University of Glasgow

Analysis of IMDB Rating with GLM

(GROUP 10)

Ruiwen Ma, Xiaoyu Zhao, Xuqin Tan,

Yola Kamalita, Ziheng Cao

20-March-2024

INSPIRING PEOPLE

CONTENTS

01 Introduction

02 Exploratory data analysis

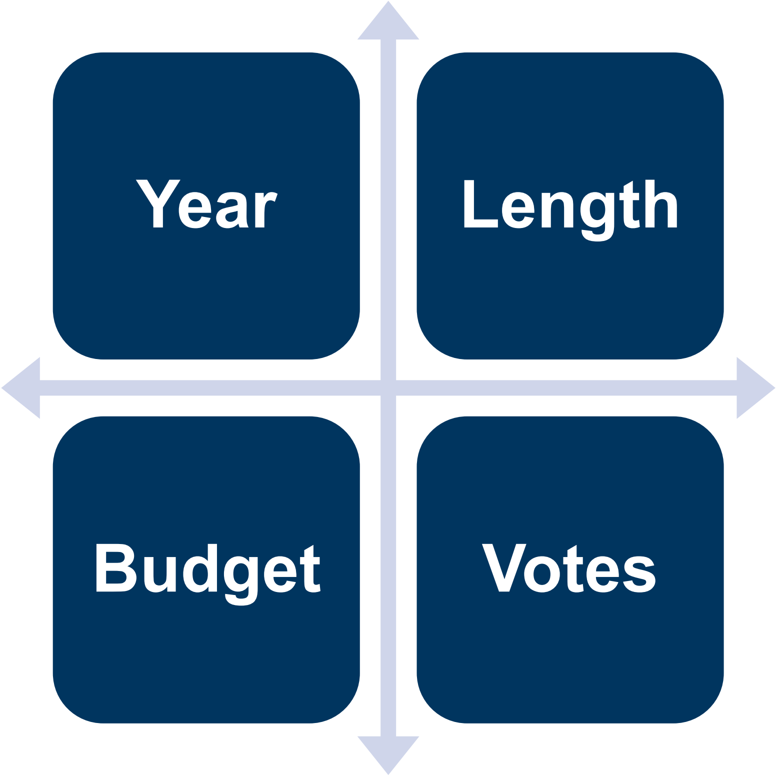
03 Statistical modelling and results

04 Conclusions

05 Future work

Introduction

The GLM for binary response variables, in this case, Logistic Regression, will be used to investigate the relationship between categorical rating and film properties.



Year

Length

Budget

Votes

Continuous variable

Genre

Categorical variable

Binary rating

(1 indicates rating >7, 0 otherwise)

Response variable

Exploratory data analysis (numeric variables)

A *weak* linear relationship to the rating is noticeable for **budget and length,** while **year and votes** show a *very weak* negative correlation to the rating.

