Prediction of movies IMDB rating greater than 7

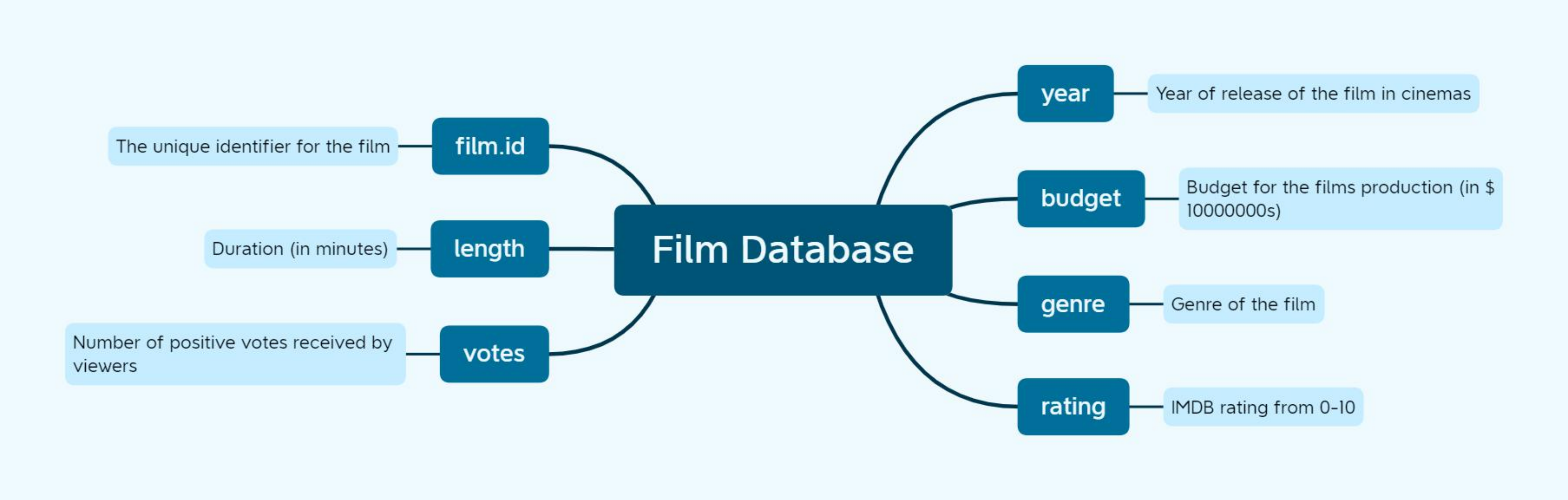
Group 5

Background

- IMDB, an online database, is the world’s most popular and authoritative source for movie, TV and celebrity content.
- The IMDB offers a rating scale that allows users to rate films on a scale of one to ten. Then the IMDB collected all the rating and put it on public for other audiences as reference.

IMDB film dataset

- The dataset contains a variety of information on all films. The data in this project is about 2387 different films, we will have access to the following variables:



Film Database
film.id
The unique identifier for the film
length
Duration (in minutes)
votes
Number of positive votes received by viewers
year
Year of release of the film in cinemas
budget
Budget for the films production (in $ 10000000s)
genre
Genre of the film
rating
IMDB rating from 0-10

Problem and Task

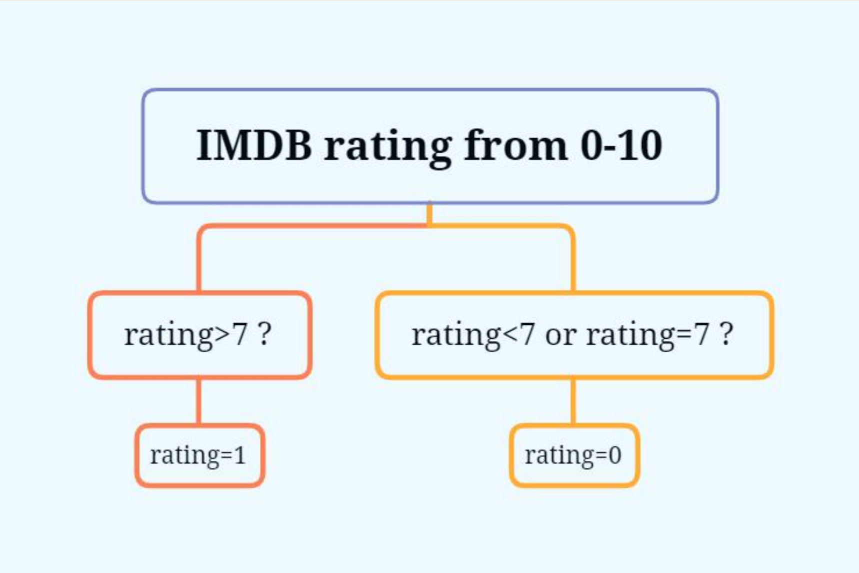
- Problem

Which properties of films influence whether they are rated by IMDB as greater than 7 or not?

- Task

Using a Generalized Linear Model (GLM) to conduct an analysis and answer the question above.

Data processing



IMDB rating from 0-10

rating>7 ?

rating<7 or rating=7 ?

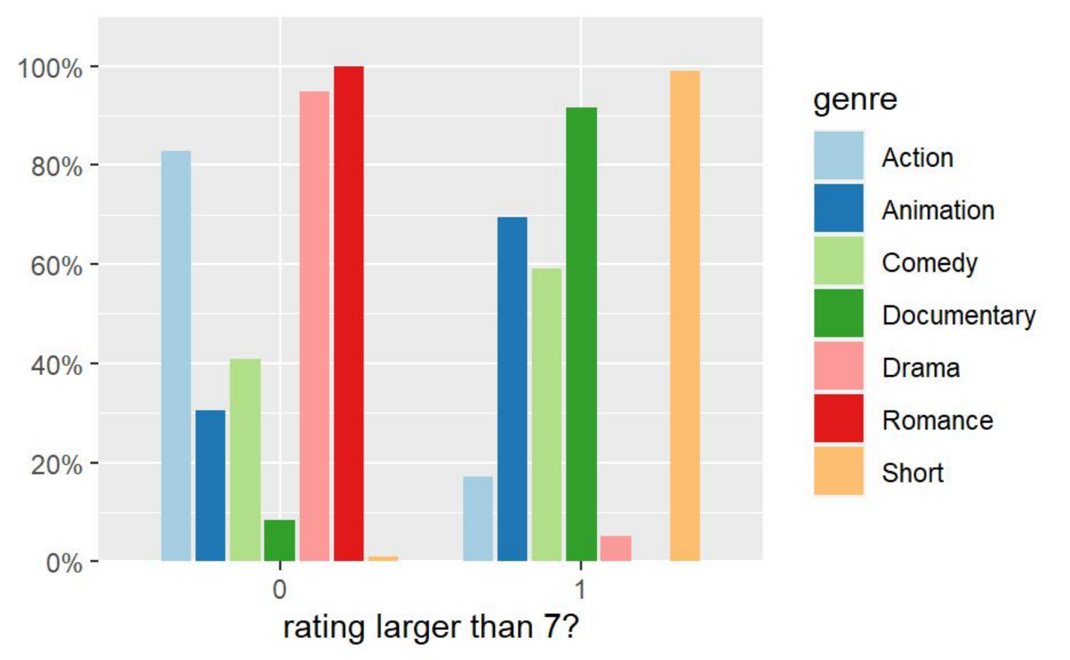
rating=1

rating=0

In this way we can divide the rating as greater than 7 or not, and the dataset can be shown as the right table:

