Report of MA678 Midterm Project

Peng Liu

2021/12/09

Abstract

Unlike other video media sites, Netflix does not have its scoring system. In addition to Netflix's fitness system, IMDb may be a good reference factor. In this report, I will find the connection between them by establishing a multi-level model of IMDb scores and other film-related factors. This report includes five parts: introduction, methods, results, and discussion.

Introduction

As a pioneer in the streaming media service industry, Netflix has become the media company with the most extensive coverage globally in recent years. However, what is interesting is that, unlike other video sites, Netflix does not have a scoring system. Except for Netflix's recommendation system, how will the public choose movies to watch? IMDb is an authoritative and well-known movie website. The score is believed to be a good reference for the audience. For this, I chose the "Latest Netflix data with 26+ connection attributes" dataset. This is a very comprehensive data set that includes sources from Netflix, Rotten Tomatoes, IMBD, posters, box office information, trailers on YouTube, and more sources using various APIs. I want to use a multi-level model to see which factors affect the IMBD scores of movies on Netflix.

Method

Data cleaning

I use the data set is Latest Netflix data with 26+ joined attributes on Kaggle. This is a vast data set containing movies released on Netflix from 2015 to 2021. This information includes APIs from data sources such as Netflix, Rotten Tomatoes, IMBD, posters, box office information, YouTube trailers, a total of 15,071 unique values. I cleaned the data. I first narrowed the scope of the data to the release time and the release time on Netflix between 2018 and 2021 and removed the missing value in the IMDb Score. Next, I processed the text part of the data set, counted the genres, tags, language, and type of the movie in the data, and removed the missing value. In the end, the data set changed from 15,071 observations and 29 columns to 3496 observations and 14 columns.

column names	explanation
Title	Title of the movie
Major.Genre	The major genres
Series.or.Movie	TV Series or More
IMDb.Score	Score from IMDb
Rotten.Tomatoes.Score	Score from Rotten Tomatoes
Metacritic.Score	Score from Metacritic
Awards.Received	The number of Awards Received
Awards.Nominated.For	The number of Awards Nominated
IMDb.Votes	Votes on IMDb

column names	explanation
num.Genre	The number of genres
num.tags	The number of tags
num.country	The number of Netflix country availability
Runtime_1hour	Runtime large than 1 hour or not
num.Lang	The number of available languages on Netflix

EDA

Regarding the EDA part, first, I made a corrplot to show the correlation between these variables. Figure 1 shows that the correlation between IMDb Score and other variables is primarily positive, except for runtime and languages. Among them, nominations, awards, IMDb Votes, and the country's number are the most influential factors. Therefore, I will visualize the relationship between IMDb Score and these variables.

