Title

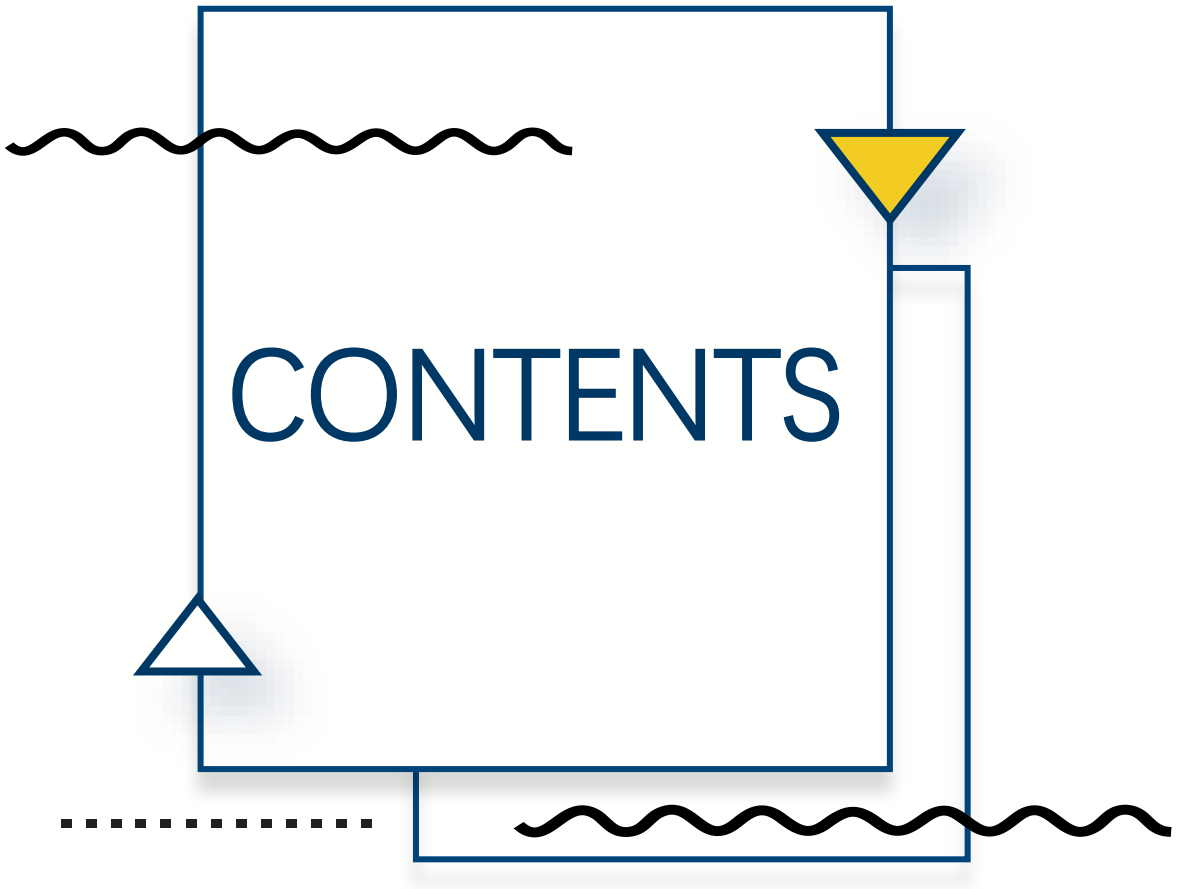
Group 4

Shiyu Lyu & Zonglin Wu & Shiqi Huang

School of Mathematics & Statistics

March 2022

University of Glasgow



PART 01

Aims of Analysis

University of Glasgow

Binary response variable

- IMDB Rating of the Film (rating out of 10)
- A new binary variable named over7
- Indicates whether over 7 or not

Numerical explanatory variable

- Year of release
- Length of Film (in minutes)
- Budget of the Film (in $1,000,000s)
- Number of positive votes (received by viewers)

Categorical explanatory variable

- Genre of the Film (7 kinds)

1. Action
2. Animation
3. Comedy
4. Documentary
5. Drama
6. Romance
7. Short

University of Glasgow

Aims

Which properties of a film influence whether a film receives an IMBD rating greater than 7 or not.

Aims and Method

Method

Fit logistic regression models with different combinations of the explanatory variables to see which variables are the most significant predictors.

PART 02

Exploratory Data Analysis

University of Glasgow

1.Numerical variables

Table 1: Summary statistics on number of films which are rating larger than 7

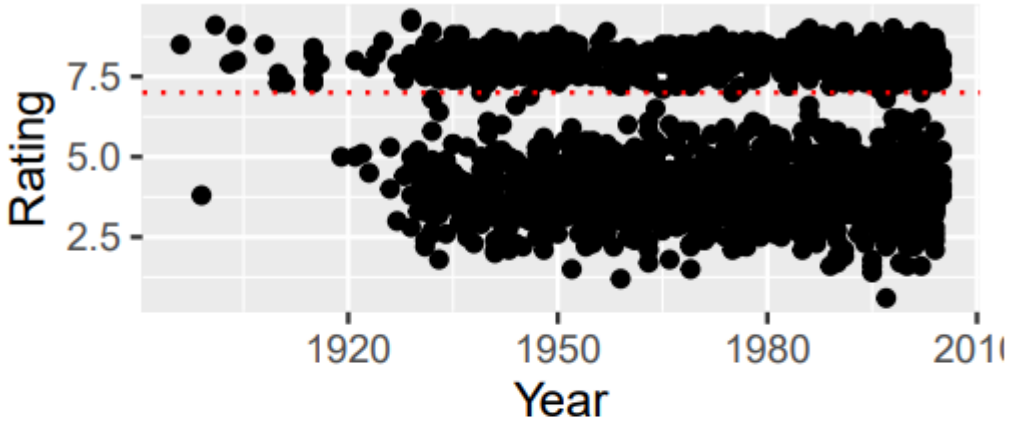
Variable	n	Mean	SD	Min	Q1	Median	Q3	Max	IQR
year	641	1974.91	26.41	1896.0	1951.0	1984.0	1999.0	2005	15.0
length	641	56.12	39.76	1.0	12.0	71.5	91.0	220	19.5
budget	641	13.08	2.84	3.7	11.1	13.0	15.1	21	2.1
votes	641	438.65	4459.97	5.0	10.0	23.0	66.0	103854	43.0