year	length	budget	votes	genre	Rating.large7
2003	75	10.9	17	Action	0
2004	120	19.6	21	Documentary	1
1959	106	12.0	14	Drama	0
1970	101	15.5	24	Comedy	1
1996	4	12.8	5	Short	1
1968	6	10.0	12	Animation	0
1967	170	8.0	12	Romance	0
…	…	…	…	…	…

The distribution of rating.large7 by genre

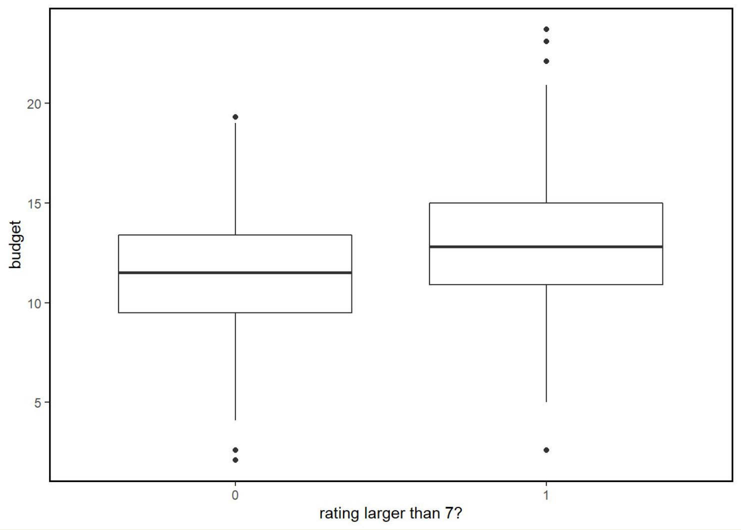


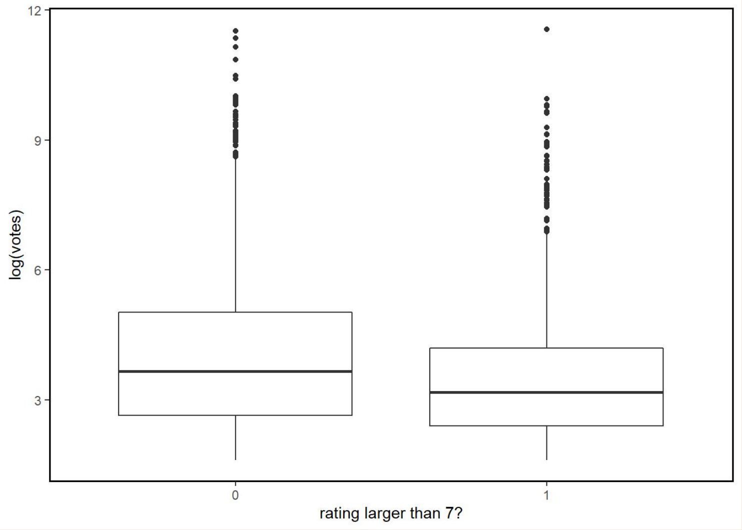
- In plot, the largest genre amongst the movies which rating is larger than 7 is Short and amongst the movies which rating is less than or equal to 7 is Romance.
- In table, Action and Drama have higher proportion which rating is less than or equal to 7, and Comedy has a higher proportion in rating larger than 7.

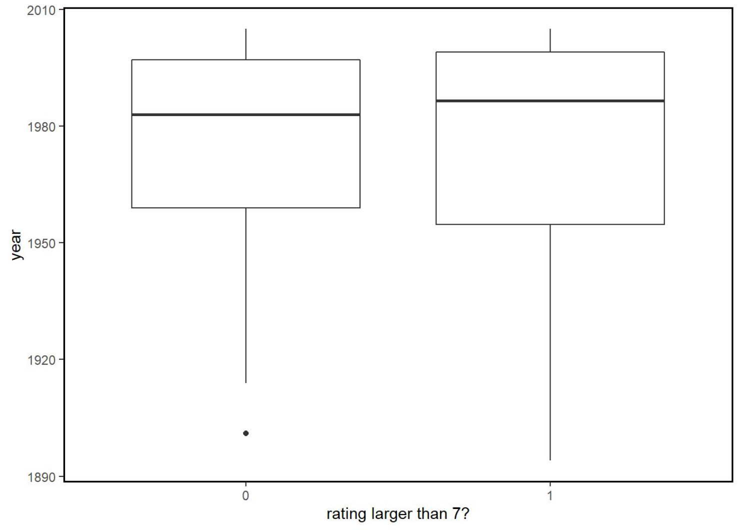
Rating.large7	Action	Animation	Comedy	Documentary	Drama	Romance	Short
0	38.0% (563)	3.3% (49)	15.1% (224)	0.7% (11)	41.8% (620)	1.0% (15)	0.1% (1)
1	14.4% (117)	13.7% (111)	40.0% (325)	14.9% (121)	4.2% (34)	0.0% (0)	12.8% (104)

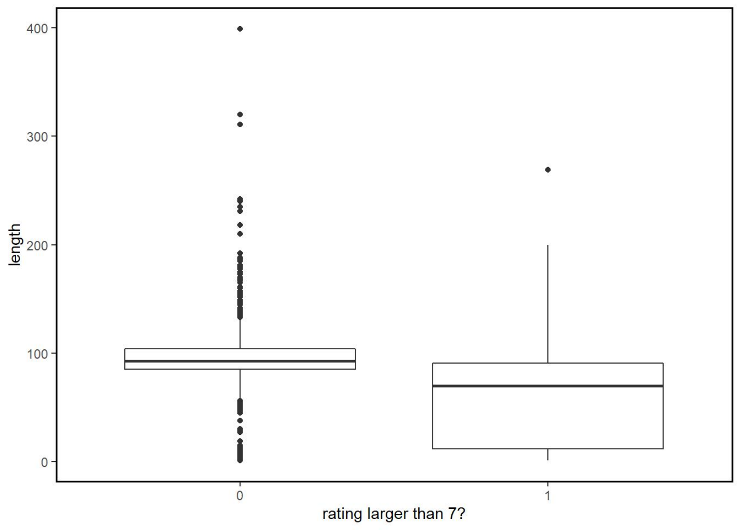
Only!

