MGSC 401: Statistical Foundations of Data Analytics
Group Project: Midterm Project
March 10th, 2025

Introduction

This project aims to predict IMDb ratings for 12 upcoming blockbuster movies using a statistical model. The model was trained on 2000 data points of previous movies, including variables such as “duration” and “number of news articles”. The main goal of this analysis is to identify patterns in historical ratings to predict the ratings of movies that have yet to be

released. The ability to predict the rating of upcoming movies can be especially useful in real world settings, particularly for critics and movie studios.

To build an accurate model, various statistical modelling techniques were performed such as correlation analysis. These tests identified the most influential variables to include in the final predictive model. Conducting such analysis is vital to create a model that is accurate and reliable, ensuring that only relevant variables are used. This not only improves the predictive performance of the model, but also reduces the computational requirement by removing useless variables.

Once the model was created, the predictive ability was tested through LOOCV and K-fold tests to ensure the most accurate results. These tests ensure the model is not overfitting to the training data and is truly able to predict future ratings of movies. This last step of the model building is essential to producing an effective predictive model that can be applied to real world scenarios. Overall, this project aims to create a robust model that can aid in predicting the IMDb scores of upcoming films, enabling potential stakeholders in the film industry to make data driven decisions.

Data Description

For the data description we will focus mostly on the variables that were used in the final model. We incorporated a blend of continuous and categorical independent variables that were either directly available in the dataset or derived via transformation. For the continuous variables, duration typically clusters between 90-120 minutes with a few films being outliers going well beyond two hours. This results in a slightly right-skewed distribution as can be seen in appendix A. The number of news articles displays a heavy right-skew; most movies only received a few news articles while a small subset garners extensive media attention as seen by the long upper whisker in the box plot and the concentration of values at the lower end of the histogram as can be seen in appendix A. Similarly, movie_meter_IMDBpro is markedly right-skewed, suggesting that although many films occupy a moderate popularity range, a few films receive extremely high ratings on this measure. Release_year is centered around the late 1990s to early 2010s with fewer films from much earlier or later years.

For the categorical predictors we created dummy variables to capture genre and stylistic features. The binary indicators for drama, horror and action denote the genre of the movies allowing for analysis of genre specific impacts on IMDB scores. Additionally, colour_film_dummy distinguishes between color and black & white films. We saw that there were very few films in black in white in comparison to color. We used a custom function to create dummy variables for the categorical variables of distributor, production_company and director. We took the top 5 highest frequencies for each of them and rated the rest as others. We also converted the **release_month** column into a factor and produced dummy variables for each month. However, after assessing predictive significance, these variables were ultimately excluded from the final model, as they did not provide additional explanatory power.

Outlier Detection:

A major problem in the dataset was outliers in certain numerical variables that, if not corrected, would have affected prediction in the model. For instance, there were clear outliers for the following: Movie length because certain movies were over 200 minutes long, yet a majority fell in the 90-120 minute range. Amount of press coverage because there are films with thousands of articles written about them, yet there are films with zero articles written about them. Ranking in IMDbPro because some movies ranked in the top 100, yet others ranked over 100,000.

To formally find and remove such outliers, we conducted a Bonferroni test, which used studentized residuals to find the outliers. We felt this was appropriate because it ensured that only the most extreme cases would be removed, allowing the data to retain its natural variations. Thus, this adjustment allowed for better generalizability since the risk of overfitting to just a few outlier situations was no longer an issue.

Heteroskedasticity:

Heteroskedasticity—unequal variance in errors across different observations—can lead to biased results and inconsistent predictions in regression models. When the model can't account for this unequal variance at different feature values, it significantly reduces its ability to predict accurately for any film.

We utilized a Non-Constant Variance (NCV) Test to check for heteroskedasticity, and it came back positive for these significant variables: budget, length, aspect ratio, number of articles, and IMDbPro rank. This result indicated that movies with larger budgets, longer runtimes, or extensive media coverage exhibited significant variability in IMDb ratings, making it difficult for a standard regression model to produce consistent predictions.

We learned that to increase the interpretability of our model, we would need to apply robust standard errors in a panel regression model, which adjusted for variations in error terms. This modification improved model stability and ensured that predictions remained unbiased regardless of differences in feature distributions (Appendix B).

