PROGETTO

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

LEONARDO MASSARO

Introduzione

Come oggetto di studio ho preso in esame i primi 15 episodi della serie televisiva **The Office**. Si tratta del remake americano dell'omonima serie cult britannica, ideata e scritta da Ricky Gervais e Stephen Merchant e trasmessa dal 2001 al 2003. La serie è stata acclamata dal pubblico e dalla critica, aggiudicandosi diversi premi, tra cui un Peabody Award nel 2006, due Screen Actors Guild Award, un Golden Globe per la performance di Steve Carell e cinque Primetime Emmy Awards, incluso uno nel 2006 per la miglior serie comedy.

Attori e personaggi

Steve Carell: Michael Scott	Rainn Wilson: Dwight Schrute	John Krasinski: Jim Halpert
Jenna Fischer: Pam Beesly	B. J. Novak: Ryan Howard	Ed Helms: Andy Bernard
Brian Baumgartner: Kevin Malone	Leslie Baker: Stanley Hudson	Kate Flannery: Meredith Palmer
Angela Kinsey: Angela Martin	Oscar Nuñez: Oscar Martinez	Phyllis Smith: Phyllis Lapin-Vance
David Denman: Roy Anderson	Melora Hardin: Jan Levinson	Mindy Kaling: Kelly Kapoor
Paul Lieberstein: Toby Flenderson	Creed Bratton: Creed Bratton	Craig Robinson: Darryl Philbin
Rashida Jones: Karen Filippelli	Amy Ryan: Holly Flax	Ellie Kemper: Erin Hannon
Zach Woods: Gabe Lewis	James Spader: Robert California	Catherine Tate: Nellie Bertram
Clark Duke: Clark Green	Jake Lacy: Pete Miller	

Come prima passo del mio studio ho estrapolato i dati necessari per la mia analisi, attraverso la visione di ogni episodio e la trascrizione dei dati a me richiesti.

DATI:

Rating: voto medio dato dagli spettatori (fonte: imdb).

Durata media delle scene: media del tempo di ogni scena all'interno di un episodio.

Numero di scene distinte: numero di scene diverse all'interno di un episodio (es. scena 1= ufficio, scena 2= ristorante).

Numero di cambi scena: numero di cambi scena all'interno di un episodio (es. scena 1= ufficio, scena 2= ristorante si avrà un solo cambio scena).

HHI battute: l'indice Herfindahl (noto anche come Herfindahl–Hirschman Index, HHI o talvolta HHI-score) è una misura della dimensione delle aziende in relazione al settore in cui si trovano e un indicatore della quantità di concorrenza tra di loro.

Nella mia analisi l'indice HHI misura la distribuzione delle battute e identifica la distribuzione omogenea o semi monopolistica.

Può variare da 0 a 1,0:

- Un H inferiore a 0,01 (o 100) indica un settore altamente competitivo
- Un H inferiore a 0,15 (o 1.500) indica un settore ben distribuito.
- Un H compreso tra 0,15 e 0,25 (o tra 1.500 e 2.500) indica una concentrazione moderata
- Un H superiore a 0,25 (sopra 2.500) indica un'alta concentrazione.

HHI scene: l'indice HHI misura la distribuzione dei minuti per scena e identifica la distribuzione omogenea o semi monopolistica.

Può variare da 0 a 1,0:

- Un H inferiore a 0,01 (o 100) indica un settore altamente competitivo
- Un H inferiore a 0,15 (o 1.500) indica un settore ben distribuito.
- Un H compreso tra 0,15 e 0,25 (o tra 1.500 e 2.500) indica una concentrazione moderata
- Un H superiore a 0,25 (sopra 2.500) indica un'alta concentrazione.

TABELLA DEI DATI

Episodi	Rating	Dur_media_scene	N_scene_distinte	N_cambi_scena	HHI_battute	HHI_scene
1	7.4	4.4	3	4	0.10	0.20
2	8.3	21.0	1	0	0.23	0.09
3	7.7	4.4	3	4	0.16	0.12
4	8.0	1.5	3	14	0.14	0.07
5	8.4	1.6	3	13	0.26	0.02
6	7.7	3.1	2	6	0.17	0.24
7	8.7	5.2	3	3	0.18	0.9
8	8.2	4.2	3	4	00.11	0.01
9	8.3	1.7	3	11	1.0	0.7
10	8.3	7.0	1	2	0.9	0.9
11	8.1	7.3	1	2	0.26	0.15
12	8.1	4.4	2	4	0.2	0.13
13	8.6	1.6	4	12	0.14	0.11
14	8.1	21.0	1	0	0.18	0.09
15	8.4	3.5	3	5	0.17	0.23

ANALISI

Strumenti utilizzati

Lo scopo dell'analisi è quello di individuare quanto le variabili sopra riportate sia influente sul rating espresso dagli utenti.