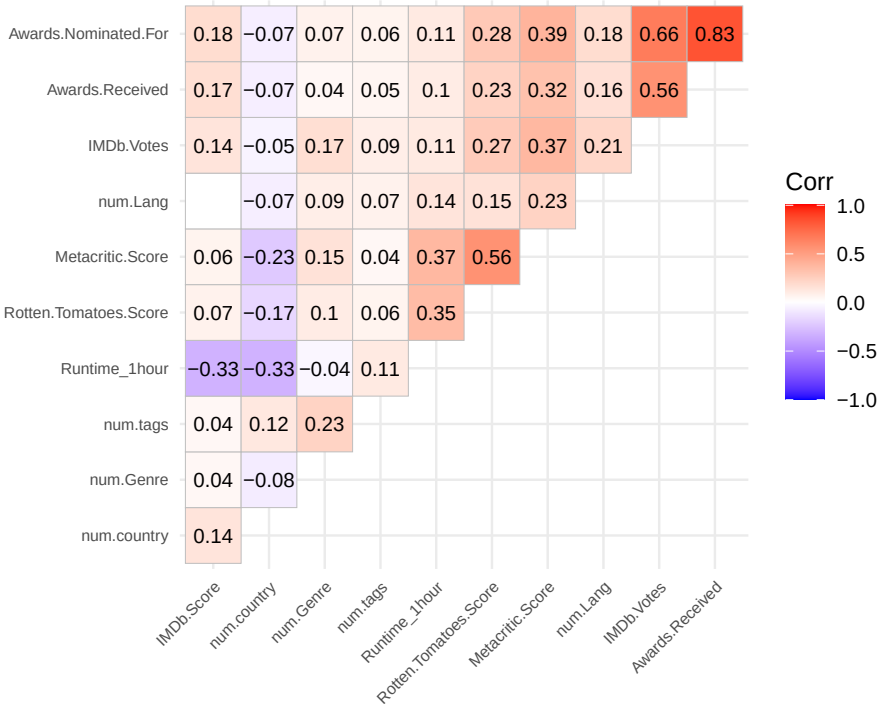
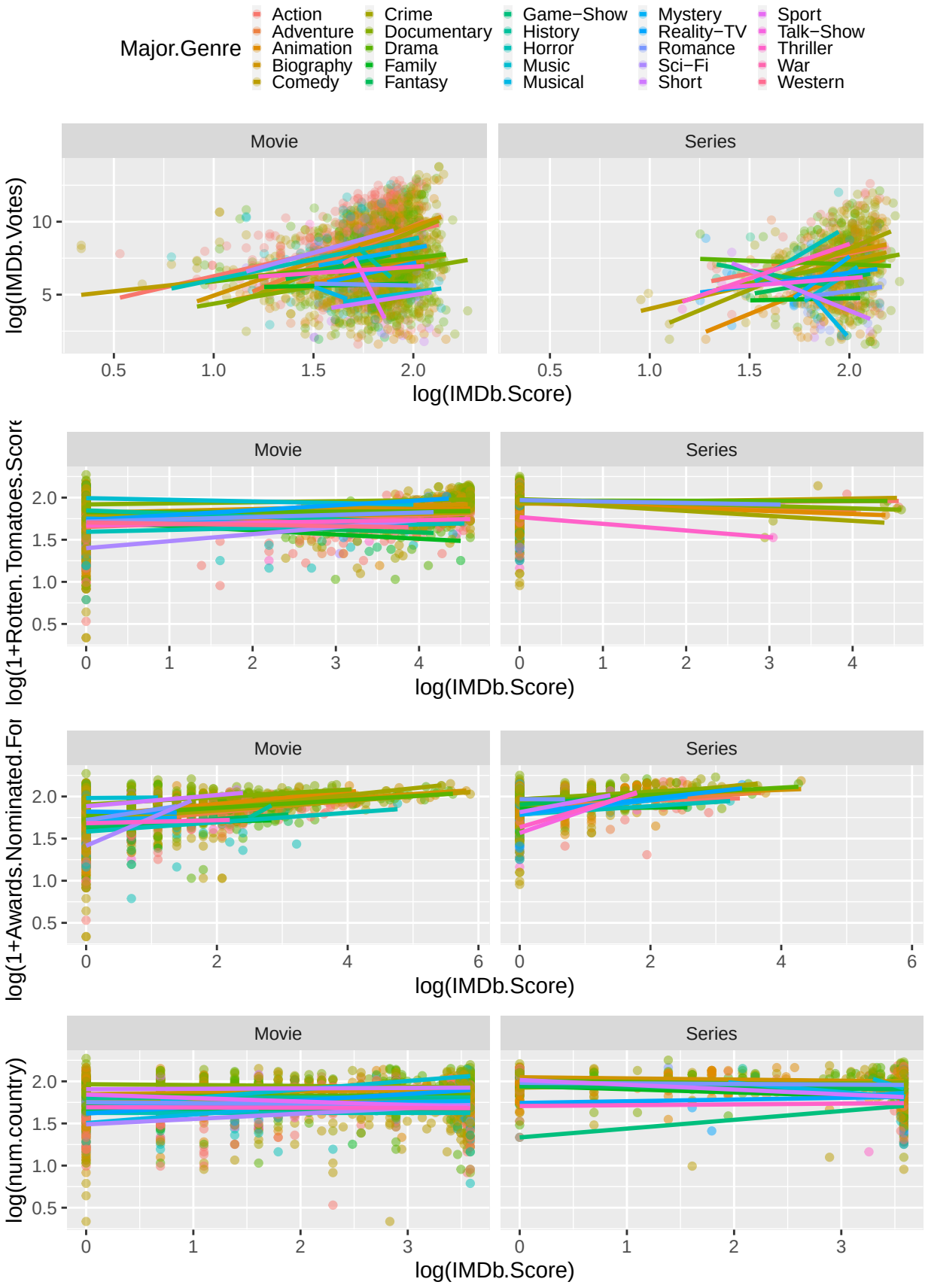