Correlation Matrix

We assessed the correlation matrix to identify the variables most correlated with the IMDb rating. The correlation matrix assessed relationships between variables/characteristics in strength. Our approach to creating our model relied on using highly correlated variables as they would be statistically the most in the most impactful potential predictors. Our findings: Run time effectively had a strong positive correlation to IMDb ratings; essentially, the longer films became, the higher they received ratings. The number of Articles reported was positively correlated to IMDb ratings, which means that films that had more critical articles written about them may have predisposed audiences with expectations to pen similar ratings in response.

Yet interestingly, release year was negatively correlated with IMDb rating, meaning that the longer films have been out, the higher the rating. This is a function either of nostalgia bias, people giving the older, classical films more of a pass or survivorship bias, only the best movies that have stood the test of time are rated and the films that were of a lesser quality are either gone or forgotten. In addition, the positive correlation between IMDbPro popularity ranking and IMDb ranking suggested that those films rated higher were not necessarily the more popular and found on the channel, meaning that just because more people rated/found it did not mean it was more liked. This supports the idea that buzz is not always the best metric for value.

Evaluating these relationships ensured that the variables selected in our model had a documented, statistical relationship with IMDb ratings ensuring more precise predictions.

Multicollinearity

Multicollinearity occurs when two (or more) independent variables are highly correlated, making it hard to determine their separate contributions in predicting the IMDb score.

To assess collinearity, we used the Variance Inflation Factor (VIF) scores which measure how much a variable is being predicted by the other variables in the equation. It was an optimistic sign that post-VIF tests, all selected variables returned VIF values far below the critical cutoff which implied that multicollinearity densification issues would not arise. This increases the probability that our subsequently generated model would be stable and not underfit (Appendix C).

Model Selection

While proceeding with the variables mentioned above, the model chosen was based on the premise of its predictive capability. To determine which of the variables should be incorporated into the final model, the ten most correlated predictors to movie score were exposed to various statistical models. After running these variables through multiple tests, it was determined that the Sci-Fi genre did not play a significant role in predicting the score of a given movie. Furthermore, the same applied for movies that were released in the month of December. Due to this phenomenon, release_month was dropped from the model altogether. This variable does not have a drastic effect on the final model being implemented because the movies that it is built to test are all released in the springtime.

After testing various specifications, we finalized a model comprising eight key predictors that demonstrated the strongest explanatory power while minimizing overfitting risks. To determine the appropriate transformations, we systematically tested each predictor using a loop that evaluated multiple functional forms, including linear, quadratic, cubic, quartic, and natural splines with varying degrees of freedom. This process allowed us to identify the best fit for each variable based on statistical significance, model performance, and residual analysis. “Duration” was transformed using a quartic polynomial to capture non-linear effects, acknowledging that while longer films generally receive higher ratings, excessively long runtimes may negatively impact audience reception. The “number of news articles” was highly right-skewed, with most films receiving minimal media coverage while a small subset garnered extensive publicity, necessitating a natural spline transformation (df =

4) to account for non-linearity. Similarly, the IMDbPro movie meter ranking exhibited a complex relationship with IMDb scores, as popularity did not always correlate with higher ratings, requiring a spline transformation (df = 5). Release year showed a negative correlation with IMDb scores, suggesting that older films tend to receive higher ratings, possibly due to nostalgia or survivorship bias, prompting the application of a natural spline transformation (df = 5). In terms of genre, drama films were associated with significantly higher IMDb scores, while horror and action films tended to receive lower ratings, possibly due to polarized audience opinions and harsher critical reviews. Lastly, a binary color film dummy indicated that black-and-white films received higher ratings on average, potentially due to their historical significance or niche audience appreciation. These transformations and selections ensured the final model effectively captured key patterns in the data while maintaining predictive accuracy and robustness.

