The Impact of film's features on IMDB Ratings: A logistic Regression approach

Yujie Tang, Jialu FU, Weiqing GUO, Bashiru Mukaila, Wanding Wang

2023-03-14

1 Introduction

The entertainment industry happens to be a competitive market and the film industry is a subset of it. The stakeholders in the film industry are interested in features that make their films successful. One key aspect of a film's success is its rating on platforms such as IMDB, which can influence audience perception and drive revenue.

The aim of this project is to investigate the relationship between the properties of films and their IMDB ratings. Specifically, we want to use a logistic regression model to know which properties of a film influence whether a film is rated by IMDB as greater than 7 or not, using variables such as year of release, length, budget, number of votes, and genre.

Section 2 speaks to the exploratory data analysis of IMDP ratings and explores the relationship between the rating and the properties of the films. Section 3 contains the results from the logistic regression model. While Section 4 gives the final remark to the research.

```
# Importing the data set
group_08 <- read.csv("dataset8.csv")

# Creating an additional column that indicate yes or no if rating is greater than 7
imdb_group_08 <- group_08 %>%
  mutate(Rating_above_7 = ifelse(rating > 7,"Yes","No"))

# Convert chr to factor
imdb_group_08$Rating_above_7 <- as.factor(imdb_group_08$Rating_above_7)
imdb_group_08$genre <- as.factor(imdb_group_08$genre)
```

2 Exploratory Data Analysis

3 Formal Data Analysis

3.1 Logistic regression with one categorical explanatory variable

An IMDB rating of over 7 indicates an excellent and highly-regarded film. Consequently, investigating which properties influence movie ratings is an intriguing research question.

Consider determining whether the categorical variable *genre* has an effect on the movie rating above seven. Display the distribution by creating a barplot of *genre* and *Rating_above_7*:

```r
# Compute the proportions of ratings
genre_counts <- as.data.frame(table(imdb_group_08$genre,
                                    imdb_group_08$Rating_above_7))
colnames(genre_counts) <- c('genre','Rating_above_7','count')

genre_counts <- genre_counts %>%
  group_by(genre) %>%
  mutate(total = sum(count)) %>%
  ungroup() %>%
  mutate(proportion = round(count / total, 2))

# Plot the proportions of Yes/No Ratings by Genre
genre_counts %>%
  ggplot(aes(genre, proportion, fill = Rating_above_7)) +
  geom_col(position = "dodge") +
  labs(title = "", x = "Genre", y = "Proportion") +
  scale_fill_manual(values = c("#BEAED4", "#A6BDDB"), labels = c("No", "Yes")) +
  theme(axis.text.x = element_text(size = 7),
        axis.text.y = element_text(size = 7),
        legend.title = element_text(size = 9),
        legend.text = element_text(size = 7),
        legend.key.size = unit(10, "pt"))
```

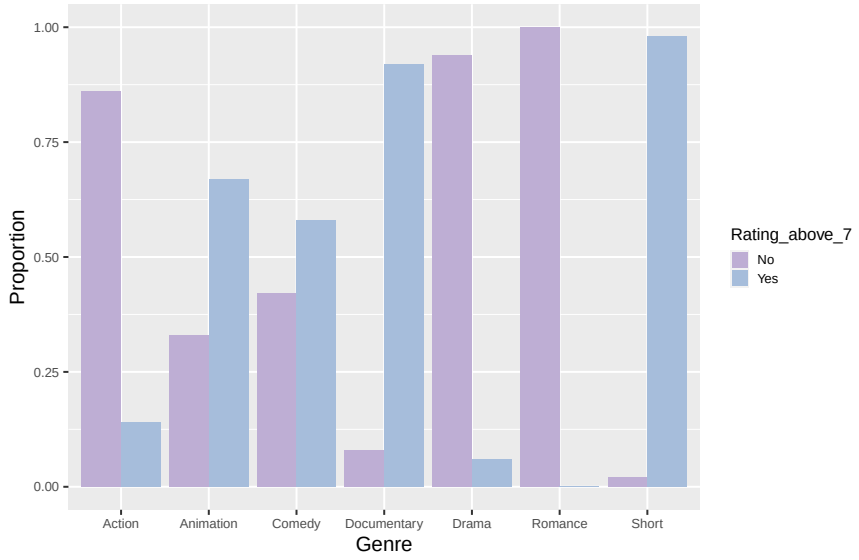

Figure 1: Proportions of Yes/No Ratings by Genre.

