流行歌曲分析與預測

廖泓傑 B07703001 財金三

杜沛慈 B08705005 資管四

丁昱寧 R10126032 圖資碩二

摘要

隨著音樂製作成本高漲，若能識別出歌曲熱門的要素並製作出深受大眾喜愛的作品，將有助於獲取龐大的商業利益。此外，利用機器學習預測歌曲熱門程度已成為可能。本研究以 Spotify API 取得 4157 首熱門歌曲與 3704 首非熱門歌曲的音樂特徵，並產生解釋型模型和預測型模型。結果顯示熱門歌曲中純朗誦和純音樂的比例較低，相對較多的是大調、音樂強度較高且傳達負面情緒等要素的歌曲，不同語言的熱門歌曲特徵也存在著差異。預測模型中羅吉斯迴歸表現最佳。這些結果可以為音樂產業從業人員提供一些靈感和啟示。

關鍵字

流行歌曲預測；Spotify；熱門歌曲

一、研究背景

根據文化內容策進院發布的《2020 年文化內容產業調查計畫 III：流行音樂產業》，指出 2019 年台灣每首單曲平均製作成本為 19.55 萬元，且歌曲的製作涉及多種相關人員的參與，包括詞曲創作者、藝人經紀、詞曲經紀、唱片製作公司、視覺設計服務、媒體廣告業者等，可見除了金錢成本，也需投入多人的時間與心力來共同完成一首歌曲。

在歌曲的銷售方面，則涵蓋多元化的途徑，包含實體銷售、音樂串流、 音樂下載、電影、廣播與電視演出、代言、KTV 伴唱帶等音樂授權。2019 年我國流行音樂產業的總營業額為 254.71 億元，若能創作出膾炙人口的歌曲，有助於從這龐大的潛在商業利益中獲利。

隨著可分析的資料增加，近年以機器學習預測熱門歌曲的熱門歌曲科學（Hit song science）又逐漸獲得關注，有多位研究者試圖以 Spotify API 建立熱門歌曲預測模型，例如 Georgieva 等（2018）、Middlebrook 與 Sheik（2019）、Raza 與 Nanath（2020）皆以美國告示牌（Billboard）選擇熱門歌曲，從 Spotify API 提取每首歌曲的音頻特徵，用以訓練模型預測熱門歌曲。Georgieva 等（2018）考量歌手的知名度，以及將歌曲分成夏日（6 月到 8 月）與假期（11 月至 1 月），觀察到流行音樂會

因發行時段而有不同特徵。另外也將數據分成五年一個時期，觀察音樂趨勢隨時間變化。Middlebrook 與 Sheik（2019）考慮了歌曲的持續時間，以及藝術家過往表現特徵，作為流行歌曲的預測評估要素。Raza 與 Nanath（2020）除了歌曲的音頻特徵，還納入歌詞的情感分析進行熱門歌曲的預測評估。在預測模型的方法上，三者皆有使用的是羅吉斯迴歸（LR），有被其中兩者使用的方法為支援向量機（SVM）、決策樹（DT）、神經網絡（NN）、隨機森林（RF），其他方法則包括最大期望演算法（EM）、高斯判別分析模型（GDA）、單純貝氏分類器（Naïve Bayes）。

基於上述的音樂產業背景與文獻，我們提出以下研究問題：

RQ1 : 是否能透過分析樂曲的音訊特徵，解釋可能影響歌曲熱門的因素？

- 目的：透過分析歌曲的特徵，建立解釋型模型，解釋可能影響歌曲熱門的因素。這種分析不僅能為音樂創作者在創作時提供有價值的參考依據，還能確保音樂製作所需的巨大成本不會白費，讓每一筆投入的成本都能獲得最大的效益。

RQ2：是否能透過分析樂曲的音訊特徵，預測歌曲成為熱門歌曲的可能性？

- 目的：若能在歌曲發行之前預測其是否會成為熱門歌曲，將有助於唱片公司在後續的音樂銷售上，獲取最大程度的利潤。這樣的預測能夠提供唱片公司有價值的市場洞察，讓他們能夠適時調整行銷策略，決定專輯要收錄的歌曲、挑選主打歌，並將資源集中在潛力歌曲的宣傳推廣。