Results

Before forecasting the IMDB movie ratings of the films released in March, we wanted to see how our model performed in predicting movie scores at random from the original dataset it was based on. We chose a sample of 5 films at random: *August: Osage County, Twilight: The New Saga, RocknRolla, Pitch Perfect, and Step Up*. The results were the following:

- August: Osage County – 7.2 predicted vs. 7.3 actual
- Twilight: The New Saga – 6.99 predicted vs. 4.6 actual
- RocknRolla – 6.6 predicted vs. 7.3 actual
- Pitch Perfect – 7.17 predicted vs. 7.2 actual
- Step Up – 7.35 predicted vs. 6.5 actual

In some cases, the model appeared to perform rather well – the results from *August: Osage County* and *Pitch Perfect* showed errors of 0.1 and 0.03, respectively. In our opinion, the model also seemed to fare decently well in its RocknRolla and *Step Up* projections – arbitrarily, we found that the difference of less than 1.0 was encouraging. On the other hand, we feel that the model did very poorly in predicting the rating for *Twilight: The New Saga*.

This led us to wonder about our model – is it good at predicting average to above-average films, but bad at predicting the score of films with bad ratings?

Now was the time to finally put our model to the test and predict the IMDB scores from the test data – in other words, a list of 12 films scheduled for release within March and April of 2025. After inputting all relevant information, this is the output of estimated IMDB movie ratings we received:

- Odessa: 5.9256 predicted
- Black Bag: 6.6156 predicted
- High Roller: 5.109 predicted
- Novocaine: 6.2375 predicted
- The Day the Earth Bley Up: 5.745 predicted
- Ash: 5.3183 predicted
- Locked: 5.5231 predicted
- Snow White: 7.129 predicted
- The Alto Knights: 6.9761 predicted
- A Working Man: 6.1928 predicted
- My Love Will Make You Disappear: 5.5476 predicted
- The Woman in the Yard: 6.1106 predicted

The model allowed us to make a handful of interesting conclusions. Take for example *High Rollers*, starring John Travolta, which scored the lowest predicted rating of the 12 films from our sample. Our model forecasted a score of 5.109. What's interesting about this film is that it is a sequel to a movie that was released in 2024 called *Cash Out*. The film was released to very poor reviews last year, scoring 4.7 on IMDB and 20% on Rotten Tomatoes, respectively. Therefore, a predicted score of 5.109 - given this information - is a very encouraging sign for the strength of our model!

Another interesting observation we gleaned from the model's output was its highest predicted score: the highly anticipated remake of *Snow White*. If this prediction holds true, it would certainly be welcome news for Disney, who had to stomach a litany of issues during production (Telegraph, 2025). This is not to mention the film's budget surpassing early planning and ballooning to $270 million USD (Forbes, 2024).

Furthermore, the model projects a relatively strong rating for *Alto Knights*, which boasts a predicted score of 6.9761 – this is the second highest rating we observed. This is something we will be keen to follow during the coming weeks. Indeed, the film features screen-acting legend Robert De Niro playing both leading characters – Italian-American crime bosses Vito Genovese and Frank Costello, in a dramatization of true events (IMDB, 2025). One film equivalent worth mentioning is 2015's *Legend*, starring Tom Hardy as the notorious Kray twins who ruled the British underworld during the 1960's (IMDB, 2025). That film scored 6.9 on IMDB, in its own right.

The final observation we thought interesting to share was the third highest predicted rating our model gave us: *Black Bag*. The Steven Soderbergh-directed project, starring Cate Blanchett and Michael Fassbender (IMDB, 2025), scored a projected rating of 6.6156.

When looking at the out of sample performance, two different statistical methods were used to measure the effectiveness of the model, ensuring the model was not overly fitted. Firstly, using the LOOCV method, the MSE was calculated to measure the predictive accuracy. The MSE obtained was 0.6758148, which is relatively low compared to the range of potential IMDb scores. The IMDb scores in the dataset go from 1.9-9.3, therefore since the MSE is considerably smaller, it suggests the model is relatively predictive with a reasonable error margin.

We also did a K-fold test to test out-of-sample performance and to ensure our results were consistent across different validation methods. We used 10 folds as this is the standard we used in class and is not computationally expensive. Running the K-fold test 5 times gave us an average MSE of 0.6783. This reiterates and confirms what the LOOCV test found, which is that our model has a relatively low error compared to the range of IMDb scores (1.9–9.3).

With these two cross validation tests implemented, the predictive power of our model can be examined, guaranteeing the model is not overfitted. If the model fails to predict

out-of-sample data points, then it cannot be applied to real life scenarios where data driven insights can be especially useful. While we cannot predict the future, we cannot help but be intrigued by the predictions we received from our model. Of course, we will only know for sure in the next few weeks!

