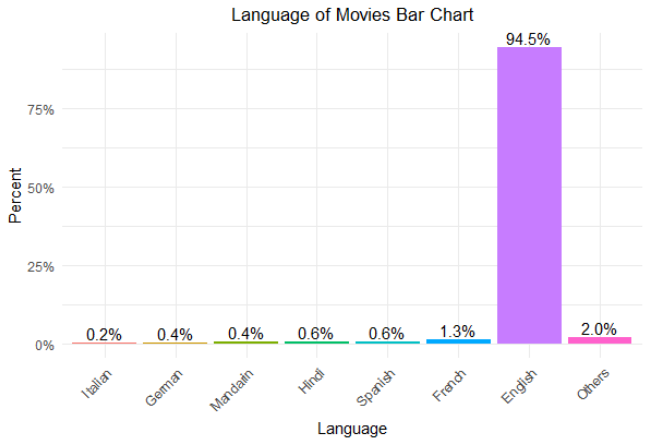


圖 37: 電影語言柱狀圖

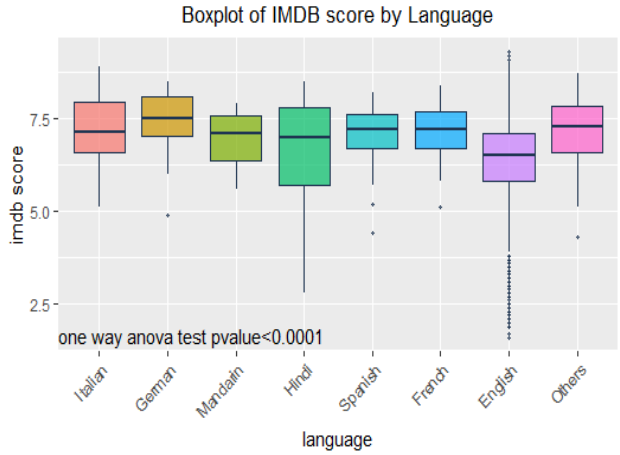


圖 38: 電影語言與 IMDB 分數盒型圖

接近 95%之電影為英語。其他(Others)包含的是在資料集中電影數量少於 10 部之語言，例如丹麥語、泰語等。總共有 7 個主要語言。

從語言與 IMDB 分數之盒形圖可以稍微推論：德語片分數分布不但較集中也較高；但是，因為德語片只有 12 部，相較於英語片的 4317 部，一同拿來比較似乎不太公平；但是與英語片之外的電影相比，的確分數也相對地高。

3.2.3 電影國家與 IMDB 分數

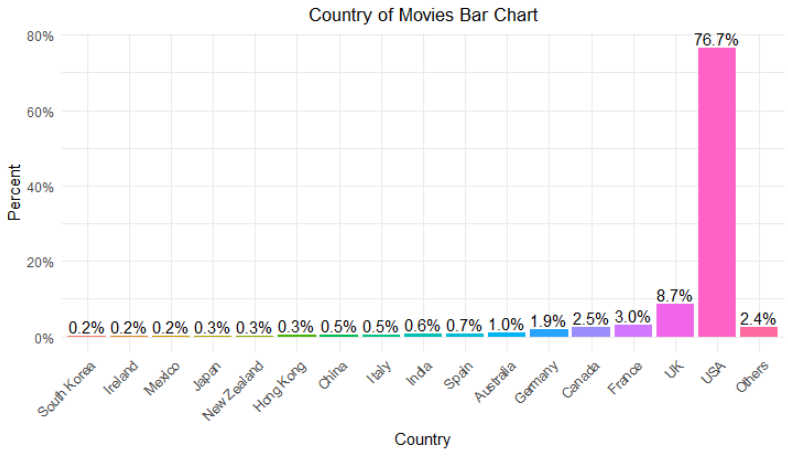


圖 39：電影國家柱狀圖

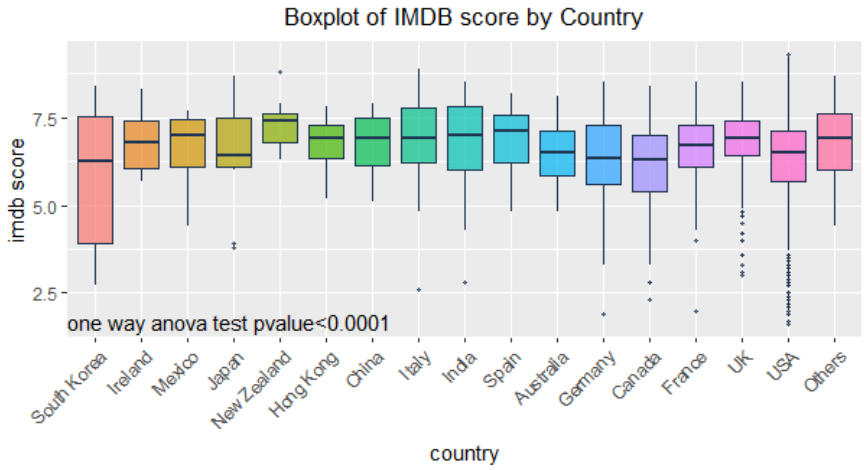


圖 40:電影國家與 IMDB 分數盒型圖

超過 4 分之 3 的電影所屬國家都為美國。其他這類合併了資料集中電影數量小於 5 部之國家，包含捷克、芬蘭等國家。總共有 16 個主要國家。

從國家與 IMDB 分數之盒型圖可以稍微推論：紐西蘭與西班牙之電影評價普遍較其他國家高，韓國的電影評價分散性很大。但是，因為數量的極端差距(紐西蘭 13 部，西班牙 30 部，英國 396 部，美國 3502 部)，所以不一定有代表性。

3.2.4 螢幕長寬比與 IMDB 分數

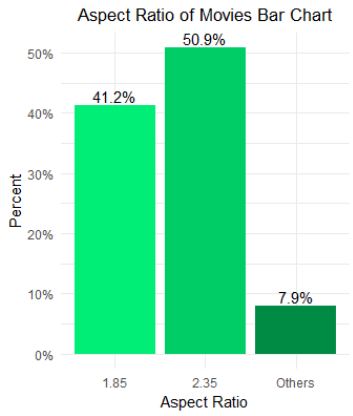


圖 41: 螢幕長寬比柱狀圖

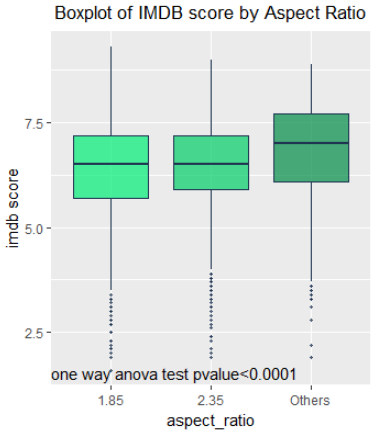


圖 42:螢幕長寬比與 IMDB 分數盒型圖

最常見的 2.35 與 1.85 佔了資料集中螢幕長寬比的 9 成以上，有 263 筆遺失值。從螢幕長寬比與電影分數的盒型圖可以得知，1.85 與 2.35 兩種規格分數差異不大，但 1.85 長寬比的電影評價相對更分散一些。

3.2.5 電影分級與 IMDB 分數

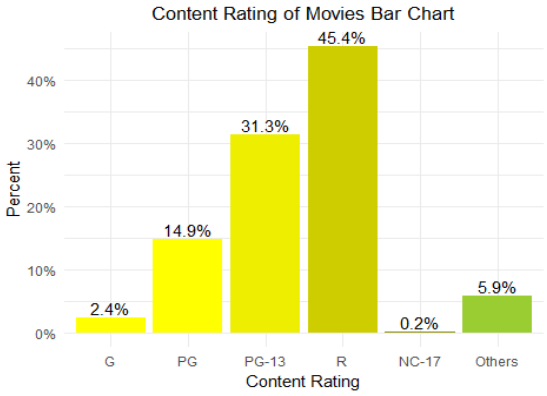


圖 43: 電影分級柱狀圖

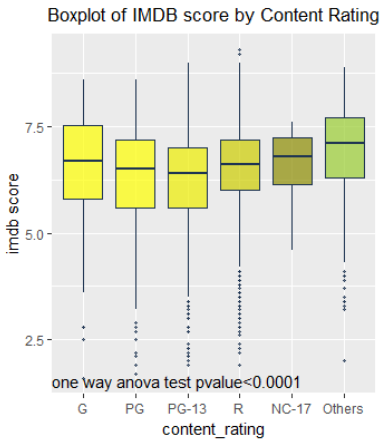


圖 44:電影分級與 IMDB 分數盒型圖

美國電影普遍分成 5 級，柱狀圖中，類別從左到右代表限制程度依序遞增。資料集中 R 級電影(限制級電影，17 歲以上才能觀賞)最多，其次是 PG-13 級(13 歲以下不能看，13-17 歲需家長陪同)。

從與 IMDB 分數盒型圖可以得知，R 級電影較 PG-13 級的電影評價高；雖然 G 級及 NC-17 級電影評價之中位數較高，但因為它們數量少，與數量多的其他分級相比，不一定準確或有代表性。

3.2.6 上映世紀與 IMDB 分數

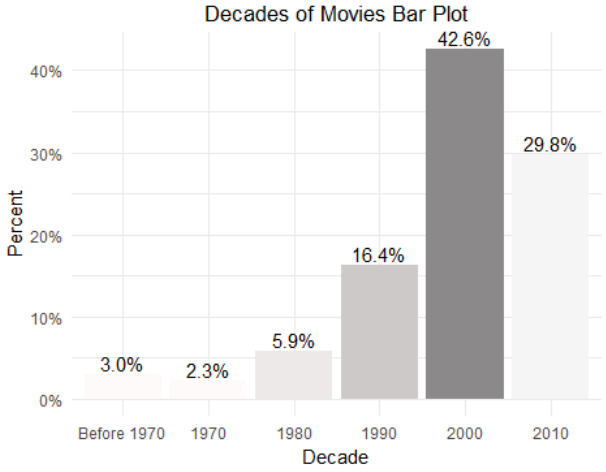


圖 45: 上映世紀柱狀圖

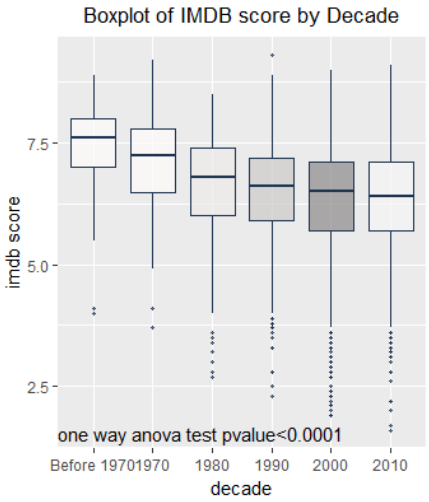


圖 46:上映世紀與 IMDB 分數盒形圖

這筆資料中，2000-2010 年的電影數量最多，其次是 2010-2016 年的電影。上映世紀與 IMDB 評價的盒型圖中，明顯發現若以 10 年為單位觀察，評價有明顯下降趨勢。雖說2000 年前以 10 年為單位之電影數量遠不如 2000-2010 年此區間，但 2010 年之後的電影數量雖然較 2000-2010 年少，評價中位數及整體分配還是較 2000-2010 年低，代表電影的品質似乎真的有隨時間下滑趨勢。