RQ3：熱門歌曲的特徵是否會隨年代有所差異？

- 目的：探究熱門歌曲特徵在各年代的差異，能夠揭示音樂在不同時代的演變趨勢，這樣的了解能夠引導音樂創作者和音樂行銷從業者，更精準地迎合現代人的喜好，創作出更受歡迎的音樂作品。

RQ4：不同語言的熱門歌曲特徵是否有所差異？

- 目的：探討受國內歡迎的不同語言熱門歌曲的特徵差異，不僅有助於國內音樂創作者更好地了解本土聽眾偏好，也能夠幫助音樂代理商找到符合聽眾偏好的國外音樂，進行代理和推廣。

二、資料蒐集與處理

首先使用 python 進行爬蟲取得熱門歌曲之「歌名」、「歌手」、「專輯名稱」等資料，資料取自台灣流行曲排行榜[1]：

1. 榜單來源

(1) KKbox 年度風雲榜(2018~2021)：「華語」、「台語」、「西洋」、「日語」、「韓語」五種語言，每種語言每年 100 首歌，共四年，總計 2000 首歌曲。

(2) KKbox 單曲月榜(2010/1~2017/12)：「華語」、「台語」、「西洋」、「日語」、「韓語」五種語言，每種語言每月 100 首歌，共 96 個月，總計 48000 首歌曲。

(3) Hit-FM 年度百首單曲(1998~2017)：不分語言每年 100 首歌，共 20 年，總計 2000 首歌曲。由於此榜單並無區分歌曲的語言，因此採手動標籤的方式標記語言欄位。

2. 去除重複歌曲

為避免熱門歌曲中各榜單有重疊的歌曲，造成模型中含多筆相同資料，因此去除「歌名」和「歌手」完全相同之資料，最後得出總計 10893 首熱門歌曲。

接著使用 spotify API 對上述排行榜歌曲進行查詢，以取得熱門與非熱門歌曲資料：

(1) 檢查缺失值：較早期的歌曲缺失「專輯」欄位。

(2) 查詢熱門歌曲：由於近半數資料缺失「專輯」欄位，本研究中使用「歌手」及「歌曲」欄位作為識別唯一歌曲之標準並使用此二條件進行查詢。

(3) 去除無搜尋結果的資料：因歌手名稱語言不同，例如排行榜中歌手為「五月天」的資料，在 Spotify 中歌手名稱顯示為「Mayday」、或因歌曲發行年較早而不存在 Spotify 曲庫中，其歌曲資料查詢結果標註為 no result 並予以去除。

(4) 整合熱門歌曲欄位：合併排行榜欄位及 Spotify 查詢、篩選後結果，共取得 4157 首熱門歌曲資料。

(5) 查詢非熱門歌曲：本研究中將非熱門歌曲定義為由熱門歌曲之歌手於同年發行但未進入排行榜的其他歌曲，因此使用熱門歌

[1] 台灣流行歌曲排行榜 https://tw-pop-chart.blogspot.com/

曲資料中的「年分」及「歌手」欄位進行查詢，去除 no result 之搜尋結果後共取得 3704 首非熱門歌曲資料，熱門與非熱門歌曲資料筆數大致平衡。

(6) 歌曲特徵值：透過 Spotify API 取得熱門與非熱門歌曲特徵值，包含曲調(key)、音量(loudness)、律動感(danceability)等共 13 項。

(7) 歌曲標籤：設熱門歌曲為陽性(label = 1)、非熱門歌曲為陰性 (label = 0)。

清理後的熱門與非熱門歌曲資料示例如下表：

年分	語言	歌手	歌曲	歌曲 ID	專輯	曲調	音量	律動感	...	標籤
2021	華語	程響	四季予你	4BGkSC	四季予你	3	-6.745	0.534		1
2010	韓語	2pm	Still	1ZjqjlM	Still 2:00pm	0	-4.404	0.644		1
2021	華語	程響	君不知	2RT85	长安伏妖	8	-8.140	0.220		0
2010	韓語	2pm	Crazy Babe	17JfICw	Crazy Babe	1	-5.369	0.471		0

