REPORT PROGETTO

Data and Web Mining ~ 2020/21 ~ Enrico Pittini 877345

INTRODUZIONE

Scopo di questo documento è quello di dare una spiegazione riassuntiva del progetto. Se si è interessati all'esposizione complessiva del progetto, consultare file README.

Innanzitutto viene data, qui nell'introduzione, una breve panoramica sull'approccio usato in questo progetto. Successivamente vengono descritti i passaggi più significativi. Infine sono mostrati brevemente i risultati finali del progetto.

L'approccio usato in questo progetto è stato il seguente. Innanzitutto si è svolta un' analisi del dataset e delle varie features. Successivamente sono state aggiunte le varie features in modo incrementale, considerando per ciascuna varie alternative di lavorazione ed estraendo diversi modelli, facendo tuning dei parametri : alla fine l'alternativa e il modello selezionati sono quelli con score cross validation minore (MSE). I modelli usati sono quattro, ma in tale report sono mostrati solo i risultati relativi a random forest (risulta sempre il modello migliore).

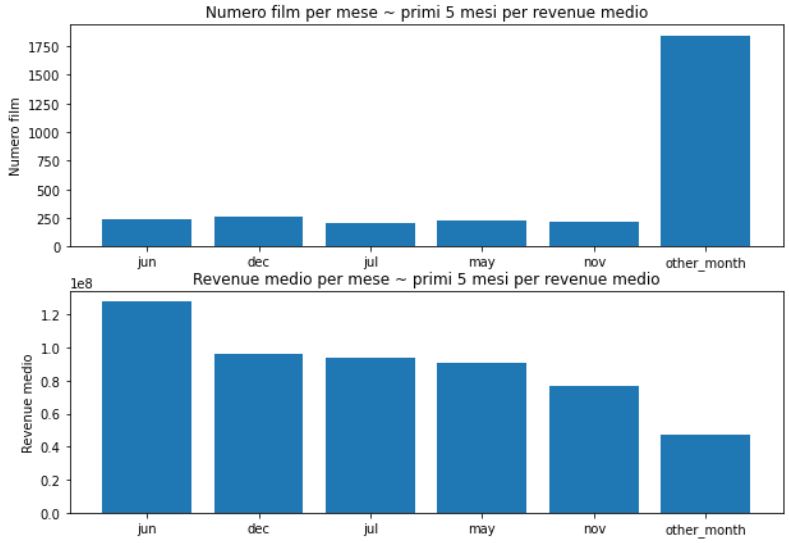
PASSAGGI PIÙ SIGNIFICATIVI

Sono mostrate ora le analisi, le alternative e le scelte più significative. I tre passaggi sono stati scelti non solo perché più rilevanti, ma anche perché strategie simili vengono usate in altri punti del progetto.

1 - MESE DI USCITA

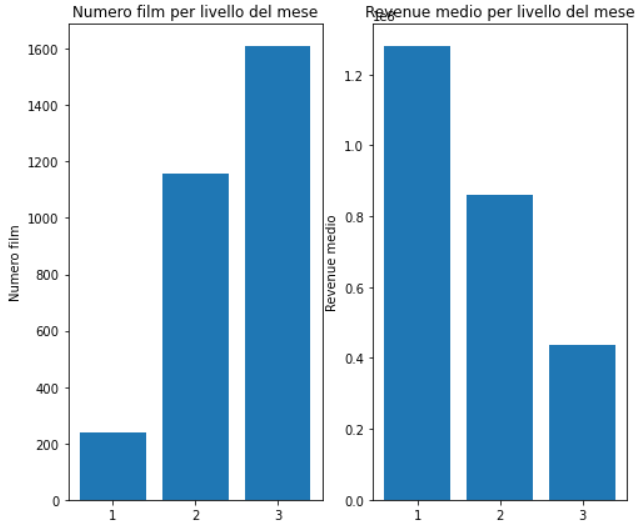
Tale informazione è estrapolata dalla feature "release_date".
Sono state proposte e valutate 6 diverse alternative circa questa feature. Ne vengono mostrate ora 3.

