IMDb Predition challenge

Introduction

IMDb is a popular platform for storing information on movies, TV shows, Awards and events, and celebrities, making it a go-to source for movie enthusiasts. People often rely on IMDb to access essential information, such as movie ratings, to decide whether a film is worth watching. However, the challenge arises when newly released movies only have limited ratings, which can fluctuate over time. Our project focuses on constructing predictive models for IMDb scores to provide valuable insights into the potential of new movies. The goal is to harness current IMDb data to create robust models for IMDb score prediction, which can benefit moviegoers and the film industry. In this project, we specifically target twelve upcoming blockbusters*. Our work unfolds in three phases:

1. **Data exploration**: We explore the data and delve into the IMDb dataset to gain insights into the variables. This initial phase helps us understand the dataset's characteristics and sets the stage for subsequent modeling.
2. **Model building**: After preprocessing and exploration, we carefully consider potentially influential variables in our models to ensure the accuracy and reliability of our predictions. This phase is crucial in crafting models that can provide IMDb score forecasts.
3. **Model selection:** We assess the performance of all the models created, focusing on choosing a model with low MSE, significant variables, and reasonable results. The process allows us to identify the model demonstrating the highest predictive accuracy and reliability.

Following the execution of our project, we have determined that the model eliminating insignificant variables delivers the most accurate predictions. Our IMDb score predictions range from 4.47 (for "Pencils vs Pixels") to an impressive 8.32 (for "Napolean"). These predictions serve as valuable tools for movie enthusiasts and industry professionals, offering a glimpse into the potential reception of upcoming blockbusters.

*Twelve movies: Pencils vs Pixels, The Dirty South, The Marvels, The Holdovers, Next Goal Wins, Thanksgiving, The Hunger Games: The Ballad of Songbirds and Snakes, Trolls Band Together, Leo, Dream Scenario, Wish, Napoleon

Data description

In this section, we will describe how our team explored the data and executed the initial data inspection and preliminary analysis before entering the model-building section:

1. **Data Cleaning and Transformation:**
    - **Remove columns:** We removed identifiers such as ***movie_title***, ***movie_id***, and ***imdb_link*** since these variables would not contribute to modelling or predictions.
    - **Character conversion:** We converted character columns to categorical type, ensuring they treated as categorical variables during analysis.
      Character variables include ***release_month, language, country, maturity_rating, distributor, director, actor1, actor2, actor3, colour_film, genres, plot_keywords, cinematographer, production_company.***
2. **Exploratory Data Analysis (EDA):**
    - **Summary statistics:** We generated summary statistics for numerical and categorical columns to understand their distributions, potential outliers, and unique values (See Table A and Table B). We observed that some numerical variables might have outlier issues (e.g., ***nb_news_articles***), and some categorical variables may have too many distinct values (e.g., ***distributor, director***). Accordingly, we conducted further analysis to evaluate the variables more deeply.
    - **IMDb scores distribution:** We visualized the distribution of IMDb scores to get initial understanding about our target variable. The distribution is close to normal distribution visually, but we still consider using log transformation to make it more standardized. (See Table C)
    - **Frequency distribution:** We visualized the distribution of ***language*** and ***colour_film*** using bar plots to understand their frequency distribution. According to the plots, English is the dominant ***language,*** and Color is the majority of ***colour_film.*** (See Table D)
    - **Overwhelming numbers of unique value:** We observed that some variables (***director, cinematographer, plot_keywords***) have notably unique values which may hinder the effectiveness of model building. (For the exact number refer to Table E)
    - **Remove columns:** To facilitate variable selection of model, we eliminated the columns like ***plot_keywords, language, release_day, release_year, director, actor1, actor2,***

actor3, colour_film, and ***cinematographer,*** based on their potential redundancy or lack of direct relevance to the target outcome.

3. **Feature Engineering:**

   To best utilize the data, we conducted below engineering to make variables more usable for building regression models:

   - **Group by:** We assumed that these variables could be potential predictors for the IMDb score. However, the values of each variable exhibit significant diversity. To address this, we used our judgment to define reasonable segments for the variables.
     (1) ***distributors:*** we grouped movies by their distributors, computing the number of movies each distributor had. The method helped in determining the prominence of certain distributors in the dataset. We introduced a new binary feature, ***distributor_dummy***, that flagged major distributors (those that distributed more than 20 movies) with a 1 and the rest with a 0.
     (2) ***production_company***: We applied similar methodology to this variable. The new binary feature is that those companies who produced more than 20 movies would be 1, and the rest are 0.
     (3) ***maturity_rating:*** We converted the column into separate binary columns for each of the common ratings like R, PG-13, and PG, while grouping the less frequent ratings under Others.
     (4) ***country:*** We created a binary column, ***country_USA***, to indicate if a movie was produced in the USA (1) or not (0), given that the majority of movies originate from the USA.
     (5) ***aspect_ratio***: We transformed the column into binary columns categorized into 2.35, 1.85, and the others, and we subsequently removed the original aspect_ratio column.
   - **Transform the release_month:** We transformed the column into separate binary columns for each month, indicating the release month for every movie.
   - **Regenerated genre columns**: To include all the genres, we regenerated the genre dummies based on the unique genres found in the ***genres*** column.
   - **Generated blockbuster_month**: We created a binary feature, ***blockbuster_month***, to pinpoint movies released during blockbuster-favored months (May, June, July, Nov, Dec), and then we analyzed its correlation with the ***imdb_score***.

