111-2 DDM Final Project

流行歌曲分析與預測

group K

財金三 廖泓傑 | 資管四 杜沛慈 | 圖資碩二 丁昱寧

Agenda

01 問題與解方

02 資料採集與處理

03 探索式資料分析

04 解釋型模型

05 預測型模型

06 結論與未來展望

07 附錄

01 問題與解方

音樂產業現況

19.55 萬元

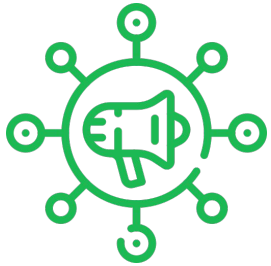
2019 年單曲平均製作成本

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

59.5 % 創作者

使用 AI 輔助創作音樂

英國 Ditto music 調查



潛在商業利益龐大

實體銷售、音樂串流、音樂下載、電影、廣播電視演出、代言、KTV伴唱帶等音樂授權

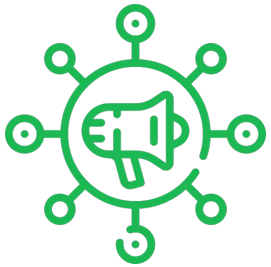
研究目的

19.55 萬元

2019 年單曲平均製作成本

59.5 % 創作者

使用 AI 輔助創作音樂



潛在商業利益龐大

解釋型模型

【音樂製作前、製作中】

找出歌曲熱門的因素

→ 避免成本浪費，輔助音樂創作

預測型模型

【音樂製作後】

預測哪些歌曲可能會熱門

→ 唱片公司簽約新人聽 demo、決定專輯收錄歌曲、主打歌

02 資料採集與處理

熱門歌曲清單

2018 ~ 2021 年榜

2010.1 ~ 2017.12 月榜

各語言 前 100 名 x 5 種語言 x 100 份排名

1998 ~ 2017 年榜

不分語言 前 100 名 x 20 份排名

手動標籤語言

熱門歌曲清單

去除重複歌曲後 .....

10893

首熱門歌曲

使用 Spotify API 搜尋冷熱門歌曲

1. 檢查熱門歌曲排行榜

較早期的歌曲缺失「專輯」欄位 (4,643 筆)

年分	語言	歌手	歌曲	專輯
2021	華語	程響	四季予你	四季予你
2019	華語	五月天	玫瑰少年	玫瑰少年
2010	韓語	2pm	Still	NaN

使用 Spotify API 搜尋冷熱門歌曲

2. 查詢熱門歌曲

使用「歌手」及「歌曲」進行 查詢

年分	語言	歌手	歌曲	專輯
2021	華語	程響	四季予你	四季予你
2019	華語	五月天	玫瑰少年	玫瑰少年
2010	韓語	2pm	Still	NaN

3. 去除無搜尋結果的資料

去除因歌手名稱語言不同或不存在 spotify 曲庫中而造成的 no result 歌曲資料

歌手	歌曲	歌曲 ID	專輯
程響	四季予你	4BGkSC	四季予你
no result			
2pm	Still	1ZjqjlM	Still 2:00pm

實際情形

Mayday	玫瑰少年	38Td9Q	玫瑰少年

使用 Spotify API 搜尋冷熱門歌曲

4. 整合熱門歌曲欄位

合併排行榜及 spotify 查詢篩選後結果的資料欄位

→ 共 4157 首熱門歌曲

年分	語言	歌手	歌曲	歌曲 ID	專輯
2021	華語	程響	四季予你	4BGkSC	四季予你
2010	韓語	2pm	Still	1ZjqjlM	Still 2:00pm

5. 查詢冷門歌曲

使用「年分」及「歌手」進行 查詢

→ 共 3704 首冷門歌曲

歌手	歌曲	歌曲 ID	專輯
程響	君不知	2RT85	长安伏妖
2pm	Crazy Babe	17JfICw	Crazy Babe

使用 Spotify API 搜尋冷熱門歌曲

6. 冷熱門歌曲特徵

透過 spotify 取得歌曲特徵值，如：曲調、音量、律動感等共 12 項

年分	語言	歌手	歌曲	歌曲 ID	專輯	曲調	音量	律動感
2021	華語	程響	四季予你	4BGkSC	四季予你	3	-6.745	0.534
2010	韓語	2pm	Still	1ZjqjlM	Still 2:00pm	0	-4.404	0.644

. . .

