A Generalized Linear Regression Model Analysis On IMDB ratings of films

Group 10: Linxuan Guo, PiuSheung TANG, Yiqing Wen, Chenghao YANG, Yanqi Zhu

1 Introduction

IMDB stands for Internet Movie Database. It is an online database that allows users to rate and review movies and TV shows. The aim of this project is to investigate the following question of interest related to IMDB: which properties of films influence whether they are rated by IMDB as greater than 7 or not? A generalised linear model (GLM) are used to fit the data in the formal analysis.

2 Data processing

Table 1 below shows the first five rows of the raw data.

Table 1: The first 5 rows of the raw data

film_id	year	length	budget	votes	genre	rating
49834	1963	107	11.4	225	Romance	3.1
53923	1984	NA	9.7	59	Comedy	2.3
30020	1992	32	15.4	6	Documentary	7.7
46364	2000	NA	11.5	69	Action	2.3
19967	1964	87	9.4	34	Comedy	5.5

Given the presence of certain non-applicable (NA) values within the dataset under scrutiny, which may have the potential to impact our findings, our initial step is to exclude these values from our analysis. Subsequently, creating a binary factor based on the "rating" column to help process the modeling. Lastly, the columns labeled as "film_id" and "rating" which are deemed to be invalid for our purposes are excluded.

According to the dataset, the character type values are transformed into factor type values. Finally, the levels of the response variable are set to make the graphs of the data visualization more understandable.

Table 2 below displays the first five rows of the data after processed. The last column "greater_than_7" was added to each row to present whether the rating of the film is greater than 7.

Table 2: The first 5 rows of the processed data

	year	length	budget	votes	genre	greater_than_7
1	1963	107	11.4	225	Romance	IMDB rate less than 7
3	1992	32	15.4	6	Documentary	IMDB rate greater than 7
5	1964	87	9.4	34	Comedy	IMDB rate less than 7
6	2003	95	13.3	22	Action	IMDB rate less than 7
7	1983	96	9.6	10	Drama	IMDB rate less than 7

3 Exploratory Data Analysis

The Scatterplot matrix **Figure 1** shows that most correlations between explanatory variables are really insignificant, therefore the initial model was built without censoring the variables.

```
##Plot to get an initial impression of the data
# Check correlations, distribution and print correlation coefficient
ggpairs(data= film.select, title="Correlation between explanatory variables",
        ggplot2::aes(color=greater_than_7))
```