The distribution of rating.large7 by other numerical variables









In the boxplot, there is a substantial overlap between two rating especially with variable of Year

think

If the variable of Year has no significant effect on rating?

Choosing which variables need to be removed.

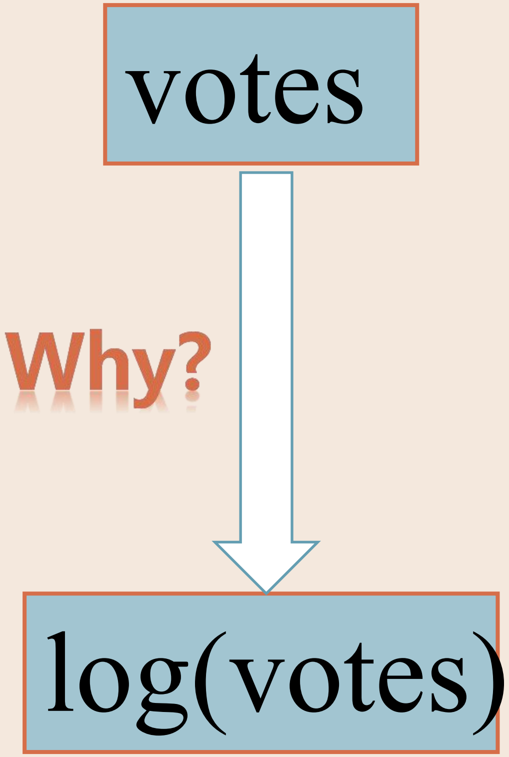
Model	Model with variables	AIC
Model 1	rating.large7 ~ length + budget + genre + log.votes + year	1304.73
Model 2	rating.large7 ~ length + budget + genre + log.votes	1303.21

Since

1303.21(Model 2 without the variable Year) < 1304.73(Model 1 with Year)

We choose the Model 2 which removes the variable Year, but obtains other four variables.

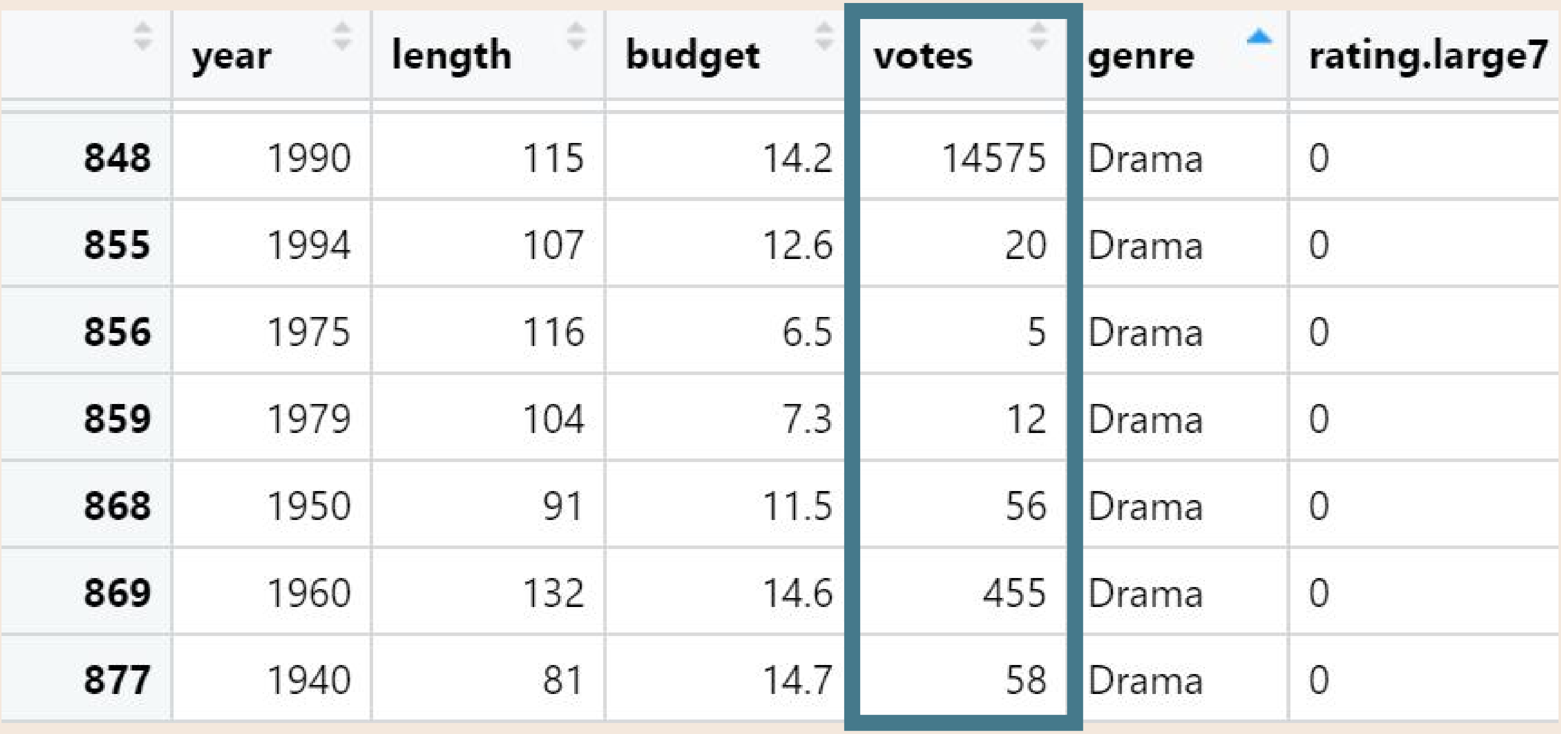
Explanation of log transformation



votes

Why?

log(votes)



	year	length	budget	votes	genre	rating.large7
848	1990	115	14.2	14575	Drama	0
855	1994	107	12.6	20	Drama	0
856	1975	116	6.5	5	Drama	0
859	1979	104	7.3	12	Drama	0
868	1950	91	11.5	56	Drama	0
869	1960	132	14.6	455	Drama	0
877	1940	81	14.7	58	Drama	0

Big gap between the value of votes, which might have an influence on the result

Comparing different link functions

Model	Link function	AIC	BIC
Model 1 (logit)	$g(p_i)=log(\frac{p_i}{1-p_i})$	1303.21	1360.6
Model 2 (probit)	$g(p_i)=\phi^{-1}(pi), pi=\phi(\frac{x_i-\mu}{\sigma})$	1328.18	1385.57
Model 3 (cloglog)	$g(p_i)=log[-log(1-p_i)]$	1402.13	1459.51