4. **Further Exploratory Analysis after feature engineering:**

   We conducted simple linear regressions to assess the strength and significance of the relationships between the target variable (***imdb_score***) and other numeric features.

5. **Potential issue detection:**

   Through detecting several issues in model building, we had below outcome, which can help to decide on potential transformations and selection.

   - **Linearity:**

Linear Variables	Non-Linear Variables
nb_faces	***duration***
actor1_star_meter	***nb_news_article***
actor2_star_meter	***movie_meter_IMDBpro***
actor3_star_meter	***movie_budget***

   (See Table G for reference)

   1. **Skewness:**

No Skewness	Moderately Skewed	Highly Skewed
movie_budget	***duration***	***nb_news_article***
	nb_faces	***actor1_star_meter***
		actor2_star_meter
		actor3_star_meter
		movie_meter_IMDBpro

   - **Heteroskedastic Variables:**
     - ***duration***
     - ***movie_budget***
     - ***nb_news_article***
     - ***movie_meter_IMDBpro***
   - **Correlation:** We produced a correlation heatmap for the numeric columns, aiming to discern potential multicollinearity, and we highlighted strong correlations between certain variables to identify potentially redundant or closely related features. Further visualizations and tests were conducted to enhance our understanding of relationships between these variables. For the correlation between ***imdb_score*** and other factors, we set the threshold for strong correlation as 0.8. In our test, we did not identify highly correlated variables (See Table H and Table I).

- **Outliers:** Regarding outliers, we detected them in the numeric columns using both the IQR (Interquartile Range) method and the 3-standard deviation method, and visualized them with boxplots. (See Table J)

Model Selection

- **Methodology:**

  When building models, we selected the following types:

  1. **<u>Polynomial regression model:</u>** We employed polynomial regression to accommodate both linear and non-linear variables. Many models were tried with different dummy varaibles to see which variables give the most significant varaibles
  2. **<u>Log-transformed IMDb score model:</u>** To address skewness in the target variable (***imdb_score***), we applied a log transformation to ***imdb_score*** and continued with polynomial regression.
  3. **<u>Spline model predictions:</u>** We explored whether a spline model could better fit the data and yield improved outcomes, applying polynomial splines for this analysis.
  4. **<u>Refinements to previous models:</u>** In our pursuit of model quality, we eliminated insignificant variables and reran the three algorithms mentioned above.

- **Rationale:**
  1. **How we select predictors:**

     Initially, we tested a model that included all factors that had undergone the feature engineering stage and were potentially significant for predicting the IMDb score. To enhance the predictive power of our model, we aimed to exclude insignificant variables. We considered factors with p-values around or below 0.05 as significant enough to retain in the model.
  2. **How we determined the degree of polynomial:**

     For models using polynomial regression, we selected the degree of polynomial that minimized the root mean square error (RMSE). In the case of models with log-transformed IMDb scores, we determined the degree by identifying the lowest mean square error (MSE) among different degree comparisons.
  3. **How we decided the number of knots in spline:**

In our spline models, knot locations were chosen based on quartiles of each predictor variable, providing a data-driven way to capture non-linear trends. These quartiles ensure knots are evenly spaced across data distributions, allowing for flexible model adaptation and improved predictive accuracy.

- **Model issues:**
  Models were trained to find the best degrees for nonlinear variables so that we do not overfit or underfit the data. Nonlinear variables Models trained had a R2 score of around 0.45.

Results

After iterative testing and reviewing the prediction results for all the models (See Table K), we obtained score estimates ranging from around 3 to 9. In addition to considering MSE, we also took the results of prediction into account when choosing the final model. In the end, we chose the **polynomial model without insignificant variables** (See Table L).

The final variables we included in the model are ***movie_budget, duration, nb_news_articles, nb_faces, movie_meter_IMDBpro, maturity_PG13, country_USA, genre_Drama, genre_Sport, genre_Horror, genre_Thriller, genre_Crime, genre_Comedy, genre_Action, genre_Mystery, genre_Family, genre_Animation, genre_Documentary.*** These predictors are primarily related to movie investment, marketing campaigns (news & posters), and movie genres.

Models	MSE
Model Predictions	0.73
Log Model	0.025 (log scale)
Spline Model	0.75
Model insignificant variables removed	0.718
Log Model insignificant variables removed	0.024 (log scale)
Spline Model insignificant variables removed	0.71

Predictions of model that will be accepted are as follows:

	Movie Names	Model Predictions insignificant variables removed
1	Pencils vs Pixels	4.47
2	The Dirty South	8.16
3	The Marvels	4.57
4	The Holdovers	8.09
5	Next Goal Wins	7.04
6	Thanksgiving	7.89
7	The Hunger Games: The Ballad of Songbirds and Snakes	7.85
8	Trolls Band Together	7.84
9	Leo	7.13
10	Dream Scenario	7.42
11	Wish	8.07
12	Napoleon	8.32

In the selected model, to ensure the significance of each predictor, we filtered the p-value of variables that are under or around 0.05 (See Table L)

In terms of the predictive power of our final model, we obtained the following numbers for the final evaluation:

- **The R-squared of the model:**
    1. Multiple R-squared: 0.4297
    2. Adjusted R-squared: 0.4219

  The R-squared indicates that our model can explain approximately 42.97% of the variation in IMDb score, which can be attributed to the variation in the factors.