Figure 1: Figure 1: correlation of variables

In order to find the correlation between IMDb Score and other variables, I drew some scatter plots to find the correlation between them. Because movies can be divided into movies and TV series, I will divide the following pictures into two groups: movies and series.


Major.Genre
Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
Game−Show
History
Horror
Music
Musical
Mystery
Reality−TV
Romance
Sci−Fi
Short
Sport
Talk−Show
Thriller
War
Western
Movie
Series
log(IMDb.Votes)
log(IMDb.Score)
log(1+Rotten.Tomatoes.Score
log(1+Awards.Nominated.For
log(num.country)

It can be seen that in these two groups, the slopes of most genres are similar, although the intercept varies due to the difference in the count of genres. The intercepts of Figure 1 and Figure 3 are positively correlated. What's interesting is that the intercepts of the Rotten Tomatoes score and IMDb score in Figure 2 are in a relatively ambiguous state in the film group, and they cannot be accurately judged. There is a correlation between factors, but there is an apparent negative correlation between the two in the Series group.

Model fitting

According to the figure in the EDA part, it can be found that although the overall trend of the data is roughly the same, there are still noticeable differences between different genres, so I chose to use a multi-level model to fit the model. I selected six variables: IMDb.Votes, num.country, Awards.Received, Awards.Nominated.For, Metacritic.Score and Rotten.Tomatoes.Score based on the corrplot chart I drew and added different types of genres. Intercept and slope. The following is my model:

```
lmer_model = lmer(log(IMDb.Score)~log(IMDb.Votes)+log(num.country)+log(1+Awards.Received)+
             log(1+Awards.Nominated.For)+log(1+Rotten.Tomatoes.Score)+log(1+Metacritic.Score)+
             (1|Major.Genre),data =netflix1)
```

Result

Coefficients

The following is the fixed effects of the model:

	Estimate	Std. Error	df	t value	Pr(>
(Intercept)	1.671816	0.024781	32.924125	67.462	< 0.0000000000000002 ***
log(IMDb.Votes)	0.010744	0.002017	3291.518154	5.327	0.00000010660308 ***
log(num.country)	0.015964	0.003478	8.932133	4.590	0.00134 ***
log(1 + Awards.Received)	0.031455	0.006108	3467.000188	5.150	0.00000027537285 ***
log(1 + Awards.Nominated.For)	0.036028	0.005104	3459.175145	7.059	0.00000000000201 ***
log(1 + Rotten.Tomatoes.Score)	-0.006127	0.002096	3470.042095	-2.923	0.00349 **
log(1 + Metacritic.Score)	-0.011671	0.002472	3471.628782	-4.720	0.00000244773816 ***

the following is the final model of :

$$log(IMDb.Score) = 1.671816 + 0.010744 \cdot log(IMDb.Votes) + 0.015964 \cdot log(num.country) + 0.031455 \cdot$$

$$log(1+Awards.Received) + 0.036028 \cdot log(1+Awards.Nominated.For) - 0.011671 \cdot log(1+Metacritic.Score)$$

$$-0.006127 \cdot log(1 + Rotten.Tomatoes.Score)$$

Through this model, we can find a fascinating phenomenon. For every 0.1 point increase in IMDb Score, Rotten Tomatoes and Metacritic scores will decrease by 1.1582% and 0.5998%, respectively.

```
##           (Intercept) log(num.country)
## Action    -0.07316735      0.006525872
```

```
## Adventure -0.04821452      0.004215991
## Animation  0.11626811     -0.011088766
## Biography  0.03373627     -0.002008224
## Comedy    -0.02905385      0.006189504
## Crime      0.01028343      0.005197512
```

Discussion

By comparing the model and the EDA part, this model is more reasonable. The IMDb score is positively correlated with the number of votes, the number of nominated awards, and the number of mentions on IMDb. As these variables increase, the movie's IMDb score will be higher. The IMDb score and the Metacritic Score of Rotten Tomatoes are negatively correlated. I did not notice this obviously during the EDA process because the intercept of most of the generators in the Rotten Tomatoes graph is close to 0, and some Positive correlation.