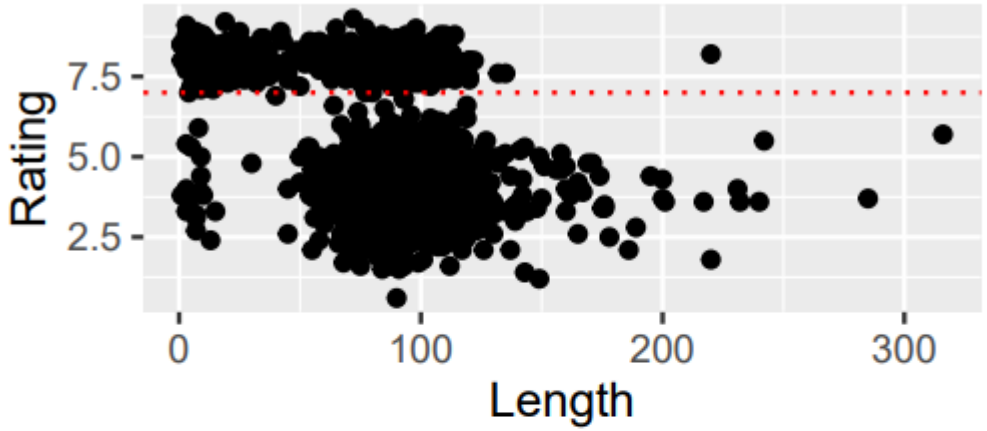
Table 2: Summary statistics on number of films which are rating smaller than 7

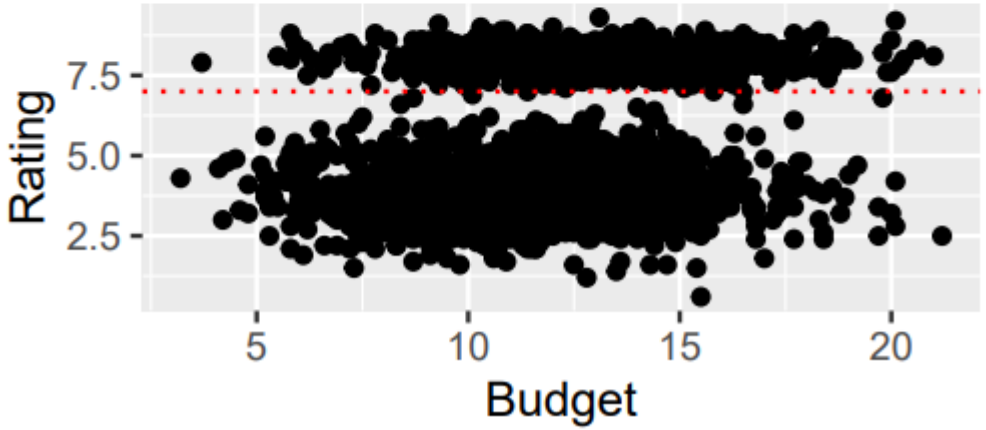
Variable	n	Mean	SD	Min	Q1	Median	Q3	Max	IQR
year	1296	1976.85	21.81	1899.0	1960.0	1981.0	1997.00	2005.0	16.00
length	1296	96.02	25.43	1.0	85.0	94.0	105.00	316.0	11.00
budget	1296	11.51	2.82	3.2	9.5	11.5	13.50	21.2	2.00
votes	1296	665.56	3581.20	5.0	14.0	37.0	158.25	89722.0	121.25

