[Spotify歌曲流行趨勢分析及預測]

|  |  |  |
| --- | --- | --- |
| 系級 | 姓名 | 學號 |
| 電機112  機械113 | 謝宗翰  操之晴 | E14084078  E14086622 |

|  |
| --- |
| Dataset 及目標 |
| * 來源：Kaggle * 數據量：232725筆 * 分類：曲風、歌手、人氣度，以及各種feature(音樂特色) * 目標：分析當前曲風及features的流行趨勢以及其中相關性，並進一步預測何種類型的音樂未來人氣度會較高。 |

# 讀取資料

|  |  |
| --- | --- |
| 資料型態 | 尋找缺失值 |
| * 11個連續型變數長條圖分析 * 7個類別型變數 散佈圖分析 | * 使用計算值總數 * 0表示無缺失資料，無須進行填補 |

# 資料視覺化與資料分析

|  |
| --- |
| 人氣密度(popularity density) |
| * 人氣密度：有多少歌曲獲得該分數 如：約1.7%(0.017)的歌曲獲得0分 * 欲觀察人氣密度分佈與走向，使用分佈圖 * 由分佈圖可得知，大約得到60分以上便有一定機率可以獲得較高人氣，進入排行榜 |

|  |  |
| --- | --- |
| 曲風與人氣度之相關性 | 拍點與人氣度之相關性 |
| * 欲觀察各類曲風的人氣高低，使用可明顯看出差距的橫向條狀圖 * pop、rap等想像中人氣很高，排行也確實位於前面 * 原以為電影原聲(Movie)會受多數人歡迎，實則不然； 以為冷門的R&B實際上分數卻相當高 | * 可以發現4/4拍的音樂是目前較受歡迎的，而1/4拍則較不為趨勢 * 黑線為誤差線 * 0/4的部分為其他類，推測是因為有些較特別的拍點有著高人氣的特例才有著明顯的誤差線。 |

|  |  |
| --- | --- |
| 調性、關係調與人氣度之相關 | 關係調與人氣度之相關 |
| * 調性：決定風格最主要的因素，如中國風、日本風等。下意識認為特定調性的分數會有明顯高低，然整體下來是平均的，沒有相差太多。 | * 調性和大小調放在一起呈現，看不出明顯差異，故將其拉出來比較。 * 可發現，小調分數略高於大調。 |

|  |  |
| --- | --- |
| 原聲與人氣度之相關 | 響度與人氣度之相關 |
|  |  |
| * 因原聲程度及響度等第過多，使用散佈圖表達其相關性。 * 然因數據量過大使其太過密集，無法清楚辨識，故將兩者前50%的資料抽出進行分佈圖呈現，以明顯看出集中範圍。 | |
|  |  |

|  |  |
| --- | --- |
| 所有特徵值之相關性 | |
|  | * 音樂的時長與響度，在現今流行中有較固定的模式。 * 單看人氣度與其他特徵值的關係可發現，要有不錯的成績與某項特徵無特別大的關係，其中最有關係的為音樂的響度。 * 我們也發現歌曲給人的正能量與響度有著很大的關聯，響度越大給人的正能量也越大。 |

# 模型訓練

## 前置作業

|  |  |
| --- | --- |
| 文字轉換 | |
|  | * 將key等類型項的資料轉換為整數項 * 其中前人氣度前2為57分，故設定57分以上的歌曲為有名的歌曲，並設定為1 |
| 資料分段 | |
| 因為只有一個資料集，故將資料集隨機抽取來當作訓練資料集，其餘做為測試資料集。 | |

## 訓練 建立四種預測模型：RandomForest、DecisionTree、KNN、回歸分析

|  |  |
| --- | --- |
| Accuracy | AUC |
|  |  |
| 不論在準確度或AUC上，Random Forest的訓練皆為最高 | |

# 模型測試

|  |  |
| --- | --- |
| Accuracy | AUC |
|  |  |
| 在測試上，Random Forest在準確度和AUC上亦為最高  且與訓練結果近乎相同 | |

# 心得

操之晴：

這次是第一次接觸機器學習，在開始前確實擔心自己不多的程式經驗無法妥善應用在其中，也擔心僅有5天的課程不能好好吸收。但實際經過這禮拜後，能明顯感受到自己在思考和實作上的進步，雖然仍未成熟，但也比預想的學到了更多。在做term project的時候，另一位組員提供了很多協助，我們也選了一個做起來相對好理解也有興趣的主題，所以整體下來是非常有趣且有意義的，感謝教授與助教這段時間內的教導和幫助。

謝宗翰：

身為一個電機系學生，之前有碰過影像處理跟邏輯設計的東西，聽說機器學習最近很紅，也感覺對這個蠻有興趣。學習過程中一直碰壁，在寫Term Projectd看完了kaggle上許多大神的文章，也一直去試圖理解他們的邏輯，但是自己去寫後才發現一堆問題，最常遇到的問題就是不知道該用哪個函式，第一次遇到這種需要大量引入函式的領域，但最後寫久了就熟悉了，撰寫速度也提升非常多。因為我們兩個人都沒有統計的背景，對於圖表或數字的分析沒那麼敏銳，常常不知用什麼圖表分析，也不太懂為什麼要那樣解釋圖表，所以在資料視覺化的解釋舊卡了蠻久的。但經過這次的磨練，我們都有進步一點點。

對於這堂課我真的學到很多，也讓我更明白我的興趣所在，未來或許也會朝這條路繼續邁進。

# 參考資料

https://www.kaggle.com/huanntran100/spotify-song-popularity-prediction?select=SpotifyFeatures.csv