References

Black Bag. IMDb. *https://www.imdb.com/fr-ca/title/tt30988739/*

Cash Out. IMDb. https://www.imdb.com/fr-ca/title/tt24131288/

Horner, A. (2025, March 6). *How Disney's Snow White Remake Became a $270 million Headache*. The Telegraph. https://www.telegraph.co.uk/films/0/how-disneys-snow-white-remake-became-a-headache/

Legend. IMDb. https://www.imdb.com/fr-ca/title/tt3569230/

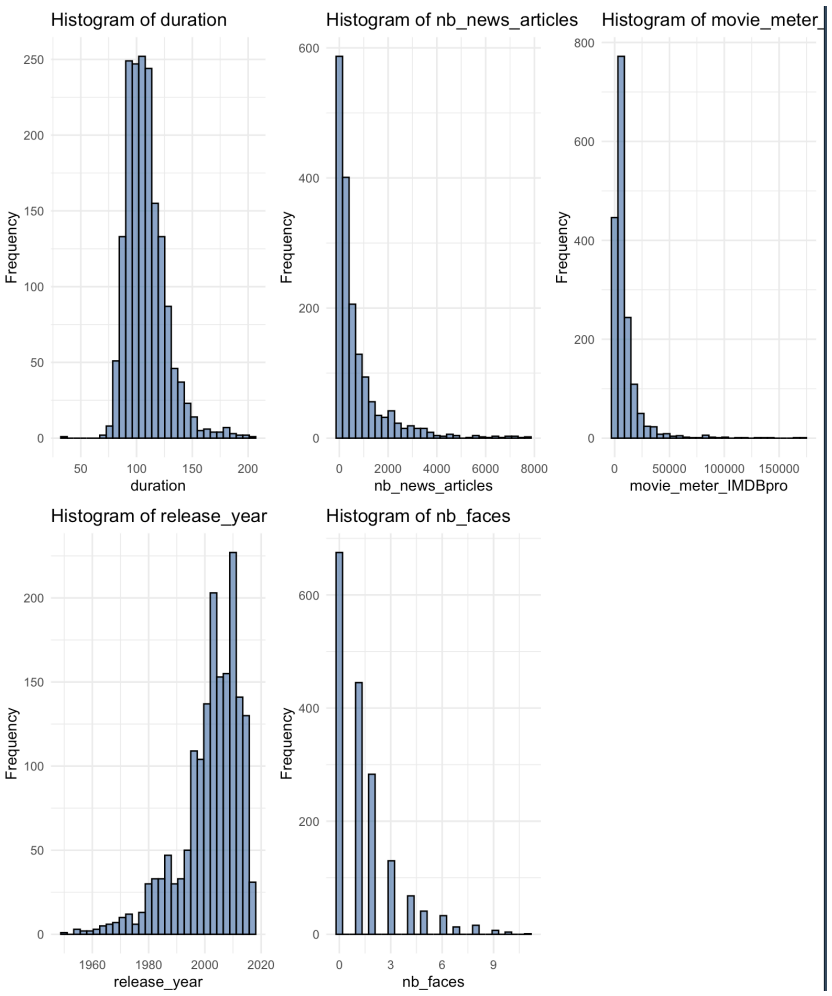
Reid, C. (2024, November 18). *Disney Reveals "Snow White" Remake is Set to Blow its Budget*. Forbes. https://www.forbes.com/sites/carolinereid/2024/11/14/disney-reveals-snow-white-remake-is-set-to-blow-its-budget/

Rotten Tomatoes. (2024, April 26). *Cash Out*. Rotten Tomatoes. https://www.rottentomatoes.com/m/cash_out

The Alto Knights. IMDb. https://www.imdb.com/fr-ca/title/tt21815562/?ref_=nv_sr_srsg_0_tt_8_nm_0_in_0_q_alto%2520knights

