Guess, which are my songs?

Team PowerTransform: Damiano Casiraghi, Giacomo Floriani

Abstract

“E’ possibile creare un algoritmo interattivo in grado in grado di capire se una canzone è di gradimento per me o per te usando shiny?”. Questa è la domanda che ci siamo posti durante la presentazione del progetto di Charlie Thompson dal titolo “fitteR happieR”. Lo scopo dell’analisi è quindi quello di creare un’applicazione shiny che sia in grado di effettuare una analisi veloce e interattiva usando i dati forniti da spotify, con lo scopo di riuscire a prevedere quali canzoni fossero di interesse per l’uno o per l’altro.

Sommario

1. Introduzione

Vista la grande passione che entrambi i componenti del gruppo hanno per la musica, abbiamo tentato di costruire un algoritmo che sappia identificare se una canzone è preferita dall’uno o dall’altro utilizzando un’app shiny. Lo scopo è quello di creare un’applicazione che possa essere usata facilmente per effettuare una analisi dello stesso tipo di quella da noi proposta ma con dati differenti.

2. Dataset

Per ottenere i dati in questione abbiamo selezionato su Spotify 500 canzoni di nostro gradimento per ciascun elemento del gruppo. Il passo successivo è stato quello di scaricare da spotify i dati inerenti a queste canzoni usando il pacchetto di Charlie Thompson “spotifyr”. In questo modo abbiamo ottenuto un dataset composto da 15 variabili di diversa tipologia, a cui ne abbiamo aggiunto altre 3 per identificare la canzone (titolo brano, album, artista) e una quarta come variabile risposta, Class.
Vediamo ora le variabili esplicative considerate:

1. Track_Name: il titolo del brano
2. Artist_name: il nome dell’artista che ha composto il pezzo
3. Album_name: il nome dell’album della canzone
4. Danceability: Quanto una traccia è adeguata per essere ballata, basandosi su una combinazione di elementi musicali che includono tempo, stabilità del ritmo, potenza dei beat e regolarità complessiva. 0 non ballabile, mentre 1 è il valore massimo
5. Energy: Misura compresa tra 0 e 1 che rappresenta la percezione di intensità e azione. Tipicamente, le canzoni energiche sono veloci, forti e rumorose
6. Key: La tonalità complessiva stimata del brano
7. Loudness: La forza complessiva di una traccia in decibel. I valori di rumorosità sono calcolati in media lungo l’intera traccia e sono utili per comparare la rumorosità di canzoni. La variabile assume valori tra -60 e 0 db
8. Mode: Modalità (maggiore o minore) della traccia, il tipo di scala dal quale il contenuto melodico è derivato. Maggiore è rappresentato da 1 e minore da 0
9. Speechiness: Determina la presenza di parole nella traccia. Più una traccia è 'parlata', più il valore si avvicina a 1. Un

valore oltre 0.66 descrive tracce interamente parlate, come audiolibri. Tracce con un valore compreso tra 0.33 e 0.66 sono riferiti a canzoni composte sia da musica che da testo

10. Acousticness: Una misura di confidenza compreso tra 0 e 1, dove 1 indica una canzone è acustica con un alto livello di confidenza
11. Instrumentalness: Ci dice se una traccia non contiene voci. Suoni come “ooh” e “aah” sono trattati come strumentali in questo contesto. Rap e canzoni parlate sono chiaramente 'vocali'. Più vicino il valore è a 1, più verosimilmente la traccia non conterrà voci. Valori intorno 0.5 si suppone rappresentino brani strumentali, ma la probabilità di esserlo aumenta quando il valore si avvicina a 1
12. Liveness: Indaga sulla presenza di pubblico durante la registrazione. Alti valori di liveness rappresentano alta probabilità che la tracccia sia live
13. Valence: Una misura da 0 a 1 che descrive la positività trasmessa da una traccia.
    Brani con un’alta valenza sembrano più positivi, mentre pezzi con valori più bassi sembrano più negativi
14. Tempo: Il tempo stimato complessivo di una traccia calcolato in beat per minuto (BPM). In musica, il tempo è la velocità o il ritmo di un certo pezzo e deriva direttamente dalla durata medie dei beat
15. Track_uri: Un HTTP URL per accedere alla analisi completa della traccia
16. Duration_ms: la durata della traccia in millisecondi
17. Time_signature: E' una stima della metrica generale della traccia. La metrica è la notazione convenzionale per specificare il numero di beat per battuta
18. Key_mode: Una stima complessiva della tonalità del brano e della sua modalità
19. Class: Variabile di interesse: 1 indica che la canzone è stata scelta da Damiano, 0 invece da Giacomo

