Progetto di Data Analytics
A.A. 2022-2023

Daniele Perrella 1038893
Simone Boldrini 1038792

13 marzo 2023

Indice

Capitolo 1

Introduzione

L'obiettivo di questo report è presentare il progetto del corso di Data Analytics finalizzato alla predizione del voto medio di un film, date le sue caratteristiche, utilizzando un dataset proveniente da MovieLens, un recommendation system per contenuti video. Il dataset contiene rating e tag per oltre 60.000 film, raccolti da più di 150.000 utenti negli anni 1995-2019. Ogni file del dataset dispone di un genoma che identifica una caratteristica del film e la sua rilevanza.

Per raggiungere questo obiettivo, abbiamo utilizzato tecniche di Machine Learning supervisionate tradizionali quali Linear Regression, SVM, NB in seguito verranno illustrate nel dettaglio; tecniche di ML basate su Reti Neurali ed infine modelli deep per Tabular Data. In particolare, il report descrive il processo di acquisizione e preparazione dei dati, l'analisi esplorativa del dataset, la selezione delle feature rilevanti e la costruzione dei modelli predittivi.

Infine, il report conclude con una valutazione critica dei modelli creati, evidenziando i loro punti di forza e di debolezza, e suggerisce possibili sviluppi futuri per migliorare ulteriormente la predizione del voto medio dei film.

Capitolo 2

Metodologia

2.1 Data Acquisition

In uno studio di Data Analysis, lo step di Data Acquisition è il primo della pipeline da seguire. In questa fase abbiamo raccolto i dati necessari dal dataset [1, MovieLens]. I file sono stati raccolti in una cartella denominata `ml-25m` contente un README.

2.2 Dataset

Movielens è un Recomendation System per contenuti multimediali, quali film, serie tv, documentari ecc. Movielens mette a disposizione degli sviluppatori un dataset opensource, generator dal database TMDB (The Movie Database). Questo datasedt contiene recensioni con voti e tag di oltre 60.000 film, raccolte da oltre 150.000 utenti durante il periodo che va dal 1995 fino al 2019.
Il dataset è composto dai seguenti file:

- *genome-scores.csv*: contiene il *relevance score* di ogni tag per tutti i film (ovvero, quanto un tag è importante per il dato film).
- *genome-tags.csv*: contiene tutti i tag presenti all'interno del dataset.
- *links.csv*: contiene gli ID di ogni film per i due database TMBD e IMDB.
- *movies.csv*: contiene i titoli dei film, con i rispettivi generi.
- *ratings.csv*: contiene più di 25.000.000 votazioni provenienti dalle recensioni degli utenti.

- *tags.csv*: contiene i tag che sono stati assegnati ai film dagli utenti nelle rispettivi recensioni.

Per il nostro caso di studio, non abbiamo fatto uso di tutti i file disponibili, ma esclusivamente quelli che abbiamo considerato adatti per lo scopo. Di seguito viene mostrata la struttura dei tali.
Il dataset ***ratings*** eèstato selezionato per poter far uso dei voti assegnati ai film dagli utenti, per poter quindi ottenere il voto mediio di tali film. La struttura è la seguente

userId	**movieId**	**rating**	**timestamp**
1	296	5.0	1147880044
1	306	3.5	1147868817
1	307	5.0	1147868828
1	665	5.0	1147878820
1	899	3.5	1147868510
...	...	...	...

Tabella 2.1: ratings.csv

Il dataset ***genome-scores*** è stato selezionato per poter identificare le relazioni che ci sono tra un dato voto ed i valori dei tag assegnati

movieId	**tagId**	**relevance**
1	1	0.028749999999999998
1	2	0.023749999999999993
1	3	0.0625
1	4	0.07574999999999998
1	5	0.14075
...	...	...

Tabella 2.2: genome-scores.csv

Infine, si è fatto uso del dataset ***movies*** per ottenere le informazioni inerenti al genere di ogni film.

movieId	title	genres
1	Toy Story (1995)	Adventur|Animation|...
2	Jumanji (1995)	Adventure|Children|...
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|...
...	...	...