From Figure 1, we can observe that the proportion of low ratings for Action, Drama, and Romance genres is relatively high, exceeding 80%, while the proportion of high ratings for Documentary and Short genres is higher, exceeding 90%. The genre of Comedy has a relatively balanced proportion of high and low ratings.

3.1.1 Model specification and estimation

Therefore, we further investigate the impact of movie genres on movie ratings. Here $p_i = Prob(Rating\_above\_7 = Yes)$, with $genre_i$ being the type of the film for $i = 1, 2, ..., 2847$. The model we will consider is of the form:

$$g(p_i) = log(\frac{p_i}{1-p_i}) = \alpha + \beta \cdot genre_i$$

and we fit it in R as follows:

```
model.genre <- glm(Rating_above_7 ~ genre,
                   data = imdb_group_08,
                   family = binomial(link = "logit"))
```

```
model.genre %>%
  summary()
```

```
##
## Call:
## glm(formula = Rating_above_7 ~ genre, family = binomial(link = "logit"),
##     data = imdb_group_08)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8921  -0.5417  -0.3409   0.4200   2.3976
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.84494    0.09865 -18.702  < 2e-16 ***
## genreAnimation     2.54653    0.18731  13.595  < 2e-16 ***
## genreComedy        2.15141    0.12537  17.160  < 2e-16 ***
## genreDocumentary   4.22875    0.30618  13.811  < 2e-16 ***
## genreDrama        -0.97113    0.18244  -5.323 1.02e-07 ***
## genreRomance     -13.72113  261.39713  -0.052    0.958
## genreShort         6.01161    0.71936   8.357  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3621.5  on 2846  degrees of freedom
## Residual deviance: 2309.4  on 2840  degrees of freedom
## AIC: 2323.4
##
## Number of Fisher Scoring iterations: 14
```

From the output, the estimated coefficients for *Action (Intercept)*, *Drama*, and *Romance* are negative, suggesting that movies of these three genres are more likely to receive ratings below 7, consistent with our previous observation. However, the coefficient for *Romance* is not significant, which may be due to a lack of observations with ratings above 7. All other estimated coefficients are significant at the 5% level of significance.

3.1.2 Model inference and interpretation

The baseline category for our binary response is *Yes*, and the baseline category for our explanatory variable is *Action*. Hence the estimates from the logistic regression model are on the log-odds scale and apply to movies with a rating above 7:

$$\log(\frac{\hat{p}}{1-\hat{p}}) = -1.84 + 0 \cdot \text{II}_{genre}(Action)$$

$$\log(\frac{\hat{p}}{1-\hat{p}}) = -1.84 + 2.55 \cdot \text{II}_{genre}(Animation)$$

...

where $\text{II}_{genre}(\cdot)$ is an indicator function, that takes a specific movie genre as input and returns 1 if the input variable matches the genre, and 0 otherwise.

The point estimate and the corresponding 95% confidence interval for the log-odds of each movie genre can be obtained, as shown in the Figure 2:

```
plot_model(model.genre, show.values = TRUE, transform = NULL,
           title = "Log-Odds (Rating above 7)", show.p = FALSE)
```