年分	語言	歌手	歌曲	歌曲 ID	專輯	曲調	音量	律動感
2021	華語	程響	君不知	2RT85	长安伏妖	8	-8.140	0.220
2010	韓語	2pm	Crazy Babe	17JfICw	Crazy Babe	1	-5.369	0.471

. . .

使用 Spotify API 搜尋冷熱門歌曲

7. 冷熱門歌曲標籤

設熱門歌曲為陽性、冷門歌曲為陰性

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

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	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

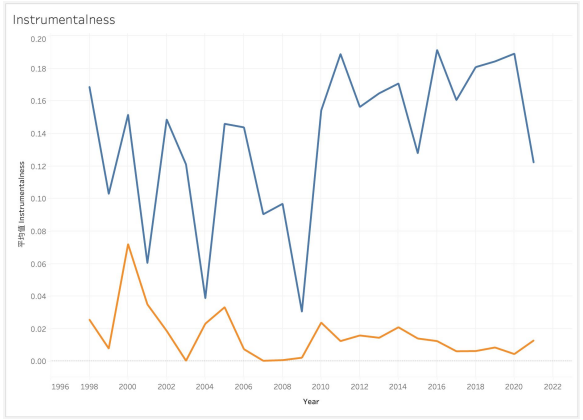
03 探索式資料分析

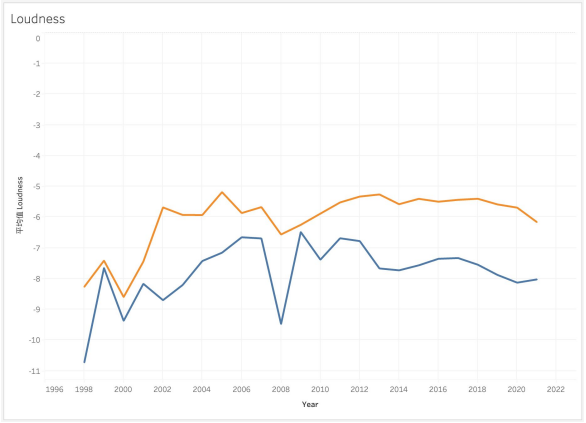
13 項 spotify audio features

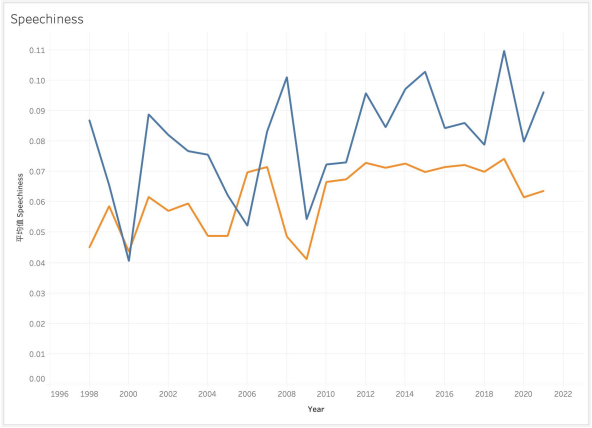
1. acousticness：原聲程度，音樂含非電子音的程度
2. danceability：律動感，適合跳舞的程度
3. duration_ms：音樂時長，單位為毫秒
4. energy：對音樂強度與活躍度的感知
5. instrumentalness：純音樂，不含人聲的佔比
6. key：曲調，0 代表 c 調，1 代表升 c，2 代表 d 調，依此類推
7. liveness：現場感，檢測錄音中是否存在觀眾
8. loudness：音量，單位為分貝
9. mode：音軌調性，1 為大調，0 為小調
10. speechiness：口說、朗誦比例
11. tempo：音軌的整體節奏速度，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
13. time_sigmature：音軌的整體拍號