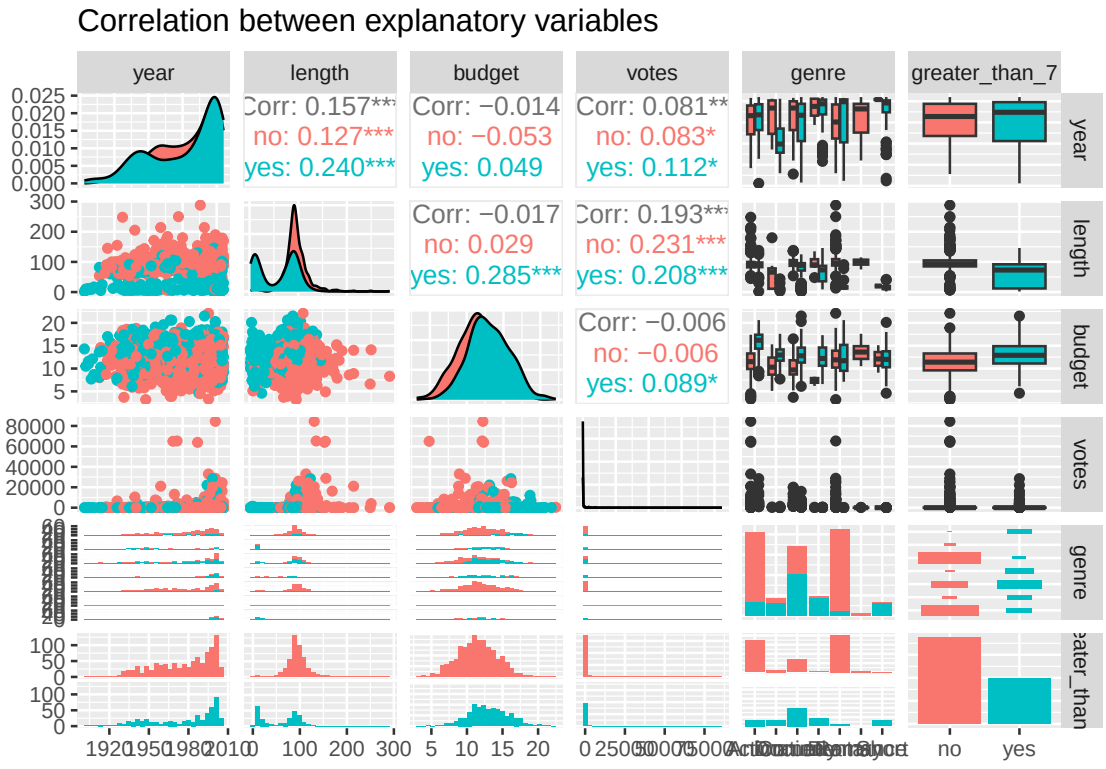

Figure 1: Correlation between explanatory variables.

Figure 2 demonstrates the relationships between the four numerical variables and the response variable Y (IMDB rate greater than 7) through multiple boxplots. These plots can be used to identify associations, trends, outliers, and non-linear relationships between the variables, thus providing valuable insights for further analysis and model development.