三、探索式資料分析

本研究共取得 13 項 Spotify 提供的音樂特徵進行分析，分別為(1) Acousticness，表示音樂含非電子音的程度，值介於 0 到 1 之間。(2) Danceability，律動感，表示適合跳舞的程度，值介於 0 到 1 之間。(3) Duration_ms，音樂時長，單位為毫秒。(4) Energy，對音樂強度與活躍度的感知，值介於 0 到 1 之間。(5) Instrumentalness，純音樂，不含人聲的佔比，值介於 0 到 1 之間，1 代表純音樂。(6) Key，曲調，0 代表 c 調，1 代表升 c，2 代表 d 調，依此類推。(7) Liveness，現場感，檢測錄音中是否存在觀眾，值介於 0 到 1 之間。(8) Loudness，音量，單位為分貝。(9) Mode，音軌調性，1 為大調，0 為小調。(10) Speechiness，口說、朗誦比例。(11) Tempo，音軌的整體節奏速度，以每分鐘節拍數(BPM)為單位。(12) Valence，音樂帶給人的正向心理感受程度，值介於 0 到 1 之間。(13) Time_signature，音軌的整體拍號。

1.熱門歌曲特徵年度變化

初步以視覺化圖表觀察自 1998 年至 2021 年間所有熱門歌曲的 13 項音樂特徵的平均值變化，發現「Loudness」和「Mode」這兩個特徵有較為明顯的差異。Loudeness 顯示近年熱門歌曲的音量較過往來得高，而 Mode 則顯示近年熱門歌曲為大調的比例較過去更高。

其他音樂特徵則無觀察出明顯的變化，雖然「Acousticness」、「Instrumentalness」、「Speechiness」這三個特徵的趨勢線有稍微明顯的斜率變化，但細看其數值的差異僅約介於 0.03 至 0.1 之間，變化不太明顯。細節可參閱附件。

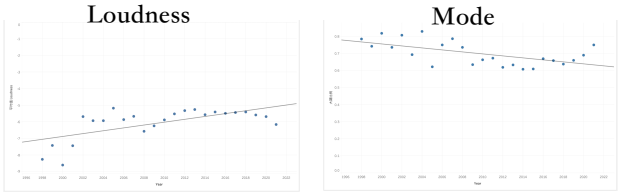


圖 1 熱門歌曲特徵年度變化較明顯者

2. 熱門與非熱門歌曲特徵差異

分析 1998 年到 2021 年所有語言的熱門歌曲與非熱門歌曲特徵，以初步的視覺化探索，橘線表示熱門歌曲，藍線表示非熱門歌曲，觀察到以下三個音樂特徵在冷熱門歌曲有較明顯的差異：