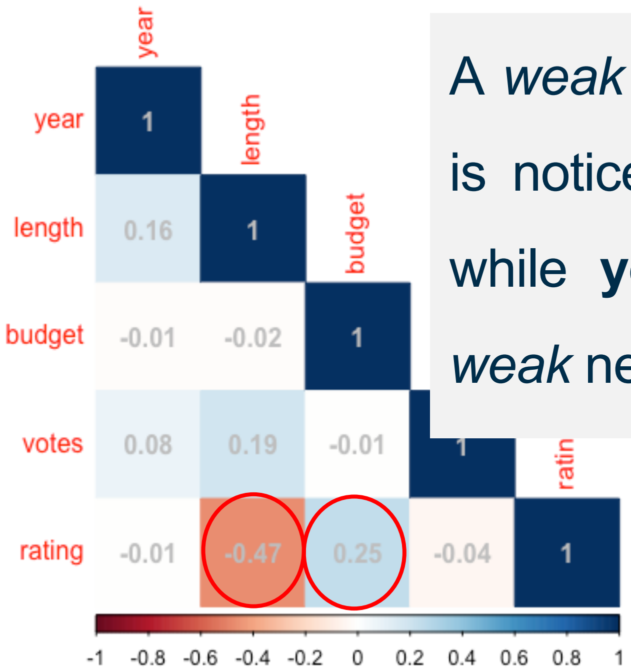


Figure1: correlation

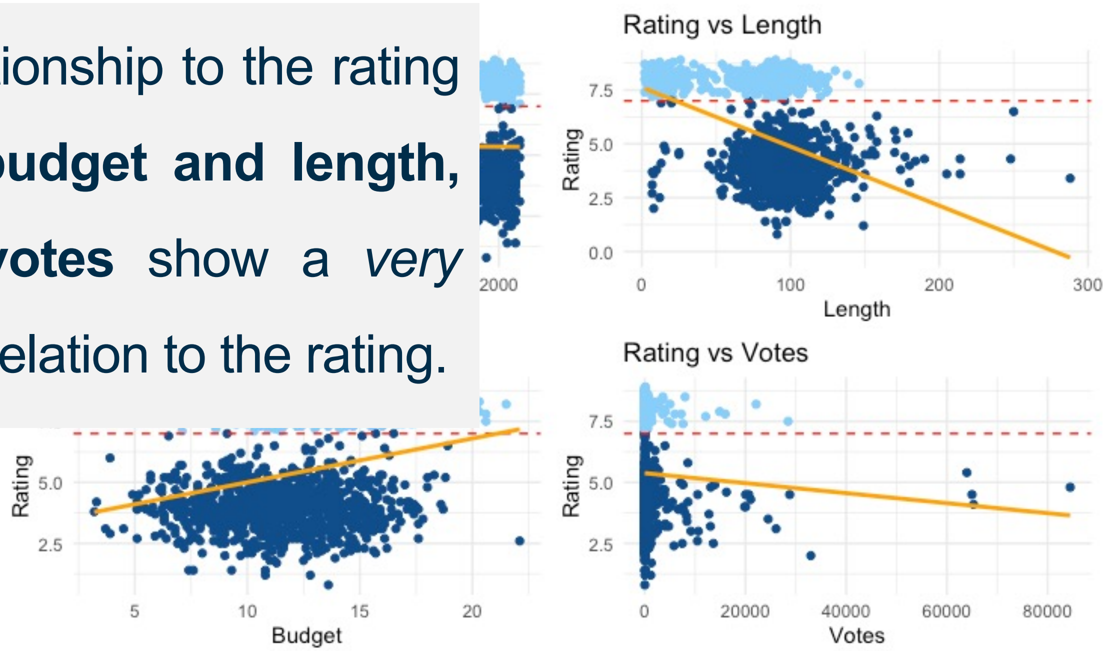


Figure2: scatterplot

Exploratory data analysis (categorical variables)

