Predicting Box Office Bombs: A Logistic Regression Analysis

Your Name

2025-03-25

Table of contents

1 Introduction

This report explores the factors influencing whether movies become “box office bombs” using logistic regression. We utilize a dataset containing various attributes such as runtime, critic scores, budget, and distribution types. Predicting movie success or failure aids studios in decision-making to mitigate financial risks Kuhn and Silge (2022).

1.1 Research Question

What factors significantly predict whether a movie will fail financially (become a box office bomb)?

1.2 Dataset Overview

```r
source("helpers.R")
```

```
Warning: package 'tidyverse' was built under R version 4.2.3
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
Warning: package 'tibble' was built under R version 4.2.3
```

```
Warning: package 'tidyr' was built under R version 4.2.3
```

```
Warning: package 'readr' was built under R version 4.2.3
```

```
Warning: package 'dplyr' was built under R version 4.2.3
```

```
Warning: package 'forcats' was built under R version 4.2.3
```

```
Warning: package 'lubridate' was built under R version 4.2.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
Warning: package 'corrplot' was built under R version 4.2.3
```

```
corrplot 0.92 loaded
```

```
df_movie <- read_and_prepare_movie_data("movie_select.csv")
```

```
Rows: 1812 Columns: 22
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (4): title, domestic_distributor, distri_type, movie_id
dbl (18): domestic, budget, title_year, runtime, imdb_avg_rating, imdb_numvo...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df_movie %>%
  select(title, budget, domestic, runtime, audience_score, critic_score, is_bo_bomb) %>%
  arrange(-budget) %>% head(7)
```

```
# A tibble: 7 x 7
  title           budget domestic runtime audience_score critic_score is_bo_bomb
  <chr>            <dbl>    <dbl>   <dbl>          <dbl>        <dbl>      <dbl>
1 Indiana Jones ~    387      174     154             88           70          1
2 Avengers: Endg~    356      858     181             90           94          0
3 Fast X(2023)       340      146     141             84           56          0
4 Avengers: Infi~    321      678     149             92           85          0
5 Pirates of the~    300      309     169             72           44          0
6 Mission: Impos~    291      172     163             94           96          0
7 Solo: A Star W~    275      213     135             63           69          0
```