Tabella 2.3: movies.csv

2.3 Data Visualization

Nella fase di *data visualization* andremo a visualizzare alcune proprietà e peculiarità del dataset.

Distribuzione dei ratings

Nella seguente figura, è mostrata la distribuzione dei voti sul dataset.

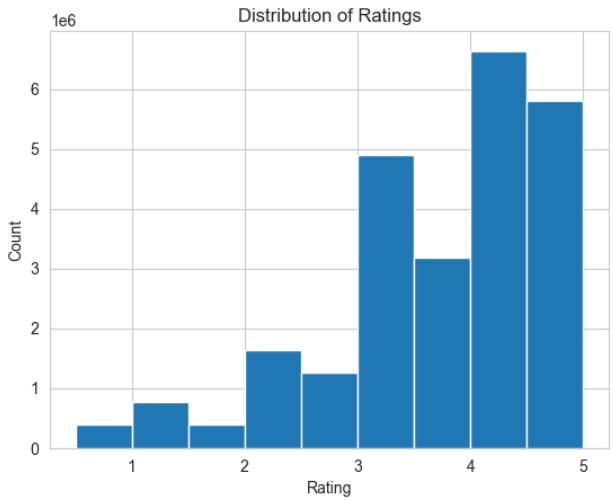

Figura 2.1: Distribution of Ratings.

Il gafico mostra la distribuzine dei vbti sul dataset, evidenziando un chiaro sbilanciamento nella distribuzione. In particolare, si può osservare una concentrzione significativamente maggiore di voti con valore di 3 o superiore rispetto ai voti inferiori. Questo sbilanciamento è ulteriormente evidenziato dal fatto che i voti con valore di 3 o superiore sono 6 o più volte maggiori rispetto ai voti inferiori. Tale distribuzione suggerisce una maggiore prevalenza di valutazioni positive rispetto a quelle negative o neutre nel dataset.

Questo risultato potrebbe essare utile per comprendere meglio le caratteristihe del dataset, ad esempio potrebbe suggerire la presenza di una tendenza positiva nelle valutazioni, o la necessità di bilancuiare meglio el categorie di voto.

Relazione tra numero di voti e voto medio

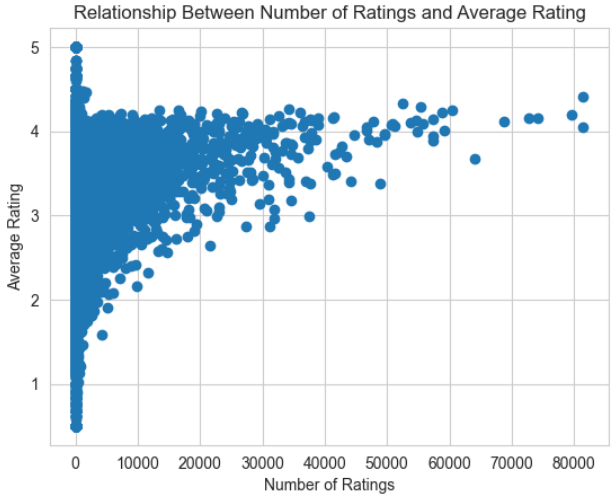

Figura 2.2: Relationship between Number of Ratings and Average Rating.