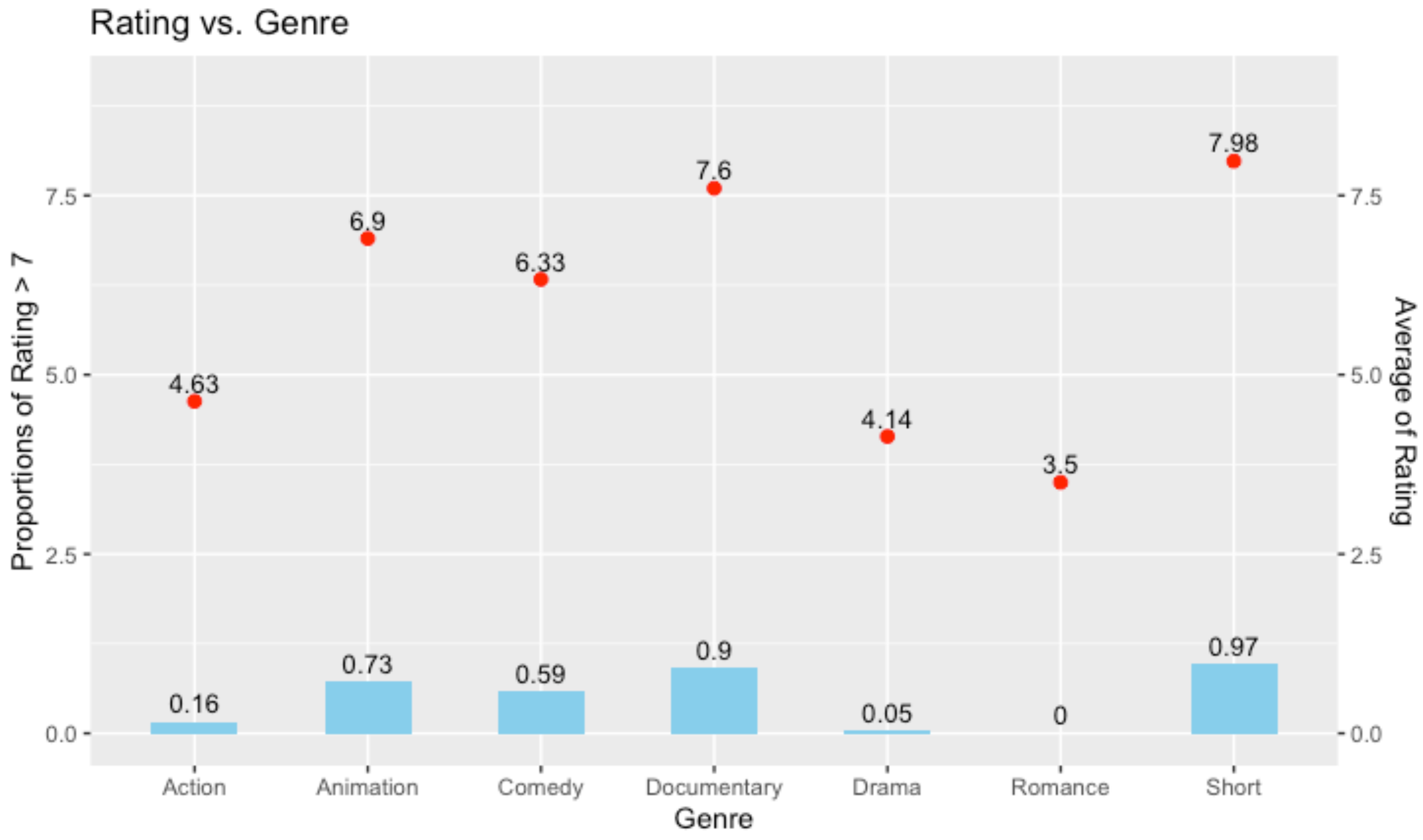


Figure3: **The proportion and average of rating for each genre**

The **Short genre** becomes the highest for both proportion and average rating. Visually, the genre helps distinguish whether the rating will be high or low.