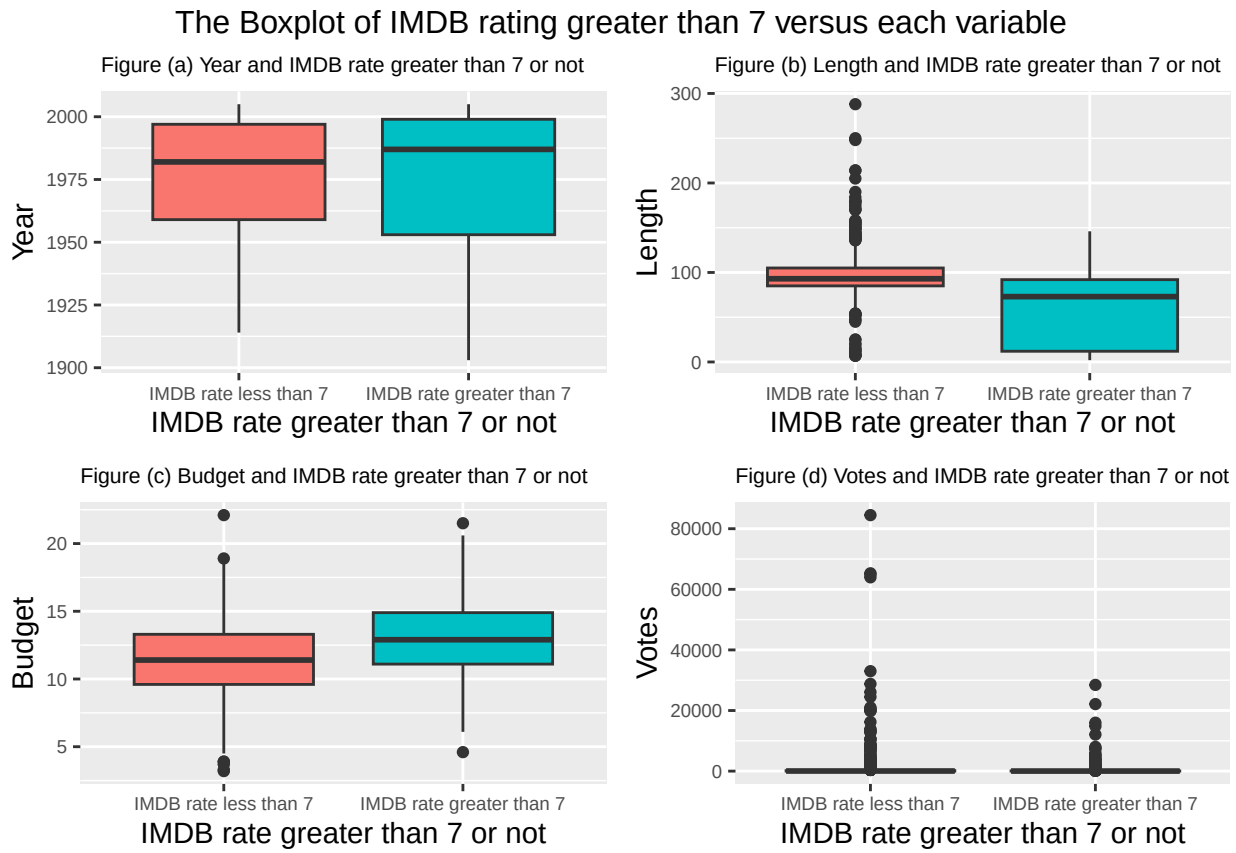

Figure 2: The Boxplot of IMDB rating greater than 7 versus each variable

-**Figure (a)**: There are **zero** potential outliers. The variable "year" has little effect on rating of the films. The distributions of the left and right bars look the same, and the median values of "year" of different rated films are basically in the same level range.

-**Figure (b)**: There are **many potential outliers**. As we can see from the graph, the "length" of films with high ratings are more densely distributed, while the "length" of films with poor ratings are spread out.

-**Figure (c)**: There are **eight** potential outliers. In terms of the "budget", popular films have slightly higher budgets, but the difference between that is not too significant.

-**Figure (d)**: There are **all potential outliers**. The distributions appear to be similar, but the distribution shown on the left clearly has more extreme values and requires further discussion.

The histogram(**Figure 3**) shows the proportions of whether IMDB rates are less or more than in different movie genres, respectively.

```
#Plot of genre against greater_than_7
ggplot(data = DataSet, aes(x= greater_than_7,  y = ..prop.., group=genre, fill=genre)) +
  geom_bar(position="dodge", stat="count") +
  labs(x = "IMDB rate greater than 7 or not", y = "Proportion",
       title = "Barplot of the proportion of whether IMDB rate greater than 7 by genre")
```