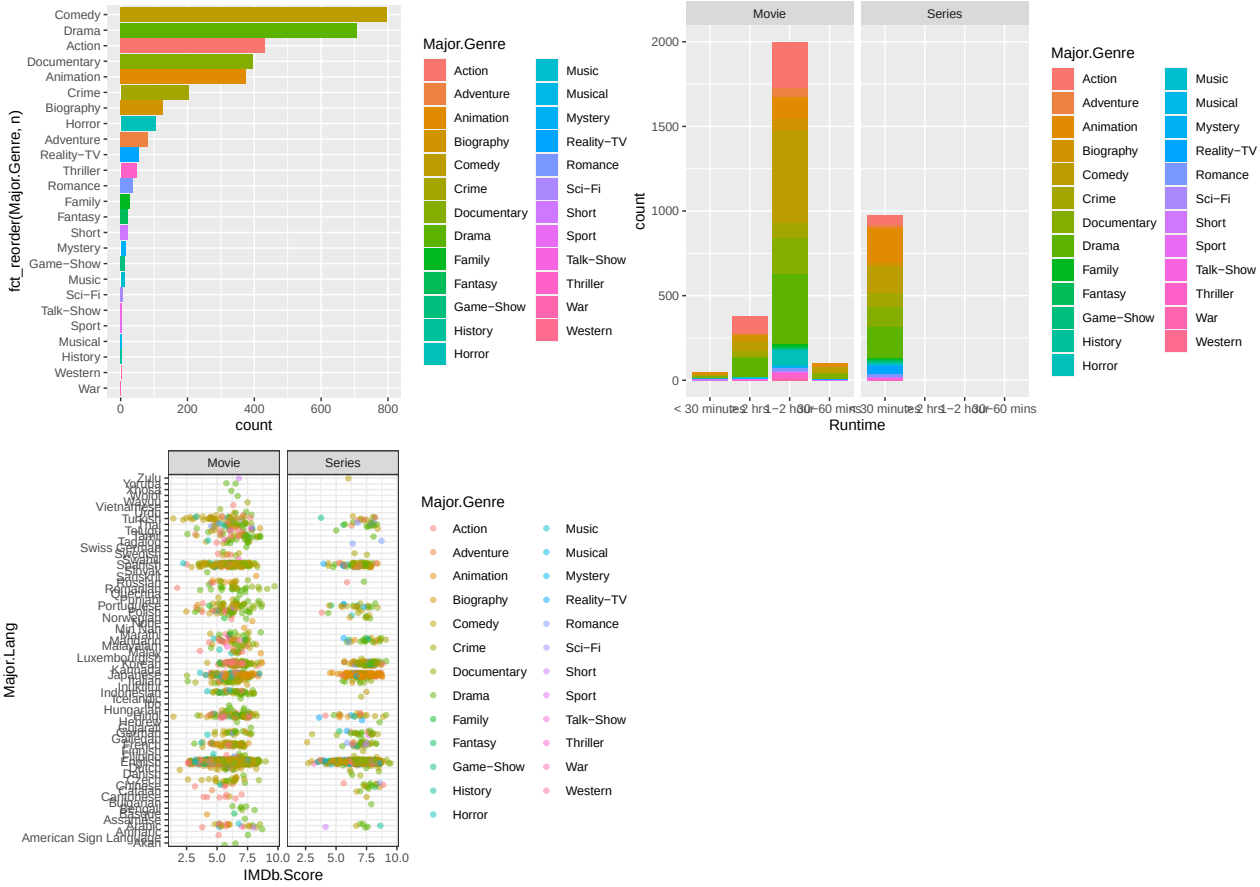
Of course, this model also has some weaknesses. I did not add all the variables. For example, some negatively correlated variables and variables with too small a correlation were removed by me. Adding these variables may change the model.