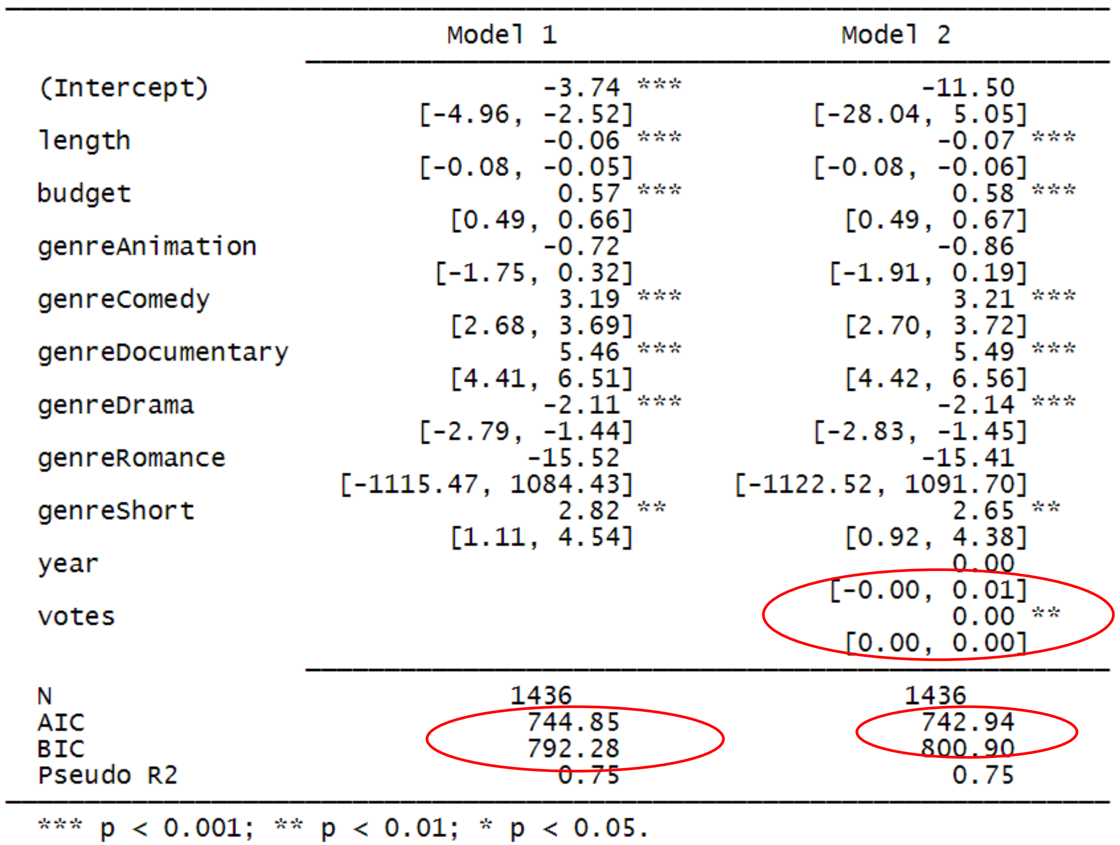
Statistical Modelling

Method: GLM

**Model : **Logistic

Reason:

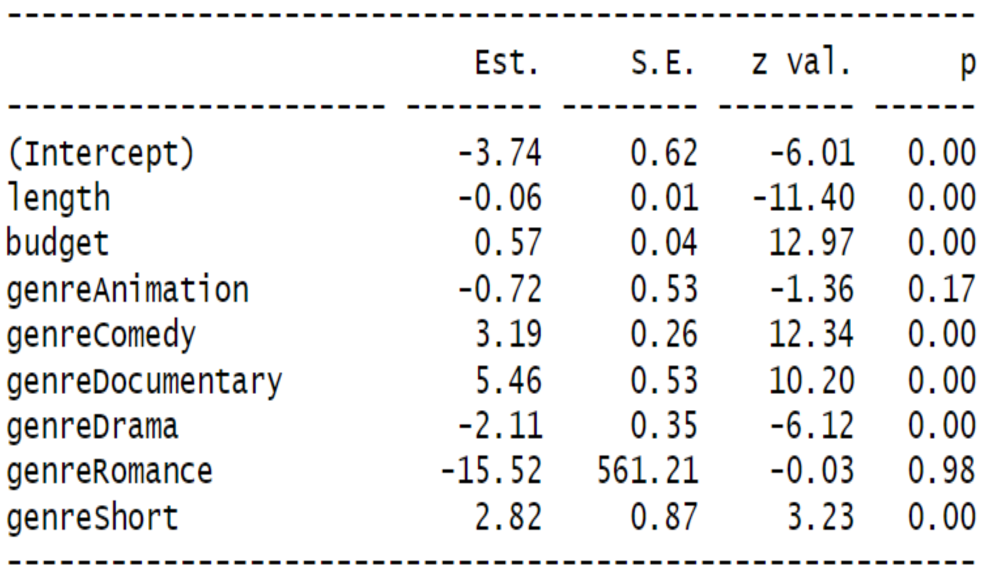
Our target response data is a binary variable, following a binomial distribution. GLM allows us to choose the appropriate link function to model the relationship between probability and explanatory variables.



```
=====================================================================
                          Model 1                 Model 2
                  ---------------------------------------------------
(Intercept)                  -3.74 ***               -11.50
                      [-4.96, -2.52]          [-28.04,  5.05]
length                       -0.06 ***                -0.07 ***
                      [-0.08, -0.05]          [-0.08, -0.06]
budget                        0.57 ***                 0.58 ***
                      [0.49,  0.66]           [0.49,  0.67]
genreAnimation               -0.72                    -0.86
                      [-1.75,  0.32]          [-1.91,  0.19]
genreComedy                   3.19 ***                 3.21 ***
                      [2.68,  3.69]           [2.70,  3.72]
genreDocumentary              5.46 ***                 5.49 ***
                      [4.41,  6.51]           [4.42,  6.56]
genreDrama                   -2.11 ***                -2.14 ***
                      [-2.79, -1.44]          [-2.83, -1.45]
genreRomance                -15.52                   -15.41
                 [-1115.47, 1084.43]     [-1122.52, 1091.70]
genreShort                    2.82 **                  2.65 **
                      [1.11,  4.54]           [0.92,  4.38]
year                                                   0.00
                                              [-0.00,  0.01]
votes                                                  0.00 **
                                              [0.00,  0.00]
                  ---------------------------------------------------
N                          1436                    1436
AIC                         744.85                  742.94
BIC                         792.28                  800.90
Pseudo R2                     0.75                    0.75
=====================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05.
```

