University of Glasgow

Analysis of IMDB Rating with GLM

Group09：

Diwen Jiang, Wei Xiang, Zilu Liu,
Wenli Zhang, Haoran Jing

21 March 2025

Table of Contents

Project Background

Why analyze IMDb ratings?

IMDb is one of the most influential film rating platforms worldwide. Its scores are widely used to assess a movie's popularity, quality, and commercial success.

A high rating often indicates:

- Better box office performance
- Stronger promotional appeal
- Higher chances of winning awards and audience approval

Therefore, filmmakers, investors, and marketers are eager to understand what drives IMDb ratings.

Research Objective

Core Research Question:

What movie characteristics significantly influence whether an IMDb rating exceeds 7?

We focus on the following predictors:

- `length` (movie duration)
- `budget` (production budget)
- `votes` (number of user votes)
- `genre` (film genre)

Objective: Apply a Generalized Linear Model (GLM) to identify the most influential factors behind high ratings.

Dataset Overview & Processing

Data Source & Variables

- The dataset is a subset of the **IMDb (Internet Movie Database)**, used for educational purposes.
- It contains **1474 movies**, each represented by multiple descriptive features.
- The data is clean and contains no missing values.

Table1： Variable Descriptions

Varibale	Type	Description
rating	Numeric	IMDb score (0–10)
length	Numeric	Duration in minutes
budget	Numeric	Production budget (in million USD)
votes	Numeric	Number of user votes
genre	Categorical	Genre of the movie (Drama, Comedy, etc.)

Dataset Overview & Processing

Variable Transformation & Preparation

Binary Target Variable:

The rating variable was transformed into a binary variable high_rating:

- Rating > 7 → 1 (high rating)
- Rating ≤ 7 → 0 (not high)

Data Preparation Summary:

- genre is converted to a factor type for categorical variable modeling
- Observations with missing values were removed (Enough data to eventually retain 1474 items)
- All variables are now suitable for use in a logistic regression model