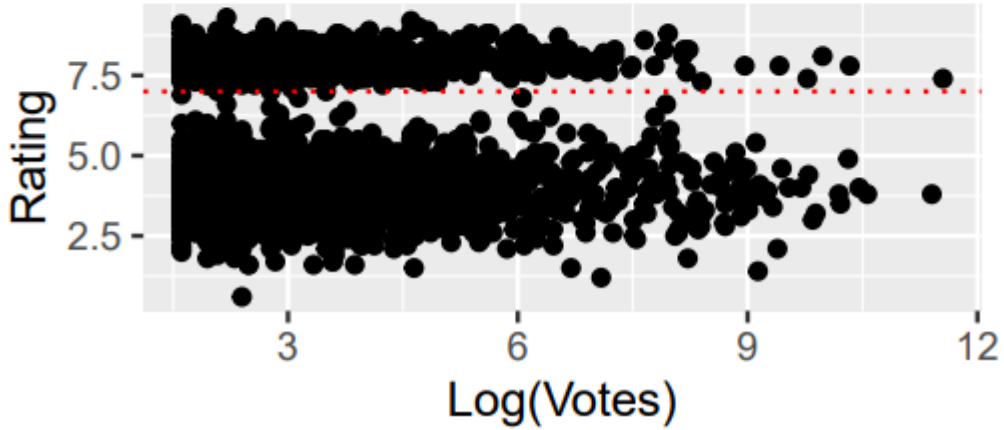
University of Glasgow

1.Numerical variables



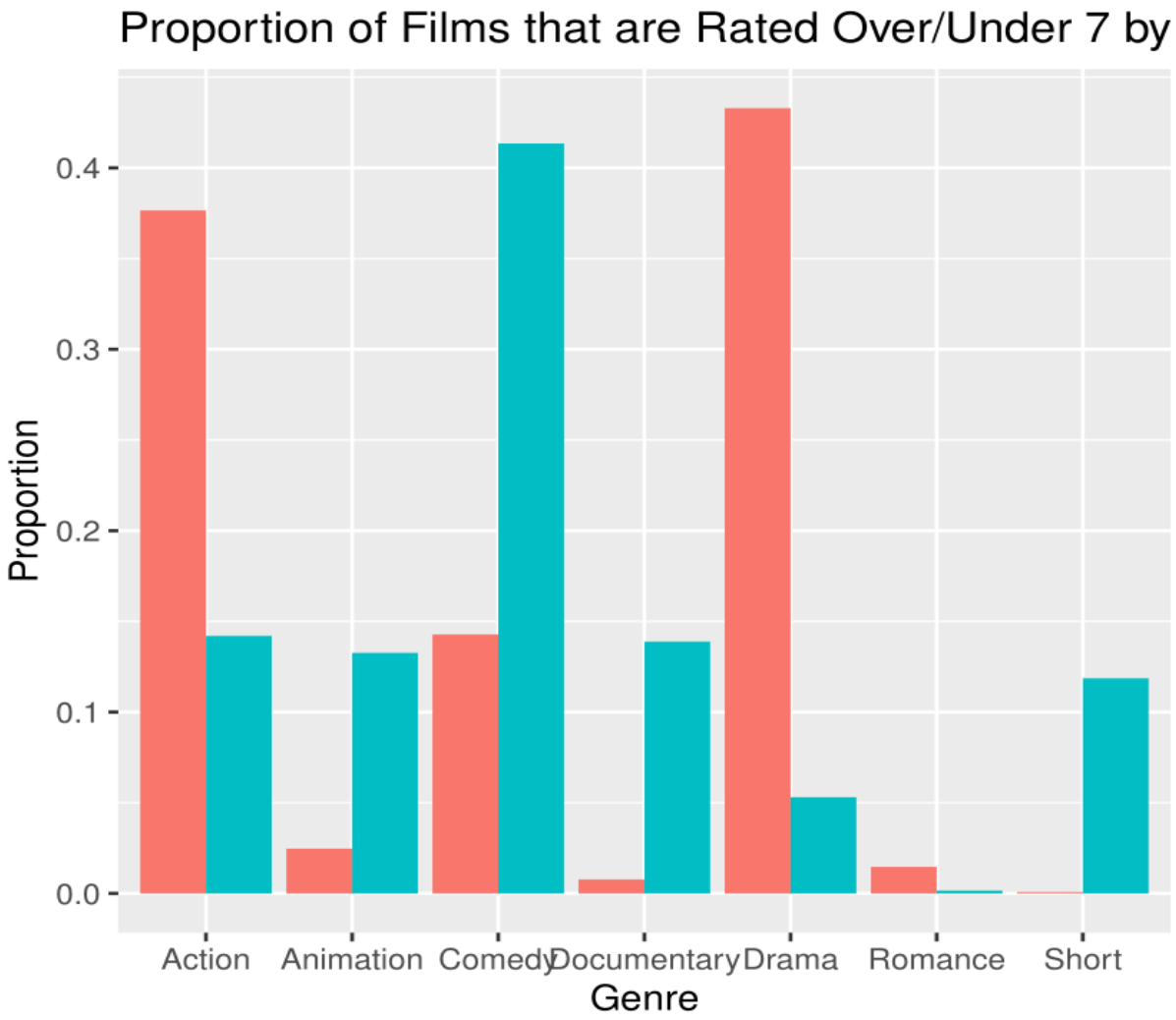






University of Glasgow

2.Categorical variables



genre	FALSE		TRUE	
Action	84.3%	(488)	15.7%	(91)
Animation	27.4%	(32)	72.6%	(85)
Comedy	41.1%	(185)	58.9%	(265)
Documentary	10.1%	(10)	89.9%	(89)
Drama	94.3%	(561)	5.7%	(34)
Romance	95.0%	(19)	5.0%	(1)
Short	1.3%	(1)	98.7%	(76)

PART 03

Statistical Modelling and Results

University of Glasgow

Case1: One numerical explanatory variable

1 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{year}$$

2 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{length}$$

3 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{budget}$$

4 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \log(\text{votes})$$

- p is the probability that the film is ranked over 7
- α is the intercept value
- β is the regression coefficient
- year, length, budget and log(votes) are numerical explanatory variables, respectively

University of Glasgow

Confident Intervals

1 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{year}$$

2 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{length}$$

3 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{budget}$$

4 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \log(\text{votes})$$

95% CI	Model 1	Model 2	Model 3	Model 4
Intercept (α)	(-1.66 , 14.51)	(2.25, 3.00)	(-3.46 , -2.53)	(-0.18 , 0.31)
regression coefficient (β)	(-0.01 , 0.00)	(-0.05, -0.04)	(0.15, 0.22)	(-0.27 , -0.14)

University of Glasgow

AIC and BIC

1 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{year}$$

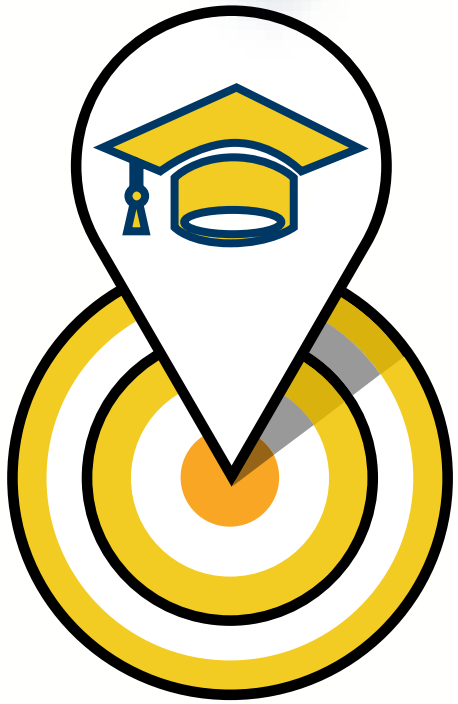
2 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{length}$$

3 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{budget}$$

4 $$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \log(\text{votes})$$

	Model 1	Model 2	Model 3	Model 4
AIC	2332.500	1764.540	2224.033	2288.244
BIC	2343.528	1775.569	2235.062	2299.272