3.2.7 上映月份與 IMDB 分數

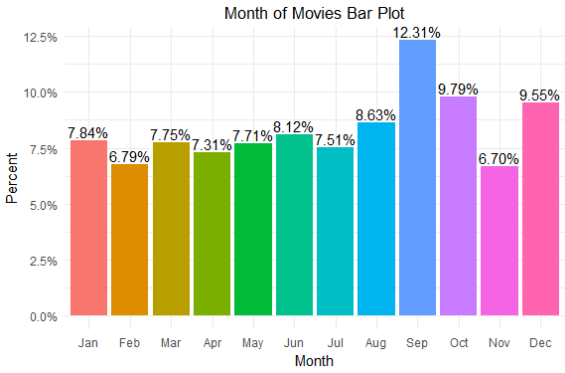


圖 47: 上映月份柱狀圖

資料集中，9 月份上映的電影最多，其次是 10 月及 12 月。各個月分上映的電影數量雖然有差異，但不致於差異太大，整體而言滿平均的。

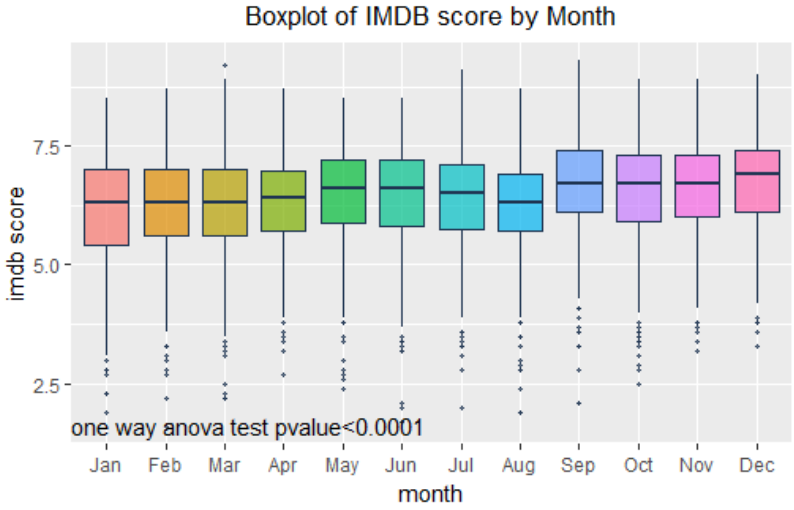


圖 48:上映月份與 IMDB 分數盒型圖

9 月及 12 月的 IMDB 分數分布相對較高，又以 12 月的分數中位數較高。這兩個月之電影數量較多，分數仍然高，可見並不會因數量多就拉低品質。IMDB 分數分布最低的月份為 1 月和 8 月，又以 8 月之分數中位數最低、1 月 IMDB 分數分散度最大。

3.2.8 電影類型與 IMDB 分數

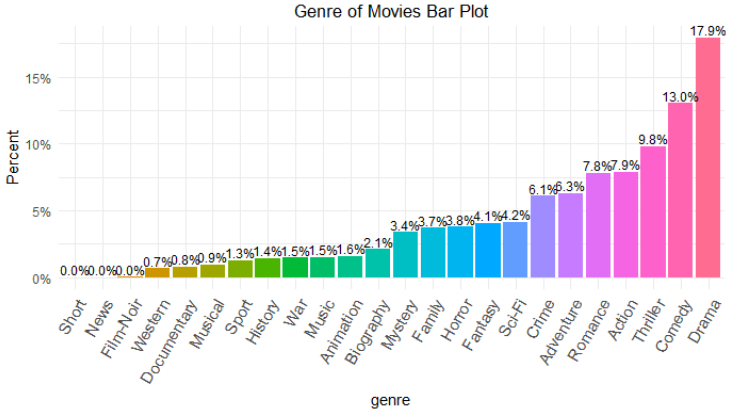


圖 49: 電影類型柱狀圖

一部電影不只屬於一個類型，多數可以同時被歸到好幾類。例如：愛情喜劇片為非常常見的類型，就同時囊括了愛情及喜劇兩類。資料集中，一部電影最多同時屬於 8 個類型。

4567 部電影中，總共有 24 個不同之電影類型。出現次數最高的為劇情片，其次為喜劇片及驚悚片。戰爭片、歷史片、歌舞劇、動畫片出現次數較少。

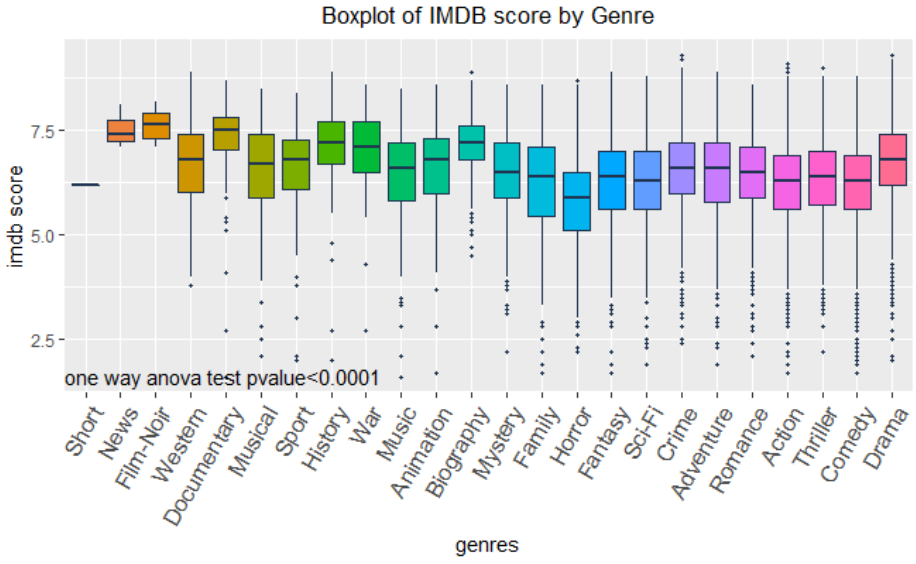


圖 50:電影類型與 IMDB 分數盒型圖

IMDB 分數分布最高的類型依序為：黑色電影(Film Noir，多指好萊塢偵探片)、紀錄片、新聞片、歷史片、戰爭片及劇情片。前五類型片數都在 200 部以下，劇情片則在有 2361 部的情況下，評價中位數及分配仍偏高，很不容易。IMDB 分數中位數最低的類型為恐怖片及科幻片，這兩類電影似乎比較難以獲得高評。

3.3 續集電影趨勢分析

有續集的電影總共有 68 部，其中有第 1、2 集的電影有 67 部；有第 2、3 集的電影有 21 部；有第 1、3 集的電影有 16 部。因此我們便以這些電影為對象，兩兩比較，繪製出 IMDB 分數盒型圖，看看續集電影分數是否有顯著下降。

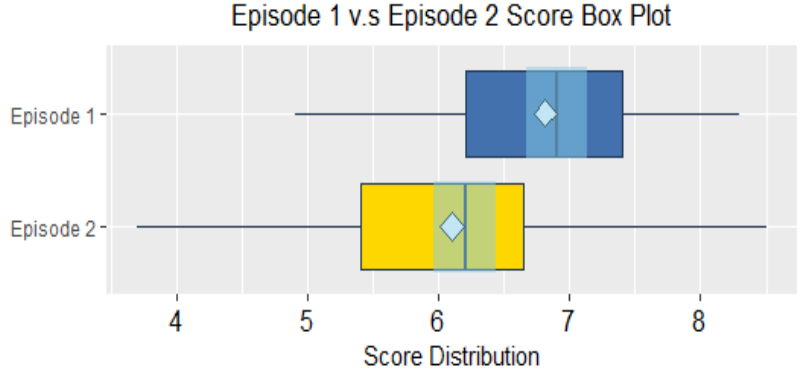


圖 51:第 1 集、第 2 集 IMDB 分數盒型圖比較

第 1 集電影 IMDB 分數分布位置明顯比第 2 集高，第一集平均為 6.81 分，第二集平均只有 6.09 分。第 1 集的分數之分散程度也比第二集低，標準差分別為 0.85 與 1。為了檢驗首部曲分數是否顯著高於二部曲，我們採用
統計檢定中，兩相依母體差的期望值μ_D之假說檢定(非常態且大樣本(n>30))。：

1. Ho: μ_D=0 Ha: μ_D >0
2. α=0.05
3. Rejection Region: z>1.645
4. Test Statistic: z= $\frac{(6.81-6.09)}{0.69/\sqrt{69}}$=8.667781
5. Conclusion: z>1.96，第 1 集平均分數顯著高於第二集平均分數。

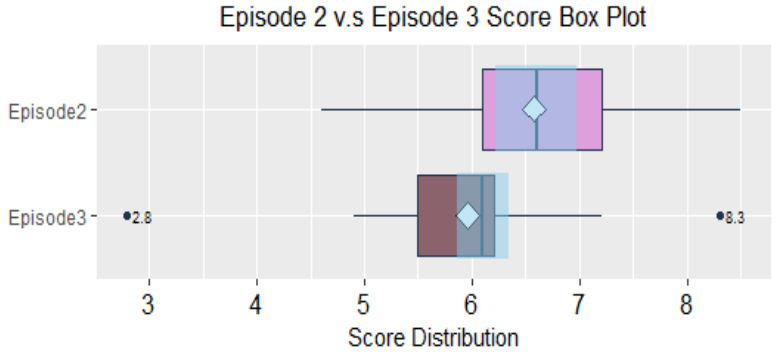


圖 52:第 2 集、第 3 集 IMDB 分數盒型圖比較

第 2 集電影分數分布更是明顯比第 3 集高，第 2 集分數之第 1 四分位數甚至接近第 3 集之中位數。中位數為 6.6 分對 6.1 分，平均為 6.57 分對 5.96 分。

為了檢定續集分數是否顯著降低，又樣本數只有 27 筆，所以我們採用無母數統計之 Wilcoxon 符號等級檢定(Wilcoxon signed-rank test)。

1. Ho:Mx=My Ha:Mx>My
2. α=0.05
3. Rejection Region: z>1.645 (n>15，大樣本近似)
4. Test Statistic: W=192, z=$\frac{(192-20(20+1)/4)}{\sqrt{20(20+1)(40+1)/24}}$=3.24
5. Conclusion: z>1.96，第 2 集分數中位數顯著高於第 3 集分數中位數。

圖 53:第 1 集、第 3 集 IMDB 分數盒型圖比較

第 1 集和第 3 集的分數分布差異最為明顯，第 1 集的最小值甚至大於第 3 集之中位數。平均分別為 7.375 分及 6.068 分，中位數分別為 7.45 及 6.1 分，兩者都相差甚遠，超過 1 分以上。這邊我們就不再作中位數是否顯著降低之無母數檢定，既然第 1 集與第 3 集的差異更明顯，又前面第 2 集與第 3 集已確認有顯著差異，因此，相信這邊的檢定結果也會發現，第 3 集分數顯著低於第 1 集分數。