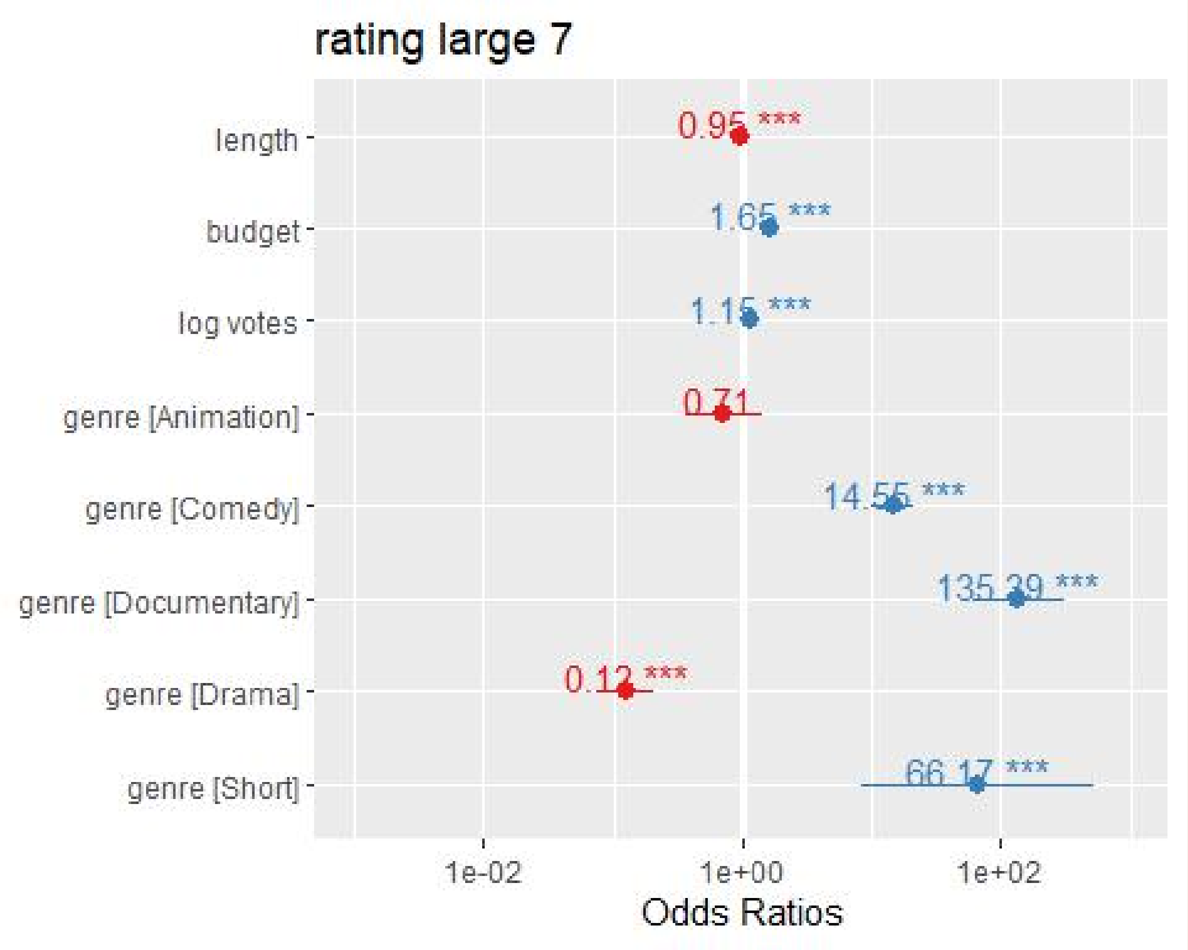
- By comparing the AIC and BIC of different link function, we choose ‘logit’ link function to fit our model with lowest AIC and BIC.
- While the probability of Romance is 1, which is unsuitable in the logit and cloglog link function. We should delete the Romance with usage of logit link function.

Residual Deviance

$$D_0 - D_1 = 2982.5 - 1283.2 = 1699.3 > \chi^2(0.95, 9) = 16.91898$$

It is clear that the difference in deviance statistics between the two models is very large, so that the model we chose is significantly better than the null model.

Odds ratios of model1



- This figure shows that ratings are negatively correlated with duration and positively correlated with budget and votes. Documentary, comedy and short is more easily praised than action films, while animation and drama is not as well praised as action films.
- The number of romantic films rated below 7 is 0, which makes some errors in model calculation, so it shows a red line on the figure. However, we have verified with code that removing it will not affect the results irrelevant to it.

Final Model

$$log(\frac{p(rating.large7 = 1)}{1 - p(rating.large7 = 1)}) =- 3.9 + (-0.05 * length_i) + 0.5 * budget_i + 0.14 * log(votes_i) - 0.34 * I(genre = Animation) + 2.68 * I(genre = Comedy) + 4.91 * I(genre = Documentary) - 2.08 * I(genre = Drama) + 4.19 * I(genre = Short)$$

$$I(genre = Animation) = \begin{cases} 1, genre = Animation \\ 0, otherwise \end{cases},$$

$$I(genre = Comedy) = \begin{cases} 1, genre = Comedy \\ 0, otherwise \end{cases},$$

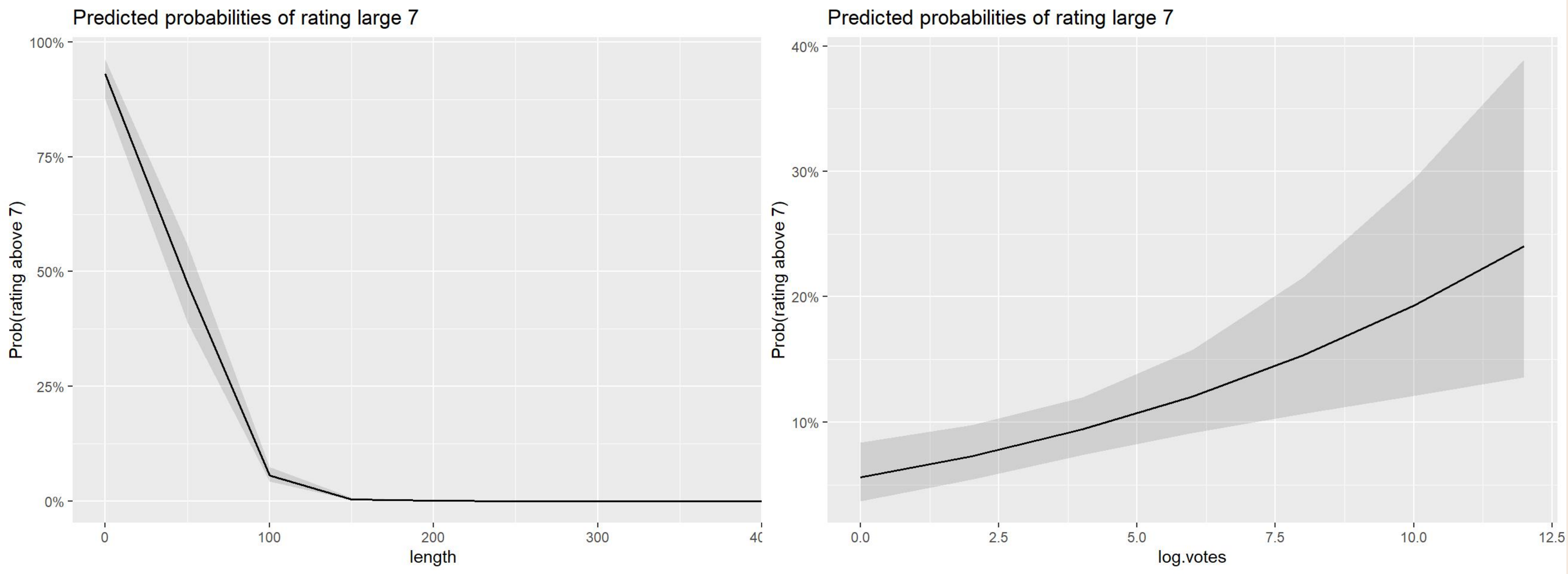
$$I(genre = Documentary \ = \begin{cases} 1, genre = Documentary \\ 0, otherwise \end{cases},$$