(1) Instrumentalness：整體而言，每年熱門歌曲的純音樂性平均比非熱門歌曲低，這意味著熱門歌曲中僅含伴奏而無演唱部分的比例相對較低。

(2) Loudness：平均而言，熱門歌曲的音量比非熱門歌曲來得大，顯示人們可能偏好音量較大的歌曲。

(3) Speechiness：熱門歌曲的朗誦比例平均比非熱門歌曲低，意味著在熱門歌曲中，僅以單純的朗誦形式呈現而無音樂伴奏的比例相對較低。

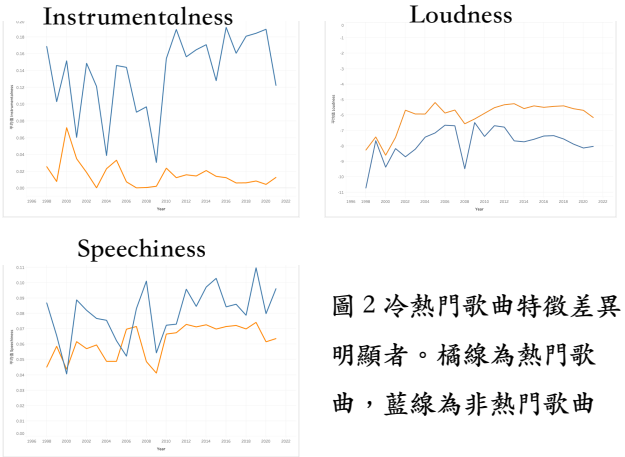


圖 2 冷熱門歌曲特徵差異明顯者。橘線為熱門歌曲，藍線為非熱門歌曲

3. 各語言熱門歌曲特徵差異

將歌曲按照華語、台語、西洋、日語和韓語五種語言進行分類，觀察不同語言熱門歌曲之間的特徵差異，我們發現在以下五個音樂特徵上，不同語言的熱門歌曲呈現較明顯的區別：

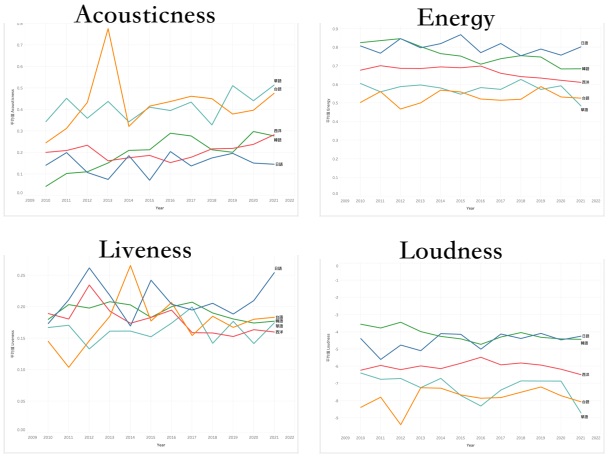
(1) Acousticness：華語、台語熱門歌曲的原聲程度最高，日語熱門歌曲最低。

(2) Energy：日語熱門歌曲的音樂強度最高，韓語次之，華語、台語最低。

(3) Liveness：日語熱門歌曲的現場感高於其他語言熱門歌曲。

(4) Loudness：日語、韓語熱門歌曲的音量較高，華語、台語音量較低。

(5) Valence：日語、韓語熱門歌曲的正向情緒較高，華語、台語較低。



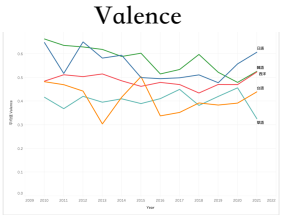


圖 3 各語言熱門歌曲特徵差異明顯者

四、解釋型模型

本研究為分析歌曲特徵是否影響歌曲成為熱門歌曲，使用羅吉斯迴歸作為解釋型模型。應變數為 0、1 變數，0 為非熱門歌曲，1 為熱門歌曲。

1. 變數篩選

檢查所有自變數變異度後，發現變數 Time_signature 變異度不足，此變數代表歌曲一小節有多少拍，因不論熱門或非熱門歌語都有極高比例一小節為四拍，因此排除變數 Time_signature。

為避免自變數之間的高度相關影響模型結果，使用相關矩陣檢視自變數之間相關性，結果如圖 4：

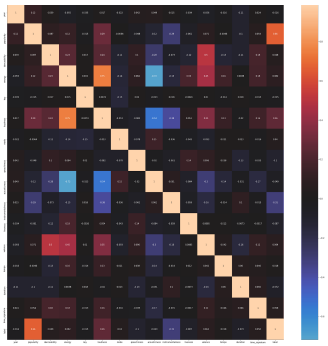


圖 4 自變數的相關矩陣

相關矩陣中，Loudness（音量）和 Energy（音樂強度）之相關係數為 0.75，兩者高度正相關；Acousticness（非電子音程度）和 Energy（音樂強度與）之相關係數為-0.72，兩者高度負相關，因此，模型中排除自變數 Loudness（音量）和 Acousticness（非電子音程度）以避免自變數之間的互相影響。

2.全體資料之羅吉斯迴歸

將篩選過後之變數放入羅吉斯迴歸模型中，得到以下結果：

```
Optimization terminated successfully.
         Current function value: 0.617129
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                  label   No. Observations:                 7861
Model:                          Logit   Df Residuals:                     7850
Method:                           MLE   Df Model:                           10
Date:                Thu, 08 Jun 2023   Pseudo R-squ.:                  0.1075
Time:                        17:56:07   Log-Likelihood:                -4851.3
converged:                       True   LL-Null:                       -5435.8
Covariance Type:            nonrobust   LLR p-value:                6.875e-245
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5168      0.215      2.401      0.016       0.095       0.939
danceability   0.3461      0.201      1.722      0.085      -0.048       0.740
key           -0.0019      0.007     -0.273      0.785      -0.015       0.011
mode           0.1259      0.052      2.410      0.016       0.024       0.228
energy         0.8818      0.138      6.384      0.000       0.611       1.153
speechiness   -3.4449      0.334    -10.315      0.000      -4.099      -2.790
instrumentalness -4.0424   0.204    -19.804      0.000      -4.443      -3.642
liveness      -1.2894      0.147     -8.747      0.000      -1.578      -1.001
valence       -0.2939      0.136     -2.164      0.030      -0.560      -0.028
tempo          0.0014      0.001      1.576      0.115      -0.000       0.003
duration      -0.0024      0.000     -5.448      0.000      -0.003      -0.002
==============================================================================
```