3.4 按照 fb 讚數選出演員與演職員表字幕出現順序演員之同異數量

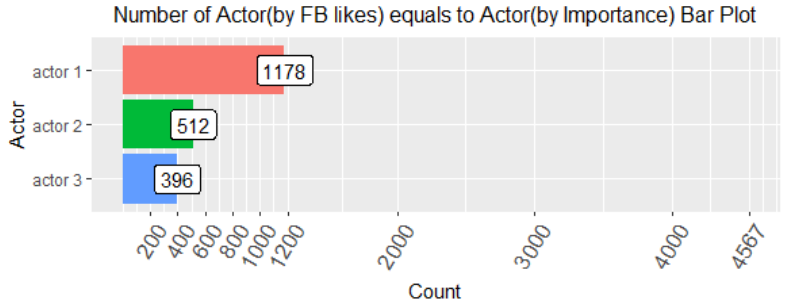


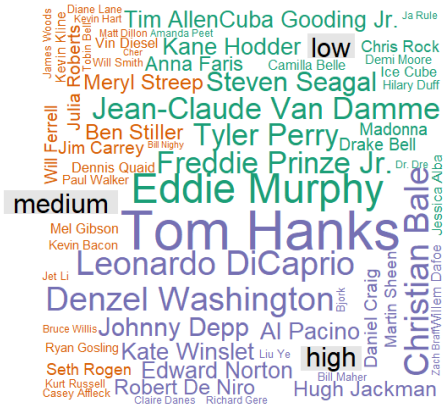
圖 54: 按照 fb 讚數選出演員與演職員表字幕出現順序演員同異數量柱狀圖

可以發現兩種方式(按照 IMDB 鏈結按讚數多寡排序、演職員表出現順序)找到的前三名演員，十分不同。演員 1 只有 25%相同，到了演員 2，就只剩約 10%，演員 3 就更少相同演員了。兩個方式找到的相同演員 3 之數量只有 396 位。

用演職員表出現順序找出的前 3 名演員，較有可能是該電影之主要演員、影壇界的巨星，他們貫穿整部電影，通常也是電影的宣傳重點。所以下個項目的文字雲分析，會以演職員表順序找到的前 3 個演員為畫圖的資料來源。

3.5 導演、演員 1,2,3、製作公司之文字雲

以 IMDB 5.5 分及 7.5 分為界線，將分數分為低、中、高三級。左邊的一律為比較文字雲(comparison wordcloud)，它呈現不同分數分級中，哪些人、公司較常出現，相對頻率較高。右邊為共通文字雲(commonality wordcloud)，它呈現哪些人、公司在不同分數分級中都有出現。文字越大，代表它出現頻率越高。



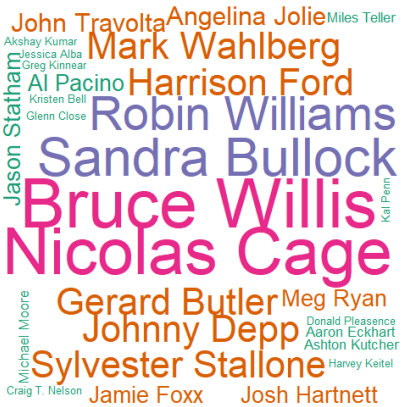


圖 55：演員 1 之比較文字雲　　圖 56：演員 1 之共通文字雲

演員 1 中，湯姆・漢克斯、李奧納多狄卡皮歐、丹佐華盛頓最常演出高分電影；艾迪·墨菲、泰勒・佩瑞，則最常演出低分電影。這份資料中，梅莉・史翠普、班史提勒、金凱瑞較常演出中分區段之電影。

布魯斯・威利、尼可拉斯・凱吉及珊卓・布拉克則在高、中、低分之電影都有演出。

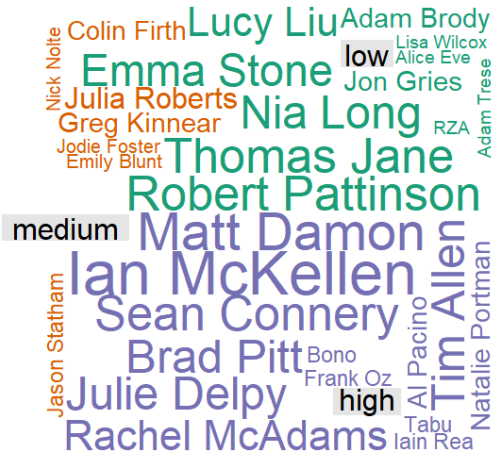


圖 57：演員 2 之比較文字雲　　圖 58：演員 2 之共通文字雲

演員 2 中，麥特・戴蒙、演出《魔戒》中甘道夫之知名演員-伊恩·麥克連及布萊德彼特是最常演出高分電影的主要演員；《暮光之城》之主演羅伯・派汀森則最常演出低分電影。茱莉亞・羅伯茲較常演出中分區段之電影。

班・金斯利、蓋瑞・歐德曼、丹尼斯・奎格及卡麥蓉・狄亞茲則頻繁地出現在高、中、低分之電影。

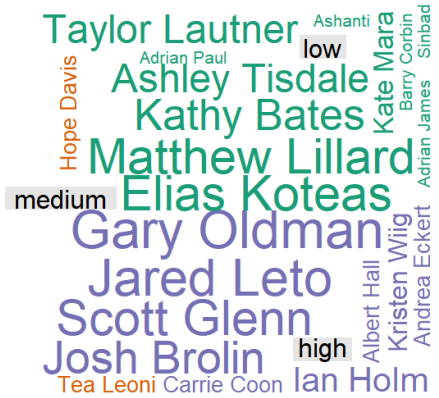


圖 59：演員 3 之比較文字雲　　圖 60：演員 3 之共通文字雲

演員 3 中，蓋瑞・歐德曼、傑瑞德・雷托最常演出高分電影；《暮光之城》之演員泰勒・洛納、馬修李勞得則最常演出低分電影。

奧斯卡最佳男主角得主-資深演員勞勃・杜瓦、神鬼奇航中的反派-比爾·奈伊在高、中、低分之電影都有演出。

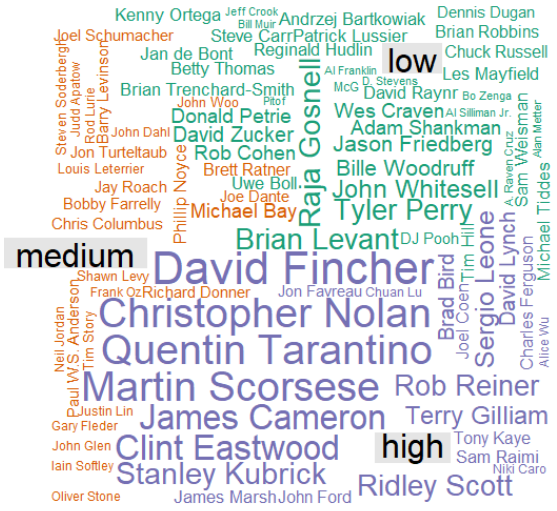


圖 61: 導演之比較文字雲

圖 62: 導演之共通文字雲

《社群網站》、《控制》之著名導演大衛・芬奇、《全面啟動》、《星際效應》之 21 世紀最受好評導演之一之克里斯多福・諾蘭、《黑色追緝令》、《追殺比爾》之擅長新黑色電影導演-昆汀·塔倫提諾較常導出高分電影。《變形金剛》導演麥可・貝的導演電影則多落於中分區段；泰勒·佩瑞、《絕地奶霸》之導演雷賈·瑞蒙·高斯奈較常導出低分電影。

獲奧斯卡終身成就獎之導演-史派克·李的作品，在高中低分都有頻繁出現。

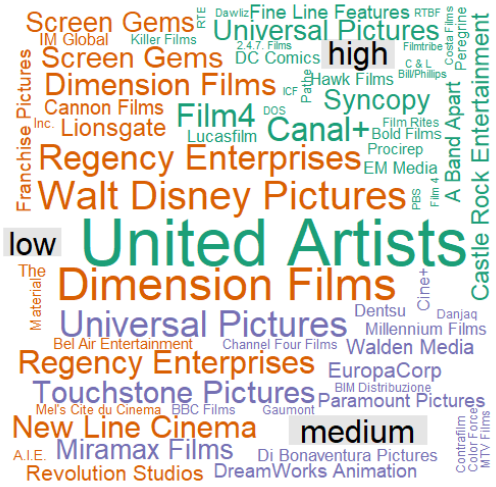


圖 63: 製片公司之比較文字雲

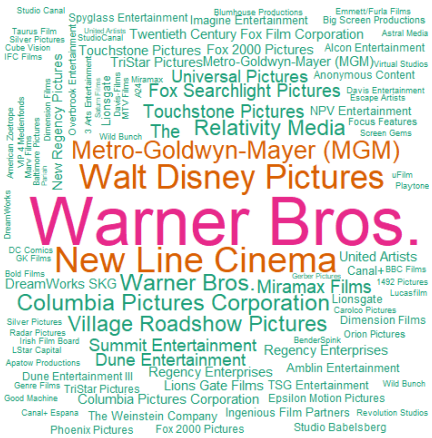


圖 64: 製片公司之共通文字雲

最常製作出高分電影之公司為聯藝電影公司（United Artists），為米高梅旗下子公司，曾製作出奧斯卡得獎片-《亂世佳人》及 007 系列電影。環球影業-美國最大的製片公司之一，製作較多中段分數之電影。溫斯坦公司(鮑柏及哈維溫斯坦兄弟創立)旗下的次元影業(Dimension Films)則製作出較多低分電影。

華納兄弟及它的子公司新線電影，以及華特迪士尼旗下之電影公司在高、中、低分數區間都有相對頻繁的製作作品。

4. 模型建立與預測

4.1 變數選取

我們只利用可以在上映前得知的特徵來建立模型及預測。例如：雖然用戶投票數與 IMDB 分數之相關性是所有屬量變數中最高的，但因為上映前無法得知，所以不予採用。

我們選取出 39 個變數之中的 16 個變數,1 個是反應變數,15 個為解釋變數。

在這 15 個變數之中，國家與語言這兩個因子的相關性非常高，Cramer’s V statistic(類別變數之相關係數)高達 0.63。主要語言有 7 個，少於主要國家的數量 16 個，在進行多元線性迴歸及多元羅吉斯迴歸時，選用國家此變數需要創造出更多虛擬變數，喪失更多自由度，又它們相關性極高，所以可以推測喪失的自由度不能解釋更多模型的變異，卻會造成模型的檢定力下降。所以我們最終選擇語言變數，刪除國家變數。

我們將發行日期分為兩個變數，年份與月份。所以我們目前的解釋變數共有 15 個，反應變數 1 個-IMDB 分數。

電影類型方面，一部電影同時有好幾種類型，為了方便分析，我們選出列出的第一類型，也就是主要類型來做分析。因此分析時，一部電影只會有一項主要類型。

最後，我們從之前資料視覺化中，觀察到續集電影分數明顯下降，所以在分析前新加了一個變數-是否為續集。此變數有 3 個水準-不是續集、是第一部續集(二部曲)、是第二部續集或更多(三部曲以上)。目前總共有 16 個自變數。

變數列表如下：

變數類型	分析變數
屬質型變數(8 個)	分級、顏色(彩色或黑白)、語言、類別、螢幕規格、上映年份、上映月份、是否為續集
屬量型變數(8 個)	電影長度、預算、海報人頭個數、演員 1 FB 讚數(按照讚數多寡)、演員 2 FB 讚數(按照讚數多寡)、演員 3 FB 讚數(按照讚數多寡)、導演 FB 讚數、卡司 FB 讚數

表 3:分析變數介紹

4.2 遺失值處理

變數	分級	顏色	語言	類別	規格	年份	月份	長度	預算
遺失數	204	12	0	0	263	0	0	6	270
變數	海報人頭數	演員 1 讚數	演員 2 讚數	演員 3 讚數	導演讚數	卡司讚數	續集與否	表 4: 遺失值個數列表	
遺失數	11	3	6	12	8	0	0		