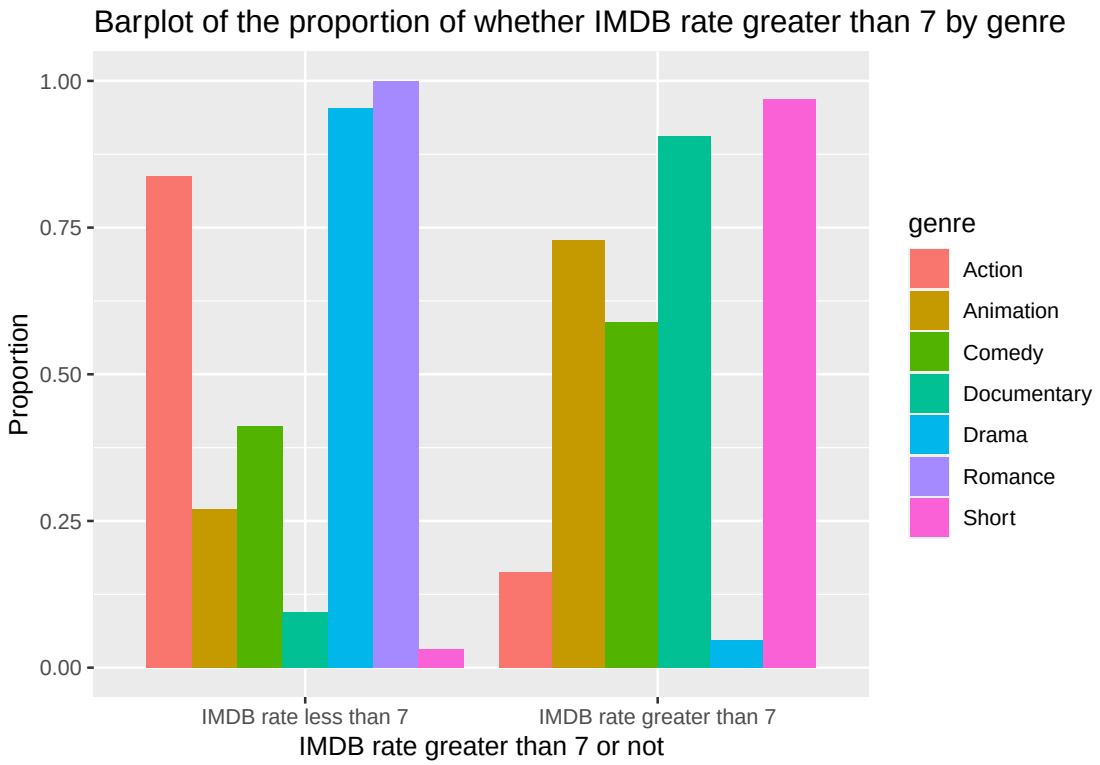

Figure 3: The Histograms of the proportion of movie categories and IMDB rate greater than 7 or not.

It is noticeable that the ratings of different genres of films vary widely, especially Action, Short films, Comedy, Drama and Romance. short films and comedies are more likely to be loved by audiences in film ratings, while Action, Drama and Romance are less likely to receive high ratings. There is a rough indication that the genre of film does have an impact on the film scores.

4 Modeling

4.1 Initial modelling

Initial modelling is carried out using the GLM function and the model results are as follows.

```
model = glm(greater_than_7 ~ year + length + budget + votes + genre, data = DataSet,
            family = binomial(link = "logit"))
model %>%
  summary()
```

```
Call:
glm(formula = greater_than_7 ~ year + length + budget + votes +
    genre, family = binomial(link = "logit"), data = DataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6363  -0.3475  -0.0829   0.2049   3.2741
```

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.150e+01  8.442e+00  -1.362  0.17326
year              4.014e-03  4.306e-03   0.932  0.35123
length           -6.783e-02  5.984e-03 -11.336  < 2e-16 ***
budget            5.810e-01  4.478e-02  12.975  < 2e-16 ***
votes             5.676e-05  2.046e-05   2.774  0.00553 **
genreAnimation   -8.599e-01  5.371e-01  -1.601  0.10940
genreComedy       3.213e+00  2.599e-01  12.360  < 2e-16 ***
genreDocumentary  5.491e+00  5.468e-01  10.042  < 2e-16 ***
genreDrama       -2.141e+00  3.505e-01  -6.107 1.02e-09 ***
genreRomance     -1.541e+01  5.649e+02  -0.027  0.97824
genreShort        2.648e+00  8.840e-01   2.995  0.00274 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1854.97  on 1435  degrees of freedom
Residual deviance:  720.94  on 1425  degrees of freedom
AIC: 742.94