$$I(genre = Drama) = \begin{cases} 1, genre = Drama \\ 0, otherwise \end{cases},$$

$$I(genre = Short) = \begin{cases} 1, genre = Short \\ 0, otherwise \end{cases},$$

When all indicative functions are 0, it is the prediction of action film. The Romance is not contained in here, because its probability is 0 and can not fit the function.

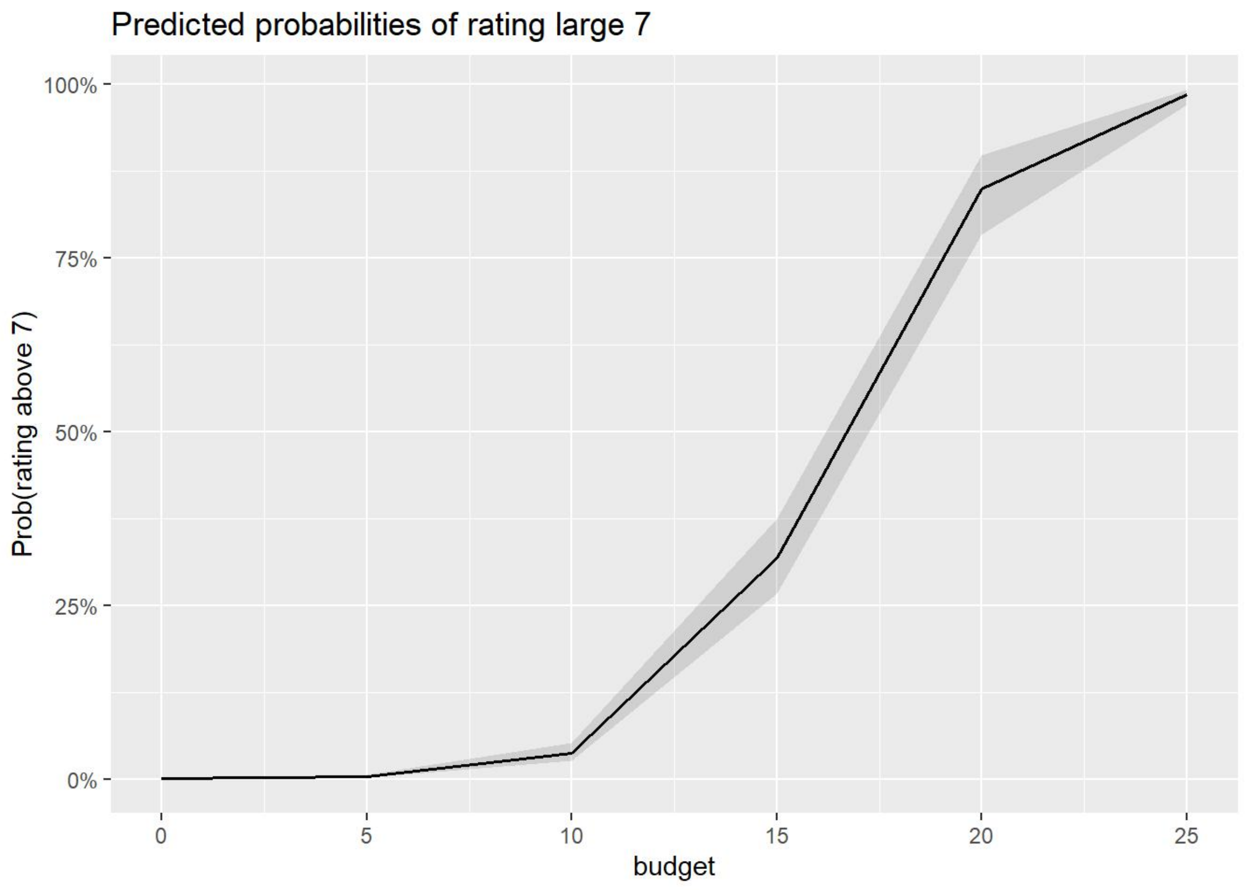
Prediction