- **Out-of-sample performance:**

  We employed K-fold cross-validation as the validation method and achieved an MSE=0.718.

The chosen model provides a low MSE and uses the most significant predictors. Logarithmic models were not considered as they affect the interpretability of model coefficients.

Appendices

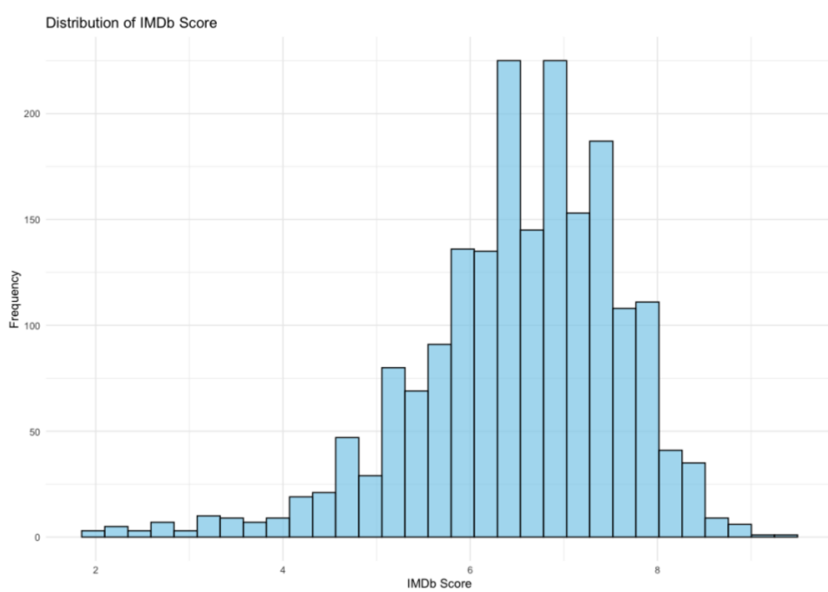
- Table A. Summary statistics for numerical variables

```
Summary Statistics for Numerical Variables:
> print(numerical_summary)
   imdb_score     movie_budget       release_day     release_year      duration      aspect_ratio    nb_news_articles
 Min.   :1.900   Min.   :  560000   Min.   : 1.00   Min.   :1936   Min.   : 37.0   Min.   :1.180   Min.   :    0.0
 1st Qu.:5.900   1st Qu.: 8725000   1st Qu.: 9.00   1st Qu.:1997   1st Qu.: 96.0   1st Qu.:1.850   1st Qu.:   78.0
 Median :6.600   Median :18000000   Median :17.00   Median :2004   Median :106.0   Median :2.350   Median :  286.0
 Mean   :6.512   Mean   :20973774   Mean   :15.95   Mean   :2001   Mean   :109.7   Mean   :2.096   Mean   :  770.6
 3rd Qu.:7.300   3rd Qu.:30000000   3rd Qu.:23.00   3rd Qu.:2010   3rd Qu.:118.0   3rd Qu.:2.350   3rd Qu.:  845.5
 Max.   :9.300   Max.   :55000000   Max.   :30.00   Max.   :2018   Max.   :330.0   Max.   :2.760   Max.   :60620.0
 actor1_star_meter actor2_star_meter actor3_star_meter    nb_faces         action          adventure           scifi
 Min.   :      9   Min.   :      3   Min.   :      8   Min.   : 0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:    505   1st Qu.:   1895   1st Qu.:   3075   1st Qu.: 0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :   1888   Median :   3986   Median :   5856   Median : 1.00   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :  21190   Mean   :  17114   Mean   :  35469   Mean   : 1.44   Mean   :0.2005   Mean   :0.1264   Mean   :0.1083
 3rd Qu.:   4665   3rd Qu.:   7667   3rd Qu.:  12250   3rd Qu.: 2.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :8342201   Max.   :5529461   Max.   :6292982   Max.   :31.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
    thriller          musical            romance          western             sport             horror           drama
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000
 Median :0.0000   Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000   Median :0.000   Median :1.0000
 Mean   :0.2979   Mean   :0.07047   Mean   :0.2451   Mean   :0.01762   Mean   :0.04819   Mean   :0.113   Mean   :0.5492
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000   Max.   :1.0000
      war             animation            crime          movie_meter_IMDBpro
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :    71
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:  2836
 Median :0.00000   Median :0.00000   Median :0.0000   Median :  5406
 Mean   :0.03627   Mean   :0.01036   Mean   :0.2161   Mean   : 11612
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 10198
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :849550
```

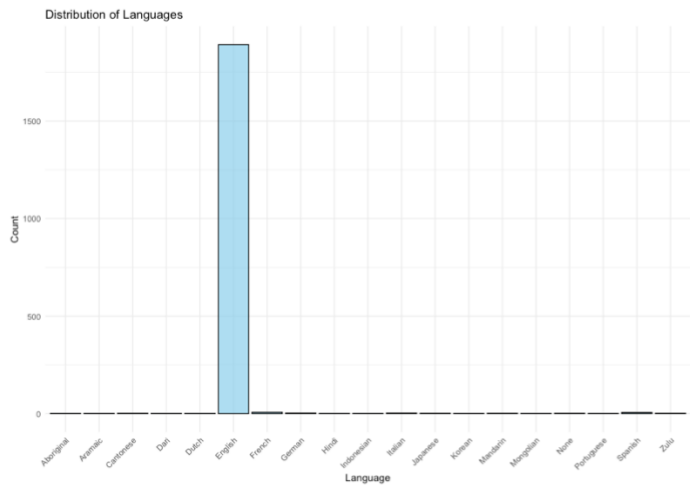
- Table B. Summary statistics for categorical variables