Case1: One numerical explanatory variable



Result

Model 2 is the best

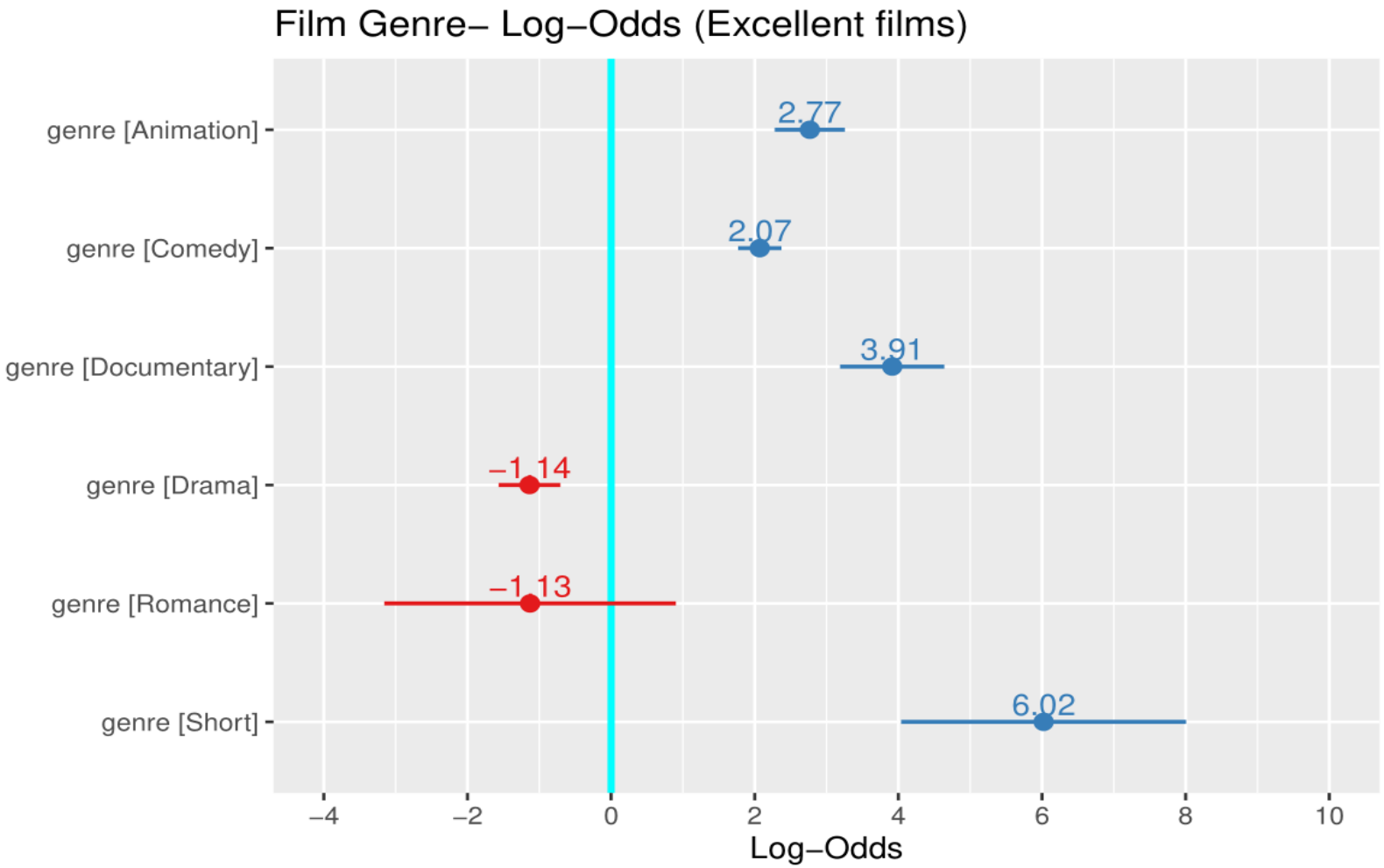
$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{length}$$

University of Glasgow

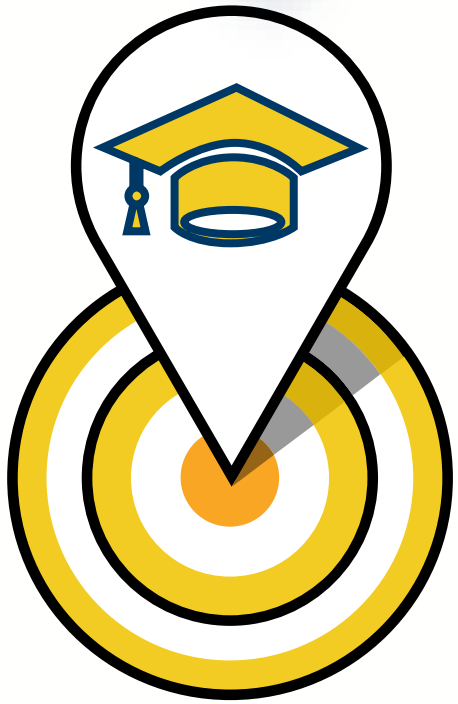
Case2: One categorical explanatory variable

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre}$$

- p is the probability that the film is ranked over 7
- α is the intercept value
- β_{genre} is the regression value of the categorical variable (Animation as the baseline)