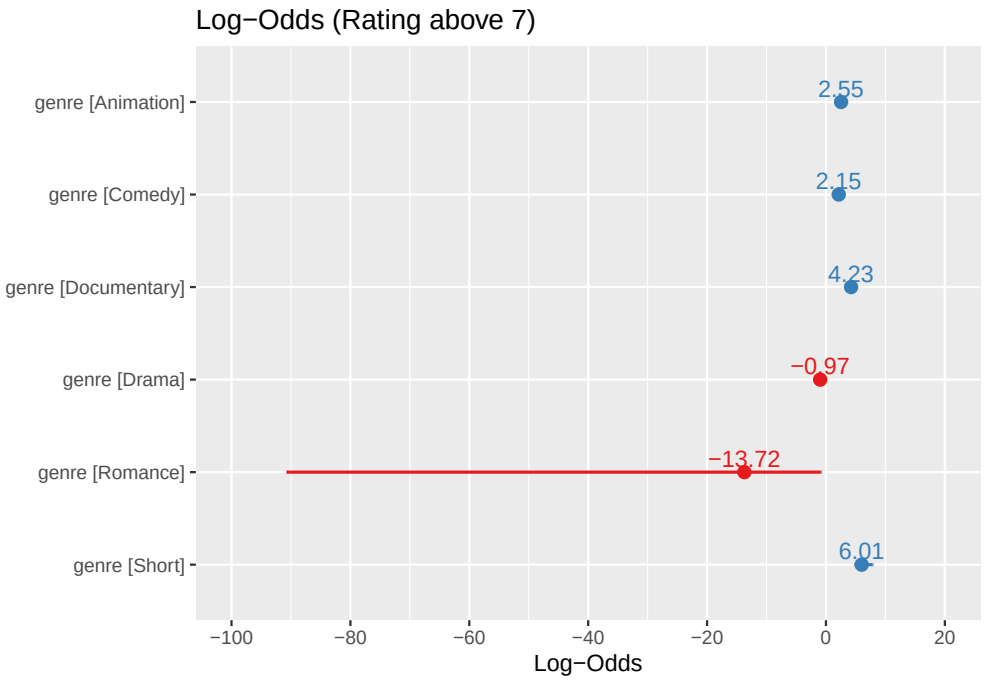

Figure 2: Log-Odds (Rating above 7).

Consider using the estimated coefficients to quantify the effect of genre, the regression coefficients on the odds scale are given by:

```
# Coefficients on the odds scale
model.genre %>%
  coef() %>%
  exp()
```

```
##      (Intercept)   genreAnimation      genreComedy genreDocumentary
##     1.580345e-01     1.276271e+01     8.596986e+00     6.863154e+01
##       genreDrama     genreRomance       genreShort
##     3.786541e-01     1.098982e-06     4.081387e+02
```

Similarly, the point estimate and the corresponding 95% confidence interval for the odds scale of each movie genre can be obtained, as shown in the Figure 3:

```
plot_model(model.genre, show.values = TRUE,
           title = "Odds (Rating above 7)", show.p = FALSE) +
  theme(axis.text.x = element_text(size = 2))
```

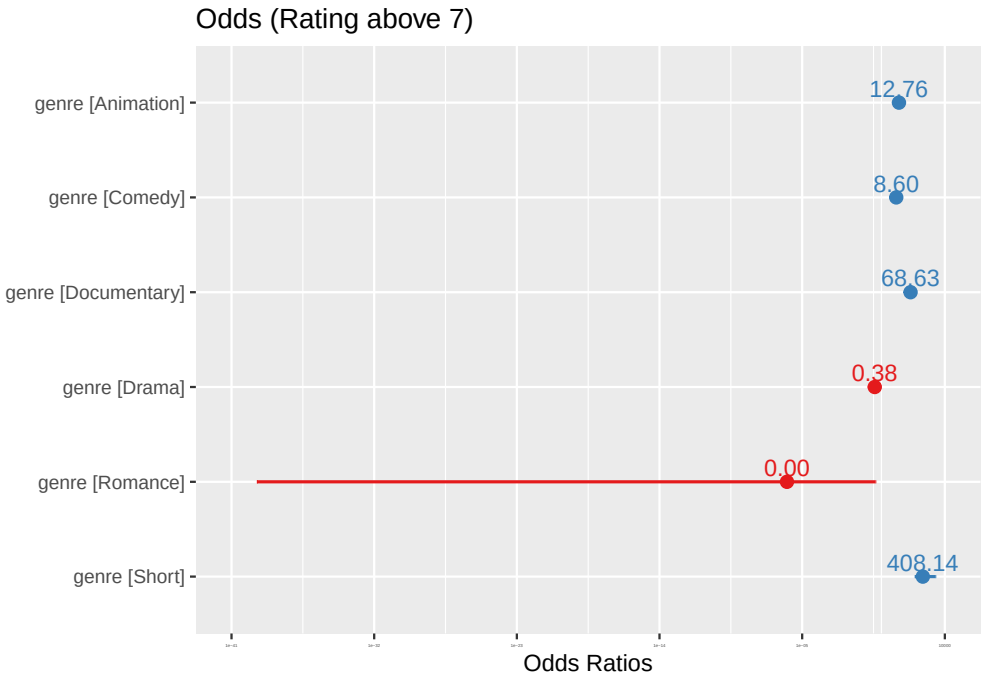

Figure 3: Odds (Rating above 7).