Reference

Latest Netflix data with 26+ joined attributes: https://www.kaggle.com/ashishgup/netflix-rotten-tomatoes-metacritic-imdb

Appendix

EDA




```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(IMDb.Score) ~ log(IMDb.Votes) + log(num.country) + log(1 +
##     Awards.Received) + log(1 + Awards.Nominated.For) + log(1 +
##     Rotten.Tomatoes.Score) + log(1 + Metacritic.Score) + (log(num.country) |
##     Major.Genre)
##    Data: netflix1
##
## REML criterion at convergence: -1747.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.8629 -0.4213  0.1392  0.6058  2.7715
##
## Random effects:
##  Groups      Name              Variance  Std.Dev. Corr
##  Major.Genre (Intercept)       0.0087637 0.093615
##              log(num.country) 0.0000757 0.008701 -0.76
```

```
##  Residual                     0.0343245 0.185269
## Number of obs: 3496, groups:  Major.Genre, 25
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    1.671816   0.024781  67.462
## log(IMDb.Votes)                0.010744   0.002017   5.327
## log(num.country)               0.015964   0.003478   4.590
## log(1 + Awards.Received)       0.031455   0.006108   5.150
## log(1 + Awards.Nominated.For)  0.036028   0.005104   7.059
## log(1 + Rotten.Tomatoes.Score) -0.006127   0.002096  -2.923
## log(1 + Metacritic.Score)      -0.011671   0.002472  -4.720
##
## Correlation of Fixed Effects:
##             (Intr) l(IMD. lg(n.) l(1+A.R l(1+A.N l(1+R.
## lg(IMDb.Vt) -0.419
## lg(nm.cntr) -0.480 -0.208
## lg(1+Awr.R) -0.009  0.003  0.036
## lg(1+A.N.F)  0.074 -0.226  0.004 -0.730
## lg(1+R.T.S)  0.042 -0.184  0.060 -0.038  -0.022
## lg(1+Mtc.S)  0.110 -0.357  0.149  0.012  -0.120  -0.334
```

```
##                    (Intercept)                log(IMDb.Votes)
##                     1.67181615                     0.01074402
##               log(num.country)       log(1 + Awards.Received)
##                     0.01596421                     0.03145497
##  log(1 + Awards.Nominated.For) log(1 + Rotten.Tomatoes.Score)
##                     0.03602775                    -0.00612739
##      log(1 + Metacritic.Score)
##                    -0.01167088
```

```
## $Major.Genre
##                (Intercept) log(num.country)
## Action      -0.0731673492    0.00652587184
## Adventure   -0.0482145166    0.00421599098
## Animation    0.1162681116   -0.01108876554
## Biography    0.0337362743   -0.00200822361
## Comedy      -0.0290538541    0.00618950377
## Crime        0.0102834323    0.00519751160
## Documentary  0.1812555900   -0.01430568133
## Drama        0.0479914423   -0.00398783529
## Family      -0.0425825821   -0.00017783159
## Fantasy     -0.0184265026    0.00093485264
## Game-Show   -0.1018418356    0.00672351827
## History      0.0096152830   -0.00045598972
## Horror      -0.1599685687    0.00459966456
## Music        0.1422486356   -0.00720221822
## Musical     -0.0107198595    0.00120875757
## Mystery     -0.0284465749    0.00214016249
## Reality-TV   0.0127192235   -0.00009664082
## Romance      0.0853618097   -0.00528600516
## Sci-Fi      -0.1089384352    0.00839020791
## Short        0.0967106807   -0.00630760654
## Sport        0.0152594124   -0.00143418463
## Talk-Show   -0.0002346155    0.00001235157
```

```
## Thriller    -0.0755796297    0.00254125388
## War         -0.0374610898    0.00253396176
## Western     -0.0168144818    0.00113737358
##
## with conditional variances for "Major.Genre"
```