Case2: One categorical explanatory variable



Result

Animation
Comedy
Documentary
Short

→ Higher score

Romance
Drama

→ Lower score

University of Glasgow

Case3: Full Models

Base research 2 3 4 5

Base reality 1

1

Full model with every explanatory variables

2

Best AIC Model

3

Optimal Model with only Significant Variables

4

Simplified Optimal Model

5

C log-log and Probit Models

University of Glasgow

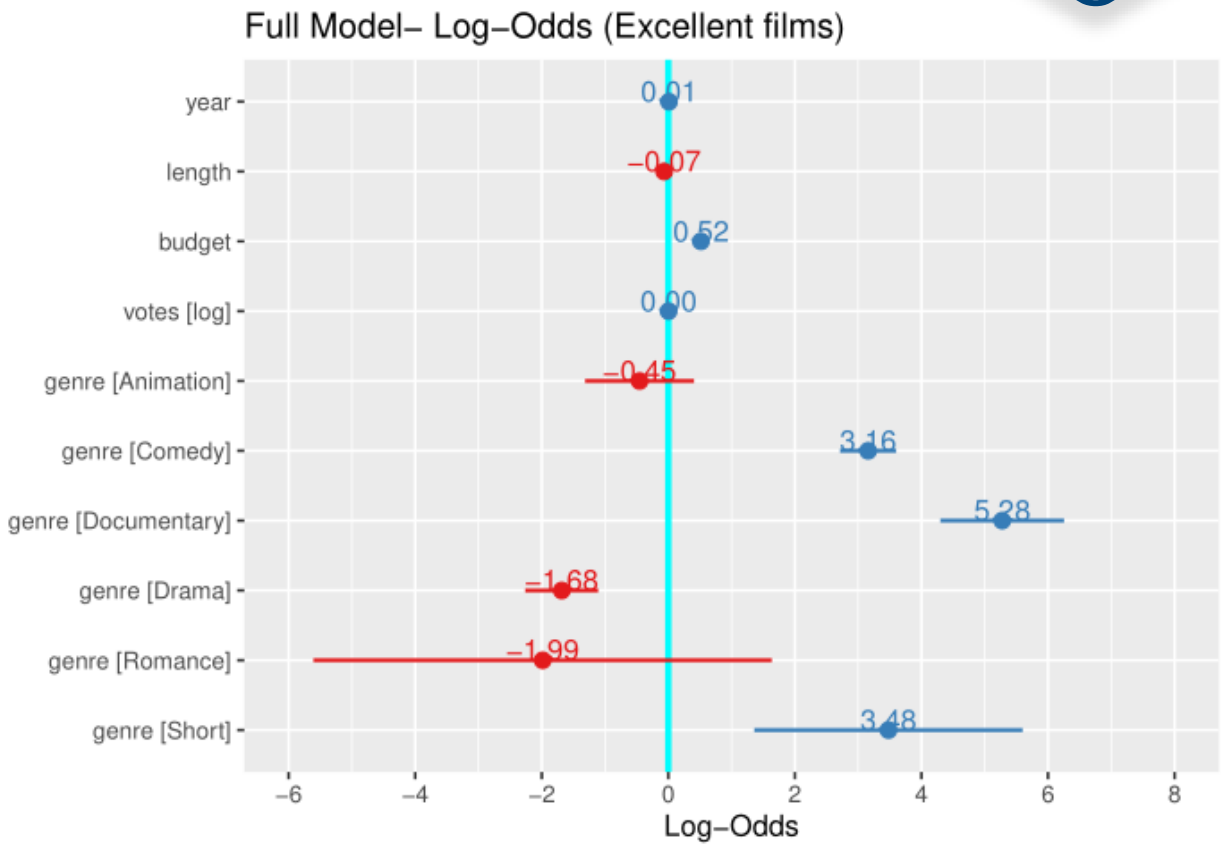
Case3: Full Model1- with every explanatory variables

1

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \log(\text{votes}) + \beta_3 \cdot \text{length} + \beta_4 \cdot \text{budget} + \beta_5 \cdot \text{year}$$

where

- p is the probability that the film is ranked over 7
- votes is the number of positve votes the film received by viewers
- genre is the genre of the film
- length is the length of the film in minutes
- budget is the budget of the film in $1000000
- α is the intercept value
- β genre is the regression value for the i genre