3. Analisi

Dopo aver scaricato i dati abbiamo iniziato l’analisi vera e propria.

Prima di ogni cosa abbiamo separato il nostro dataset in Train e test set (rispettivamente 70%-30% delle osservazioni) per poi lavorare ovviamente sul primo per l’implementazione del modello.

Per prima cosa abbiamo fatto un’analisi esplorativa escludendo le variabili Artist_name, Track_name, Album_name e track_uri poiché non sarebbero mai state utili in termini implementare algoritmi. Dopo questa selezione abbiamo osservato la correlazione tra tutte le variabili esplicative e la nostra risposta Class. Ciò che n’è emerso è che le due variabili maggiormente correlate con Class sono “danceability” e “valence”; abbiamo anche osservato la correlazione tra le due per vedere se essa fosse molto alta, poiché avrebbe portato problemi di multicollinearità, ma non era così. In seguito abbiamo pensato di ottenere delle conferme dal punto di vista grafico alle conclusioni tratte osservando la correlazione. Guardando i box-plot condizionati della variabile Class rispetto alle variabili esplicative ancora una volta “danceability” e “valence” erano le due variabili che discriminavano maggiormente, anche altre variabili hanno mostrato una buona discriminazione motivo per cui non tutte verranno escluse nei modelli. Posteriormente ai box-plot abbiamo osservato il grafico a dispersione tra le due variabili di maggiore influenza mettendo in evidenza quali osservazioni fossero per l’uno o per l’altro componente del gruppo; da questo la separazione tra individui era forte e netta, indice che queste variabili svolgono realmente un ruolo importante.

Finita l’analisi esplorativa abbiamo cominciato con l’implementazione dei modelli, testandone diversi tipi (RandomForest, XGBoost, LASSO, GLM) con diversi gruppi di variabili. Alcune di queste tecniche restituiscono come default un grafico che indica l’importanza delle variabili (tutti tranne GLM), che ci ha aiutato moltissimo nella selezione dei modelli migliori. La nostra valutazione di ciò è stata fatti in termini di AUC, poiché osservare solo la percentuale di previsioni corrette sul test set

potrebbe essere malamente influenzata da un overfitting. Infine il modello che risultava essere migliore per quanto riguarda questo indice è la RandomForest con uno score del 96,38%.

4. Conclusioni

In conclusione possiamo dire che la risposta alla nostra domanda è "Si". Siamo soddisfatti dei risultati ottenuti, sia per quanto riguarda l'analisi vera e propria, che ci ha permesso di raggiungere una percentuale di corrette osservazioni del 90%, sia a livello di applicazione. Quest'ultima si è rivelata infatti veloce e semplice da utilizzare, motivo per cui potrebbe essere utilizzare facilmente su nuovi dati.
Il nostro lavoro si è basato sul classificare le canzoni di una o dell'altra persona poiché si trattava di un lavoro di gruppo, ma il problema potrebbe essere tradotto in un altro modo: "Quali canzoni possono piacermi, quali invece no?". Questo problema potrebbe essere molto utile ad esempio quando decidiamo di ascoltare un nuovo gruppo, in questo modo infatti potremmo avere un'idea iniziale di quali canzoni potrebbero essere di nostro gradimento e quali invece no.

5. Sitografia

- Fitter happier, by Charlie Thompson
- Spotify API
- Get audio features from Spotify
- SportifyR
- Shiny:
  - Video tutorial
  - Written tutorial
  - Projects
  - Widget
  - Shiny dashboard