16 個自變數中，年份、月份、語言、類別、卡司讚數、續集與否沒有遺失值，剩下的變數有 6 到 270 個遺失值，總遺失值有 795 個。

為後續分析，我們插補所有的遺失值，選用 R 的 mice 套件。插補原理如下：

要插補 X1 變數，就把 X1 當應變數，剩下變數當自變數建模，用模型預測該變數之值當作插補值。我選用的模型為決策樹，因為它沒有限制反應變數型態。

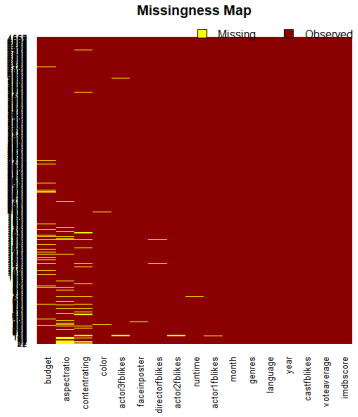


圖 65: 插補前遺失值分布

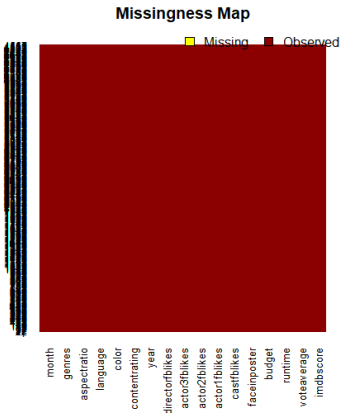


圖 66: 插補後遺失值分布

4.3 相關性分析及主成分轉換

插補完，需考慮變數之間的相關性，若相關性很大，在迴歸分析中會有很嚴重的共線性問題，造成估計不準確。因此，先繪製 15 個自變數間的相關係數圖。

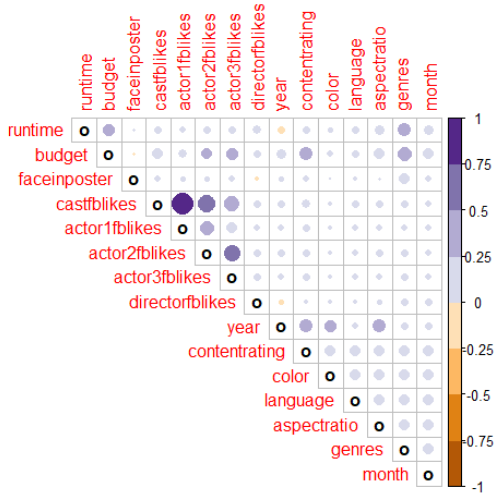


圖 67: 15 個自變數相關係數圖

類別變數間採 Cramer’ s V Statistic、類別與屬量變數間採實驗設計檢定之 R square 開根號、屬量變數間則是最常見的 Pearson Correlation Coefficient。可以發現，卡司 FB 讚數及演員 1,2,3 FB 讚數彼此之間相關性非常大，相關係數全部都介於 0.5 到 1 之間，可能會對之後的迴歸分析造成影響。

因此，我決定利用相關係數矩陣 R 做主成分分析，將這 4 個變數用較少主成分解釋。這樣做的好處不僅可以讓新變數之間彼此獨立，也可以盡量解釋 4 個變數的最大變異。

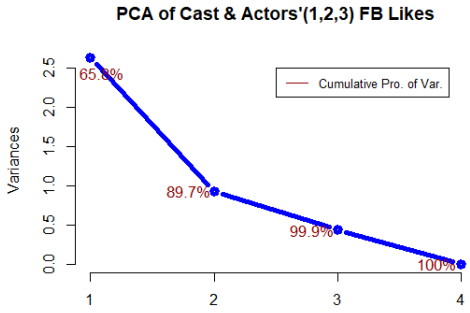


圖 68: 主成分解釋變異及解釋累積變異百分比圖

從上圖中，可以發現，第 1 主成分就已解釋超過 65%以上之變異，第二主成分則解釋了 23.9%變異，兩個主成分加起來已解釋接近 90%的變異。另外，以經驗法則看(選變異在 1 以上的)，前 2 個主成分的變異數約在 1 以上。因此，我們選取前兩個主成分，取代我們的新變數。前兩主成分與原變數的關係圖如下：

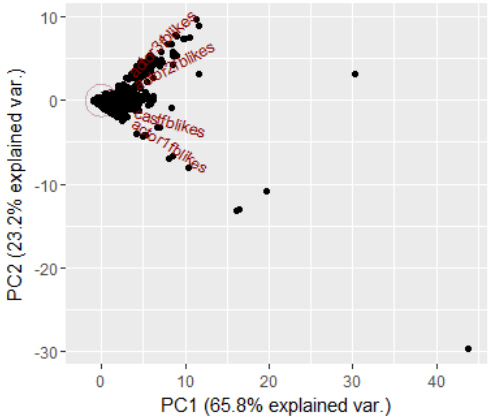


圖 69: 前兩主成分與原變數的關係圖

主成分	PC1	PC2
卡司讚數	0.59	-0.28
演員 1 讚數	0.51	-0.55
演員 2 讚數	0.47	0.40
演員 3 讚數	0.39	0.66

表 5:主成分 1,2 對原變數之轉換係數

從左圖及右表中可得知，第 1 主成分是結合 4 種讚數的綜合指標、第 2 主成分代表演員 2,3 與卡司、演員 1 的相對讚數，值越大代表演員 2,3 讚數越多。從之前 4 變數的相關性就可大概理解這樣的結果，因為卡司與演員 1 相關性最高，演員 2,3 之間相關性也相對較高。演員相關之變數從 4 個變 2 個。最後我們總共有 14 個自變數。

4.4 預測 IMDB 分數

一開始要先決定訓練組資料及測試組資料。我們隨機抽取約 90%資料為訓練組，剩下約 10%資料為測試組。訓練組有 4111 筆資料，測試組有 456 筆資料。

4.4.1 多元線性迴歸(Multiple Linear Regression)

IMDB 分數為應變數，剩下 14 個變數均為自變數，先配適完全線性迴歸模型。因為之後還會用變數選取方式，選出子模型為最終模型，所以在此先不列出係數及 p value 表。下一步，我們需進行殘差分析，看需不需要對變數進行轉換。

雖然常態性跟變異數為常數之假設，對以預測為目的的迴歸模型不重要，也不必要。但因為模型建立後，我們仍想探討影響顯著的因子，牽涉到統計檢定，這時候這兩個假設就必須成立了，所以我們還是對殘差進行分析。

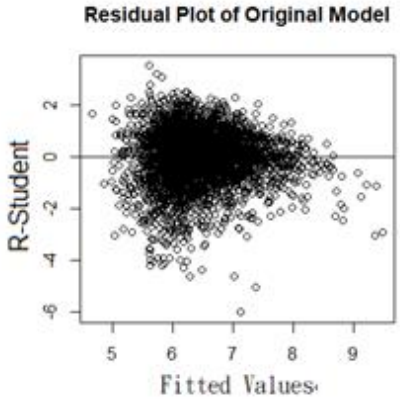


圖 70:殘差與模型配適之反應值圖

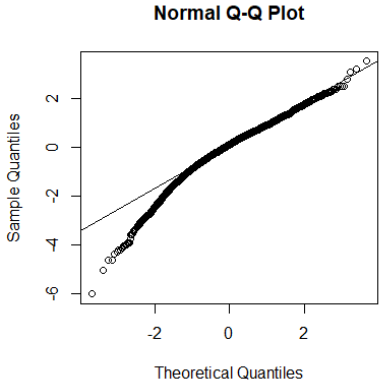


圖 71:殘差之常態機率圖

R-Student 並無在+2 到-2 之間均勻分配，也有很嚴重的左偏分布。

因此，需對變數做轉換。我們用 Box-Cox 決定如何轉換：

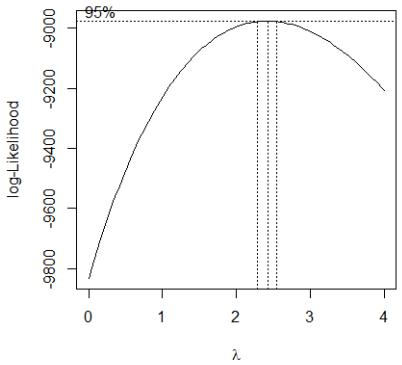
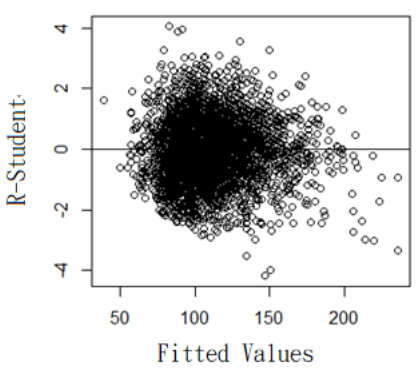


圖 72: λ與對數概似估計值圖

可以發現，在對反應變數做 2.5 次方轉換後，λ的概似函數值為最大(殘差平方和最小)。因此我們採用此轉換，再次配適迴歸模型，模型診斷如下：



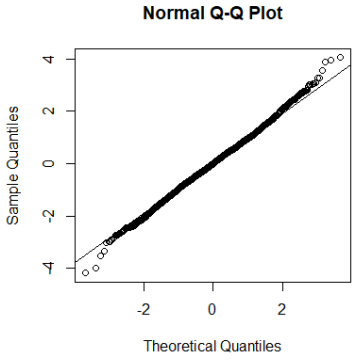


圖 73:轉換後殘差與模型配適之反應值圖　圖 74:轉換後殘差之常態機率圖

可以看到 R-Student 相較於之前，更為均勻分布了，常態機率圖的偏態情況也減輕很多。所以，接下來我們進入選取模型階段，我們採用迴歸分析中常用的逐步迴歸法(Stepwise Regression)。

變數選取方法	選取模型	AIC	表 6: 迴歸模型 選模表
原始模型	囊括全部 14 個自變數	-492.6866	
逐步迴歸法	-螢幕規格、-第 2 主成分 FB 讚數	-496.4672	

針對選取後的變數再次配適迴歸模型，結果如下：

解釋變數	p value	解釋變數	p value	解釋變數	p value
截距項	<0.00001***	語言 西語	0.607	類型 驚悚	0.004**
長度	<0.00001***	語言 華語	0.438	類型 西部	0.389
預算	0.0011**	語言印度語	0.004**	月份 二月	0.439
海報人頭數	<0.00001***	類型 動作	0.080	月份 三月	0.350
導演 FB 讚數	<0.00001***	類型 冒險	0.608	月份 四月	0.616
年份	<0.00001***	類型 動畫	0.696	月份 五月	0.192
分級 G	0.874	類型 自傳	0.748	月份 六月	0.214
分級 PG	0.027*	類型 喜劇	0.211	月份 七月	0.114
分級 PG-13	0.031*	類型 犯罪	0.799	月份 八月	0.747
分級 R	0.049*	類型 紀錄	0.010**	月份 九月	0.0015***
分級 NC-17	0.886	類型 劇情	0.784	月份 十月	0.043*
顏色 彩色	<0.00001***	類型 家庭	0.446	月份十一月	0.13
語言 英語	<0.00001***	類型 奇幻	0.359	月分十二月	0.091
語言 法語	0.125	類型 恐怖	0.009**	第 1 主成分	<0.00001***
語言 德語	0.078	類型 懸疑	0.675	續集 1	0.119
語言 義語	0.815	類型 科幻	0.360	續集 2	0.048*