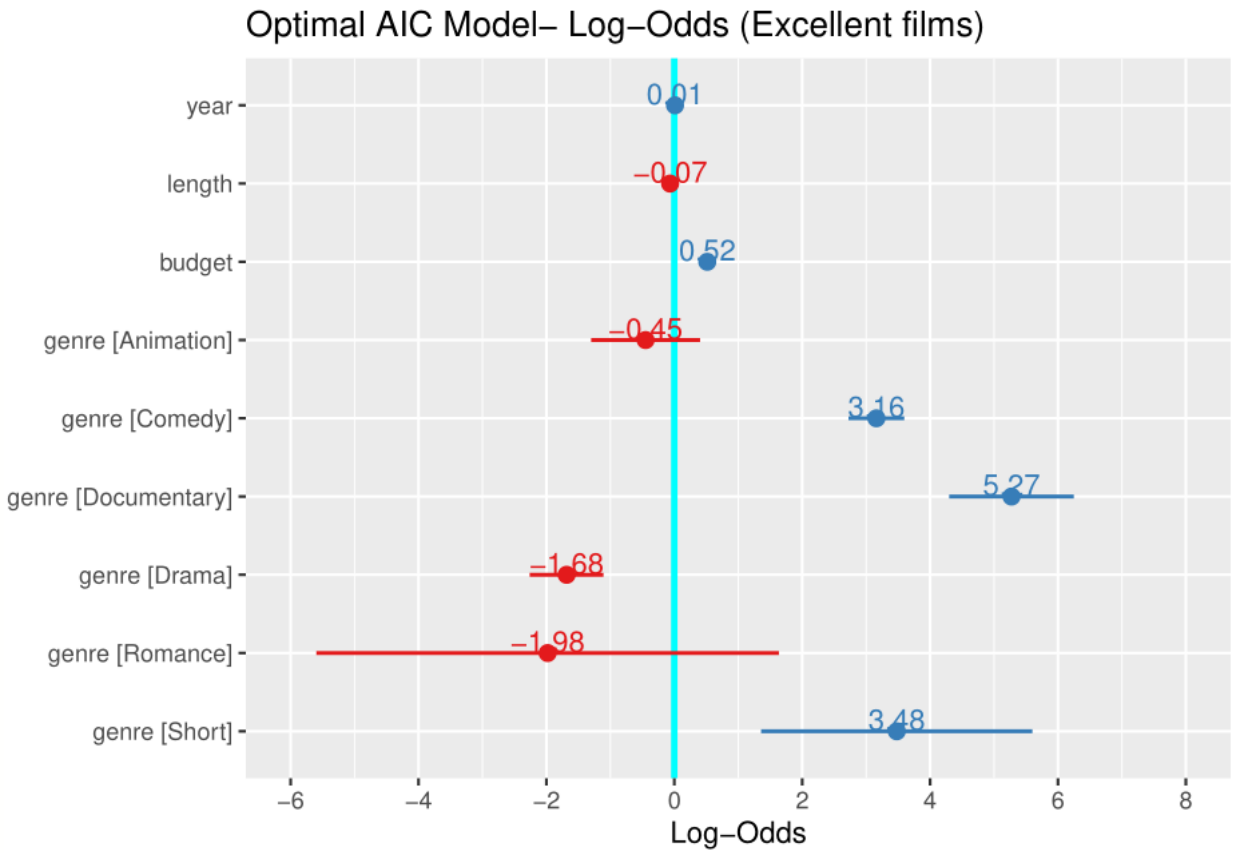


University of Glasgow

Case3: Full Model2- Best AIC Model

2

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \text{length} + \beta_3 \cdot \text{budget} + \beta_4 \cdot \text{year}$$



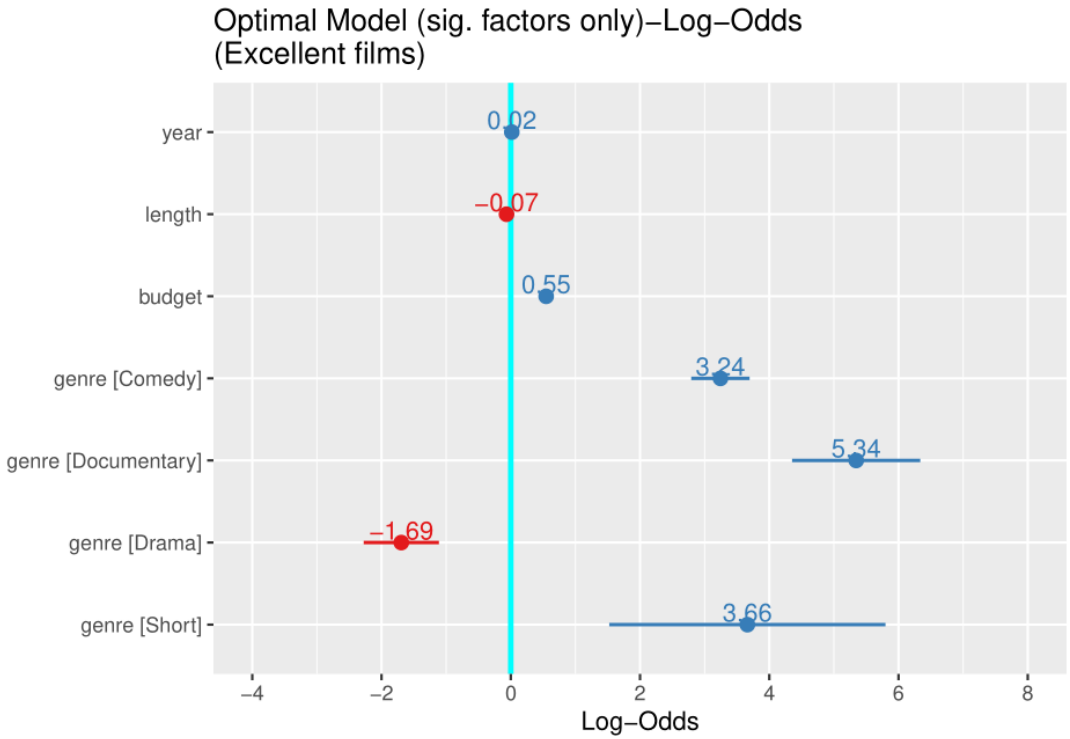
University of Glasgow

Case3: Full Model2- Optimal Model with only Significant Variables

3

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \text{length} + \beta_3 \cdot \text{budget} + \beta_4 \cdot \text{year}$$

Genre is the genre of the film without Romance and Animation



University of Glasgow

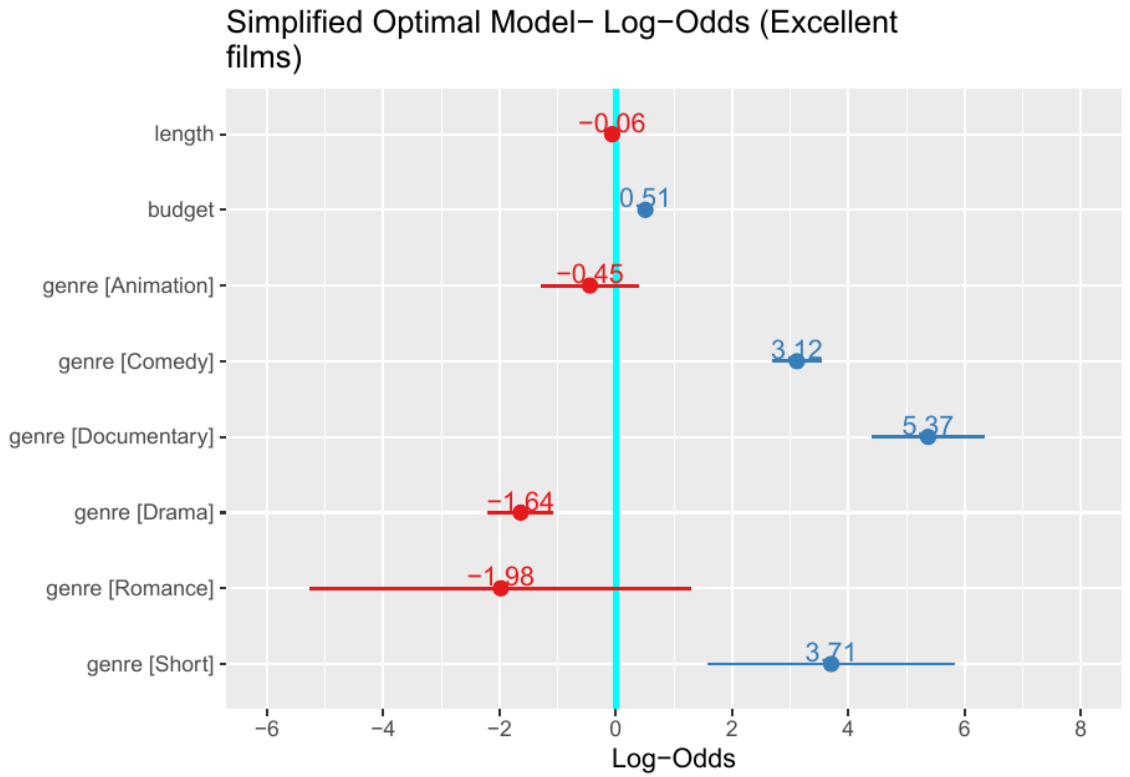
Case3: Full Model4- Simplified Optimal Model