本研究以 P 值=0.05 作為閥值，若 P<0.05，則代表該自變數對應變數是否熱門的影響足夠顯著，

得出足夠顯著的自變數有七個，分別為 Mode, Energy, Speechiness, Instrumentalness, Liveness, Valence 和 Duration，整理表格如下。其中自變數除了 Mode 和 Duration 有較精確的含意和單位外，其餘變數皆多為計算出的綜合評分，因此係數大小僅供參考，係數的正負（變數對結果的影響為正面或負面）更為觀察重點。

變數	係數	Odds Ratio
mode	0.1259	1.134169
energy	0.8818	2.415243
speechiness	-3.4449	0.031908
instrumentalness	-4.0424	0.017555
liveness	-1.2894	0.275436
valence	-0.2939	0.745351
duration(s)	-0.0024	0.997603

其中 Mode（0 為小調、1 為大調）的係數為 0.1259，可換算出 odds ratio 為 1.134，代表大調的歌曲相較小調的歌曲成為熱門歌曲的機會高出 1.134 倍；Energy(音樂強度)的係數為 0.8818，代表一般來說，強度越高越有可能成為熱門歌曲。

Speechiness(口說朗誦比例)係數為-3.4449，表示音頻中含有無音樂、純朗誦的比例越高，越不可能成為熱門歌曲；同時，Instrumentalness(純音樂比例)係數為-4.0424，表示音頻中純音樂無人聲的比例越高，也越不可能成為熱門歌曲。兩者雖然為看似相反的概念，但在檢查相關係數，發現兩者並不相關，相關係數僅-0.061，因此我們推測，綜合 Speechiness 和 Instrumentalness 兩變數，一首歌曲要成為熱門歌曲，應盡量使音頻同時包含人聲及音樂。

Liveness（現場感）的係數為-1.2894，顯示音頻中收錄到的現場觀眾聲越少，越容易成為熱門歌曲；Valence（正向心理感受）的係數為-0.2939，顯示描述負面情緒的歌曲，更容易成為熱門歌曲。

Duration（音樂時長）的係數為-0.0024，換算出的 odds ratio 為 0.9976，在模型中有納入之歌曲的正常時長範圍，約兩分半至四分半內，時間每短一秒約多出 0.24%的機會成為熱門歌曲。

3.不同語言歌曲分析

在全體資料集中有中「華語」、「台語」、「西洋」、「日語」、「韓語」五種語言的歌曲資料，為分析不同語言歌曲自變數的影響有何變化，因此分別針對各語言歌曲資料作羅吉斯迴歸，並觀察不同語言的係數高低評估影響力的大小及正面或負面。需注意的是下列圖表使用「係數」作圖，且多數變數沒有明確的定義和單位，因此關注其正負值和相對高低。

(1) Mode（大小調）：

如圖 5 所示，日語是所有語言中，唯一小調歌曲更有可能成為熱門歌曲者，其餘語言均以大調歌曲更有可能成為熱門歌曲，影響力最大的為華語。

(2) Energy（音樂強度）：

圖 5 中，華語是所有語言中，唯一音樂強度越小，越有可能成為熱門歌曲者，其餘語言均以音樂強度大的歌曲更有可能成為熱門歌曲，影響力為最大的為日語。其結果與他人研究發現華語熱門歌曲多為抒情歌吻合。