La regressione lineare è lo strumento utilizzato per l'analisi per prevedere il valore di una variabile in base al valore di un'altra variabile.

la variabile che si desidera prevedere viene chiamata "Variabile Dipendente" (rating), invece la variabile che si utilizza per prevedere il valore dell'altra variabile si chiama "Variabile Indipendente".

Questa forma di analisi stima:

- i coefficienti dell'equazione lineare
- implica una o più variabili indipendenti che meglio predicono il valore della variabile dipendente.

L'output dell'analisi corrisponde a una linea retta o a una superficie che minimizza le discrepanze tra i valori di output previsti ed effettivi.

Esistono semplici calcolatrici di regressione lineare che usano un metodo detto dei "minimi quadrati" per trovare la retta ottimale per una serie di dati accoppiati. Quindi, si calcola il valore di X (variabile dipendente) da Y (variabile indipendente).

È possibile eseguire il metodo di regressione lineare in vari programmi e ambienti, che includono:

R: è un linguaggio di programmazione e un ambiente software per il calcolo statistico e la grafica, essa è una multipiattaforma, funziona infatti su diverse piattaforme UNIX, Windows e MacOS inoltre utilizza l'interfaccia a riga di comando, sebbene siano disponibili diverse GUI

Python: è un linguaggio di programmazione di più "alto livello" rispetto alla maggior parte degli altri linguaggi, orientato a oggetti, adatto, tra gli altri usi, a sviluppare applicazioni distribuite, scripting, computazione numerica e system testing.

Prime analisi

Per poter individuare, quale delle variabili sopra indicate fosse più correlata con il rating, ho utilizzato l'enumerazione completa, cioè, effettuare tutte le possibili combinazioni di regressori mettendo poi a confronto i valori ottenuti in output.

Il valore più importante ottenuto è l'r-squared (coefficiente di determinazione), cioè, in statistica, un indice che misura il legame tra la variabilità dei dati e la correttezza del modello statistico utilizzato.

Può variare da 0 a 1 dove:

- 0 corrisponde a un legame nullo tra le variabili
- 1 a un legame perfetto.

Analisi in R

Come primo linguaggio per l'analisi utilizzeremo R.

La libreria che ci permette di effettuare la regressione lineare in R è chiamata states, al cui interno possiamo trovare il comando lm.

Come primo passo ho inserito i dati tramite CSV creando una tabella da analizzare, successivamente attraverso lm ho creato le 32 regressioni lineari da mettere a confronto.

Esempio di script:

```
#####################regressione con un solo regressore##################

mod_scene_distinte <- lm(Rating~n_scene_distinte, data=dati_episodi2)
plot(mod_scene_distinte)
summary(mod_scene_distinte)
```

In questo esempio mod_scene_distinte è la variabile da creare con all'interno la regressione lineare, mettendo in correlazione il rating con il numero di scene distinte. Poi effettuando il summary otteniamo l'e-square.

Una volta effettuate tutte le regressioni andiamo a visualizzare tutti gli r-square a confronto

1 regressore

dur_media scene	0.00048
n_scene_distinte	0.01457
n_cambi scena	0.0242
HHI_battute	0.05081
HHI_scene	0.1162

2 regressori

n_cambi scena+dur_media scene	0.05717
n_cambi_scena+HHI_scene	0.1576
n_cambi_scena+HHI_battute	0.06961
n_cambi_scena+n_scene_distinte	0.02452
HHI_battute+HHI_scene	0.1164
HHI_battute+dur_media_scene	0.05234
HHI_battute+n_scene_distinte	0.08154
n_scene_distinte+HHI_scene	0.1384
n_scene_distinte+dur_media_scene	0.04038

HHI_scene+dur_media_scene	0.1204

3 regressori

n_cambi_scena+dur_media_scene+HHI_battute	0.1044
n_cambi_scena+dur_media_scene+HHI_scene	0.2528
n_cambi_scena+dur_media_scene+n_scene_distinte	0.07119
n_cambi_scena+HHI_battute+HHI_scene	0.162
n_cambi_scena+HHI_battute+n_scene_distinte	0.0817
n_cambi_scena+HHI_scene+n_scene_distinte	0.1578
dur_media_scene+HHI_battute+HHI_scene	0.1206
dur_media_scene+HHI_battute+n_scene_distinte	0.1545
dur_media_scene+HHI_scene+n_scene_distinte	0.2086
HHI_battute+HHI_scene+n_scene_distinte	0.1388

4 regressori

n_cambi_scena+dur_media_scene+HHI_battute+HHI_scene	0.2794
n_cambi_scena+dur_media_scene+HHI_battute+n_scene_distinte	0.1644
n_cambi_scena+dur_media_scene+HHI_scene+n_scene_distinte	0.2883
n_cambi_scena+HHI_battute+HHI_scene+n_scene_distinte	0.1631
dur_media_scene+HHI_battute+HHI_scene+n_scene_distinte	0.2164

5 regressori

n_cambi_scena+dur_media_scene+HHI_battute+HHI_scene+n_scene_distinte	0.2942

Analisi Python

Le librerie che ci permettono di attuare la regressione lineare in Python sono pandas, numpy e sklearn.