Distribution of IMDB Ratings

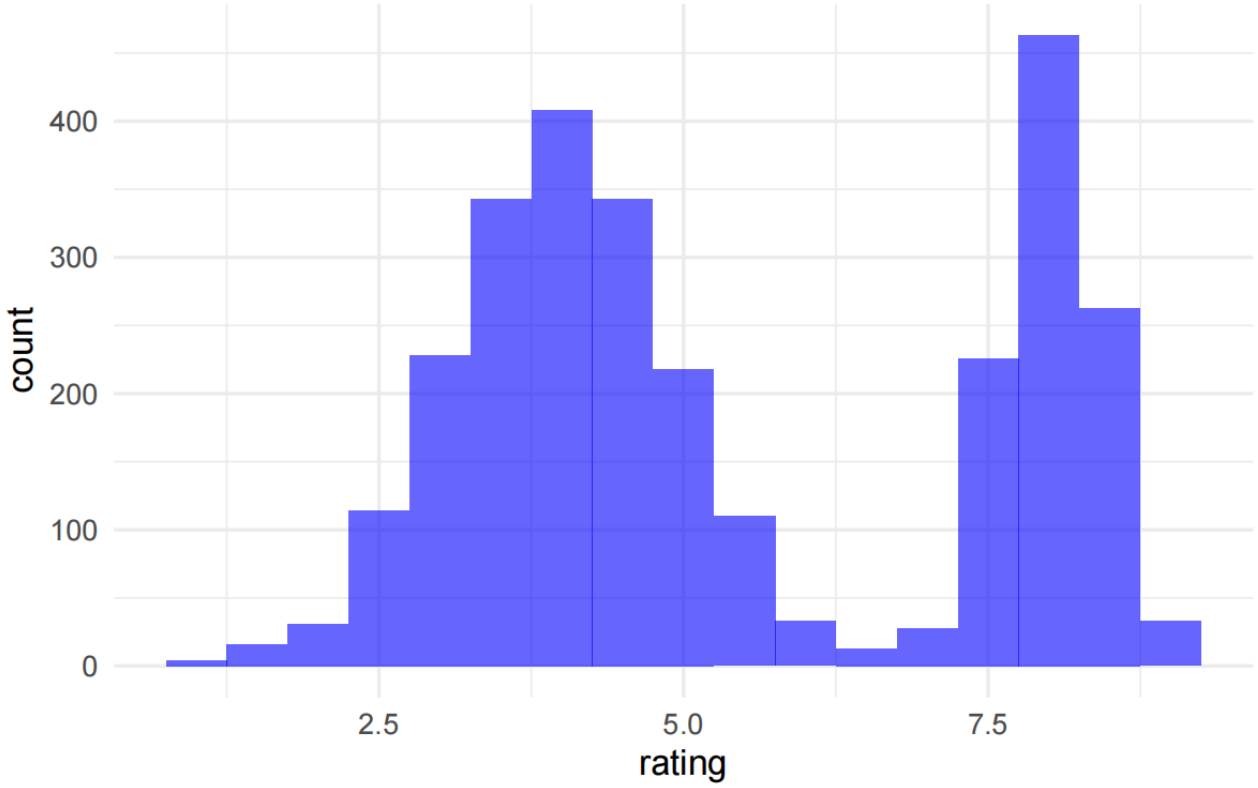

Figure 1: Distribution of IMDB Rating

Result:

- Most IMDB ratings range between **4–5** and **7–8.**
- The IMDb ratings exhibit a **bimodal distribution**.
- Around 30–40% of movies have ratings greater than 7, supporting the binary classification setup.