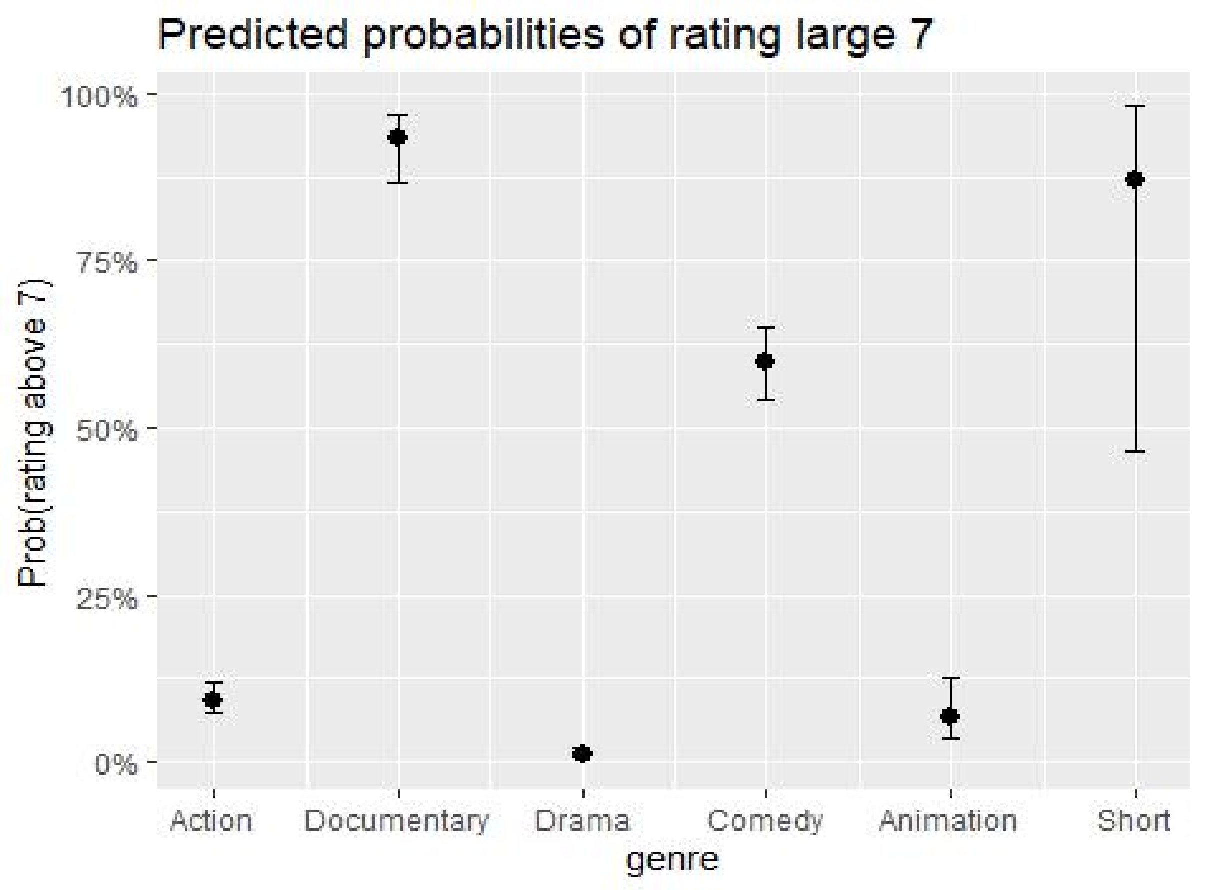


Predicted probabilities of rating large 7

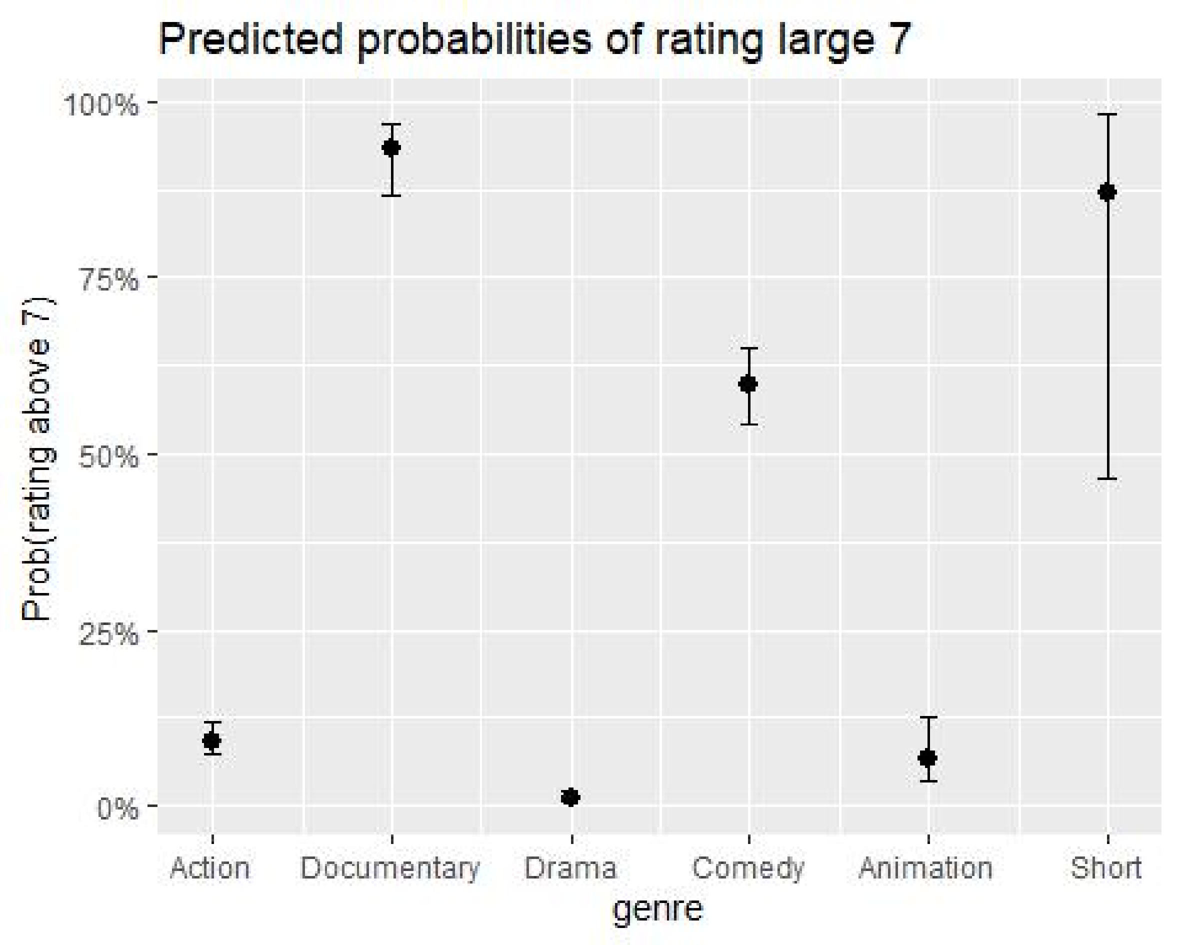
Predicted probabilities of rating large 7

Prediction


Predicted probabilities of rating large 7
Prob(rating above 7)
100%
75%
50%
25%
0%
0
5
10
15
20
25
budget


Predicted probabilities of rating large 7
Prob(rating above 7)
100%
75%
50%
25%
0%
Action
Documentary
Drama
Comedy
Animation
Short
genre