Come primo passo ugualmente come è stato fatto per R ho inserito il dataset come CSV in forma tabellare, successivamente attraverso due codici diversi uno per la regressione con un regressore e uno per due o più regressori ho completato la mia analisi.

Esempi di script

Un regressore

```
X = dati['dur_media_scene']
y = dati['Rating']

from sklearn.linear_model import LinearRegression

X=X.to_numpy()
X=X.reshape(-1,1)

lr = LinearRegression().fit(X,y)


lr.score(X,y)
```

Duo o più regressori

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
x = dati.iloc[:,[4,5]]
y = dati.iloc[:,1]
model = LinearRegression()
model.fit(x, y)
model = LinearRegression().fit(x, y)
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
```

Una volta effettuate le regressioni analizziamo gli r-square:

1 regressore

dur_media scene	0.0004879441267181539
n_scene_distinte	0.014573032669469965
n_cambi scena	0.02420398357229736
HHI_battute	0.05080978088282051
HHI_scene	0.11617383733457431

2 regressori

n_cambi scena+dur_media scene	0.057167599079076314
n_cambi_scena+HHI_scene	0.15764497890845686
n_cambi_scena+HHI_battute	0.06961384151468075

n_cambi_scena+n_scene_distinte	0.02452257360548049
HHI_battute+HHI_scene	0.11636170977873328
HHI_battute+dur_media_scene	0.05233834748744959
HHI_battute+n_scene_distinte	0.08154279450109314
n_scene_distinte+HHI_scene	0.13839725382768542
n_scene_distinte+dur_media_scene	0.04037879470561334
HHI_scene+dur_media_scene	0.12037181061020341

3 regressori

n_cambi_scena+dur_media_scene+HHI_battute	0.10440500199761338
n_cambi_scena+dur_media_scene+HHI_scene	0.25276953034140104
n_cambi_scena+dur_media_scene+n_scene_distinte	0.07118925993545078
n_cambi_scena+HHI_battute+HHI_scene	0.16203005523818026
n_cambi_scena+HHI_battute+n_scene_distinte	0.08170432047157117
n_cambi_scena+HHI_scene+n_scene_distinte	0.1577795571367837
dur_media_scene+HHI_battute+HHI_scene	0.12058667729252004
dur_media_scene+HHI_battute+n_scene_distinte	0.15453263879005108
dur_media_scene+HHI_scene+n_scene_distinte	0.20856297156418668
HHI_battute+HHI_scene+n_scene_distinte	0.13884819087375277

4 regressori

n_cambi_scena+dur_media_scene+HHI_battute+HHI_scene	0.27936203664311154
n_cambi_scena+dur_media_scene+HHI_battute+n_scene_distinte	0.16436522838822165
n_cambi_scena+dur_media_scene+HHI_scene+n_scene_distinte	0.28834555446477594
n_cambi_scena+HHI_battute+HHI_scene+n_scene_distinte	0.16306657116850287
dur_media_scene+HHI_battute+HHI_scene+n_scene_distinte	0.21637557436755284

5 regressori

n_cambi_scena+dur_media_scene+HHI_battute+HHI_scene+n_scene_distinte	0.294222801471294

CONCLUSIONE

I risultati ottenuti non sono sicuramente quelli sperati avendo dei valori molto bassi di r-square, però possiamo comunque osservare molte cose.

Nelle prima regressioni con un solo regressore la variabile più correlata con il rating è l'HHI scene, questo perché all'interno della serie ci sono 4 personaggi principali (Steve Carell: Michael Scott Rainn Wilson: Dwight Schrute John Krasinski: Jim Halpert Jenna Fischer: Pam Beesly) infatti, si può notare che negli episodi dove c'è maggior presenza di questi attori il rating è più alto di altri episodi.

Andando avanti con le regressioni si nota che l'HHI battute ha sempre valori più bassi quindi non utile alla nostra analisi, al contrario l'HHI scene e tutte le altre caratteristiche delle scene continuano ad essere le più correlate, questo ci fa intuire che il rating per una buona parte è influenzato:

- quali personaggi si trovano nelle scene
- quante scene diverse ci sono in un episodio
- quante volte si cambia da una scena ad un'altra
- quanto durano mediamente le scene in un episodio.

Avendo poche variabili da correlare con il rating ho riscontrato un basso r-square, infatti, ci sono molti altri fattori che possono influenzare lo spettatore come, su che sito di streaming guarda la serie, attori più o meno famosi, in che momento della sua vita sta guardando la serie ecc.…

In conclusione, possiamo dire che le variabili utilizzate influiscono sul rating espresso per un valore di r-square pari a 0.30 ma per un'analisi più corretta avremmo bisogno di molte più variabili che ci spieghino l'andamento dei dati.