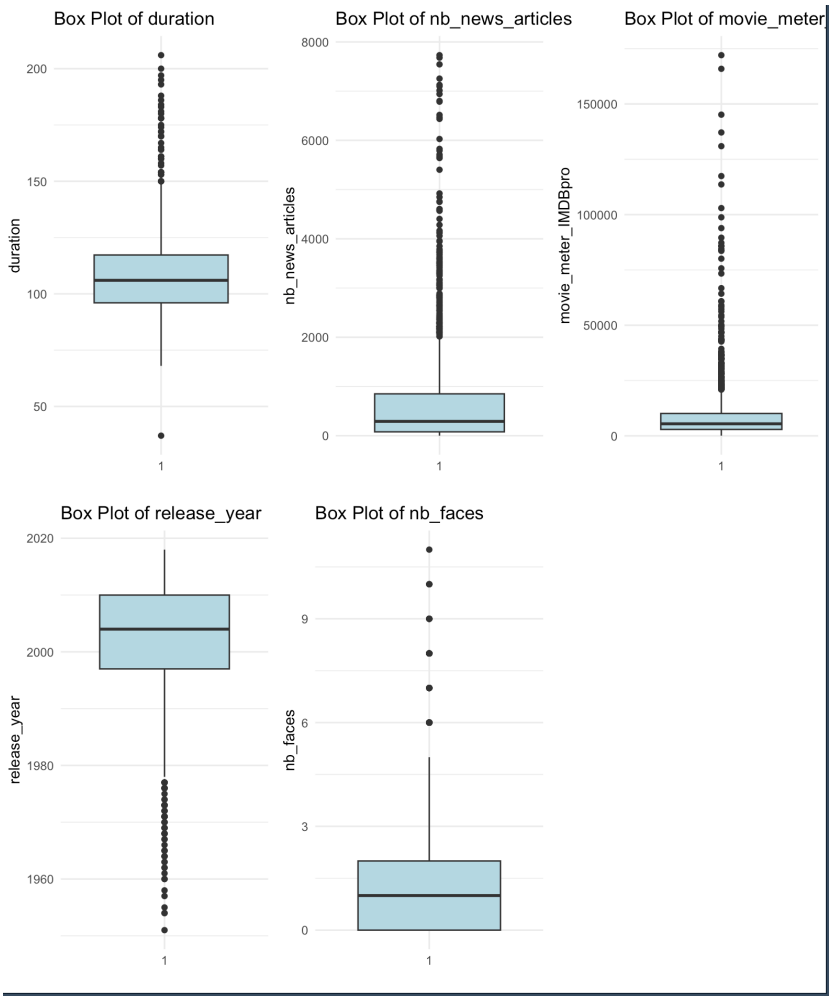
Appendix

Appendix A



Histogram of Each Quantitative Variable

Appendix A part 2



Boxplot of Each Quantitative Variable

Appendix B

```
> # Check and print significant heteroskedasticity results
> if (nrow(ncv_results) > 0 && any(ncv_results$P_Value < 0.05)) {
+   print("Significant He ..." ... [TRUNCATED]
[1] "Significant Heteroskedasticity:"
              Variable      P_Value
2         movie_budget 6.525640e-07
5             duration 2.951504e-17
6         aspect_ratio 3.096897e-07
7     nb_news_articles 4.555358e-02
9    actor2_star_meter 1.902600e-02
10   actor3_star_meter 2.812964e-04
11            nb_faces 1.948951e-02
12 movie_meter_IMDBpro 6.214082e-20
```

Heteroskedasticity Results

Appendix C

```
> print(vif_values)
         duration    nb_news_articles movie_meter_IMDBpro        release_year           nb_faces
         1.084374            1.155879            1.094145            1.114711            1.018223
```