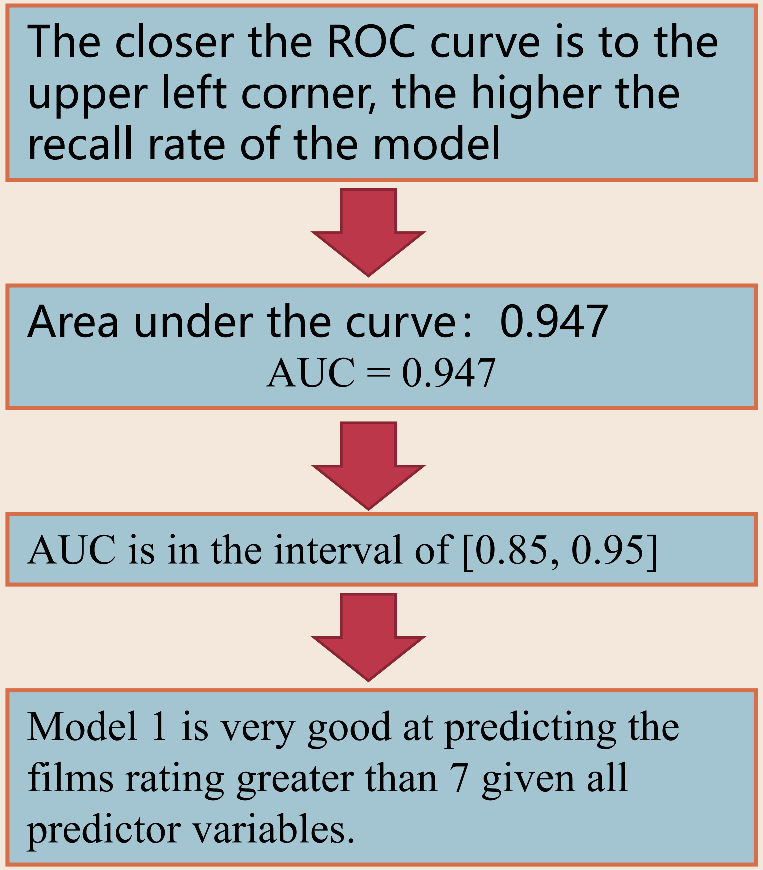
Prediction



- From the figure, the genre of Documentary has the the highest possibility of rating above 7, then the genre of Short. The genre of Action, Drama and Animation has a low possibility of rating above 7.
- For the genre of Romance, all rating of film in this dataset in lower than 7. This may be caused by too few Romance samples in this data set.

Model checking and diagnostic

ROC curve and AUC

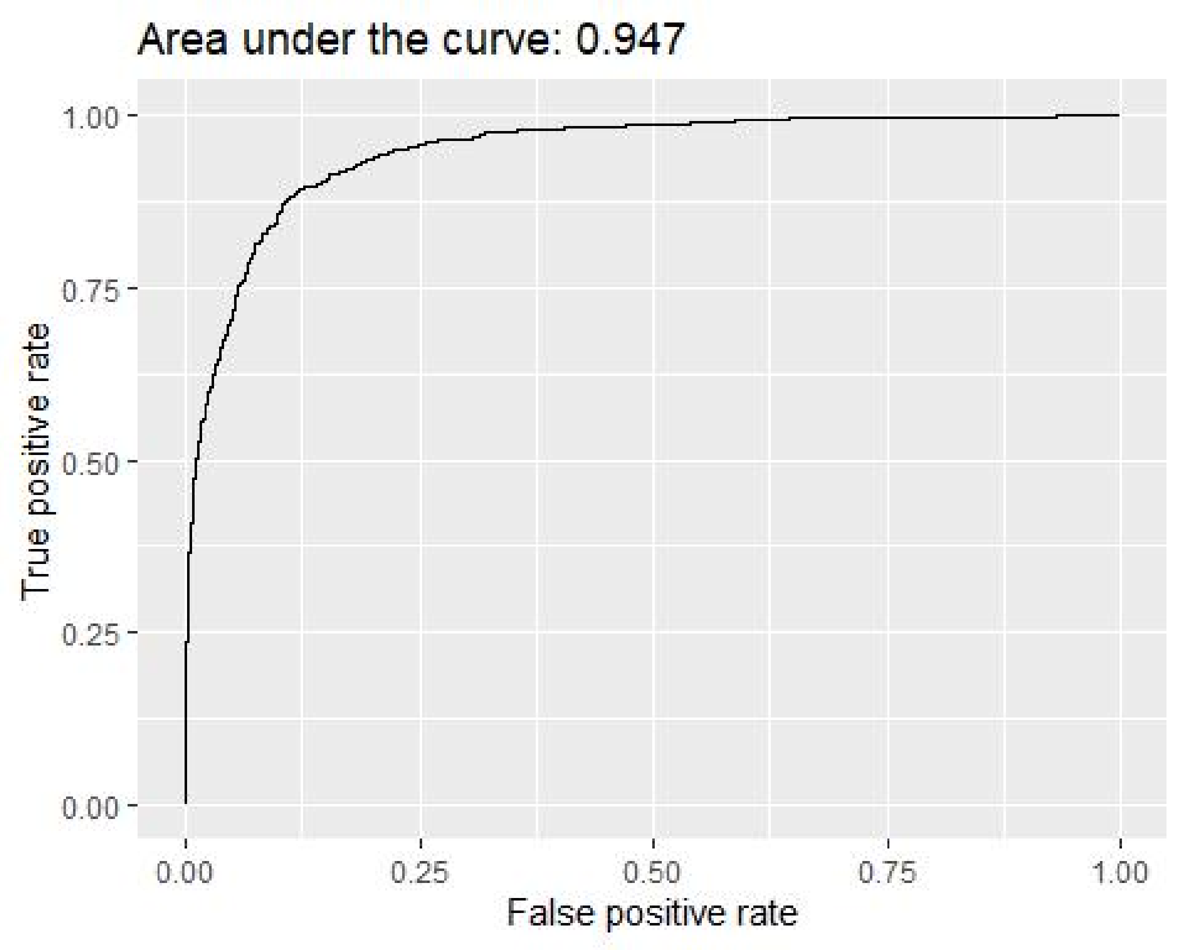


The closer the ROC curve is to the upper left corner, the higher the recall rate of the model

Area under the curve： 0.947
AUC = 0.947

AUC is in the interval of [0.85, 0.95]

Model 1 is very good at predicting the films rating greater than 7 given all predictor variables.

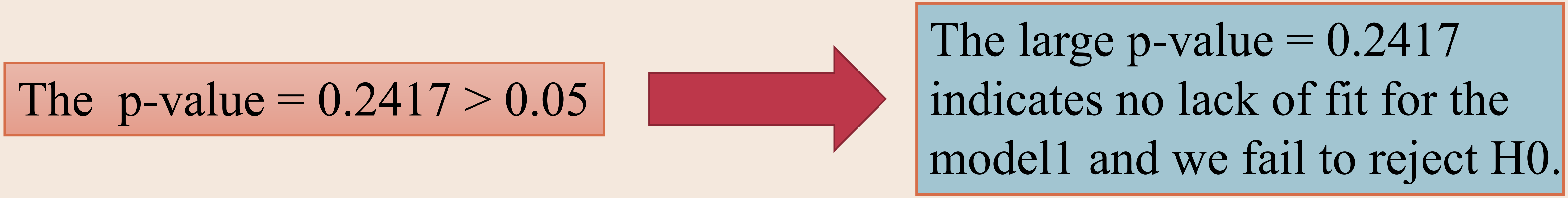


Model checking and diagnostic

Hosmer-Lemeshow goodness of fit test

H0: Model1 fits the data well

H1: Model1 is not a good fit for the data



The p-value = 0.2417 > 0.05

The large p-value = 0.2417 indicates no lack of fit for the model1 and we fail to reject H0.