Number of Fisher Scoring iterations: 15
```

Hence, the fitted model for "the probability of IMDB rate greater than 7" is given by:

$$\ln\left(\frac{p}{1-p}\right) = -11.5 + 0.004 \cdot \text{year} - 0.07 \cdot \text{length} + 0.58 \cdot \text{budget} + 0.00005 \cdot \text{votes} - 0.86 \cdot \mathbf{I}_{(Animation)} + 3.21 \cdot \mathbf{I}_{(Comedy)} + 5.49 \cdot \mathbf{I}_{(Documentary)} - 2.14 \cdot \mathbf{I}_{(Drama)} - 15.41 \cdot \mathbf{I}_{(Romance)} + 2.65 \cdot \mathbf{I}_{(Short)}$$

where $\mathbf{p}$ is the probability of IMDB rate greater than 7, and $\mathbf{1-p}$ is the probability of IMDB rate less than 7.

From the results it is clear that:

• The point estimate of odds for variable "Year" is estimated to be positive, which means that, all else being equal, films released later are more likely to receive a higher rating. Considered its high p-value in parametric test, the variable "year" has a high probability of being a non-significant variable.

• The point estimate of odds for "length" is negative, suggesting that, all else being equal, the longer the film the more likely it is to receive a poor rating.

• A positive odds estimate for "budget" means that, all else being equal, the higher the budget, the more likely it is that the film will have a high rating.

• A positive odds point estimate for the "votes" variable suggests that, all else being equal, the higher the number of votes, the more likely the film is to receive a high rating.

• Different "genre" of films all have different effects on film ratings. For example, when all else is equal, films in the Animation/ Drama/ Romance category are less likely to be well-received; films in the Comedy/ Documentary/ Short category are more likely to be well-recognised.

4.2 Model optimisation

Based on the initial modelling results, the variable "year" looks more likely to be a non-significant variable, so this variable will be removed and the model will be reconstructed using the remaining variables, resulting in the following:

```
model1 = glm(greater_than_7 ~ length + budget + votes + genre, data = DataSet,
             family = binomial(link = "logit"))
#Summarize the second model
model1 %>%
  summary()
```

```
Call:
glm(formula = greater_than_7 ~ length + budget + votes + genre,
    family = binomial(link = "logit"), data = DataSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6107  -0.3395  -0.0820   0.2024   3.2722

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.653e+00  6.256e-01  -5.839 5.24e-09 ***
length          -6.661e-02  5.814e-03 -11.456  < 2e-16 ***
budget           5.796e-01  4.469e-02  12.970  < 2e-16 ***
votes            5.784e-05  2.026e-05   2.855  0.00431 **
genreAnimation  -8.336e-01  5.359e-01  -1.556  0.11979
genreComedy      3.221e+00  2.602e-01  12.379  < 2e-16 ***
genreDocumentary 5.550e+00  5.418e-01  10.243  < 2e-16 ***
genreDrama      -2.137e+00  3.486e-01  -6.131 8.76e-10 ***
genreRomance    -1.549e+01  5.587e+02  -0.028  0.97788
genreShort       2.750e+00  8.782e-01   3.132  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1854.97  on 1435  degrees of freedom
Residual deviance:  721.81  on 1426  degrees of freedom
AIC: 741.81

Number of Fisher Scoring iterations: 15
```

Therefore, the fitted generalised linear model is:

$$\ln\left(\frac{p}{1-p}\right) = -3.65 - 0.07 \cdot \text{length} + 0.58 \cdot \text{budget} + 0.00005 \cdot \text{votes} - 0.83 \cdot \mathbf{I}_{(Animation)} + 3.22 \cdot \mathbf{I}_{(Comedy)} + 5.55 \cdot \mathbf{I}_{(Documentary)} - 2.14 \cdot \mathbf{I}_{(Drama)} - 15.49 \cdot \mathbf{I}_{(Romance)} + 2.75 \cdot \mathbf{I}_{(Short)}$$

5 Conclusions

After adjustment and re-modelling, it is found that most of the variables were significant, but some variables in genre were still non-significant. Predicting the results of the model allows us to calculate the probability of an IMDB score greater than 7.

Assuming a probability greater than 0.5, the predicted outcome is classified as a rating greater than 7 and vice versa. This gives a prediction accuracy of 90%, which is relatively close to 1. Therefore, this model is considered to have a good performance in terms of explanatory power and level of prediction.