The *Action (Intercept)* gives us in the Action movie genre, the probability of a high rating is approximately 0.16 times that of a low rating. And the odds of the rating being above 7 for *Animation* are 12.76 times greater than the odds if the movies are *Action*. For the movie genres *Documentary* and *Short*, the probability of having a high rating is 68.63 and 408.14 times higher than that of *Action*, respectively.

Next, we calculate the estimated probability of a rating above 7 for each movie genre using the following formula:

$$\hat{p} = \frac{\exp(\hat{\alpha} + \hat{\beta} \cdot \mathrm{II}_{genre}(\cdot))}{1 + \exp(\hat{\alpha} + \hat{\beta} \cdot \mathrm{II}_{genre}(\cdot))}$$

The estimated probabilities of each type of movie reaching a score of 7 or higher were obtained, and the probabilities were compared with the frequencies of the data to produce the following graphs.

```
plot_model(model.genre, type = "pred", title = "",
           axis.title = c("Genre", "Prob. of ranting being above 7"))
```

```
## $genre
```

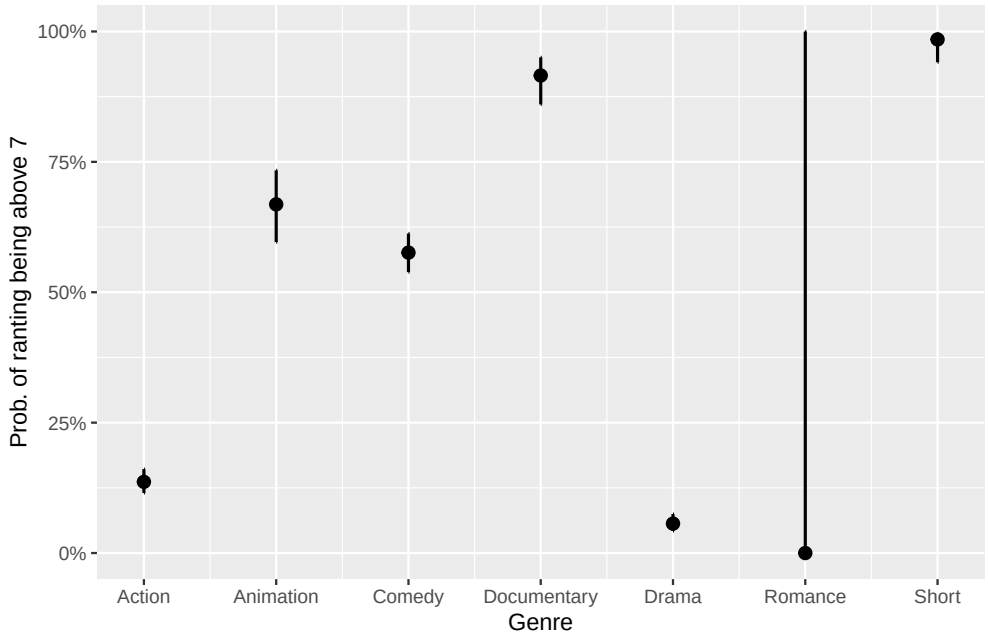

Figure 4: Estimated probability of ranting being above 7 by genre.

```r
# Probability
mod.genre.coef.logodds <- model.genre %>%
  summary() %>%
  coef()
action_prob <- plogis(mod.genre.coef.logodds["(Intercept)", "Estimate"])

# Vector of genre names and probabilities
genres <- c("Animation", "Comedy", "Documentary", "Drama", "Romance", "Short")
genre_prob <- c(action_prob)

for (g in genres) {
  prob <- plogis(mod.genre.coef.logodds["(Intercept)", "Estimate"] +
                   mod.genre.coef.logodds[paste0("genre", g), "Estimate"])
  genre_prob <- append(genre_prob, prob)
}

# Create a data frame with genre and probability columns
genres <- c("Action", genres)
genre_prob_df <- data.frame(genre = genres, probability = genre_prob)

genre_freq_df <- genre_counts %>%
  filter(Rating_above_7 == "Yes")

genre_freq_df <- genre_freq_df[ , -c(2:4)]
colnames(genre_freq_df)[2] <- "frequency"

genre_prob_df <- genre_prob_df %>%
  left_join(genre_freq_df, by='genre')
```

```r
# Plot the frequency and estimated probability by Genre
genre_prob_df_long <- genre_prob_df %>%
  pivot_longer(cols = c("frequency","probability"),
               names_to = "Fitted_vs_True",
               values_to = "value")

genre_prob_plot <- genre_prob_df_long %>%
  ggplot(aes(genre, value, fill = Fitted_vs_True)) +
  geom_col(position = "dodge") +
  labs(title = "", x = "Genre", y = "Value") +