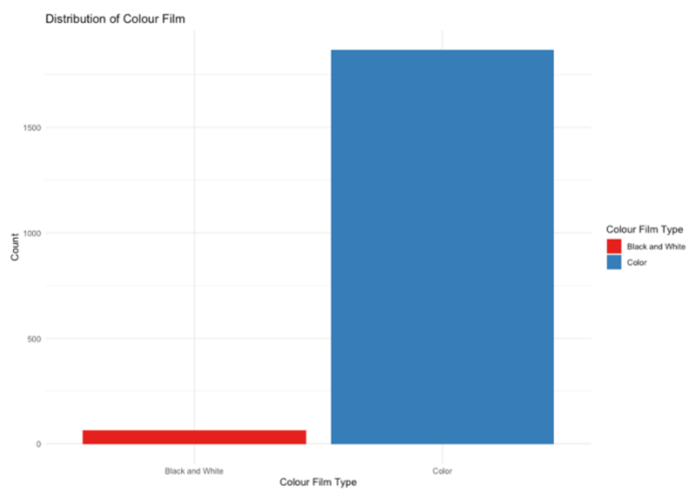
```
Summary Statistics for Categorical Variables:
> print(categorical_summary)
 release_month      language          country      maturity_rating                      distributor                  director
 Oct     :216   English  :1892   USA       :1555   R        :1013   Warner Bros.                 : 169   Woody Allen       :  18
 Jan     :205   French   :   7   UK        : 177   PG-13    : 582   Universal Pictures           : 146   Steven Spielberg  :  12
 Sep     :187   Spanish  :   6   France    :  40   PG       : 255   Paramount Pictures           : 138   Clint Eastwood    :  11
 Aug     :172   German   :   3   Canada    :  38   G        :  34   Twentieth Century Fox        : 126   Spike Lee         :  11
 Apr     :169   Italian  :   3   Germany   :  34   Approved:  21   Columbia Pictures Corporation: 113   Steven Soderbergh:  10
 Mar     :157   Cantonese:   2   Australia:  23   X        :   8   New Line Cinema              :  73   Martin Scorsese   :   9
 (Other):824   (Other)  :  17   (Other)  :  63   (Other) :  17   (Other)                      :1165   (Other)           :1859
             actor1                 actor2                  actor3              colour_film                    genres
 Robert De Niro:  30   Morgan Freeman  :    9   Hope Davis     :    6   Black and White:  63   Drama                 :  92
 Bill Murray   :  17   Charlize Theron :    7   Ben Mendelsohn:    5   Color          :1867   Comedy|Drama|Romance:  87
 J.K. Simmons  :  17   Brad Pitt       :    6   John Heard     :    5                          Comedy                :  85
 Kevin Spacey  :  17   Chazz Palminteri:    6   Robert Duvall :    5                          Comedy|Romance        :  80
 Jason Statham :  15   Demi Moore      :    6   Steve Carell  :    5                          Comedy|Drama          :  74
 Harrison Ford :  14   Meryl Streep    :    6   Thomas Lennon :    5                          Drama|Romance         :  58
 (Other)       :1820   (Other)         :1890   (Other)        :1899                          (Other)               :1454
                                                                          plot_keywords        cinematographer
 10 year old|dog|florida|girl|supermarket                                         :   1   multiple      :  79
 12 year time span|coming of age|domestic abuse|growing up|separated parents:   1   Roger Deakins:  18
 13 year old|13th birthday|30 year old|wish|year 1987                              :   1   Mark Irwin   :  17
 13 year olds|adolescence|friend|peer pressure|teacher                            :   1   John Bailey  :  16
 14 year old|boat|bounty hunter|boy|river                                          :   1   Andrew Dunn  :  13
 14th century|king|knight|sword duel|time travel                                   :   1   Jack N. Green:  13
 (Other)                                                                           :1924   (Other)      :1774
                       production_company
 Universal Pictures            : 110
 Paramount Pictures            :  99
 Columbia Pictures Corporation:  96
 Warner Bros.                  :  76
 New Line Cinema               :  75
 Twentieth Century Fox         :  70
 (Other)                       :1404
```

- Table C. IMDb scores distribution



- Table D. Frequency distribution (Language & Colour Film Type)