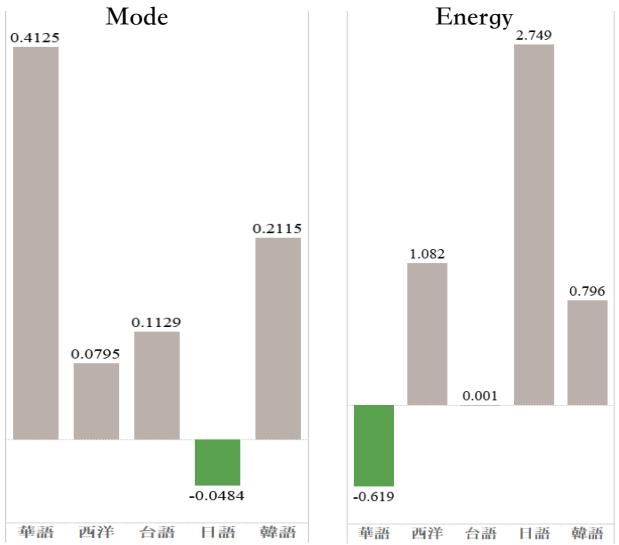


圖 5 各語言熱門歌曲的 Mode 和 Energy 差異

(3) Speechiness（口說、朗誦比例）：

在圖 6 所有語言均以口說朗誦比例較低者更可能成為熱門歌曲，其中影響力以台語歌曲較小。

(4) Instrumentalness（純音樂比例）：

圖 6 顯示所有語言均以純音樂比例較低者更可能成為熱門歌曲，其中影響力以台語歌曲較其他語言大。

綜合對 Speechiness 和 Instrumentalness 的觀察，所有語言歌曲都以音頻盡量同時有人聲及音樂聲更有可能成為熱門歌曲。

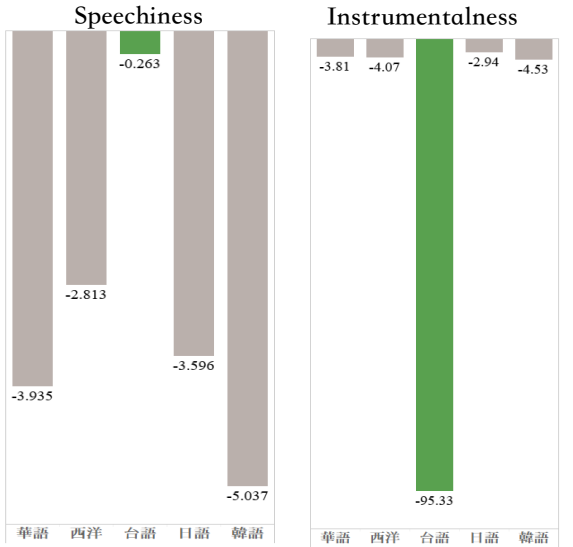


圖 6 各語言熱門歌曲的 Speechiness 和 Instrumentalness 差異

(5) Liveness（現場感）：

如圖 7 所示，所有語言均以現場音較少者更可能成為熱門歌曲，其中負面影響以台語歌曲較小於其他語言。

(6) Valence（音樂情緒）：

從圖 7 觀察到韓語是所有語言中，唯一 Valence 的係數為正者，代表韓語歌以正向快樂的歌曲更有可能成為熱門歌曲，其餘語言均以負面憂鬱之歌曲更有可能成為熱門歌曲，負面對歌曲熱門度影響力最大者為台語歌曲。