Voto medio per Genere

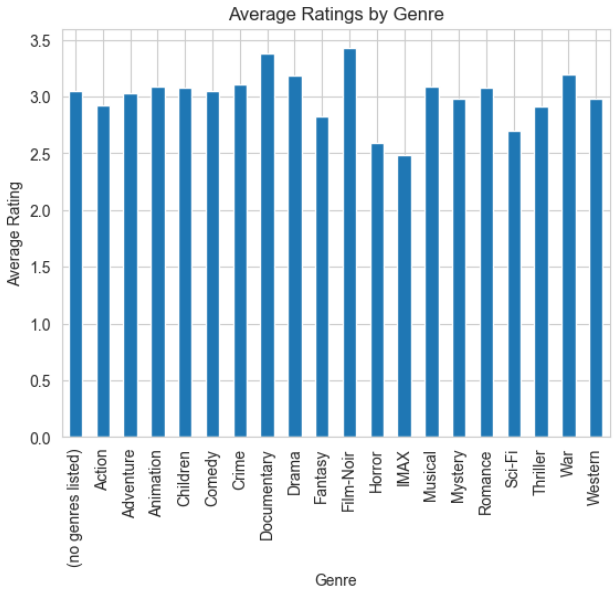

Figura 2.3: Average Ratings by Genre.

2.4 Tabular Data

Capitolo 3

Implementazione

3.1 Data Preprocessing

La fase di preprocessing dei dati consiste nel preparare i dati per l'addestramento del modello. Una volta acquisiti i dati e averne studiato le peculiaritá e le caratteristiche, si procede con la pulizia dei dati. Abbiamo usato il metodo *pivot_table* di Pandas per creare una tabella pivot che contiene i genome-score degli utenti per ogni film che lo possiede(andando cosi a ridurre la cardialita' da 60k a 13.816).
In seguito a ció verrá effettuata una merge con un i generi dei film applicando una One-Hot Encoding, attraverso la funzione *get_dummies()*. Ci siamo inoltre andati a focalizzare sui voti da parte degli utenti, raggruppandoli per film ci abbiamo calcolato la media dei voti. Quest'ultimo capo sará il nostro *target* per l'addestramento del modello.

3.2 Modeling

La nostra fase di modellazione include uno studio di regressione basato sulla predizione del voto medio di un film dati i suoi genome-score(ed in seguito anche i generi). Abbiamo inoltre applicato una tecnica di *PCA* per ridurre la cardinalitá dei dati, ma non abbiamo ottenuto risultati soddisfacenti, quindi abbiamo deciso di non applicarla.

Eseguiamo il train-test split con un rapporto 80-20 ed in seguito addestreremo il nostro modello.

3.2.1 Tecniche di ML Supervisionate con Approccio non-Deep

In questa sezione verranno descritte le tecniche di ML supervisionate con approccio non-Deep utilizzate per la predizione del voto medio di un film.

- **Linear Regression**: La regressione lineare é una tecnica di ML supervisionata che permette di predire il valore di una variabile dipendente a partire da una o piú variabili indipendenti.
  - –
  - **Lasso**: Lasso é un algoritmo di ML supervisionato che permette di effettuare la regressione di un dato mediante la creazione di un modello lineare.
- **Random Forest**: Random Forest é un algoritmo di ML supervisionato che permette di effettuare la classificazione o la regressione di un dato mediante la creazione di piú alberi (Metodo di Emsable Learning).
- **SVR**: Support Vector Regression é un algoritmo che cerca di trovare una funzione che minimizzi la distanza tra i dati di training e una fascia di tolleranza definita dall'utente.
- **KNN**: K-Nearest Neighbors é un algoritmo di ML supervisionato che permette di effettuare la classificazione o la regressione di un dato cerca di stimare il valore di una variabile dipendente su nuovi dati in base alla loro vicinanza ad altri dati di training.
- **NB Gaussian**: Naive Bayes é un algoritmo di ML supervisionato che permette di effettuare la classificazione o la regressione di un dato mediante la creazione di un modello probabilistico.

Per ogni modello precedentemente citato é stato applicato il Tuning dei parametri per cercare di ottimizzare il modello. Attraverso la funzione *RandomizedSearchCV* di Scikit-Learn abbiamo cercato di trovare i migliori parametri per ogni modello.

3.2.2 Tecniche di ML Supervisionate con Reti Neurali

Tabular Data

Capitolo 4

Risultati

4.1 Performance Evaluation

4.2 Performance Evaluation on Tabular Data

Bibliografia

[1] GroupLens Research. Movielens. http://www.grouplens.org/node/73, 1998. Accessed: 2012-11-01.