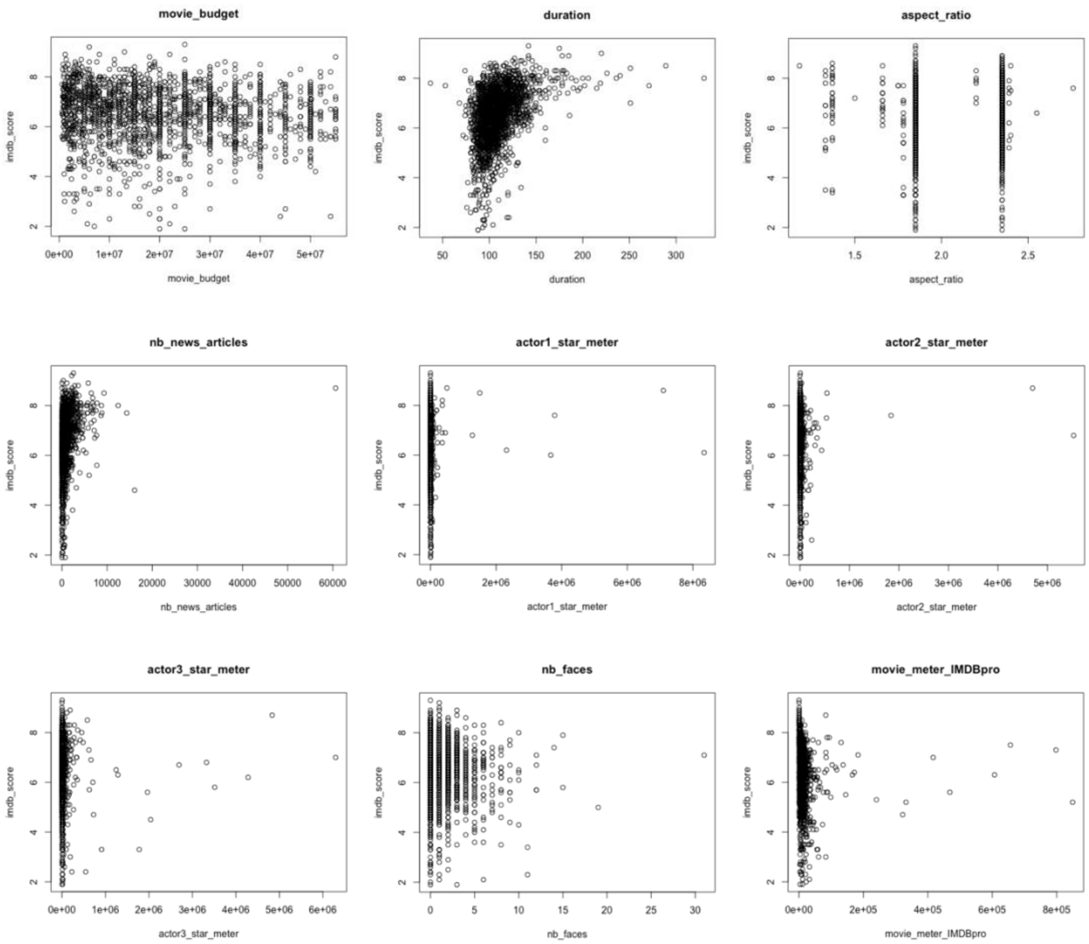




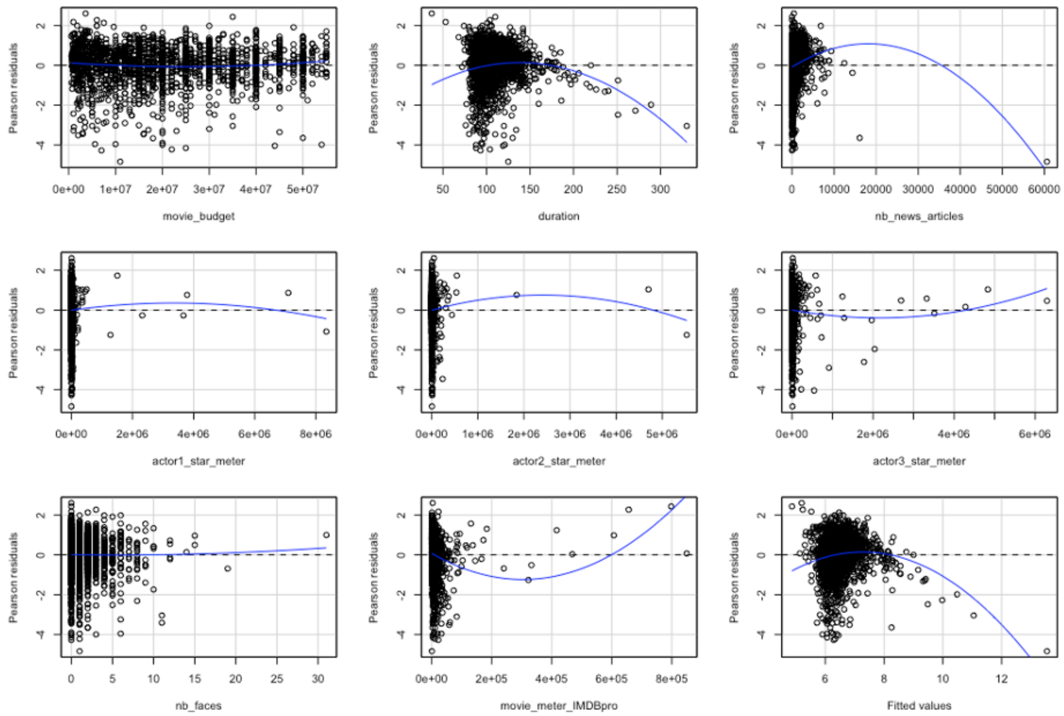
- Table E. Overwhelming numbers of unique value

	The number of unique values
director	1115
cinematographer	737
plot_keywords	4330
Total observations	1930