4

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \text{length} + \beta_3 \cdot \text{budget}$$

where

- p is the probability that the film is ranked over 7
- genre is the genre of the film
- length is the length of the film in minutes
- budget is the budget of the film in $1000000
- α is the intercept value
- βgenre is the regression value for the i genre

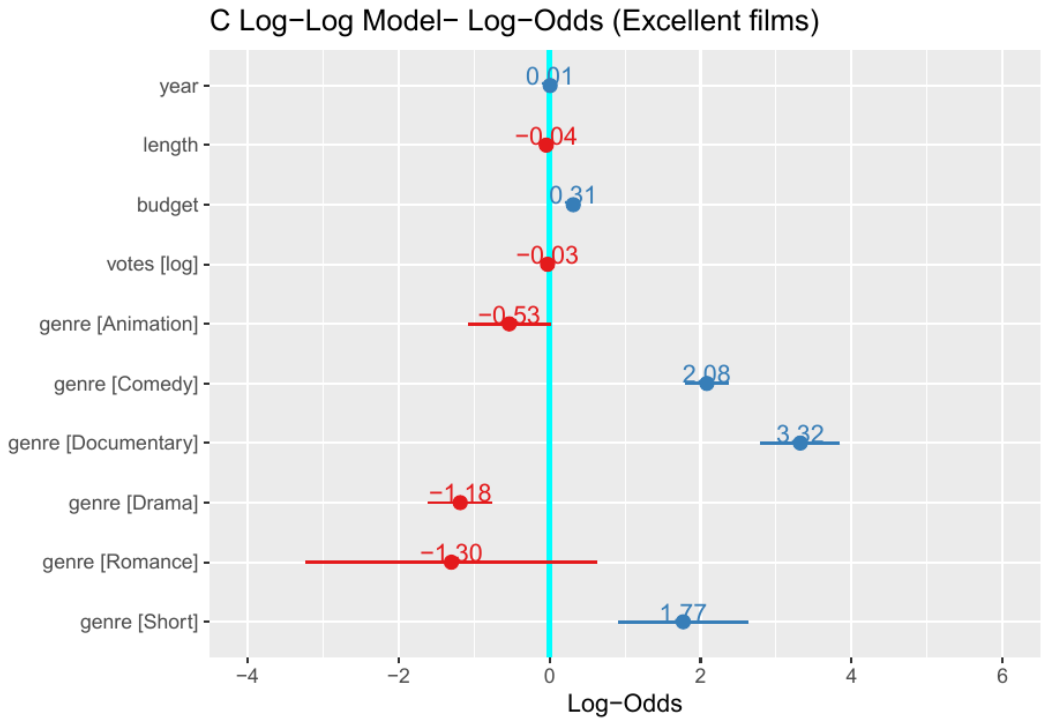


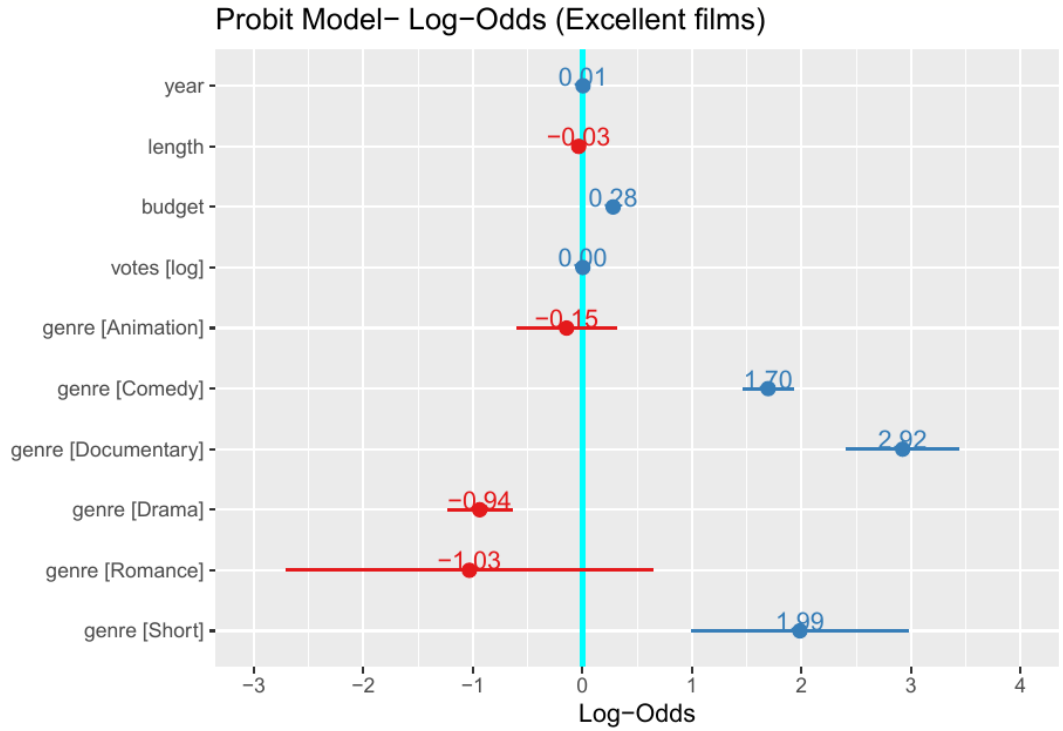
University of Glasgow

Case3: Full Model5- C log-log and Probit Models

5

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \log(\text{votes}) + \beta_3 \cdot \text{length} + \beta_4 \cdot \text{budget} + \beta_5 \cdot \text{year}$$