冷熱門歌曲特徵差異

冷門 熱門







Instrumentalness

熱門歌曲的純音樂性較低

Loudness

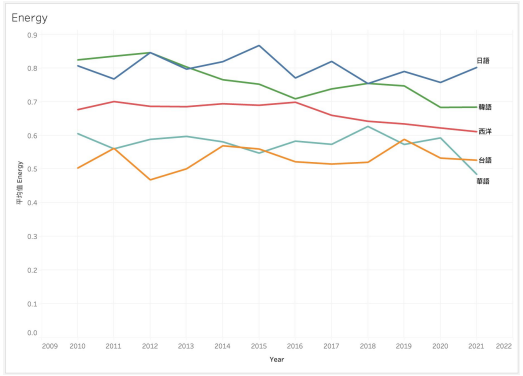
熱門歌曲的音量較大

Speechiness

熱門歌曲的朗誦比例較低

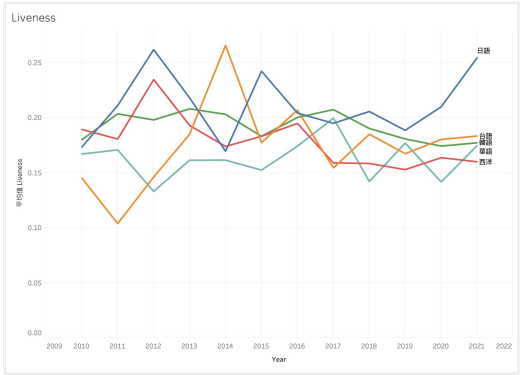
各語言熱門歌曲特徵差異

日語 台語 西洋 華語 韓語



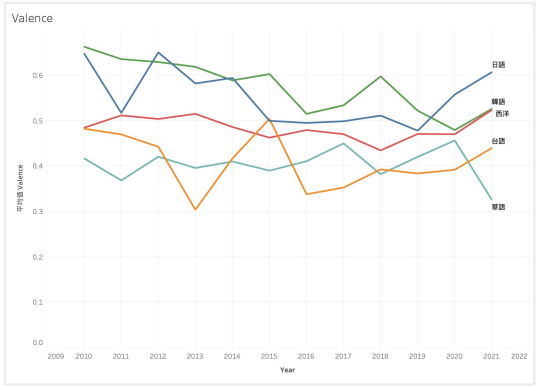
Energy

日語熱門歌曲的音樂強度最高，韓語次之，華語、台語最低



Liveness

日語熱門歌曲的現場感高於其他語言熱門歌曲

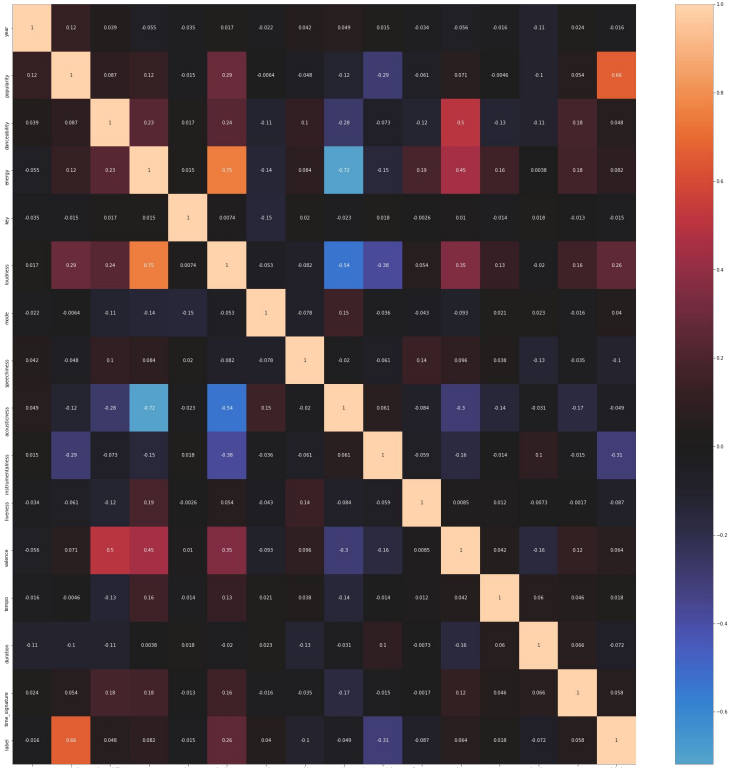


Valence

日語、韓語熱門歌曲的正向情緒較高，華語、台語較低

04 解释型模型

解釋型模型 變數篩選



1

Time_signature 變異度不足

2

Loudness, Energy 高度正相關 (0.75)

Acousticness, Energy 高度負相關 (-0.72)

→ 模型排除變數 Time_signature, Loudness, Acousticness

解釋型模型 羅吉斯迴歸 - 總體

取閥值 = 0.05, P < 0.05 代表足夠顯著

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

解釋型模型 羅吉斯迴歸 - 總體

怎樣的歌曲更有可能成為 Hot song?