```
plot_model(model1, show.values = TRUE, show.p = FALSE)
```

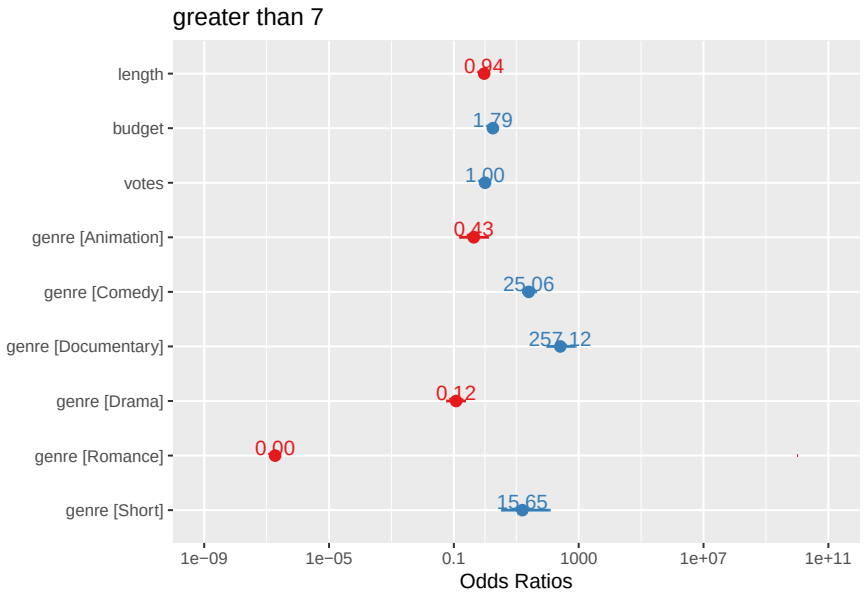

Figure 4: Odds of IMDB rate greater than 7

Figure 4 above shows the odds ratios of each variable.A red number indicates that the variable it corresponds to has a negative impact on the movie rating, and conversely, the variable corresponding to a blue number has a positive impact on the movie.

The value of odds can be interpreted as follows:

- For two films which differ by one unit in length, the Longer film is 0.97 times more likely to have a rating above 7 than the shorter one.
- For two films which differ by one unit in the budget, the film with a larger budget is 1.79 times more likely to have a rating above 7 than that with less budget.