Nella prima alternativa (alternativa 3 del progetto) si tengono come mesi possibili solo i primi 5 mesi con associato revenue medio maggiore : tutti gli altri mesi li considero come "other_month". Da questi 6 valori creo 6 features dummy, che segnalano se il film ha o no quel mese di uscita.



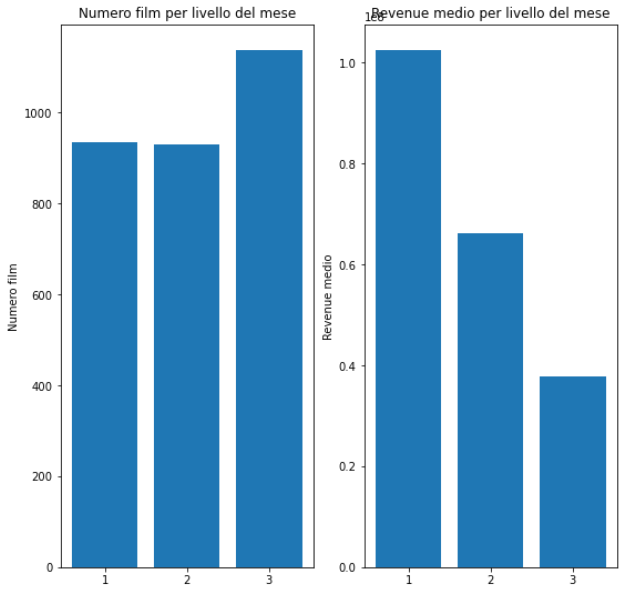
Come si vede, i 6 valori risultano sbilanciati e poco significativi rispetto al numero film.

Nella seconda alternativa (alternativa 5 del progetto) sono creati 3 livelli : il primo mese per revenue medio ; i successivi 5 mesi per revenue medio ; i restanti 6 mesi. Rimangono 3 valori → creo le corrispondenti 3 features dummy.



Anche qua i valori risultano sbilanciati per numero film.

Nella terza alternativa (alternativa 6 del progetto) sono creati 3 livelli : primi 4 mesi migliori rispetto a revenue medio; successivi 4 mesi migliori ; ultimi 4 mesi. Dunque sempre 3 livelli, ma questa volta più bilanciati. E sono comunque discriminanti rispetto al revenue.
Creo le corrispondenti 3 features dummy.



Si sono osservati i seguenti score.

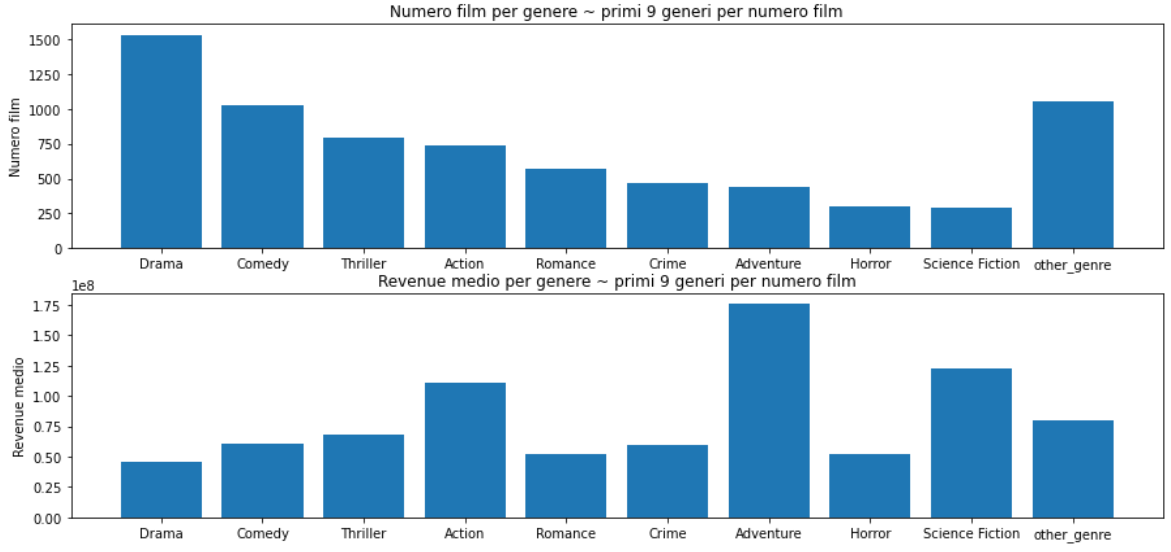
- Alternativa 1.
  MSE minore : 0.002686. Random forest con 41 alberi.
- Alternativa 2.
  MSE minore : 0.002688. Random forest con 30 alberi.
- Alternativa 3.
  MSE minore : 0.002672. Random forest con 39 alberi.