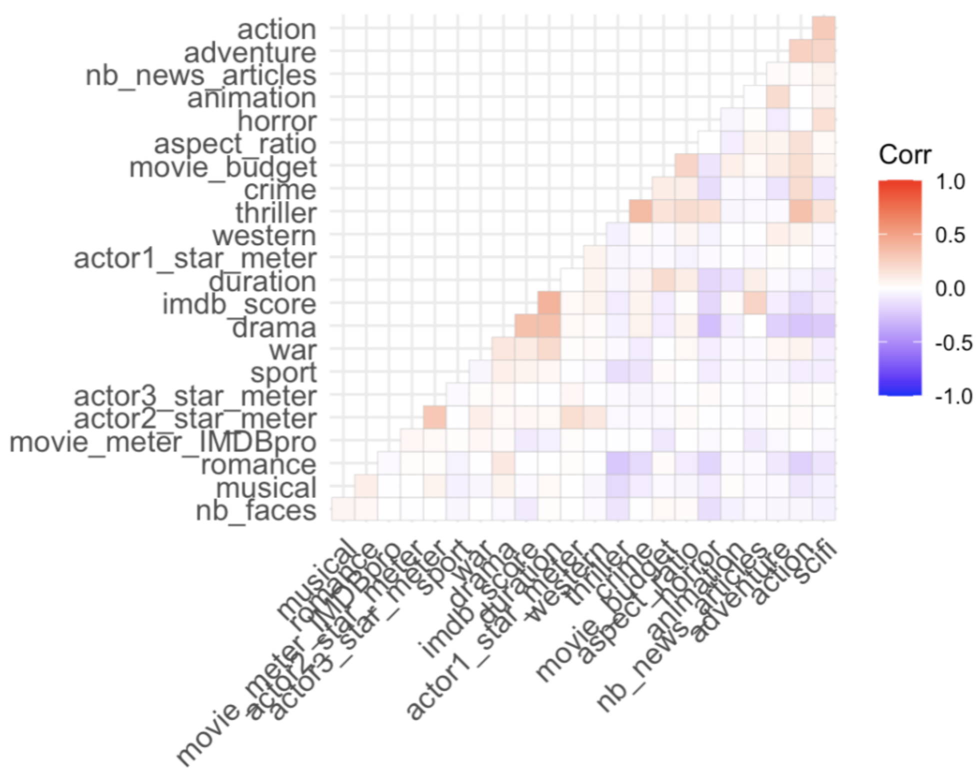
- Table F. Factors scatter plots (X = factors, Y = imdb_score)


movie_budget
duration
aspect_ratio
nb_news_articles
actor1_star_meter
actor2_star_meter
actor3_star_meter
nb_faces
movie_meter_IMDBpro
imdb_score

- Table G. Residual plot for linearity checking:



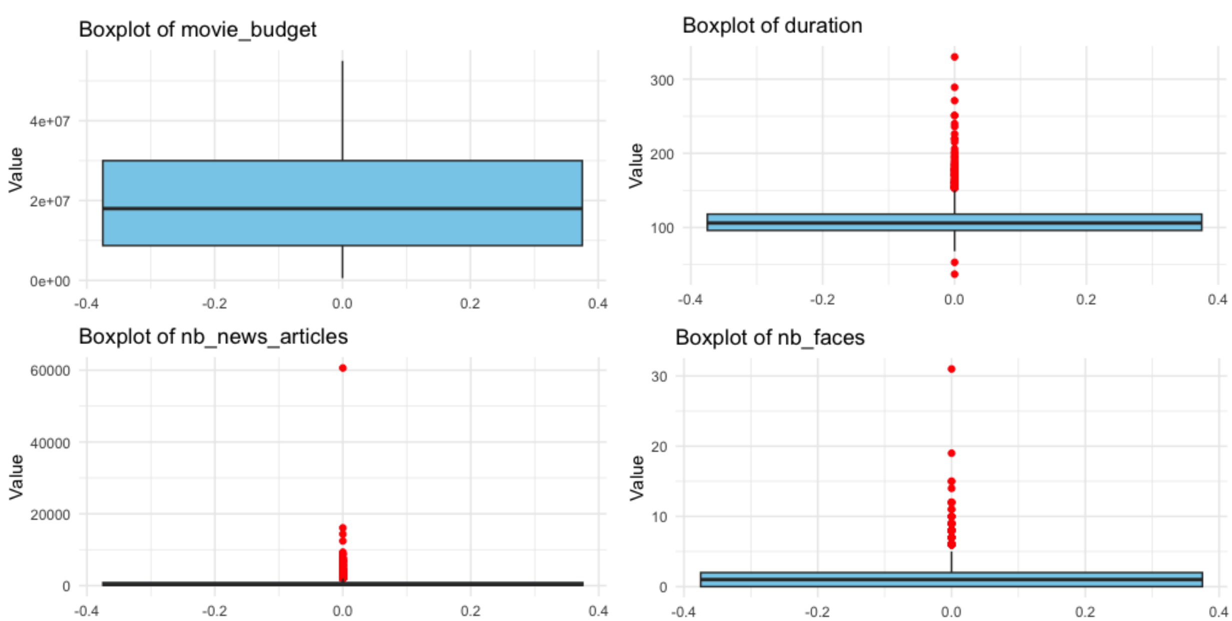
- Table H. Correlation between factors

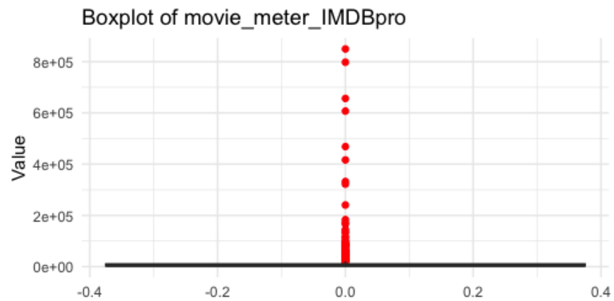


- Table I. Correlation between imdb_score and factors

	Correlation	Strength	Direction
movie_budget	-0.078669355	Weak	Negative
duration	0.410643860	Moderate	Positive
aspect_ratio	0.011146360	Weak	Positive
nb_news_articles	0.225451564	Weak	Positive
actor1_star_meter	0.028927852	Weak	Positive
actor2_star_meter	0.038274993	Weak	Positive
actor3_star_meter	-0.004070924	Weak	Negative
nb_faces	-0.089400348	Weak	Negative
action	-0.159057615	Weak	Negative
adventure	-0.066890417	Weak	Negative
scifi	-0.093802929	Weak	Negative
thriller	-0.080035686	Weak	Negative
musical	-0.022655065	Weak	Negative
romance	-0.014883144	Weak	Negative
western	0.065532975	Weak	Positive
sport	0.055001449	Weak	Positive
horror	-0.166071401	Weak	Negative
drama	0.338203870	Moderate	Positive
war	0.108544288	Weak	Positive
animation	0.016579825	Weak	Positive
crime	0.061444283	Weak	Positive
movie_meter_IMDBpro	-0.089732043	Weak	Negative

- Table J. Boxplots for initially checking outliers





- Table K. Potential models: predictions for the 12 movies by all models

	Movie Names	Model Predictions	Log Model Predictions	Spline Model Predictions	Model insignificant variables removed	Log Model Predictions insignificant variables removed	Spline Model Predictions insignificant variables removed
1	Dream Scenario	7.28	7.41	7.33	7.42	7.26	7.48
2	Leo	7.03	7.91	7.74	7.13	6.14	7.05
3	Napoleon	8.35	8.01	8.30	8.32	7.81	8.50
4	Next Goal Wins	7.13	7.81	8.06	7.04	7.07	7.97
5	Pencils vs Pixels	4.51	4.85	4.61	4.47	3.39	2.83
6	Thanksgiving	7.88	7.98	8.17	7.89	8.21	8.20
7	The Dirty South	7.86	6.02	6.09	8.16	8.41	6.41
8	The Holdovers	7.80	8.60	8.63	8.09	8.31	9.25
9	The Hunger Games: The Ballad of Songbirds and Snakes	7.75	7.22	7.72	7.85	7.38	8.01
10	The Marvels	4.32	4.29	4.65	4.57	4.99	4.80
11	Trolls Band Together	7.81	8.32	8.07	7.84	6.71	7.21
12	Wish	8.10	8.28	8.32	8.07	7.17	7.34

- Table L. Summary of the chosen model

```
> summary(final_model1)

Call:
glm(formula = imdb_score ~ poly(movie_budget, 2) + poly(duration,
    2) + poly(nb_news_articles, 4) + nb_faces + poly(movie_meter_IMDBpro,
    3) + maturity_PG13 + country_USA + genre_Drama + genre_Sport +
    genre_Horror + genre_Thriller + genre_Crime + genre_Comedy +
    genre_Action + genre_Mystery + genre_Family + genre_Animation +
    genre_Documentary, data = scaled_dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3608  -0.3845   0.0664   0.5202   3.2059


Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.74711    0.07242  93.166  < 2e-16 ***
poly(movie_budget, 2)1        -7.47950    0.95274  -7.851 7.00e-15 ***
poly(movie_budget, 2)2         3.56559    0.83734   4.258 2.16e-05 ***
poly(duration, 2)1            13.56035    0.98974  13.701  < 2e-16 ***
poly(duration, 2)2            -4.26213    0.86569  -4.923 9.27e-07 ***
poly(nb_news_articles, 4)1     8.83648    0.87294  10.123  < 2e-16 ***
poly(nb_news_articles, 4)2    -6.30880    0.84736  -7.445 1.48e-13 ***
poly(nb_news_articles, 4)3     0.71359    0.84451   0.845 0.398235
poly(nb_news_articles, 4)4    -2.31717    0.83799  -2.765 0.005747 **
nb_faces                      -0.76209    0.20106  -3.790 0.000155 ***
poly(movie_meter_IMDBpro, 3)1 -4.10170    0.85086  -4.821 1.55e-06 ***
poly(movie_meter_IMDBpro, 3)2  5.11026    0.88023   5.806 7.55e-09 ***
poly(movie_meter_IMDBpro, 3)3 -5.90124    0.86714  -6.805 1.36e-11 ***
maturity_PG13                 -0.26650    0.04539  -5.871 5.13e-09 ***
country_USA                   -0.15078    0.04991  -3.021 0.002552 **
genre_Drama                    0.28703    0.04840   5.931 3.60e-09 ***
genre_Sport                    0.22188    0.09208   2.410 0.016069 *
genre_Horror                  -0.47132    0.07118  -6.621 4.67e-11 ***
genre_Thriller                -0.12271    0.05533  -2.218 0.026697 *
genre_Crime                    0.10945    0.05242   2.088 0.036939 *
genre_Comedy                  -0.08664    0.05174  -1.675 0.094175 .
genre_Action                  -0.24467    0.05698  -4.294 1.85e-05 ***
genre_Mystery                  0.13847    0.06886   2.011 0.044487 *
genre_Family                  -0.29695    0.07855  -3.780 0.000162 ***
genre_Animation                1.06189    0.20061   5.293 1.35e-07 ***
genre_Documentary              0.78084    0.37230   2.097 0.036099 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Table M. Summary table for all other models

(1) Polynomial model:

```
> summary(final_model)

Call:
glm(formula = imdb_score ~ poly(movie_budget, 2) + poly(duration,
    2) + poly(nb_news_articles, 4) + actor1_star_meter + actor2_star_meter +
    actor3_star_meter + nb_faces + poly(movie_meter_IMDBpro,
    3) + distributor_dummy + production_company_dummy + maturity_R +
    maturity_PG13 + maturity_PG + maturity_Others + country_USA +
    month_Jan + month_Feb + month_Mar + month_Apr + month_May +
    month_Jun + month_Jul + month_Aug + month_Sep + month_Oct +
    month_Nov + month_Dec + genre_Drama + genre_Biography + genre_Sport +
    genre_Horror + genre_Thriller + genre_Crime + genre_Comedy +
    genre_Adventure + genre_Action + genre_Fantasy + genre_Mystery +
    genre_Family + genre_Animation + genre_Documentary + blockbuster_month +
    aspect_ratio_2_35 + aspect_ratio_1_85 + aspect_ratio_others,
    data = scaled_dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1538  -0.3808   0.0633   0.5095   3.0430
```

(2) Model with log transformation:

```
> summary(final_model_log)

Call:
glm(formula = log_imdb_score ~ poly(movie_budget, 2) + poly(duration,
    2) + poly(nb_news_articles, 4) + actor1_star_meter + actor2_star_meter +
    actor3_star_meter + nb_faces + poly(movie_meter_IMDBpro,
    1) + distributor_dummy + production_company_dummy + maturity_R +
    maturity_PG13 + maturity_PG + maturity_Others + country_USA +
    month_Jan + month_Feb + month_Mar + month_Apr + month_May +
    month_Jun + month_Jul + month_Aug + month_Sep + month_Oct +
    month_Nov + month_Dec + genre_Drama + genre_Biography + genre_Sport +
    genre_Horror + genre_Thriller + genre_Crime + genre_Comedy +
    genre_Adventure + genre_Action + genre_Fantasy + genre_Mystery +
    genre_Family + genre_Animation + genre_Documentary + blockbuster_month +
    aspect_ratio_2_35 + aspect_ratio_1_85 + aspect_ratio_others,
    data = scaled_dataset)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.08226  -0.05289   0.01921   0.08970   0.51381
```

(3) Spline model:

```
> summary(final_spline_model)

Call:
glm(formula = imdb_score ~ bs(movie_budget, degree = 3) + bs(duration,
    degree = 2) + bs(nb_news_articles, degree = 2) + bs(movie_meter_IMDBpro,
    degree = 1) + actor1_star_meter + actor2_star_meter + actor3_star_meter +
    nb_faces + distributor_dummy + production_company_dummy +
    maturity_R + maturity_PG13 + maturity_PG + maturity_Others +
    country_USA + month_Jan + month_Feb + month_Mar + month_Apr +
    month_May + month_Jun + month_Jul + month_Aug + month_Sep +
    month_Oct + month_Nov + month_Dec + genre_Drama + genre_Biography +
    genre_Sport + genre_Horror + genre_Thriller + genre_Crime +
    genre_Comedy + genre_Adventure + genre_Action + genre_Fantasy +
    genre_Mystery + genre_Family + genre_Animation + genre_Documentary +
    blockbuster_month + aspect_ratio_2_35 + aspect_ratio_1_85 +
    aspect_ratio_others, data = scaled_dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0719  -0.3910   0.0691   0.5340   3.3004
```

(4) Model with log transformation (w/o insignificant variables):

```
> summary(final_model_log2)

Call:
glm(formula = log_imdb_score ~ poly(movie_budget, 2) + poly(duration,
    2) + poly(nb_news_articles, 2) + nb_faces + poly(movie_meter_IMDBpro,
    3) + maturity_PG13 + country_USA + blockbuster_month + genre_Drama +
    genre_Sport + genre_Horror + genre_Crime + genre_Action,
    data = scaled_dataset)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.09226  -0.05555   0.02081   0.08784   0.48922
```

(5) Spline model (w/o insignificant variables):

```
> summary(final_spline_model2)

Call:
glm(formula = imdb_score ~ bs(movie_budget, degree = 3) + bs(duration,
    degree = 2) + bs(nb_news_articles, degree = 2) + nb_faces +
    production_company_dummy + bs(movie_meter_IMDBpro, degree = 1) +
    maturity_PG13 + country_USA + blockbuster_month + genre_Drama +
    genre_Sport + genre_Horror + genre_Family + genre_Crime +
    genre_Action, data = scaled_dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2279  -0.4115   0.0742   0.5329   3.2135
```

- Picture sources:
  https://pixabay.com/photos/stethoscope-medical-health-doctor-2617701/
  https://seeklogo.com/free-vector-logos/data-science
  https://brand.imdb.com/imdb
  https://www.imdb.com/
  https://www.flaticon.com/free-icon/clapperboard_10351880?term=movie&page=1&position=17&origin=search&related_id=10351880