According to all the analysis and the model presented above, the factors that influence whether a film rating is greater than 7 are: length, budget, votes, and genre.

Figure 5 above shows that shorter film are more likely to be predicted as a highly rated film, this may due to some correlation between "length" and "short" in the film genres. By implementing visualizations on the data, most of the short films received positive reviews.

Figure 6: Similar to the findings of the previous exploratory data analysis, it can be observed that a higher "budget" is positively correlated with the likelihood of a film being well-received in the future.

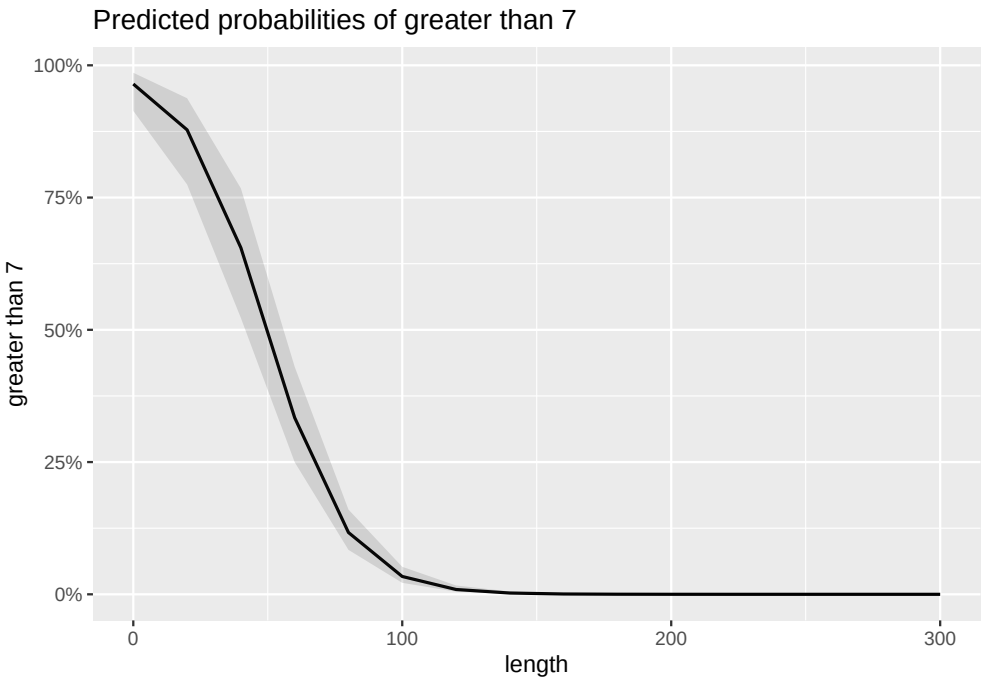

Figure 5: Length:predicted probabilities of greater than 7