Table 1: Hypothesis Testing and Goodness of fit.

Modelling result



```
---------------------------------------------------------
                          Est.      S.E.    z val.      p
---------------------- -------- -------- -------- ------
(Intercept)              -3.74      0.62    -6.01   0.00
length                   -0.06      0.01   -11.40   0.00
budget                    0.57      0.04    12.97   0.00
genreAnimation           -0.72      0.53    -1.36   0.17
genreComedy               3.19      0.26    12.34   0.00
genreDocumentary          5.46      0.53    10.20   0.00
genreDrama               -2.11      0.35    -6.12   0.00
genreRomance            -15.52    561.21    -0.03   0.98
genreShort                2.82      0.87     3.23   0.00
---------------------------------------------------------
```

Table2: Regression Model Coefficients.

$$log(\frac{P_i}{1-P_i}) = \alpha + \beta_{budget} \cdot budget_i + \beta_{length} \cdot length_i + \beta_{\text{Animation}} \cdot \mathbb{I}_{\text{Animation}}(i) + \beta_{\text{Comedy}} \cdot \mathbb{I}_{\text{Comedy}}(i) +$$

$$\beta_{\text{Documentary}} \cdot \mathbb{I}_{\text{Documentary}}(i) + \beta_{\text{Drama}} \cdot \mathbb{I}_{\text{Drama}}(i) + \beta_{\text{Romance}} \cdot \mathbb{I}_{\text{Romance}}(i) + \beta_{\text{Short}} \cdot \mathbb{I}_{\text{Short}}(i)$$

$$\mathbb{I}_{\text{Genre}}(x) = \begin{cases} 1 & \text{xth observation is part of the genre,} \\ 0 & \text{Otherwise.} \end{cases}$$

Variable explain

The log-odds coefficients for budget is positive indicating the higher their values then the rating will be more likely to be higher than 7. Furthermore, the log-odds coefficients for length is negative, and it indicates that the lower the values, the rating to be higher than 7 is more likely.