表 7:最終迴歸模型變數顯著性表格

類別型變數部分，有「其他」這個水準的變數以此為基礎水準；顏色以「黑白」；月份以「一月」；續集則以「非續集」為基礎水準。

影響最顯著的因素包括電影長度、年份、海報人頭數，導演 FB 讚數、顏色、第一主成分之人員 FB 讚數等。R 平方為 0.3216，調整 R 平方為 0.3137，不是特別高，代表還有許多分數之變異沒有被模型解釋；但相對於原作者所做的 R 平方 0.20，已有顯著進步。

變數	VIF	變數	VIF
時間	1.12	顏色	1.15
預算	1.82	語言	1.19
海報人頭數目	1.09	類型	2.35
導演 FB 讚數	1.07	月份	1.29
年份	1.45	第 1 主成分讚數	1.15
分級	1.93	續集	1.05

表 8:最終迴歸模型解釋變數之 VIF

可以發現，變異數膨脹係數(Variance Inflation Factor)都很小，沒有高度共線性問題。原本要嘗試的 Lasso 以及脊回歸(Ridge Reg.)，優點就是在可以在有共線性的情況下，預測及分析能力較好；既然資料沒有這樣的特性，因此就作罷，直接進入機器學習方法。

最後，用上述之選取模型對測試組資料做預測。需注意：預測完的分數為原分數的 2.5 次方，需轉回原始單位。結果如下:

RMSE	0.977

表 9:迴歸模型預測均方根誤差

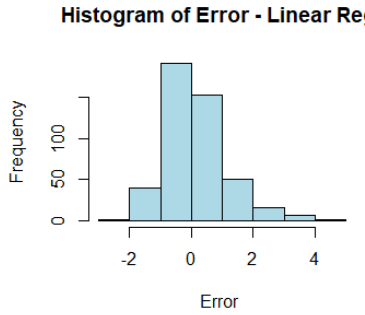


圖 75:迴歸模型預測誤差直方圖

多數預測值與真實值差距在 1 分以內，預測與真實值平均差異為 0.977。

4.4.2 決策樹(Decision Tree)

在建立決策樹模型之前，我們不做特徵篩選，因為決策樹本就會挑選重要的變數進行分枝(embedded selection)，且決策樹的成本複雜修剪參數

(cost-complexity parameter)本身就會對選取的變數做限制。

一開始，我們調整成本複雜修剪參數 cp，選擇可以讓交叉驗證誤差值最低之 cp 值。這裡的交叉驗證是採用十折交叉驗證(ten-fold cross validation)。

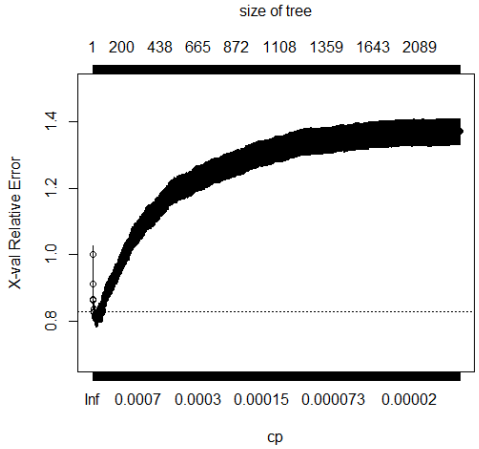


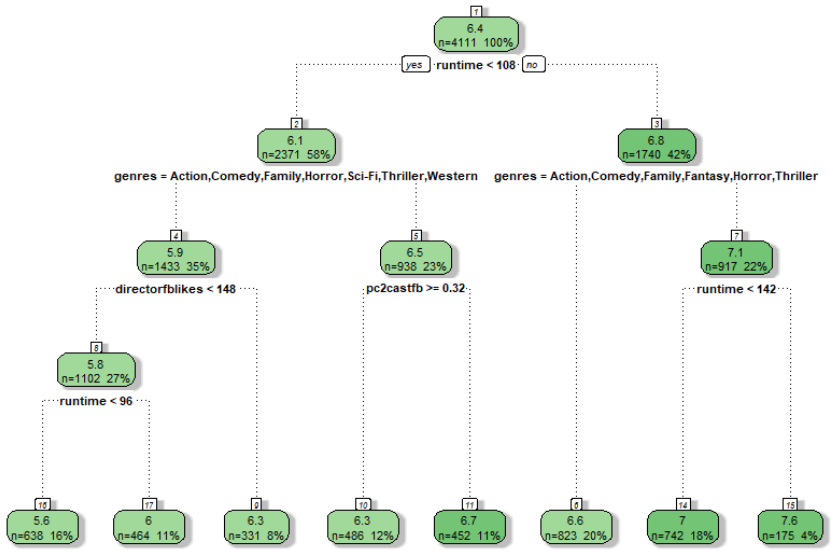
圖 76:cp 對相對交叉驗證誤差值

表 10:cp 列表	cp	分枝數	相對誤差	絕對誤差	誤差變異
最小誤差	0.00472	13	0.748	0.792	0.02335
With 1 sd.	0.00844	7	0.785	0.813	0.02361

這邊的相對誤差是指與根節點(第一個分枝節點)的誤差比值。我們選用的分枝準則為 CART 演算法的變異數縮減，採用二元分枝。

依經驗法則，我們選擇誤差在最小誤差一倍標準差以內，較大的 cp 值，避免過度配適的情況發生。

校調完 cp 後，我們用我們選定的 cp=0.00844 此值，做決策樹，圖示如下：



Rattle 2018-二月-03 15:58:29 user

圖 77:IMDB 分數模型之決策樹

總共有 7 個分枝，8 個葉節點，以該節點所有值之平均為預測值。最重要的變數為電影長度，再來是電影類別。我們以此模型對測試組資料做預測。

RMSE	1.012973

表 11:決策樹模型預測均方根誤差

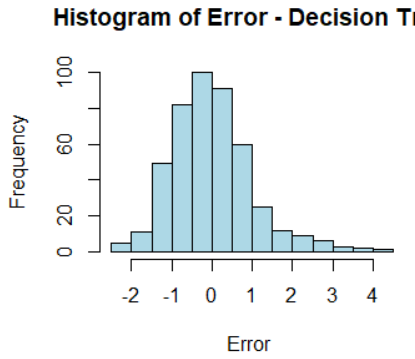


圖 78:決策樹預測誤差直方圖

多數預測與實際誤差在 1 分內，預測與實際分數之平均差異約為 1.01 分。

4.4.3 隨機森林(Random Forest)

隨機森林較需要校調的參數為每次選取的變數個數 m。

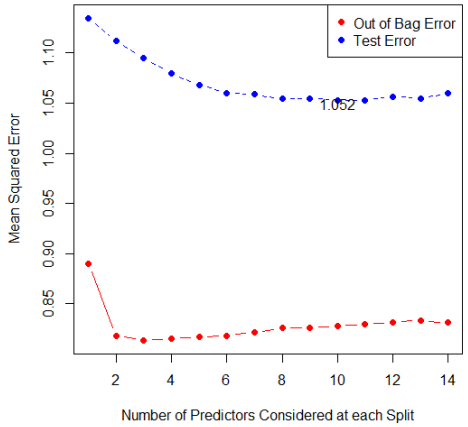


圖 79:抽取變數個數之預測均方差(OOB & test MSE)

藍線為測試組資料的均方差，紅線為 Out of Bag MSE(OOB MSE)，就是每次沒有被抽取到的樣本誤差平均。因為最終還是以測試組資料為預測對象，而 m=10 時測試組均方差最小，所以採用 m=10，種 n=2000 顆樹。

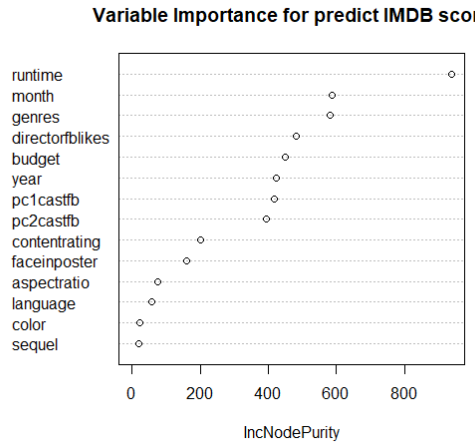


圖 80:隨機森林變數重要性圖

最重要的變數為放映長度，再來是月份、類別、導演 FB 讚數等等，可以看到放映長度的重要性明顯高於其他變數。

RMSE	1.026339

表 12:隨機森林模型預測均方根誤差

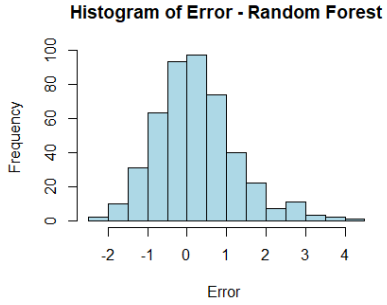


圖 81:隨機森林預測誤差直方圖

多數預測與實際誤差在 1 分內，預測與實際分數之平均差異約為 1.025 分。

4.4.4　支撐向量迴歸(Support Vector Regression(SVR))

我們也不對支撐向量迴歸做特徵篩選，因為 SVR 演算法中的成本正規項 cost 就已為防止過度配適的懲罰項。成本越高，允許犯錯的數量就越少。

在此我們採用徑向基函數核(radial basis function kernel)，它是最被推薦的函數核選擇。它會自動將資料投影到無限高維，以找到線性分割超平面；而且它靈活性很高，不用預先假設資料型態，又是全域核函數，可找到全域最佳解；某種程度上來說，它包含了線性內核(linear kernel)及多項式函數核(polynomial kernel)。通常，徑向基核函數在變數個數不是極多時(變數個數很多時，不用投影到無限高維)的預測表現是所有函數核中最好的。

因為徑向基核函數會考慮歐幾里得距離，所以我們需將資料標準化；要不單位範圍較大的變數會被過度重視。針對順序尺度以上變數，我們直接標準化；針

對名目尺度之變數，我們創造虛擬變數代表。如下圖所示：

	blue	red	yellow
sample 1	1	0	0
sample 2	0	0	1
sample 3	0	1	0
sample 4	0	0	1

圖 82:虛擬變數創造示意圖

接下來我們開始調整徑向基函數核之參數 Gamma α 及 Cost c，以 10 折交叉驗證均方根誤差為指標，選出最佳的組合。

α	0.5	1	2	3	4
c	0.1	1	10	100	1000

表 13:嘗試組合列表

總共嘗試 25 種組合，表現最好的組合是在 cost=1, α=0.5 時，均方根誤差約為 0.92。我們以此參數對訓練組資料建置 SVR 模型，並對測試組資料做預測。

RMSE	1.085449

表 14:支撐向量迴歸模型預測均方根誤差

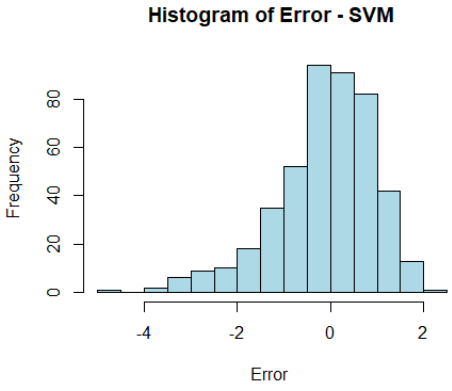


圖 83: 支撐向量迴歸預測誤差直方圖

多數預測與實際誤差在 1 分內，相對於其他方法，有較高的比例落差在 1-2 分之間，預測與實際分數之平均差異約為 1.08 分。

4.4.5 預測表現比較

模型	多元線性迴歸	決策樹	隨機森林	支撐向量迴歸
均方根誤差	0.9775	1.013	1.026	1.085

表 15:不同模型預測 IMDB 分數均方根誤差比較

多元線性迴歸模型的表現最好，是唯一一個預測與實際的平均差異在 1 分以內的方法；再來是決策樹的 1.013 分，接著是隨機森林的 1.026 分，表現最差的則是支撐向量迴歸，均方根誤差來到 1.08 分左右。