大調的歌曲(1.13倍)

純朗誦比例低

純音樂比例低

有能量的歌曲

歌曲長度短(短1秒+0.24%)

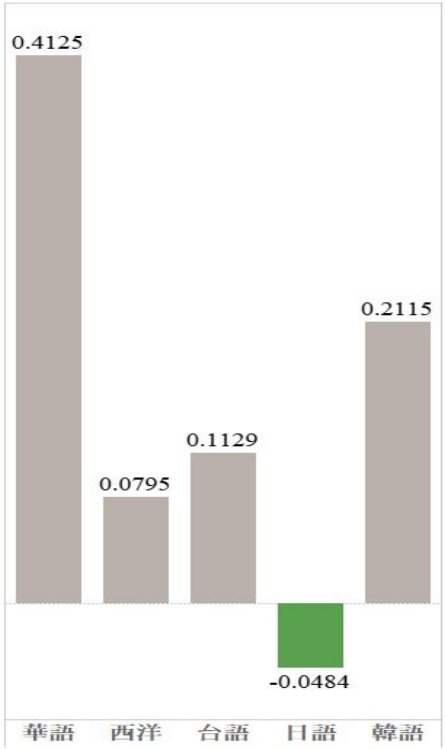
傳達負面情緒

減少現場觀眾聲

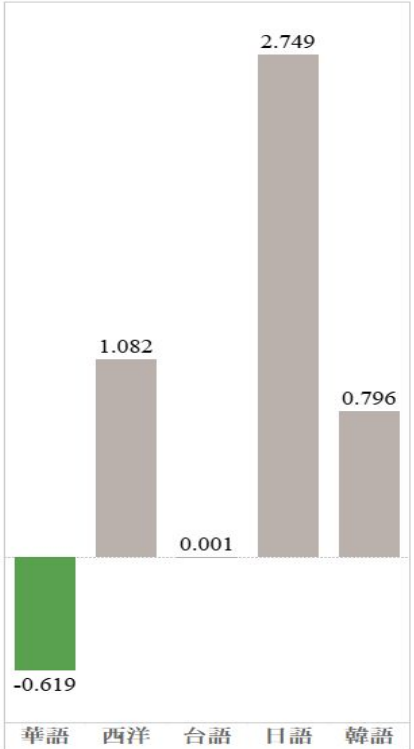
解釋型模型 羅吉斯迴歸 - 不同語言

* 使用係數觀察相對影響力

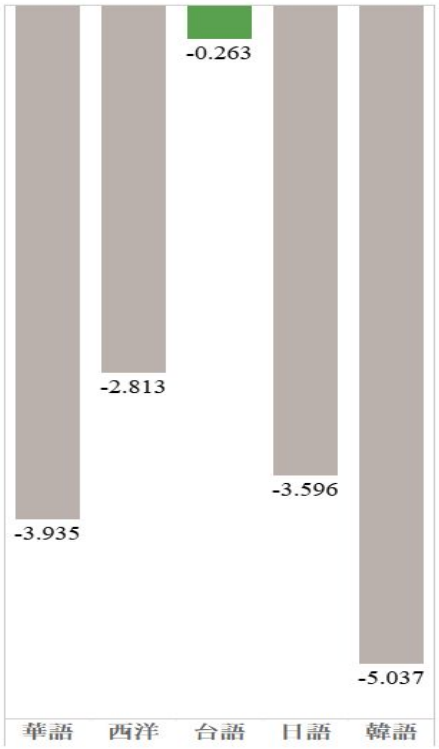
Mode



Energy



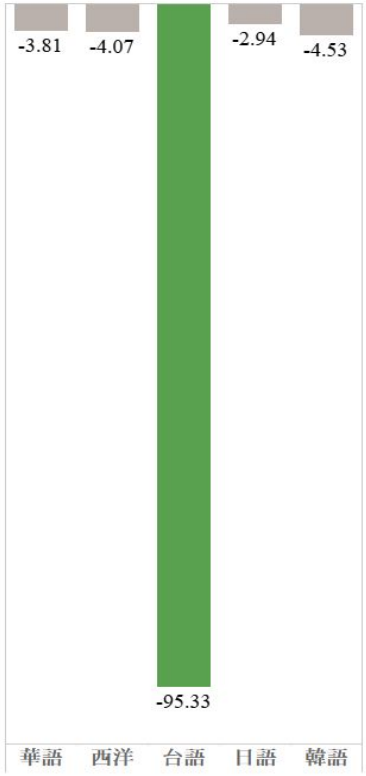
Speechiness



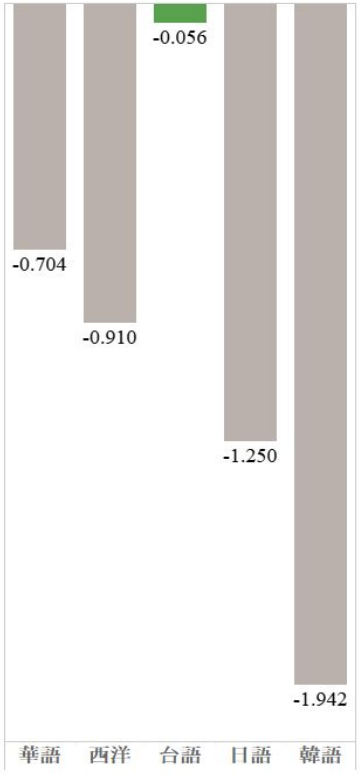
解釋型模型 羅吉斯迴歸 - 不同語言

* 使用係數觀察相對影響力

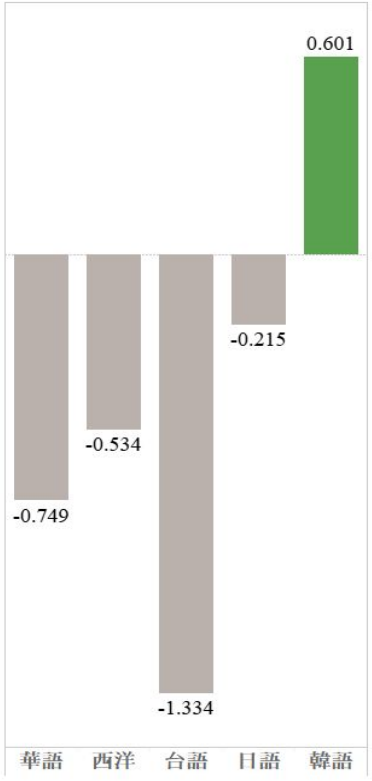
Instrumentalness



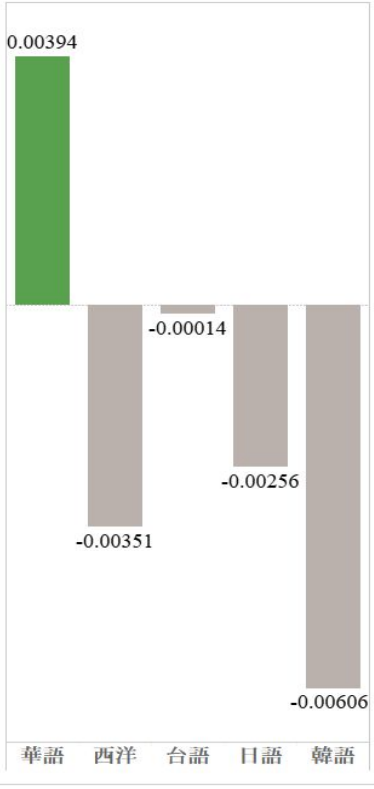
Liveness



Valance



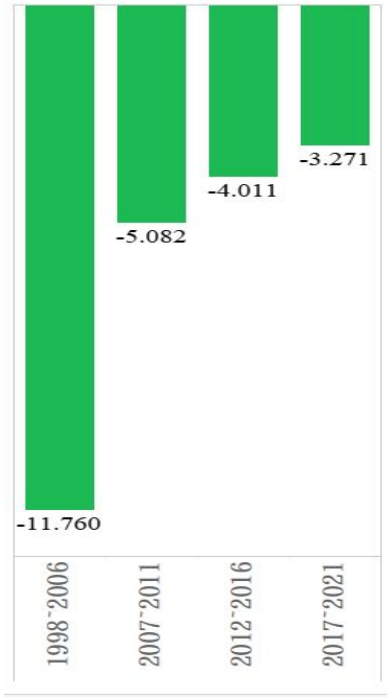
Duration