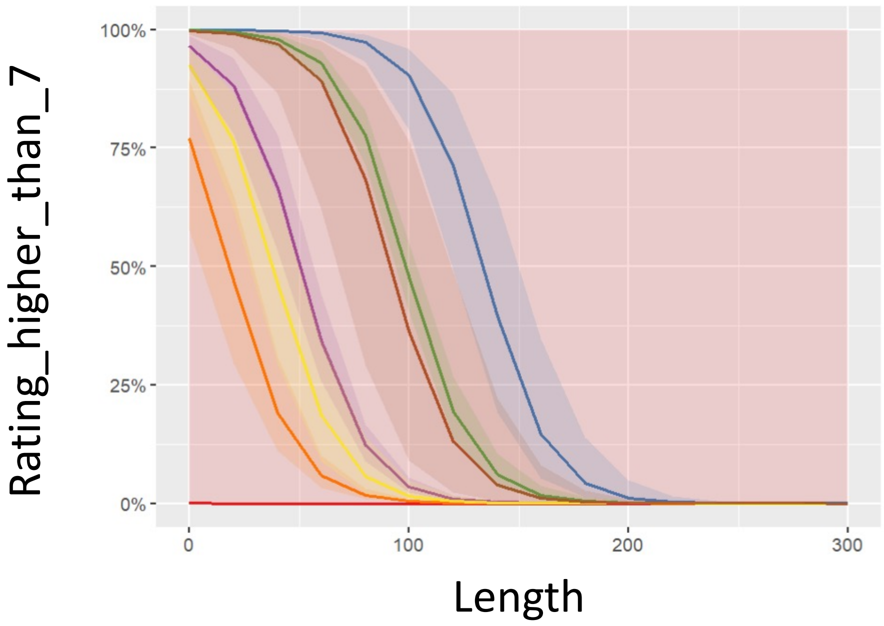
Predicted probabilities of having rating higher than 7

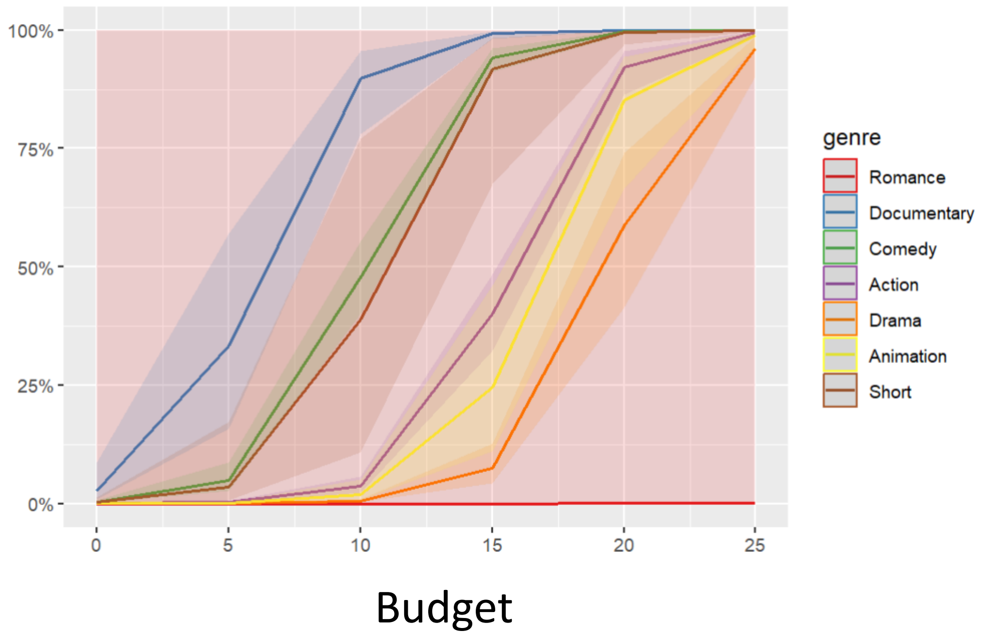
Length

The probability of a rating higher than 7 gradually diminishes with longer runtimes.

Budget

As the budget increases, the most of movies receiving high ratings also tends to increase.





Conclusions

1 There are two variables about genres that are not significant in this model, but if the genre is removed, AIC will increase significantly.

2 The budget, length, and genre significantly influence whether an IMDB rating is greater than 7. These predictors have been selected as optimal, resulting in a slight increase in AIC while substantially reducing the BIC.

3 The hypothesis on Hosmer-Lemeshow also shows no evidence for lack of fit.

Future Work

As the Animation’s (genre) odds is 0.49 times to the baseline category’s and this does not align with the fact that the Animation’s proportion (having a rating >7) is bigger than the baseline’s, future studies should consider **interaction terms between movie length and genre** which provide insights into the nuanced effects these variables have on ratings, potentially describing the data more accurately.

Thank You