線性迴歸預測表現比決策樹好的可能原因為：反應變數-IMDB 分數與自變數之間的關係實際上是很接近線性的，另外，決策樹是以該葉節點所有反應變數之平均值為預測值，當此節點有許多極端值時，用該節點之平均預測落於該節點之測試組資料，並沒有辦法預測得很準確。

這筆資料中，決策樹的表現又比隨機森林好，代表樹本身並沒有遭遇很大之變異不穩定，需要靠種很多樹，也就是隨機森林，來降低變異。另外，因為我們的自變數個數不多，對變數做抽樣，可能會導致特定變數之間的關係有時會被忽略，影響整體表現。

4.4.6 發現

模型	前 5 重要變數				
迴歸分析	長度、導演 FB 讚數、年份、語言(英語)				第 1 主成分 FB 讚數
決策樹	長度	類型	導演 FB 讚數	第 2 主成分 FB 讚數	
隨機森林	長度	月份	類型	導演 FB 讚數	預算

表 16:不同模型找到之前 5 重要變數

迴歸分析中，前四個因子的 p value 均小於 $2*10^{-16}$次方；決策樹中，最重要的長度因子被用來當分枝變數 3 次；隨機森林中，長度的重要性遠勝於其他變數，月份、類型與剩下兩項因子的重要性也有一段差距。

這三個模型第一重要的均為長度，電影放映時間越長，評價傾向越高；另外，導演 FB 讚數也在三個模型之前 5 重要變數中都有出現，可見導演的人氣與名聲也可能影響電影 IMDB 分數。類型在決策樹與隨機森林中，都是前 3 重要的因子。初始決定拍哪一類型的電影-紀錄片分數傾向較高、恐怖片分數傾向較低等等，一般而言，似乎對分數有影響力。但是，恐怖片有時也會有驚喜之作，帶出高評價，這邊闡述的只是一個平均的現象。年份、語言、演員 FB 讚數、預算，都有出現在前 5 因子中。年份越晚，分數越低；演員 FB 讚數越高，分數越高；預算越高，分數越高。演員 FB 讚數出現在前 5 重要變數 2 次，迴歸模型選中的是第 1 主成分；決策樹則選第 2 主成分，它不是最重要的因子，但仍可見演員的人氣似乎也會影響電影評價好壞。

4.5 預測會大於或小於 IMDB 分數分布之中位數-6.5 分

我們將連續型的 IMDB 分數資料，區分為兩類，大於中位數或小於中位數；此資料中的 IMDB 分數中位數為 6.5 分，我們以此為分割點。訓練組及測試組資料跟之前一樣。我們嘗試了許多針對類別型變數進行分類或建模的方法。

4.5.1 多元羅吉斯迴歸(Multiple Logistic Regression)

一開始先進行選模，我們採用向後刪除法(Backward Elimination)及逐步迴歸法(Stepwise Regression)選取囊括進入模型的最終變數，兩個方法之結果一樣。列表如下：

變數選取方法	選取模型	表 17: 羅吉斯迴歸模型選模表
原始模型	囊括全部 14 個自變數	
逐步迴歸法、向後刪除法	-預算	

針對選取後的變數再次配適羅吉斯迴歸模型，結果如下：

解釋變數	p value	odds	解釋變數	p value	odds	解釋變數	p value	odds
截距項	<0.001***		語言印度語	0.006**	0.2	類型西部	0.4	0.3
長度	<0.001***	1.0	螢幕 1.85	0.3	0.8	月份 二月	0.1	0.7
海報人數	0.01**	0.9	螢幕 2.35	0.04*	0.7	月份 三月	0.5	0.9
導演讚數	<0.001***	1.0	類型動作	0.1	0.2	月份 四月	0.8	1.0
年份	<0.001***	0.9	類型冒險	0.6	0.6	月份 五月	0.5	1.1
分級 G	0.7	0.8	類型動畫	0.6	1.5	月份 六月	0.2	1.2
分級 PG	0.01**	0.6	類型自傳	0.5	1.7	月份 七月	0.1	1.2
分級 PG13	0.05**	0.5	類型喜劇	0.2	0.3	月份 八月	0.1	0.7
分級 R	0.9	1.0	類型犯罪	0.5	0.5	月份 九月	0.05	1.3
分級 NC17	0.9	0.9	類型紀錄	0.09	5.4	月份 十月	0.09	1.3
顏色彩色	0.002**	0.5	類型劇情	0.6	0.6	月份 11 月	0.07	1.4
語言英語	<0.001***	0.2	類型家庭	0.3	0.2	月分 12 月	0.5	1.1
語言法語	0.9	0.9	類型奇幻	0.5	0.5	第 1 主成分	<0.001***	1.0
語言德語	0.09*	0.2	類型恐怖	0.03*	0.1	第 2 主成分	0.15	0.9
語言義語	0.6	0.6	類型懸疑	0.9	1.0	續集 1	0.1	0.4
語言西語	0.9	1.0	類型科幻	0.2	0.2	續集 2	0.04*	1.1
語言華語	0.7	0.8	類型驚悚	0.03*	0.08	AIC: 4725.5 (Odds 紀錄至小數第一位)		
Null dev.: 5699 on 4110 df Res.dev.:4625.5 on 4061 df								

表 18:羅吉斯迴歸模型變數顯著性表格

「螢幕規格」此變數以「其他」為基準水準，其他類別型變數的基礎水準則如上面的多元線性迴歸分析方法所定。

影響最顯著的變數為電影長度，電影長度增加 10 分鐘，IMDB 分數大於中位數之勝算為原來的 1.37 倍。第二重要變數為年份，年份增加 10 年，分數大於中位數的勝算為原來的 8 成左右。第三重要變數則為導演 FB 讚數。

接下來我們用 Likelihood Ratio Test 檢驗-是否模型中至少有 1 自變數對反應變數有顯著影響，檢定步驟如下：

1. Ho:所有參數係數均為 0

   Ha:至少有 1 參數係數不為 0

2. $\boldsymbol{\alpha}$=0.05

3. Rejection Region:79.49（ $x^2_{0.05,49}$約為 79.49）

4. LR stat=Null Deviance-Residual Deviance=5699-4625.5=1073.5

5. Conclusion:拒絕虛無假設，至少有 1 參數係數不為 0，代表至少有 1 變數影響顯著，代表此模型對分類是有解釋力的。

最後，我們以此模型預測分數是否會大於或小於中位數，結果如下：

Accuracy	0.715	AUC	0.765
False Positive Rate(FP Rate)	0.259	Positive Predictive Rate	0.706
False Negative Rate(FN Rate)	0.313	Negative Predictive Rate	0.722

表 19:羅吉斯迴歸預測結果表格

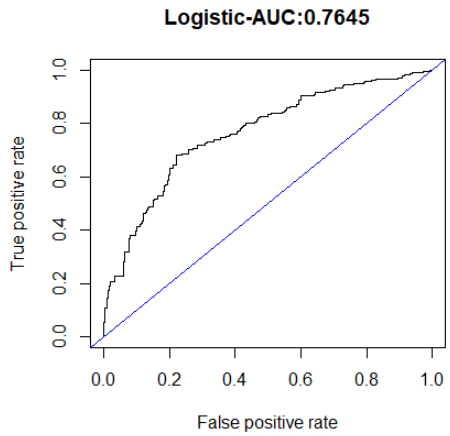


圖 84: 羅吉斯迴歸 R.O.C Curve

整體準確率在 71.5%左右，實際上分數小於中位數的電影，被預測到大於中位數的機率約為 25%;分數大於中位數的電影,被預測到小於中位數的機率較高，為 31.3%。預測大於 6.5 分的電影中，有 70.6%實際分數真正大於 6.5 分的；預測為小於 6.5 分的電影中，則有 72.2%實際分數小於 6.5 分。AUC 為 0.765。

4.5.2 線性判別分析(Linear Discriminant Analysis(LDA))

我們將羅吉斯迴歸篩選出的變數，運用在 LDA 及 QDA 上。自變數的線性判別函數(Linear Discriminant Function)簡要如下:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

每筆資料的自變數之數值帶進去上述函數後，都會有一個值。下圖為將大於 6.5 分之電影、小於 6.5 分之電影之自變數帶進去判別函數之值的對照圖。

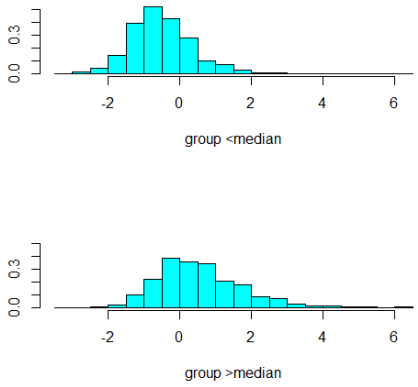


圖 85: 兩類電影分數線性判別函數值分布圖

可以發現，小於 6.5 分之電影，線性判別函數值分布較偏左；大於 6.5 分之電影，線性判別函數值較高。

我們用 LDA 在訓練組資料上，對測試組資料做分類，結果如下：

Accuracy	0.711	AUC	0.759
False Positive Rate(FP Rate)	0.251	Positive Predictive Rate	0.707
False Negative Rate(FN Rate)	0.332	Negative Predictive Rate	0.713

表 20:LDA 預測結果表格

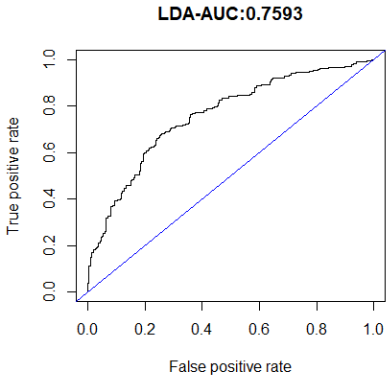


圖 86: LDA R.O.C Curve

整體準確率在 71.1%左右，實際上分數小於中位數的電影，被預測到大於中位數的機率約為 25.1%；分數大於中位數的電影，被預測到小於中位數的機率較高，為 33.2%。預測大於 6.5 分的電影中，有 70.7%實際分數真正大於 6.5 分；預測為小於 6.5 分的電影中，則有 71.3%實際分數小於 6.5 分。AUC 為 0.7593。

4.5.3 二次判別分析(Quadric Discriminant Analysis(QDA))

QDA 跟 LDA 雖然都假設自變數是來自於多元常態分配，不同的地方是，QDA 假設分數大於中位數及小於中位數中，自變數的變異數矩陣是不一樣的，也就是兩分類中之自變數本身的分散程度不一樣。用 QDA 的預測結果如下:

我們用 LDA 在訓練組資料上，對測試組資料做分類，結果如下：

Accuracy	0.640	AUC	0.712
False Positive Rate(FP Rate)	0.142	Positive Predictive Rate	0.720
False Negative Rate(FN Rate)	0.600	Negative Predictive Rate	0.612