  scale_fill_manual(values = c("#A6BDDB", "#8FB87D"),
                    labels = c("Frequency", "Probability")) +
  theme(axis.title.x = element_text(size = 9),
        axis.text.x = element_text(size = 7),
        axis.title.y = element_text(size = 9),
        axis.text.y = element_text(size = 7),
        legend.title = element_text(size = 9),
        legend.text = element_text(size = 7),
        legend.key.size = unit(10, "pt"),
        plot.title = element_text(size = 12,hjust = 0.5))
genre_prob_plot
```

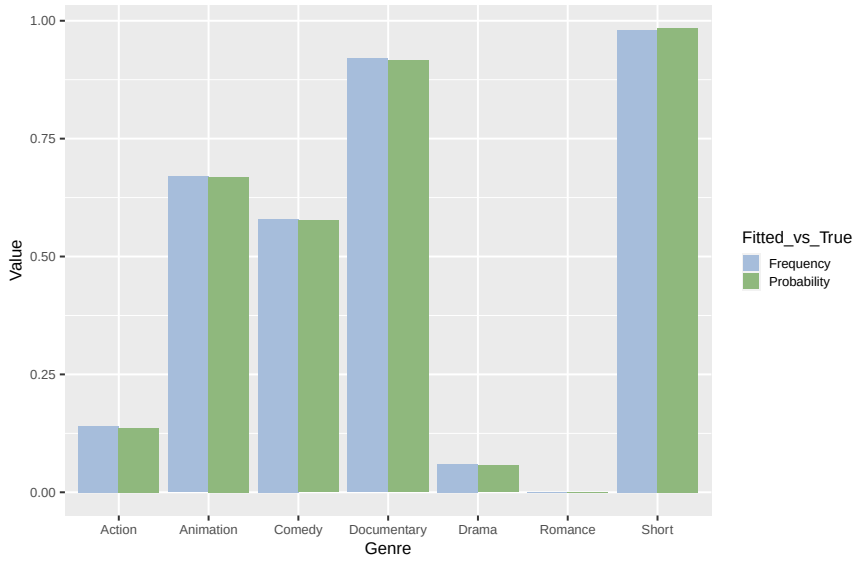

Figure 5: Frequencies and estimated probabilities by Genre.

The Figure 5 shows that the estimated probabilities are consistent with the frequencies, indicating that the model has a good fit to the data.

3.2 Logistic regression with several explanatory variables

3.2.1 Model selection and estimation

We now consider a generalized linear model that takes into account multiple variables. To identify the factors that influence movie ratings, we will use a stepwise approach and select the best model based on AIC as the selection criterion.

```
# Full model
predictor_list <- paste(colnames(imdb_group_08[,2:6]),collapse="+")
f <- paste(c("Rating_above_7 ~ ", predictor_list), collapse="")

full.model <- glm(f, data = imdb_group_08, family = binomial(link = "logit"))

# Stepwise selection
step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)
```

```
##
## Call:
## glm(formula = Rating_above_7 ~ year + length + budget + genre,
##     family = binomial(link = "logit"), data = imdb_group_08)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7225  -0.3744  -0.1209   0.1962   3.4412
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.864236   5.767450  -4.138 3.51e-05 ***
## year               0.010238   0.002940   3.483 0.000497 ***
## length            -0.056869   0.003537 -16.077  < 2e-16 ***
## budget             0.509979   0.030117  16.933  < 2e-16 ***
## genreAnimation    -0.078708   0.320139  -0.246 0.805793
## genreComedy        3.110781   0.179174  17.362  < 2e-16 ***
## genreDocumentary   5.604906   0.442282  12.673  < 2e-16 ***
## genreDrama        -1.556649   0.239136  -6.509 7.54e-11 ***
## genreRomance     -14.607631 391.828859  -0.037 0.970261
## genreShort         4.005562   0.796669   5.028 4.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3470.8  on 2715  degrees of freedom
## Residual deviance: 1456.6  on 2706  degrees of freedom
##   (   131    )
## AIC: 1476.6
##
## Number of Fisher Scoring iterations: 15
```