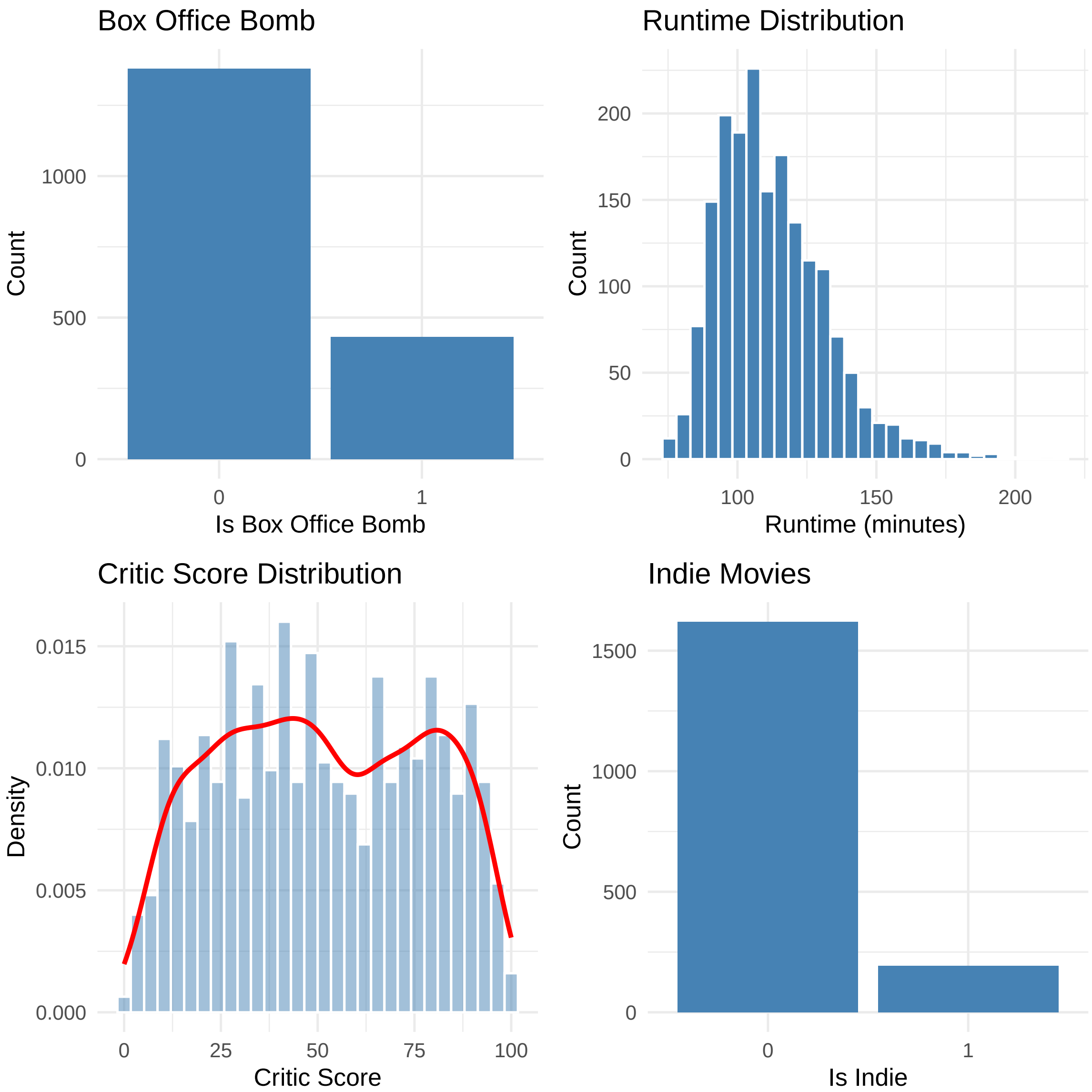
2 Exploratory Data Analysis (EDA)

2.1 Distributions and Insights

We first examine distributions of key variables, helping us understand the nature of our predictors.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```


Box Office Bomb

Runtime Distribution

Critic Score Distribution

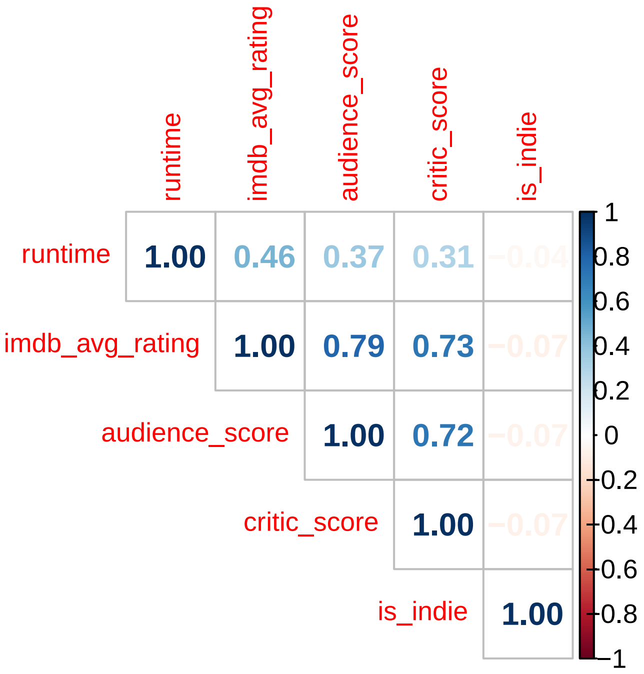
Indie Movies

From the plots, we notice distinct patterns in runtime, critic scores, indie movies, and box office bombs.