表 21: QDA 預測結果表格

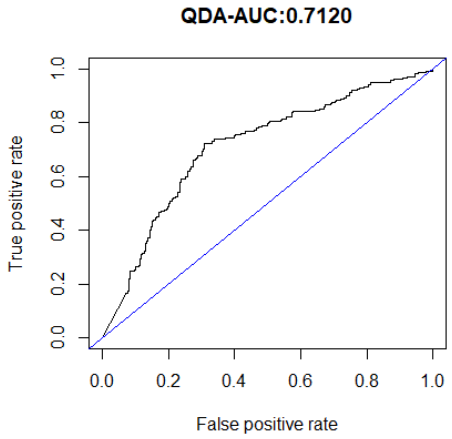


圖 87: QDA R.O.C Curve

整體準確率在 64%左右，實際上分數小於中位數的電影，被預測到大於中位數的機率約為 14.2%；分數大於中位數的電影，被預測到小於中位數的機率相對高很多，為 60%。預測大於 6.5 分的電影中，有 72%實際分數真正大於 6.5 分；預測為小於 6.5 分的電影中，則有 61.2%實際分數小於 6.5 分。AUC 為 0.712。

4.5.4 決策樹(Decision Tree)

我們一樣不做特徵篩選，因為決策樹本就會挑選重要的變數進行分枝，且決策樹的成本複雜修剪參數(cost-complexity parameter)會限制選取變數個數。

一開始，我們調整成本複雜修剪參數 cp，選擇可以讓交叉驗證誤差值最低之 cp 值。這裡的交叉驗證是採用十折交叉驗證(ten-fold cross validation)。

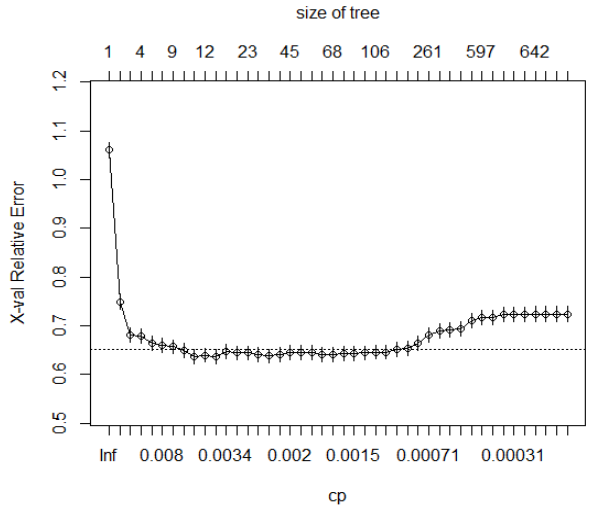


圖 88:cp 對交叉驗證相對誤差值

表 22:cp 列表	cp	分枝數	相對誤差	絕對誤差	誤差變異
最小誤差	0.00488	10	0.58224	0.6369	0.01457
With 1 sd.	0.00586	9	0.58809	0.64959	0.01464

這邊的相對誤差是指與根節點(第一個分枝節點)的誤差比值。我們選用的分枝準則為 CART 演算法的 Gini 係數指標，採用二元分枝。

依經驗法則，我們選擇誤差在最小誤差一倍標準差以內，較大的 cp 值，避免對訓練組資料過度配適的情況發生。

校調完 cp 後，我們用我們選定的 cp=0.00586 此值，做決策樹，圖示如下：

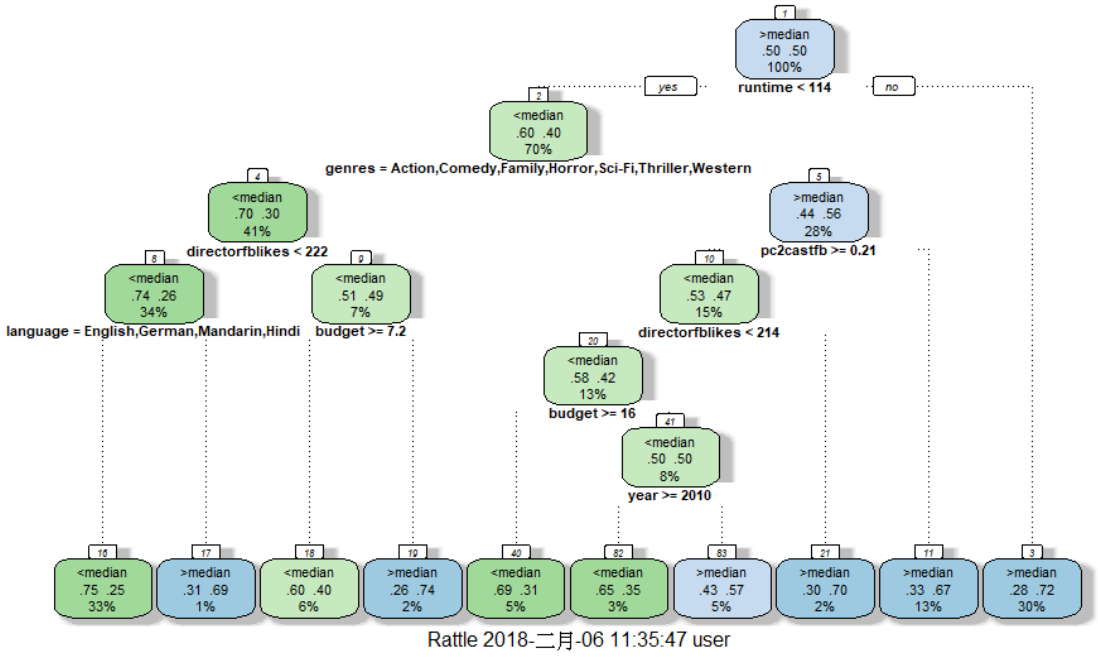


圖 89:IMDB 分數模型之決策樹

總共有 9 個分枝及 10 個葉節點，以該節點比例較高的反應類別為預測類別。最重要的變數為電影長度、電影類別、導演 FB 讚數、第二主成分演員 FB 讚數等。

我們以此模型對測試組資料做分類。

Accuracy	0.671	AUC	0.705
False Positive Rate(FP Rate)	0.339	Positive Predictive Rate	0.646
False Negative Rate(FN Rate)	0.318	Negative Predictive Rate	0.696

表 23: 決策樹預測結果表格

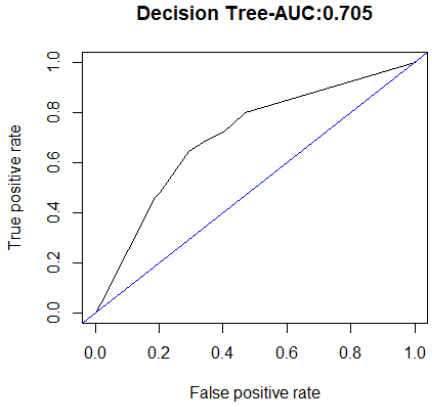


圖 90: 決策樹 R.O.C Curve

整體準確率在 67.1%左右，實際上分數小於中位數的電影，被預測到大於中位數的機率約為 33.9%；分數大於中位數的電影，被預測到小於中位數的機率較低，為 31.8%。預測大於 6.5 分的電影中，有 64.6%實際分數真正大於 6.5 分；預測為小於 6.5 分的電影中，則有 69.6%實際分數小於 6.5 分。AUC 為 0.705。

4.5.5 隨機森林(Random Forest)

隨機森林較需校調的參數為每次選取的變數個數 m。

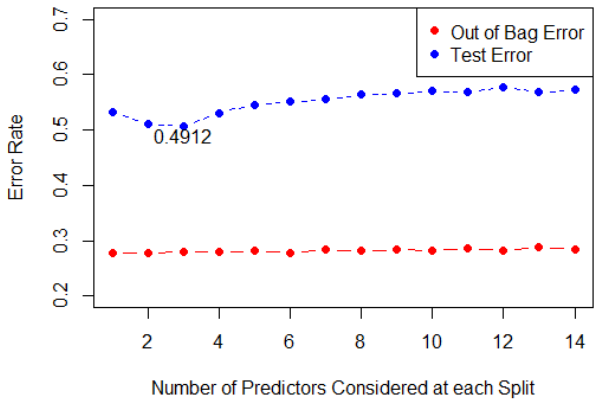


圖 91:抽取變數個數之分類錯誤率圖

藍線為測試組資料的分類錯誤率，紅線為 Out of Bag Error(OOB Error)-每次沒有被抽取到的樣本之分類錯誤率。因為最終還是以測試組資料為預測對象，而 m=3 時測試組分類錯誤率最小，所以採用 m=3，種 n=5000 顆樹。

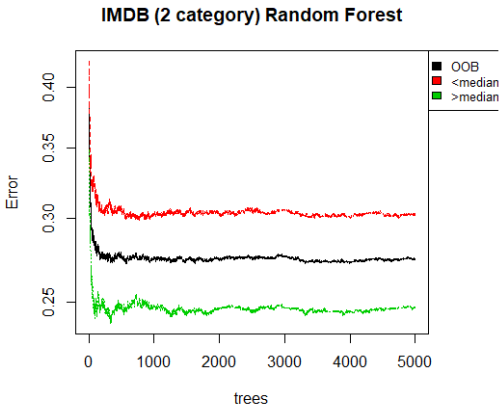


圖 92:隨機森林之 type1, type2, 整體之 OOB Error

整體的 OOB(Out of Bag)Error 為 27.44%，型一錯誤率(FP rate)約為 30%，型二錯誤率約為 25%。可以看到隨著樹的數目上升，錯誤率也漸趨穩定。

我們以此模型對測試組資料做分類。

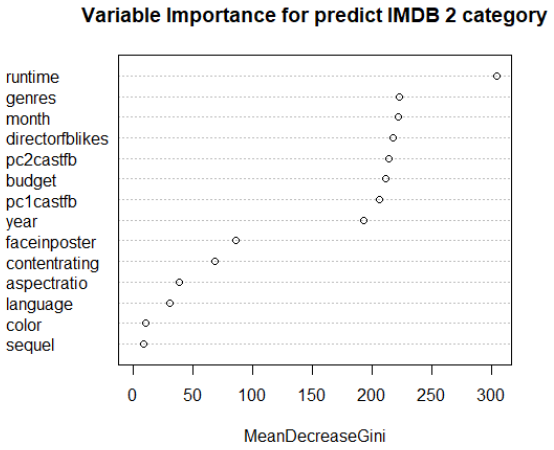


圖 93:隨機森林分類之重要變數圖

跟直接對分數做預測的結果差不多，重要性明顯高於其他變數之因子為長度，再來是類別、月份、導演 FB 讚數、第二主成分之演員讚數等等。前七個重要變數與剩下的變數重要度也有明顯差距。

Accuracy	0.518	AUC	0.743
False Positive Rate(FP Rate)	0.891	Positive Predictive Rate	0.496
False Negative Rate(FN Rate)	0.032	Negative Predictive Rate	0.788

表 24: 隨機森林預測結果表格

整體準確率才比隨機猜高一點，為 51.8%，型一錯誤率更高達 89.1%！但 AUC 有到不錯的 0.743。所以可猜想是分割點的設置問題。因為以中位數為分割，所以預測分割點之門檻值為 0.5。因為問題為型一錯誤率太高，所以應該提高被分類於高於 6.5 分這類的門檻，我們提升至以 0.7 為分割點再次做分類。結果如下：

Accuracy	0.706	AUC	0.743
False Positive Rate(FP Rate)	0.243	Positive Predictive Rate	0.709
False Negative Rate(FN Rate)	0.350	Negative Predictive Rate	0.704

表 25: 隨機森林預測結果表格(預測機率以 0.7 為門檻值)