(7) Duration（音樂時長）：

僅有華語歌時間越長越有可能成為熱門歌曲，其餘語言均以時間短為佳，時間短對成為熱門歌曲影響最顯著者為韓語歌曲。

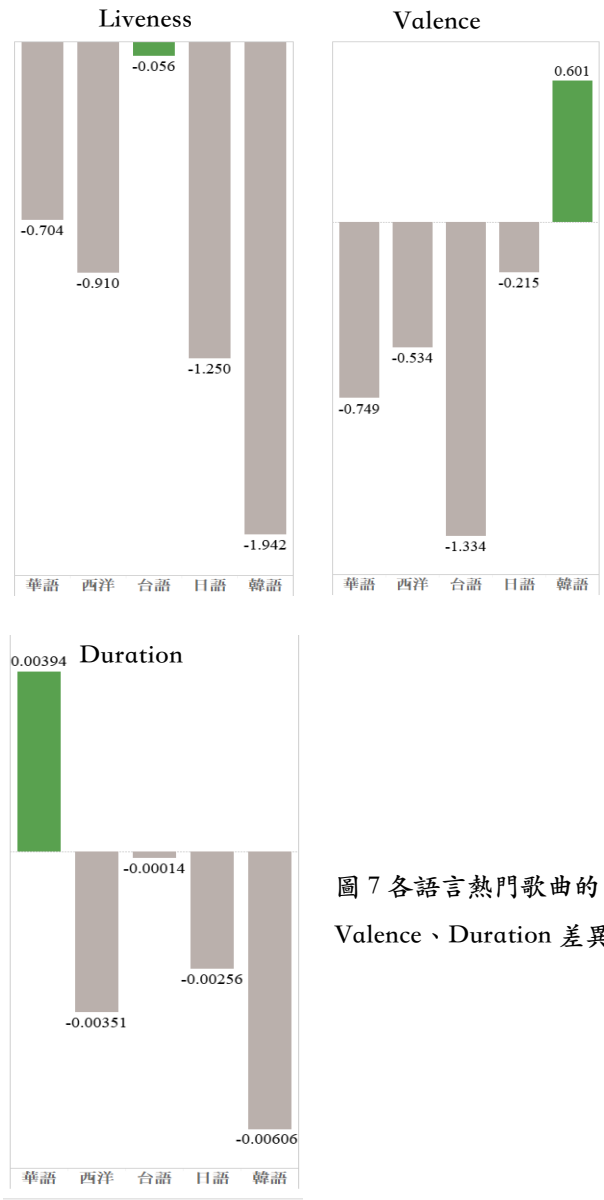


圖 7 各語言熱門歌曲的 Liveness、Valence、Duration 差異

4. 不同年代歌曲分析

在全體資料集中有中共含有 1998~2021 年之歌曲資料，為分析不同年代歌曲自變數的影響有何變化，因 1998 年至 2006 年的資料量較少，故除了第一個年代外，皆以五年為一單位分群，分為「1998~2006」、「2007~2011」、「2012~2016」、「2017~2021」四個年代，並分別針對各年代歌曲資料作羅吉斯迴歸，觀察不同年代的係數高低評估影響力的大小及正面或負面，以及是否有隨年代變化之明顯趨勢。

需注意的是下列圖表使用「係數」作圖，且多數變數沒有明確的定義和單位，因此關注其正負值和相對高低。以下三者為有較明顯趨勢之自變數。

(1) Instrumentalness（純音樂比例）：

從圖 8 可觀察到雖然在所有年代，純音樂比例高之歌曲均更不容易成為熱門歌曲，但隨年代發展，近代純音樂比例的負面影響逐年代降低，推測可能與近年電音熱門歌曲的成長有關。

(2) Speechiness（口說朗誦比例）：

如圖 8 所示，Speechiness 和 Instrumentalness 相反，隨年代發展，近代純口說朗誦比例的負面影響逐年代降增高，顯示越近代音頻中僅有人聲卻無音樂段落，對歌曲成為熱門歌曲之負面影響逐年代增高。

(3) Valence（音樂情緒）：

從圖 8 可見，Valence 對歌曲成為熱門歌曲之負面影響逐年降低，甚至在最近期(2017~2021)之影響為正面影響，顯示正向快樂的歌隨年代更受歡迎，負面憂鬱的歌隨年代發展越不易成為熱門歌曲。