L'alternativa migliore dunque risulta essere la 3.

2 - GENERE

Sono state proposte e valutate 2 alternative. Qua ne si presenta solo una.

Si tengono come generi solo i primi k generi rispetto al numero di film: tutti gli altri generi li considero come "other_genre". Da questi k+1 valori creo k+1 features dummy, che segnalano se il film ha o no quel genere.
Si possono valutare diversi valori di k. (Nel grafico esempio con k=9)



Si hanno provato i seguenti valori di k : 12, 9, 6. Ecco i risultati.

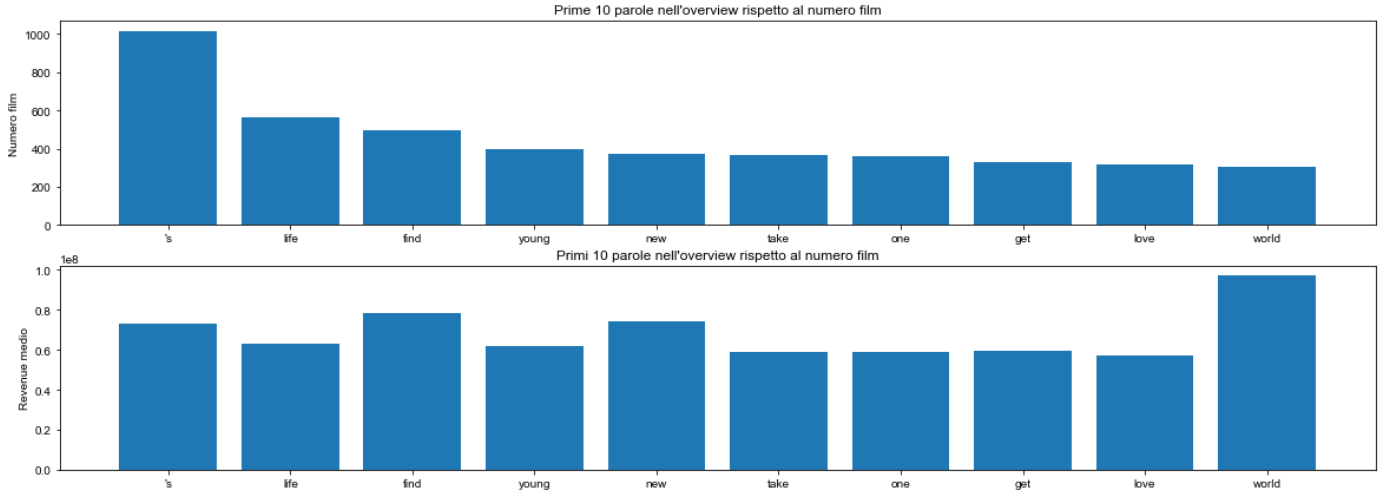
- k=12
  MSE minore : 0.002658. Random forest con 41 alberi.
- k=9
  MSE minore : 0.002630. Random forest con 49 alberi.
- k=6
  MSE minore : 0.002646. Random forest con 49 alberi.

L'alternativa migliore dunque risulta essere quella con k=9.