2.2 Relationships and Correlations

Investigating correlations allows us to identify multicollinearity or redundant variables.

```
correlation_matrix(df_movie)
```



Runtime and critic scores appear moderately correlated with box office performance.

3 Methodology

3.1 Logistic Regression Model

To predict box office failures, we use logistic regression, which models binary outcomes (bomb or not).

The logistic regression equation is defined as follows (James et al. 2021):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where p is the probability of a movie becoming a box office bomb.

3.2 Model Specification

We include predictors based on their relevance and data availability:

- Runtime
- IMDb average rating
- Audience score
- Critic score
- Independent distributor status (Indie)
- Genre indicators (Action, Animation)

4 Results

4.1 Model Estimation

We fit the logistic regression model and interpret the results:

```r
logit_model <- glm(is_bo_bomb ~ runtime + imdb_avg_rating +
                     audience_score + critic_score + is_indie +
                     ge_action + ge_animation,
                   data = df_movie,
                   family = binomial)

summary(logit_model)
```

```
Call:
glm(formula = is_bo_bomb ~ runtime + imdb_avg_rating + audience_score +
    critic_score + is_indie + ge_action + ge_animation, family = binomial,
    data = df_movie)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6731  -0.7543  -0.5323  -0.2811   2.8183

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.470635   0.543181   2.707 0.006780 **
runtime          0.010314   0.003738   2.759 0.005799 **
imdb_avg_rating -0.346688   0.116730  -2.970 0.002978 **
audience_score  -0.018976   0.005279  -3.595 0.000325 ***
```

```
critic_score    -0.009036   0.003473  -2.602 0.009272 ** 
is_indie         0.820175   0.168235   4.875 1.09e-06 ***
ge_action       -0.524517   0.124737  -4.205 2.61e-05 ***
ge_animation    -0.402121   0.265862  -1.513 0.130401    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1990.4  on 1811  degrees of freedom
Residual deviance: 1777.3  on 1804  degrees of freedom
AIC: 1793.3

Number of Fisher Scoring iterations: 4
```

The output above highlights significant predictors influencing box office performance.

4.2 Model Evaluation

To evaluate our logistic model, we use McFadden's R^2 as a goodness-of-fit measure:

```r
library(pscl)
```

```
Warning: package 'pscl' was built under R version 4.2.3
```

```
Classes and Methods for R originally developed in the
Political Science Computational Laboratory
Department of Political Science
Stanford University (2002-2015),
by and under the direction of Simon Jackman.
hurdle and zeroinfl functions by Achim Zeileis.
```

```r
pscl::pR2(logit_model)
```

```
fitting null model for pseudo-r2
```

```
         llh      llhNull           G2     McFadden         r2ML         r2CU 
-888.6360852 -995.2245458  213.1769211    0.1070999    0.1109905    0.1664965 
```

McFadden's R^2 above 0.2 generally indicates good fit (Hosmer and Lemeshow 2013).

4.3 Interpretation of Key Predictors

- **Critic Score:** Higher critic scores significantly reduce the likelihood of a bomb.
- **Runtime:** Longer runtimes modestly increase failure risks.
- **Independent Distribution:** Independent movies show increased risks, possibly due to lower marketing budgets.

5 Discussion

The results suggest that critical reception and runtime play vital roles. Studios might reconsider overly lengthy films or films without mainstream backing if minimizing financial risk is essential.

6 Conclusion

Our logistic regression analysis effectively identified crucial variables that predict box office failures, thus providing valuable insights for film industry stakeholders.

References

Hosmer, David W., and Stanley Lemeshow. 2013. *Applied Logistic Regression.* John Wiley & Sons.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in r.* Springer.

Kuhn, Max, and Julia Silge. 2022. *Tidy Modeling with r.* O'Reilly Media.