Progetto Machine Learning

Spotify award

Gianluca Giudice - 830694

Marco Grassi - 829664

Contents

Chapter 1

Introduzione

Introduzione

1.1 Descrizione del problema

Vogliamo sapere se una canzone sarà di successo

1.1.1 Spotify

1.1.2 Premi per i singoli musicali

Metrica oggettiva per il successo di una canzone

1.1.3 Approccio al problema

1.1.4 Struttura del codice

Chapter 2

Dataset

2.1 Acquisizione del dataset

2.1.1 Dataset spotify - Kaggle

2.1.2 Premi canzoni - Wikipedia

Scraping

2.2 Descrizione del dataset

2.3 Normalizzazione

OpenRefine con regex per le date

2.4 Data integration

2.4.1 Modellazione struttura documentale

2.4.2 Record linkage con MongoDB

Una roba veloce giusto per spiegare

1. Import dati mongodb
2. Creazione indici
3. Preprocessing lowercase
4. Unfold campo artista
5. Record linkage -¿ Join basandosi su titolo e artisti

6. Lista artisti come stringa
7. Dump del database in un file .csv come risultato del dataset per i modelli

2.5 Analisi esplorativa

2.5.1 Distribuzione dei valori

Variabili continue

Variabili categoriche

2.5.2 Artisti nelle canzoni

Frequenza artisti

Wordcloud

2.5.3 Correlazione features

2.5.4 Principal component analysis

Prima di PCA normalizzazione del dataset

NB: Non facciamo il plot delle prime due componenti principali in quanto varianza spiegata dalle prime due componenti molto bassa

2.5.5 Creazione dataset bilanciato

Undersampling

2.6 Scelta delle features

2.6.1 Normalizzazione e standardizzazione

Conversione in lowecase Converto in variabili numeriche e factor + label corretta

2.6.2 Dataset proiettato componenti principali

2.6.3 Variabili categoriche

Key -¿ Trasformazione a intero

2.6.4 Artisti in una canzone - BOW

Threshold freqeunza

Viene usato un threshold a 2 per non avere delle canzoni con artisti completamente sconosciuti. Assumiamo quindi di fare previsioni su canzoni cantante da artisti un minimo conosciuti.

Questa assunzione non è molto restrittiva, ci immaginiamo infatti che per vincere un premio una canzone deve essere cantata da artisti non completamente sconosciuti

Inoltre a causa dell'undersampling è possibile avere diverse canzoni cantate da artisti "sconosciuti" e non avendo scelto una particolare strategia per l'undersampling adottiamo a questo punto l'utilizzo di un threshold

2.6.5 Popularity

Chiaramente questa feature viene scartata per non creare modelli biased (Non si conosce questa feature quando una canzone esce)

Chapter 3

Campagna sperimentale

3.1 Approccio

3.1.1 10-folds cross validation

3.1.2 Training set e test set

3.1.3 Model selection

Ottimizzazione iperparametri

Grid search

3.2 Misure di performance

3.2.1 Accuracy

3.2.2 Precision, Recall e F-measure

3.2.3 Curve ROC

3.3 Support Vector Machine

3.3.1 Kernel

3.4 Decision Tree

3.5 Esperimenti

3.5.1 Performance

3.5.2 Modelli a confronto

Chapter 4

Conclusioni