DRatings vs Key Predictors

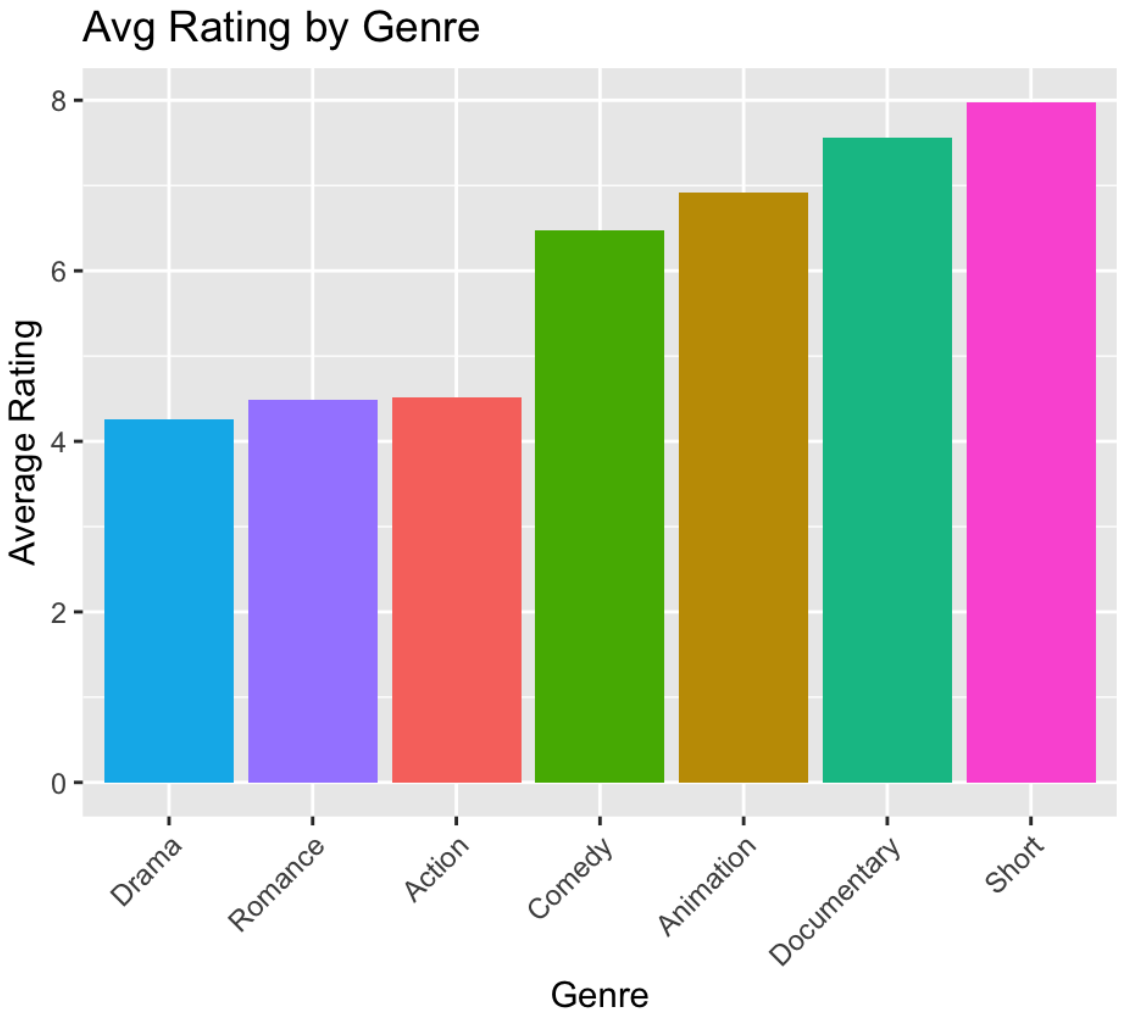

Figure2: Avg Rating by Genre

Result:

- **Short films** and **Documentaries** receive the **highest average ratings** (above 7.5).
- **Animation** and **Comedy** follow with solid mid-to-high ratings.
- **Drama**, **Action**, and **Romance** have **lower average scores** (below 5).
- **Movie genre significantly affects IMDb ratings and should be included in the model.**