The best model removed the *votes* variable from the full model, indicating that its impact on the rating was

not significant when considering other variables, which resulted in a relatively poor model fit. Additionally, due to 131 missing values in *length*, it was removed from the analysis. The model we obtained is of the form:

$$g(p_i) = log(\frac{p_i}{1-p_i}) = \alpha + \beta_1 \cdot year_i + \beta_2 \cdot length_i + \beta_3 \cdot budget_i + \beta_4 \cdot genre_i$$

From the output, the estimated coefficients for *length* is negative, suggesting that movies of longer duration are more likely to receive ratings below 7. The estimated coefficients for *year* and *budget* are positive, indicating that as the year of the movie's release becomes more recent and the budget increases, the movie's rating also tends to be higher. All three numerical variables were significant at the 5% level of significance, while the categorical variable *genre* showed that the 'Animation' and 'Romance' genres were not significant.

3.2.2 Model inference and interpretation

The estimates from the logistic regression model are on the log-odds scale and apply to movies with a rating above 7:

$$g(\hat{p}_i) = log(\frac{\hat{p}_i}{1-\hat{p}_i}) = -23.86 + 0.01 \cdot year_i - 0.06 \cdot length_i + 0.51 \cdot budget_i + \hat{\beta}_4 \cdot \mathrm{II}_{genre}(\cdot)$$

The point estimate and the corresponding 95% confidence interval for the log-odds can be obtained similarly, as shown in the Figure 6:

```
plot_model(step.model, show.values = TRUE, transform = NULL,
           title = "Log-Odds (Rating above 7)", show.p = FALSE)
```

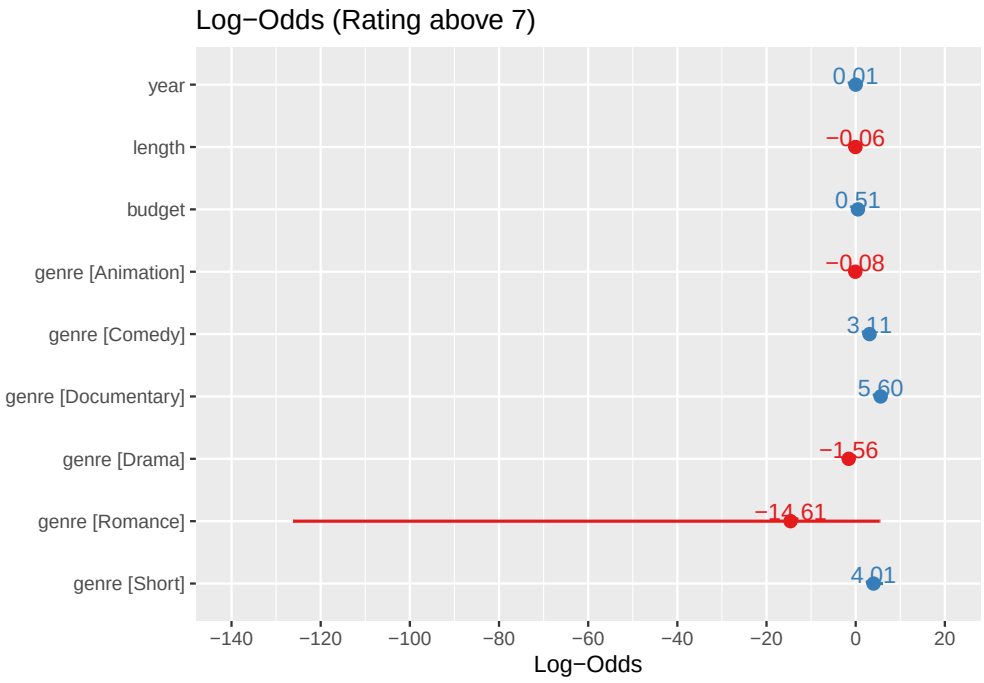

Figure 6: Log-Odds (Rating above 7) for the step model.