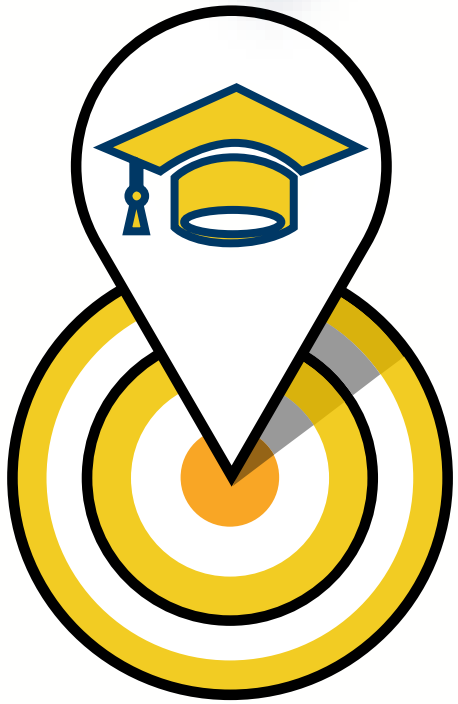


University of Glasgow

Case3: Full Model-Comparison

	Full Model	C Log-log Model	Pobit Model
AIC	958.8695	1004.9769	967.8979
BIC	1019.526	1065.634	1028.555

Case3: Full Model



Result

Choose Full model with every explanatory variables to explain the research question

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \log(\text{votes}) + \beta_3 \cdot \text{length} + \beta_4 \cdot \text{budget} + \beta_5 \cdot \text{year}$$

PART 04

Conclusion and Future Works

University of Glasgow

Conclusion

One numerical explanatory variable → Length

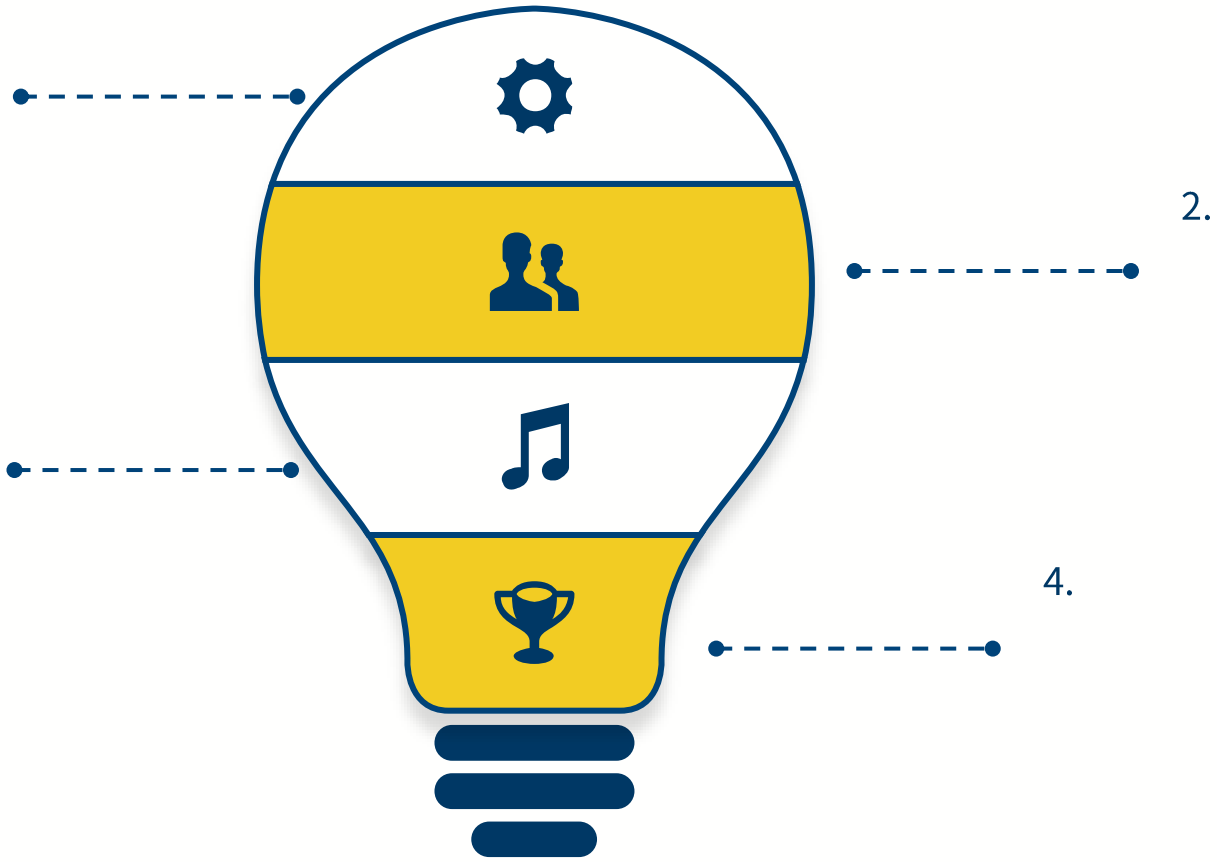
One categorical explanatory variable → Animation, Comedy, Documentary, Short

Full model → year of release, length of film, budget of film, positive votes, genre of the film → IMBD rating greater than 7

University of Glasgow

Future Works

1.



2.

3.

4.

University of Glasgow

THANK YOU