Correlation Between Predictors

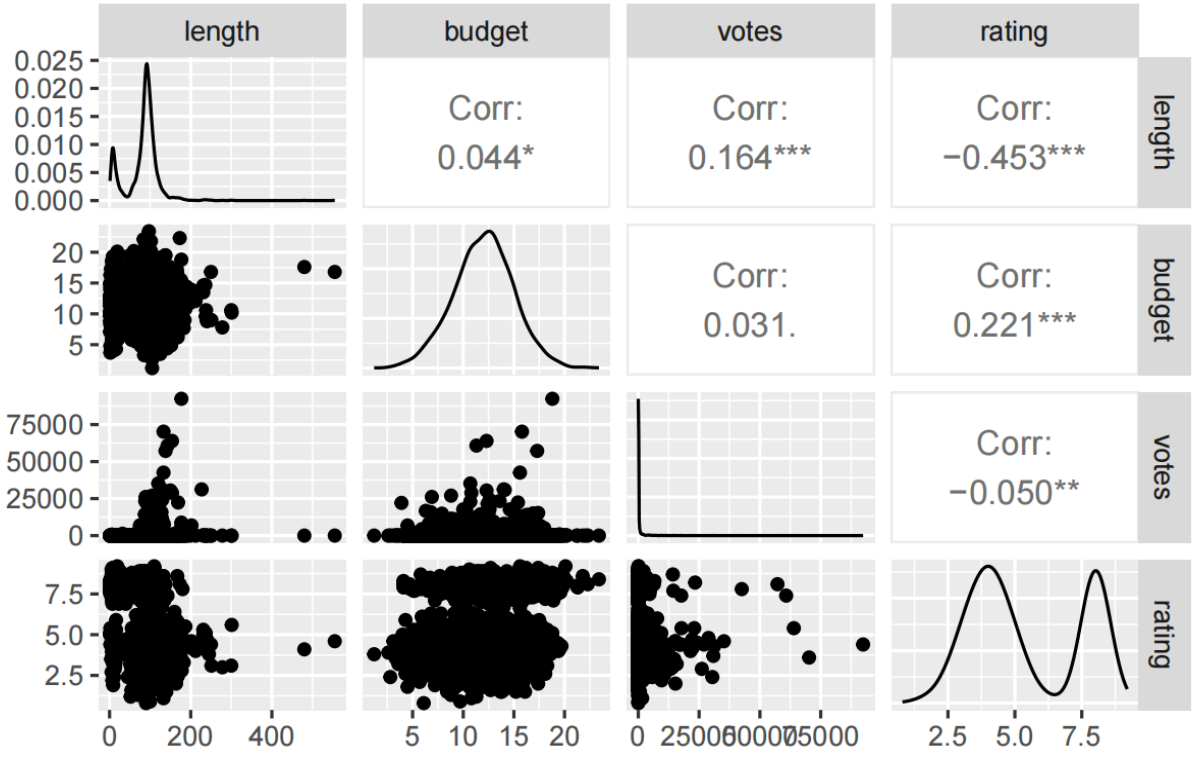

Figure 3: Relationships Between Movie Length, Budget, Votes, and Rating

Result:

- Weak correlations among all predictors support variable independence.
- The most notable relationship is a negative correlation between length and rating (-0.453).
- Other variables are weakly correlated, allowing their inclusion in the model.

The exploratory analysis confirms that the predictors are independent and show low multicollinearity, supporting their use in logistic regression modeling.

Statistical Modelling

Method: GLM

Model: Logistic

Our binary response variable follows a binomial distribution. GLM lets us use a suitable link function to model its relationship with explanatory variables.

term	estimate	std.error	statistic	p.value
(Intercept)	-3.152	0.466	-6.762	0.000
years_since	-0.006	0.003	-2.209	0.027
length	-0.061	0.004	-16.701	0.000
budget	0.515	0.030	17.443	0.000
votes	0.000	0.000	3.031	0.002
genreAnimation	-0.425	0.334	-1.272	0.203
genreComedy	3.346	0.180	18.570	0.000
genreDocumentary	5.266	0.383	13.741	0.000
genreDrama	-1.396	0.229	-6.094	0.000
genreRomance	-0.486	0.871	-0.557	0.577
genreShort	3.908	0.888	4.399	0.000

Table1:Regression Model Coefficients

Results:

- **Factors like budget, length, release year, number of votes play a significant role in whether a movie gets a rating above 7 on IMDb.**
- **Certain genres (such as comedy, documentary, short films, and drama) are also sigificant.**