Consider using the estimated coefficients to quantify the effect of these four variables, the regression coefficients on the odds scale are given by:

```
# Coefficients on the odds scale
step.model %>%
  coef() %>%
  exp()
```

```
##      (Intercept)             year           length           budget
##     4.324085e-11     1.010290e+00     9.447176e-01     1.665257e+00
##   genreAnimation      genreComedy genreDocumentary       genreDrama
##     9.243095e-01     2.243857e+01     2.717564e+02     2.108414e-01
##     genreRomance       genreShort
##     4.528834e-07     5.490269e+01
```

The point estimate and the corresponding 95% confidence interval for the odds scale can be obtained, as shown in the Figure 7:

```
plot_model(step.model, show.values = TRUE,
           title = "Odds (Rating above 7)", show.p = FALSE) +
  theme(axis.text.x = element_text(size = 1))
```

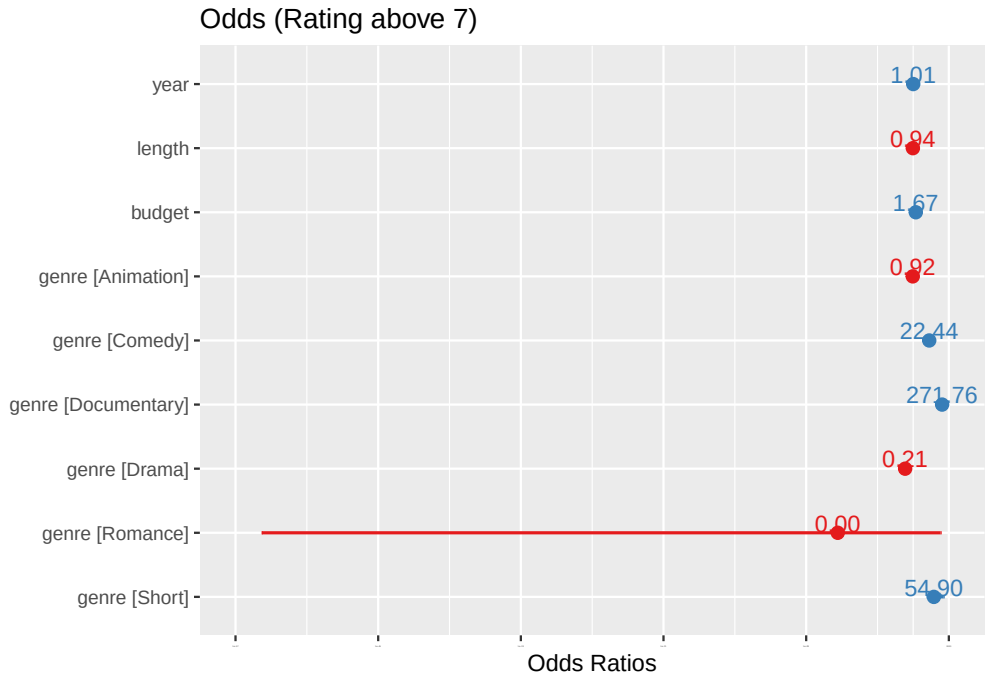

Figure 7: Odds (Rating above 7) for the step model.

This suggests that for two movies released one *year* apart, the odds of a new movie having a rating above 7 is 1.01 times that of an old movie. Likewise, for every additional minute (*length*) in the duration of a movie, the odds of the movie having a rating above 7 decreases by a factor of 0.94, indicating a slight decrease. As for the *budget*, for every additional one million US dollars invested in the production of a movie, the odds of the movie having a rating above 7 increase by a factor of 1.67. The influence of the genre is similar to what was discussed earlier.

4 Conclusion