3 - OVERVIEW

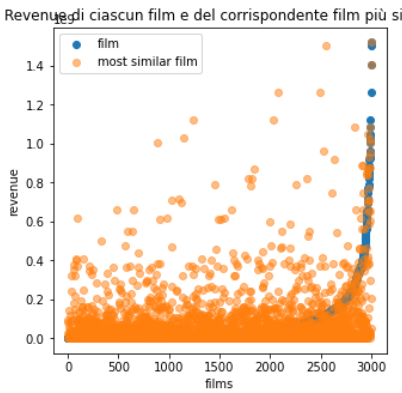
E' una feature di tipo testuale.
Sono state proposte e valutate 3 alternative.

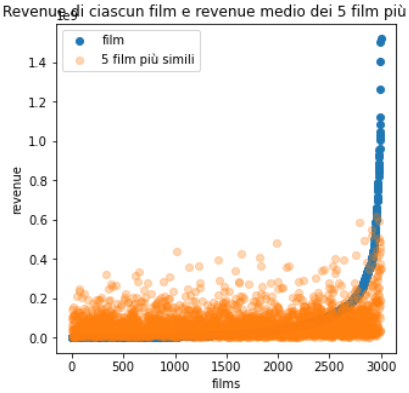
Nella prima alternativa, ogni overview è trasformato nell'insieme delle sue parole : tali parole sono state trovate usando lemmatize e togliendo le stopwords. Non si considera dunque la molteplicità.
A questo punto sono create k features dummy, relative alle prime k parole presenti in più overview.
Ci sono vari possibili k.



Nella seconda alternativa si considera la molteplicità. E' costruito il vector space, scalato con l'inverse document frequency. Sono create k features, relative alle prime k parole più importanti nel vector space (importanza di una parola è somma della sua colonna). La feature aggiunta relativa ad una data parola è semplicemente la sua colonna del vector space.
Ci sono vari possibili k.

Infine, nella terza alternativa si costruisce matrice delle similarità tra film usando tale vector space. Si aggiunge un'unica feature, che è il revenue medio dei primi k film più simili al dato film.
Ci sono vari possibili k.
(Nei seguenti grafici i film sono ordinati per revenue crescente)





(Non sembra essere molto significativa, ma è valutata comunque)

Si sono osservati i seguenti score (vengono mostrati solo alcuni).

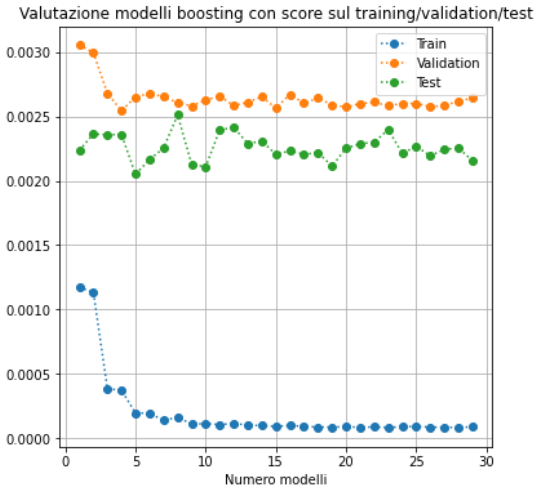
- Alternativa 1, k=3
  MSE minore : 0.002574. Random forest con 79 alberi.
- Alternativa 2, k=3
  MSE minore : 0.002600. Random forest con 82 alberi.
- Alternativa 3, k=1
  MSE minore : 0.002636. Random forest con 40 alberi.

Nel complesso, si ha ottenuto lo score migliore usando l'alternativa 1 con k=3

RISULTATI FINALI

Alla fine dell'analisi e dell'aggiunta incrementale delle features si ha ottenuto un dataset con 62 attributi totali.
Effettuando la recursive feature elimination cross validation sono state eliminate 14 features, rimanendo dunque con 48.

Modello migliore è risultato random forest con 79 alberi. Tale modello ha Bias come componente maggioritaria del suo Error : è stato dunque applicato il boosting (Ada Boost) per migliorarlo ulteriormente. In questo modo lo score cross validation è stato ulteriormente abbassato.



Modello migliore dunque è Ada Boost con 4 random forest con 79 alberi.
Lo score finale ottenuto sul proprio test set è 0.002224. (Test set ottenuto splittando il dataset fornito)