Statistical Modelling

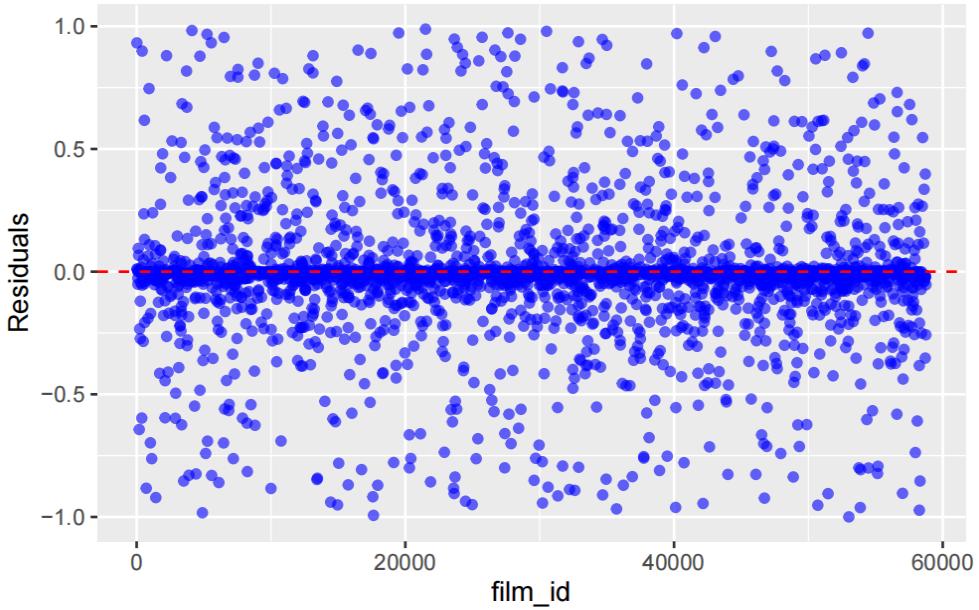

Figure 4: Residuals vs Fitted

The plot suggests that the residuals are randomly distributed without clear trends or autocorrelation, indicating that the model is likely reasonable overall.

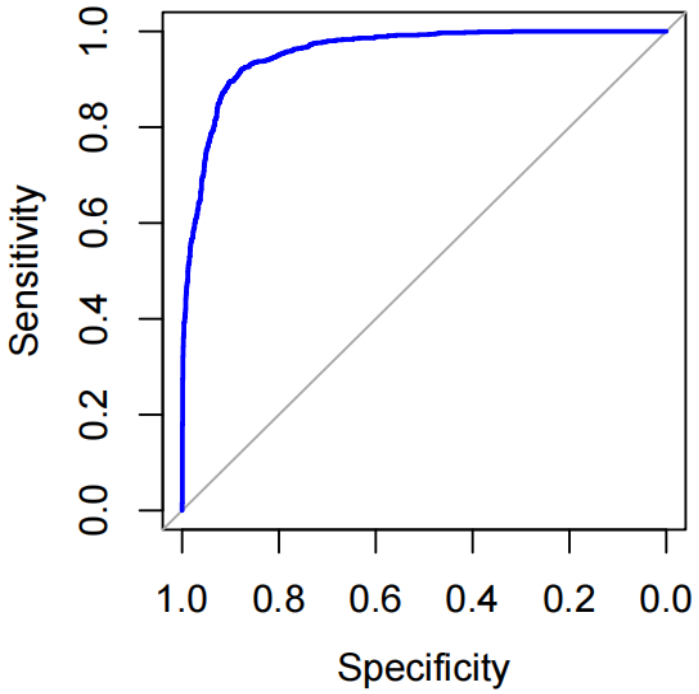

Figure 5: ROC Curve

the AUC value is 0.9566, indicating that the model performance is good, and it can accurately classify movies with IMDb scores above 7.

Confusion Matrix and Statistics

Confusion Matrix		
Prediction	0	1
0	1733	169
1	134	838

Table 2: Confusion Matrix

Statistics	Value
Accuracy	89.46%
Sensitivity	92.82%
Specificity	83.22%
Balanced Accuracy	88.02%
Prevalence	64.96%

Table 3: Part of Statistics

The confusion matrix shows that the model performs consistently in predicting high ratings (1) and low ratings (0), with particularly strong performance in predicting the negative class (sensitivity of 92.82%). The accuracy reaches 89.46%, which is significantly higher than the no information rate (64.96%), indicating that the model has strong predictive power.

Conclusions

- In conclusion, this study analyzed factors influencing whether a movie receives an IMDB rating above 7 using logistic regression. Movie **length, budget, release year,** and the number of **votes** are **key indicators.**
- Certain genres (such as comedy, documentary, short films, and drama) are also sigificant.
- The model achieved **high accuracy (0.8946) and AUC (0.9566)** demonstrating a strong classification performance.
- It effectively identified high-rated movies (0.9282 sensitivity) and showed good accuracy for low-rated movies (0.8322 specificity).

Future Work

- Some genre variables (e.g., genreAnimation) show inconsistent effects compared to the actual rating distribution. Adding interaction terms such as genre × length may better capture context-specific effects and explain contradictions.

- Rating dynamics could be more complex. Using machine learning techniques (e.g., Random Forest) may capture more hidden patterns. Comparing different modeling approaches is encouraged to improve robustness and performance.

Thank You for Your Attention!