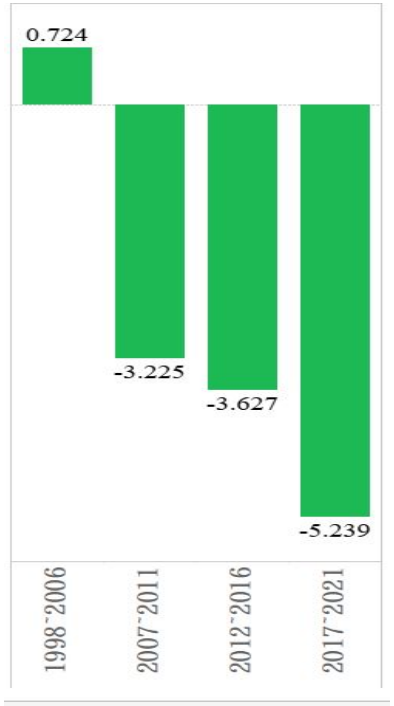
解釋型模型 羅吉斯迴歸 - 不同時間

* 使用係數觀察相對影響力

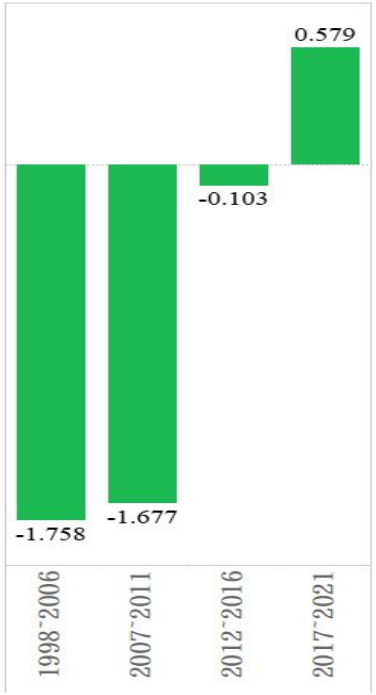
Instrumentalness



Speechiness



Valence



05 預測型模型

模型建構

- **變數選擇**
  使用 12 項歌曲特徵為自變數，冷熱門標籤為應變數

- **切分資料集**

- **候選模型**

	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

Logistic Regression 模型在 F1 score 表現最佳

Logistic	真實熱門	真實冷門
預測熱門	763	464
預測冷門	24	215
C = 1 / threshold = 0.35		

Random Forest	真實熱門	真實冷門
預測熱門	744	436
預測冷門	43	243
max_depth = 7 / threshold = 0.4		

XGBoost	真實熱門	真實冷門
預測熱門	744	436
預測冷門	43	243
alpha = 1 / threshold = 0.4		