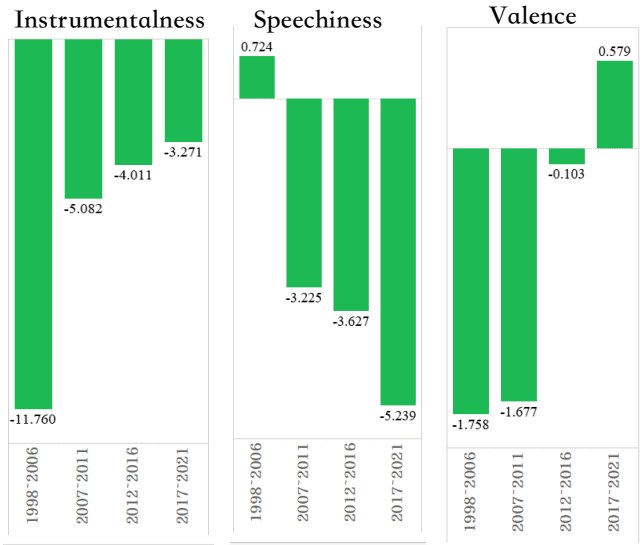


圖 8 不同年代熱門歌曲特徵差異明顯者

五、預測型模型

1. 模型建構、調整與選擇

(1) 變數選擇：使用所有 13 項歌曲特徵為自變數，熱門、非熱門標籤為應變數

(2) 切分資料集：將 1998 年至 2020 年之資料切分為 80%訓練集與 20%驗證集

(3) 定義模型好壞判斷標準：設熱門為陽性，非熱門為陰性。因為沒有明確資訊判斷偽陰性及偽陽性何者較重要，因此選用 F1 score 為評斷模型好壞的標準。

(4) 候選模型：本研究中嘗試了羅吉斯迴歸(Logistic regression)、隨機森林(Random forest)、極限梯度提升(eXtreme Gradient Boosting)三種模型，並調整其 regularization 相關參數與預測陽性的門檻值組合如下：

	Logistic Regression	Random Forest	XGBoost
regularization	C = 0.01, 0.1, 1, 10, 100	max_depth = 1, 2, … , 10	aplha = 0.01, 0.1, 1, 10, 100
threshold	0.2, 0.25, 0.3, … , 0.7	0.2, 0.25, 0.3, … , 0.7	0.2, 0.25, 0.3, … , 0.7

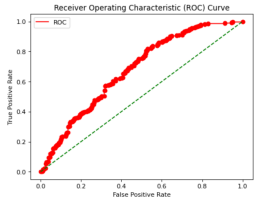
模型驗證各指標結果如下，以 Logistic Regression 模型在 regulation penalty 為 1、threshold 為 0.35 時，有最佳的 F1 score 約為 0.758。

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.667	0.662	0.970	0.758
Random Forest	0.673	0.631	0.945	0.756
XGBoost	0.673	0.630	0.944	0.756

2. 模型預測結果與評估

使用前述羅吉斯迴歸模型與最佳之 regulation penalty、threshold 組合，以 1998 年至 2020 年之資料為訓練集、2021 年資料為測試集，進行熱門、非熱門歌曲分類預測，預測結果之混淆矩陣如下所示，其 F1 score 約為 0.727，並繪製 ROC 曲線，AUC 約為 0.677。

	真實熱門	真實冷門
預測熱門	270	193
預測冷門	10	58
C = 1 / threshold = 0.35		



六、總結與未來展望

總結而言，本研究透過解釋型模型，了解隨年代變化與不同語言中影響歌曲成為熱門的因素，能夠協助歌曲創作並降低相關成本。並藉由預測型模型推測歌曲是否可能受到歡迎，有助於歌曲

簽約、收錄、行銷等流程，提升商業效益。此外也期待未來能夠進一步優化模型，包含納入歌曲主題類型等其他因子、進行歌詞文本分析，以及考量外部變因如歌手知名度、社交媒體的影響等，皆是值得深入思考與研究的議題。

參考資料

[1] 蔡振家、李家瑋、葉家含、陳容姍、林耀盛（2017）。為何華語流行樂壇以情傷歌曲為主？試析抒情歌曲的療癒潛質。本土心理學研究， 47，371 - 420。https://doi.org/10.6254/2017.47.371

[2] 文化內容策進院（2020）。2020 台灣文化內容產業調查報告 III：流行音樂產業。 https://taicca.tw/article/26bcd207

[3] Georgieva, E., Suta, M., & Burton, N. (2018). Hitpredict: Predicting hit songs using spotify data. Standford computer science 229: Machine learning. https://cs229.stanford.edu/proj2018/report/16.pdf

[4] Middlebrook, K., & Sheik, K. (2019). Song hit prediction: Predicting billboard hits using spotify data. arXiv. https://arxiv.org/pdf/1908.08609.pdf

[5] Raza, A. H., & Nanath, K. (2020). Predicting a hit song with machine learning: Is there an apriori secret formula? [Conference presentation]. 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia. https://doi.org/10.1109/DATABIA50434.2020.9190613