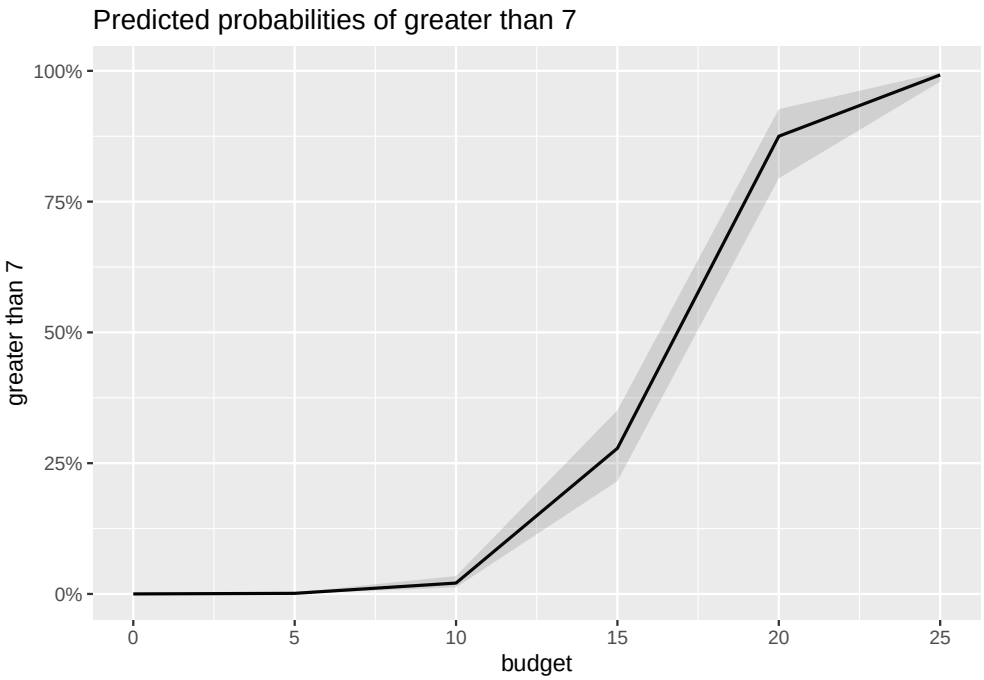

Figure 6: Budget:predicted probabilities of greater than 7

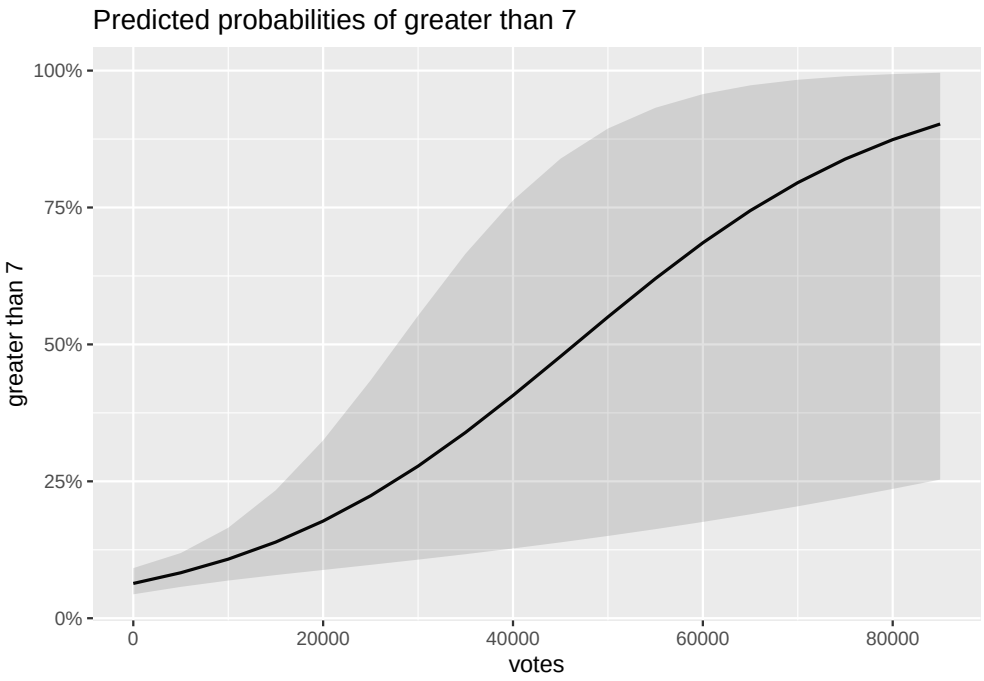

Figure 7: Votes:predicted probabilities of greater than 7

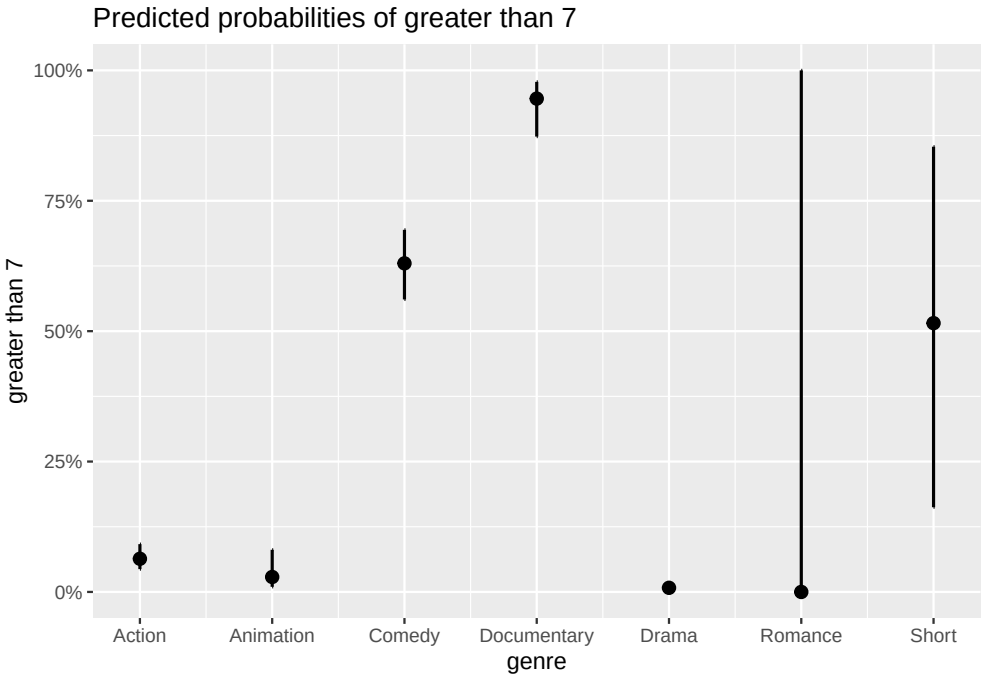

Figure 8: Genres:predicted probabilities of greater than 7

Figure 7: The variable "votes" play a role in determining a film's rating. However, their predictive power is limited for the reason that as the number of votes increases, the range of predicted outcomes becomes wider.

Figure 8 above shows that Documentary is predicted to have a high probability to be rated greater than 7. Comedy and short films are predicted to have a moderate probability to be rated above 7. The probability of other films receiving a rating above 7 is low.