	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

模型預測

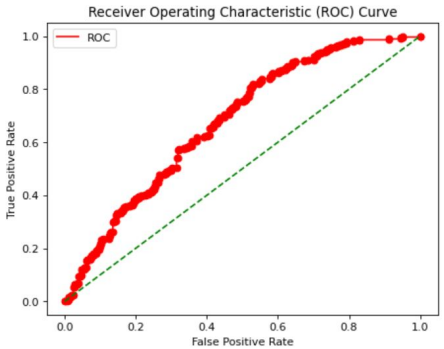
- 切分資料集

訓練集 | 測試集

1998 ~ 2020 年 | 2021 年

- 模型預測結果

Logistic	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		



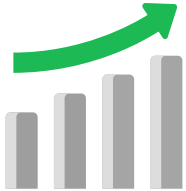
F1 score = 0.727

AUC = 0.677

06 結論與未來展望

總結

透過解釋型模型，了解總體趨勢與不同市場中影響成為熱門歌曲的因素，能夠協助歌曲創作並降低相關成本。



透過預測型模型，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。

未來展望

- 考量其他外部變因，如：歌手知名度、社交媒體的影響等
- 結合歌曲本身的因素，如：歌曲主題類型、歌詞的文本分析等，進一步優化模型
- 取得更完整、多元的資料，如：不同國家的歌曲排行榜等，可比較不同音樂市場的差異

Thanks for listening