準確率上升許多，來到 70.6%。型一錯誤率(FP Rate)，也下降到 24.3%，明顯的比以預測機率 0.5 當分界值來得好。最後選定以 0.7 為分割值的結果作為隨機森林之預測結果。

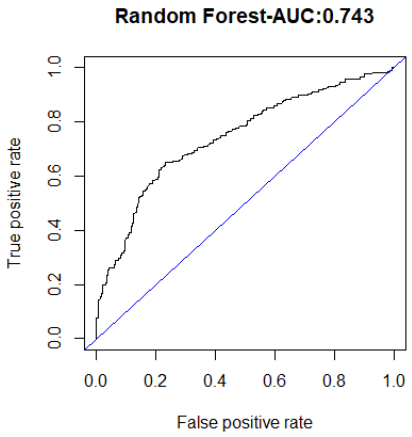


圖 94: 隨機森林 R.O.C Curve

整體準確率在 70.6%左右，實際上分數小於中位數的電影，被預測到大於中位數的機率約為 24.3%；分數大於中位數的電影，被預測到小於中位數的機率較高，為 35%。預測大於 6.5 分的電影中，有 70.9%實際分數真正大於 6.5 分；預測為小於 6.5 分的電影中，則有 70.4%實際分數小於 6.5 分。AUC 為 0.743。

4.5.6 支撐向量機(Support Vector Machine(SVM))

我們不對支撐向量迴歸做特徵篩選，因為 SVR 演算法中的成本正規項 cost 就已是防止過度配適的懲罰項。成本越高，允許犯錯的數量就越少。

在此我們採用徑向基函數核(radial basis function kernel)，原因與前面以支撐向量迴歸預測 IMDB 分數一樣。

因徑向基核函數會考慮歐幾里得距離，我們需將資料標準化；要不單位範圍較大的變數會被過度重視。針對順序尺度以上變數，我們直接標準化；針對名目尺度之變數，我們創造虛擬變數代表，與之前支撐向量迴歸前處理方式一樣。

接下來我們開始調整徑向基函數核之參數 Gamma α 及 Cost c，以 10 折交叉驗證分類錯誤率為指標，選出最佳組合。

α	0.5	1	2	3	4
c	0.1	1	10	100	1000

表 26:SVM 參數嘗試組合列表

總共嘗試 25 種組合，表現最好的組合是在 cost=1, α=0.5 時，分類錯誤率為 37.8%。我們以此參數對訓練組資料建置 SVM 模型，並對測試組資料做分類。

Accuracy	0.612	AUC	0.653
False Positive Rate(FP Rate)	0.536	Positive Predictive Rate	0.568
False Negative Rate(FN Rate)	0.226	Negative Predictive Rate	0.694

表 27:SVM 預測結果表格

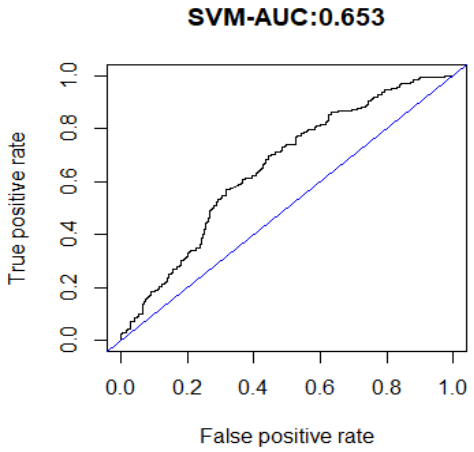


圖 95: SVM R.O.C Curve

整體準確率為 61.2%，實際上分數小於中位數的電影，被預測到大於中位數的機率約為 53.6%，非常高；分數大於中位數的電影，被預測到小於中位數的機率較低，為 22.6%。預測大於 6.5 分的電影中，有 56.8%實際分數真正大於 6.5 分;預測為小於 6.5 分的電影中,則有 69.4%實際分數小於 6.5 分。AUC 為 0.654。

我們也有嘗試提高域值，但預測表現更差，所以還是以此為最終結果。

4.5.7 預測表現比較

Classifier	Logistic	LDA	QDA	Dec.Tree	RF(0.7split)	SVM
Accuracy	0.715	0.711	0.640	0.671	0.706	0.612
AUC	0.765	0.759	0.712	0.705	0.743	0.653
FP Rate	0.259	0.251	0.142	0.339	0.243	0.536
FN Rate	0.313	0.332	0.600	0.318	0.350	0.226
Precision	0.706	0.707	0.720	0.646	0.709	0.568
Recall	0.722	0.713	0.612	0.696	0.704	0.694

表 28:不同分類方法之預測表現比較

深黃色為該項目表現最好的；淺黃色為第二好；接近白色的淡黃色第三好。

羅吉斯迴歸表現最好，不管是準確度或是 AUC 都是最高的，表現次好的分類方法為 LDA，準確率、AUC 及 Recall Rate(預測小於 6.5 分中，實際小於 6.5 分

之比例)都是次佳，與羅吉斯迴歸相去不遠。表現第三好的模型為隨機森林，整體準確率還是有在 70%以上，準確率、AUC 跟 Recall Rate 都是第三高。表現最差的方法為 SVM，整體準確率約些微大於 60%。

羅吉斯迴歸與 LDA 表現最好，代表實際上的決策分界(Decision Boundary)應該是接近線性，而非不規則的，要不然 QDA 應該更可以捕捉不規則邊界。又羅吉斯迴歸表現比 LDA 好，代表各類中的自變數分布應該不是接近多元常態(LDA 對自變數分配之假設)。LDA 比 QDA 好，代表自變數在 2 類的變異不會相差太大。SVM 可能因為將資料標準化後，喪失一些資訊(ex.原始變數的變異程度)等，表現因此受影響，但標準化可以避免給單位尺度較寬的變數較大的權重，所以還是必須執行。

QDA 的型一錯誤率最低，但相對它將很多分數並非小於中位數的電影歸類於大於中位數，所以型二錯誤率非常高；SVM 也是同樣情況，型二錯誤率最低，但是型一錯誤率卻異常高。

4.5.8 發現

模型	1st var.	2nd var.	3rd var.	4th var.	5th var.
羅吉斯迴歸	長度	年份	導演 FB 讚數	語言	第 1 主成分 演員 FB 讚數
決策樹	長度	類型	導演 FB 讚數	第 2 主成分 演員 FB 讚數	語言
隨機森林	長度	類型	月份	導演 FB 讚數	第 2 主成分 演員 FB 讚數

表 29: 不同分類模型找到之前五重要變數

三種方法最重要的變數均為電影長度，電影長度越長，IMDB 分數傾向越高。決策樹及隨機森林都認為第二重要的變數為電影類型，根據決策樹的結果，類型為動作、喜劇、家庭、恐怖、科幻以及西部片的電影，IMDB 分數較低；導演 FB 讚數在 3 個方法中都是前 4 重要變數。前 5 重要變數也都有囊括演員的 FB 讚數，羅吉斯迴歸選的是第 1 主成分，決策樹與隨機森林則選第 2 主成分，可見導演與演員的人氣名望，似乎對電影之 IMDB 分數有一定的影響。

5 結論與未來展望

我們利用資料視覺化初探各因子與分數的關係，並做了許多有趣的分析，帶來許多洞見，例如：我們發現續集電影評價顯著下降；我們也找到哪些演員、導演們演導的電影較常獲得高評價或低評價；也發現了預算與票房關係高，但與利潤關係低等等。這些發現是在後續預測時沒有觸及的，個人覺得非常珍貴。

我們不但直接預測 IMDB 分數，也預測分數會大於或小於中位數之分類預測；充分利用四年來學過的各種分析方法，也對需要不同類型答案的電影相關人士提供多元結果。直接預測分數方面，我們使用線性迴歸、決策樹、隨機森林、支撐向量迴歸模型。分析結果打破大家多認為機器學習預測能力較好的看法，由線性迴歸模型勝出，均方根誤差為 0.977，在 1 分以內。找到的最重要變數為電影長度及電影類型。在分類預測方面(預測會大於還是小於中位數 6.5 分)，我們嘗試羅吉斯迴歸、線性判別分析、二次判別分析、決策樹、隨機森林及支撐向量機。結果由傳統的羅吉斯迴歸勝出，預測準確率最高，高達 71.49%；找到的最重要變數一樣是電影長度，重要性遠勝其他變數，再來也是電影類型。

參考的 5 篇文章，有 3 篇預測電影 IMDB 分數，1 篇預測利潤，1 篇預測票房。預測 IMDB 分數文章中，2 篇直接預測，1 篇也分為大於或小於中位數做分類預測；直接預測部分，我們的均方根誤差小於其中 1 篇文章的結果；分為兩類預測部分，我們的最高準確率和該文章差不到 0.5%。雖然資料來源及型態不同，不能直接以預測指標高低為優劣，但足以顯示我們模型的預測能力有一定的準確度。

未來可以改進的方向主要有 3 項。第 1:雖然視覺化時有做，但分析部分沒有運用斷詞分析；未來希望能運用，將可以提供更多資訊。第 2:社群媒體因子可以再囊括更多，應有助於提高預測準確度。其中 1 篇參考的論文，預測均方根誤差比我們的結果低，其實它只運用社群媒體的幾個指標-如 Youtube 預告片的點擊、轉發、按讚次數等。第 3:演員、導演等過去獲得的大小獎項次數，對 IMDB 分數是重要指標，但此份資料中沒有囊括，可能因為這項資訊須從每個演員的專頁上才抓取，較難蒐集。但是若可考慮這部分的影響，一定可以再提升預測能力。

整體而言，我們的研究顯示-運用電影上映前可以得知的因子，確實可以做到一定程度的評價預測。從此份分析結果來看，如果要打造高評價電影，長度不能太短、類型要審慎挑選、上映的月份也要規劃，預算與評價的關係不大，不一定要灑大錢才能獲得高評及得獎，導演及演員的人氣名聲也需考慮。

雖然現在大家喜歡將所有產品、現象，藉由數據分析提前預測以評估，電影也不例外；但，常常最令人驚嘆的電影是統計上的離群值。如果過度仰賴演算與分析，會讓投資者趨於保守，不願冒風險，而你我可能會錯過影響一生的好電影。

參考文獻及網站

[1] Jeffrey Ericson and Jesse Grodman.(2007).A Predictor for Movie Success.
Available at:http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf.

[2] Michael T. Lash and Kang Zhao.(2016).Early Predictions of Movie Success: the Who, What, and When of Profitability.*Journal of Management Information Systems*

[3] Srikanth S V, Ayesha Rahman, Motoki Saito, Rangasayee Lalgudi Chandrasekaran, and Anand Agrawa.(2011).Predicting Indian Movie Ratings on IMDB.
Available at:
http://www.galitshmueli.com/data-mining-project/predicting-indian-movie-ratings-imdb.

[4] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke.(2012). Predicting IMDB Movie Ratings Using Social Media. *Conference Paper: Proceedings of the 34th European conference on Advances in Information Retrieval Pages 503-507.*

[5] Nithin Vr,and Sarath Babu Pb.(2014).Predicting Movie Success Based on IMDB Data. *Conference Paper:International Journal of Data Mining Techniques and Applications.*

[6] https://nycdatascience.com/blog/student-works/machine-learning/movie-rating-prediction/

[7] https://www.kaggle.com/epfreed/tidydata-movie-dataset-exploration/notebook

[8] https://www.kaggle.com/bpali26/which-actor-to-look-for-your-kind-of-genre

[9] https://www.kaggle.com/gsdeepakkumar/movie-mania-exploring-the-movie-database