Conclusion

Following the result of model checking and diagnostics, Model 1 fits the data well, so we can say our final model is what we want.

$$\begin{aligned} log(\frac{p(rating.large7 = 1)}{1 - p(rating.large7 = 1)}) \\ =- 3.9 + (-0.05length_i) + 0.5 * budget_i + 0.14 * log(votes_i) + \\ - 0.34 * I(genre = Animation) + 2.68 * I(genre = Comedy) \\ + 4.91 * I(genre = Documentary) - 2.08 * I(genre = Drama) \\ + 4.19 * I(genre = Short) \end{aligned}$$

$$I(genre = Animation) = \begin{cases} 1, genre = Animation \\ 0, otherwise \end{cases},$$

$$I(genre = Comedy) = \begin{cases} 1, genre = Comedy \\ 0, otherwise \end{cases},$$

$$I(genre = Documentary) = \begin{cases} 1, genre = Documentary \\ 0, otherwise \end{cases},$$

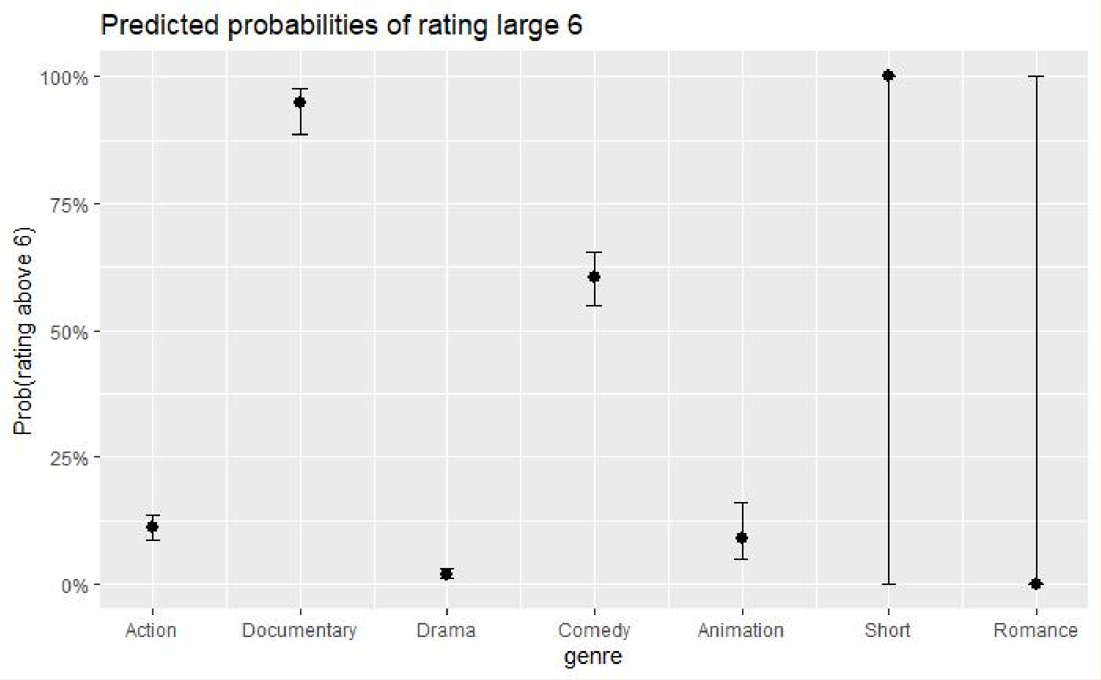
$$I(genre = Drama) = \begin{cases} 1, genre = Drama \\ 0, otherwise \end{cases},,$$

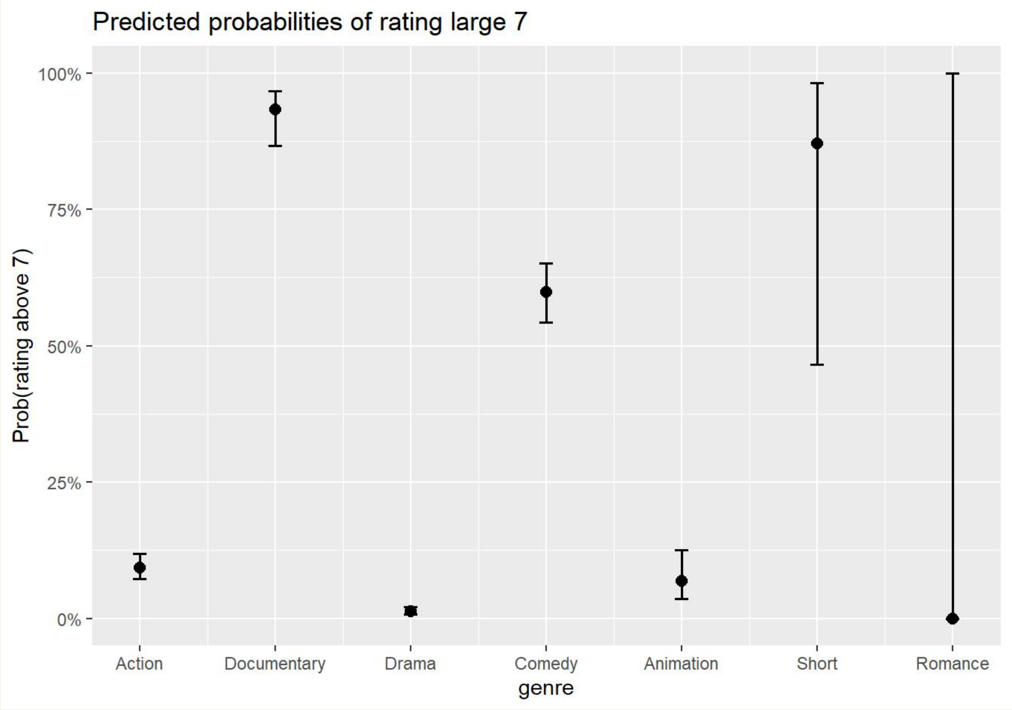
$$I(genre = Short) = \begin{cases} 1, genre = Short \\ 0, otherwise \end{cases},$$

Rating.large7	Action	Animation	Comedy	Documentary	Drama	Romance	Short
0	38.0% (563)	3.3% (49)	15.1% (224)	0.7% (11)	41.8% (620)	1.0% (15)	0.1% (1)
1	14.4% (117)	13.7% (111)	40.0% (325)	14.9% (121)	4.2% (34)	0.0% (0)	12.8% (104)

We can see there are too few Romance samples in this data set and this may lead to problems with the coefficient of Romance. This has aroused our interest in future research.

Future work





During the project, we try changing the rating to 6 and make a comparison. As we can see from the figure, the probabilities of genre of Romance that rating above 6 is still zero, instead, the probabilities of genre of Short that rating above 6 is 100%, that means in this dataset, all the Short films are rating above 6, however, all the Romance films are rating below 6. This also raises our interest in romance film research.

Future work

As we mentioned earlier, there may be too few samples of romance in this data set. So in the future, we hope to explore more data sets especially for Romance films, so as to explore the influence of the genre of romance on rating above 7. We want to know whether Romantic films have such a bad impact on rating as this data set, and we hope that with the future research, we can improve the final formula we have obtained.

VIA VERITAS VITA

UofG

Thank you!