VIF Values

Appendix D

```
Call:
lm(formula = imdb_score ~ poly(duration, 4) + ns(nb_news_articles,
    df = 4) + ns(movie_meter_IMDBpro, df = 5) + ns(release_year,
    df = 5) + drama + horror + action + colour_film_dummy, data = imdb_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2817 -0.3813  0.0982  0.5146  2.3308

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      8.36968    0.29758  28.126  < 2e-16 ***
poly(duration, 4)1               8.28913    0.95377   8.691  < 2e-16 ***
poly(duration, 4)2              -1.15356    0.84246  -1.369 0.171093
poly(duration, 4)3              -1.96995    0.82561  -2.386 0.017139 *
poly(duration, 4)4               2.17325    0.82814   2.624 0.008762 **
ns(nb_news_articles, df = 4)1    0.24270    0.08400   2.889 0.003911 **
ns(nb_news_articles, df = 4)2    0.96890    0.16404   5.907 4.21e-09 ***
ns(nb_news_articles, df = 4)3    0.88292    0.19785   4.463 8.63e-06 ***
ns(nb_news_articles, df = 4)4    0.61559    0.24955   2.467 0.013731 *
ns(movie_meter_IMDBpro, df = 5)1 -0.76653    0.12740  -6.017 2.17e-09 ***
ns(movie_meter_IMDBpro, df = 5)2 -1.09715    0.14248  -7.700 2.29e-14 ***
ns(movie_meter_IMDBpro, df = 5)3 -1.45149    0.25690  -5.650 1.88e-08 ***
ns(movie_meter_IMDBpro, df = 5)4 -1.65408    0.34146  -4.844 1.39e-06 ***
ns(movie_meter_IMDBpro, df = 5)5 -0.02986    0.42167  -0.071 0.943554
ns(release_year, df = 5)1        -0.97869    0.24216  -4.042 5.55e-05 ***
ns(release_year, df = 5)2        -0.93934    0.27708  -3.390 0.000715 ***
ns(release_year, df = 5)3        -1.07575    0.18799  -5.722 1.24e-08 ***
ns(release_year, df = 5)4        -1.95785    0.57735  -3.391 0.000712 ***
ns(release_year, df = 5)5        -0.90530    0.18475  -4.900 1.05e-06 ***
drama                             0.52113    0.04679  11.137  < 2e-16 ***
horror                           -0.36499    0.06532  -5.587 2.68e-08 ***
action                           -0.29161    0.05222  -5.584 2.74e-08 ***
colour_film_dummy                -0.47229    0.13078  -3.611 0.000313 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8184 on 1693 degrees of freedom
Multiple R-squared:  0.4376,    Adjusted R-squared:  0.4303
F-statistic: 59.88 on 22 and 1693 DF,  p-value: < 2.2e-16
```

Regression Results of Final Model

Appendix E

```
> glue("Odessa's Projected Score is {round(odessa_pred,4)}")
Odessa's Projected Score is 5.9256
> glue("Black Bag's Projected Score is {round(blackbag_pred,4)}")
Black Bag's Projected Score is 6.6156
> glue("High Roller's Projected Score is {round(highrollers_pred,4)}")
High Roller's Projected Score is 5.109
> glue("Novocaine's Projected Score is {round(novocaine_pred,4)}")
Novocaine's Projected Score is 6.2375
> glue("the Day the Earth Blew Up's Projected Score is {round(tdtebu_pred,4)}")
the Day the Earth Blew Up's Projected Score is 5.745
> glue("Ash's Projected Score is {round(ash_pred,4)}")
Ash's Projected Score is 5.3813
> glue("Locked's Projected Score is {round(locked_pred,4)}")
Locked's Projected Score is 5.5231
> glue("Snow White's Projected Score is {round(snowwhite_pred,4)}")
Snow White's Projected Score is 7.129
> glue("The Alto Knights's Projected Score is {round(ak_pred,4)}")
The Alto Knights's Projected Score is 6.9761
> glue("A Working Man's Projected Score is {round(awm_pred,4)}")
A Working Man's Projected Score is 6.1928
> glue("My Love will Make you Disappear's Projected Score is {round(mlwmyd_pred,4)}")
My Love will Make you Disappear's Projected Score is 5.5476
> glue("The Woman in the Yard's Projected Score is {round(twity_pred,4)}")
The Woman in the Yard's Projected Score is 6.1106
```

Predicted Score of Each Movie

Appendix F

$$\hat{\text{IMDB Score}} = 8.36968 + 8.28913 \cdot \text{poly(duration, 4)}_1 - 1.15356 \cdot \text{poly(duration, 4)}_2 - 1.96995 \cdot \text{poly(duration, 4)}_3 + 2.17325 \cdot \text{poly(duration, 4)}_4 + 0.24270 \cdot \text{ns(nb\_news\_articles, 4)}_1 + 0.96890 \cdot \text{ns(nb\_news\_articles, 4)}_2 + 0.88292 \cdot \text{ns(nb\_news\_articles, 4)}_3 + 0.61559 \cdot \text{ns(nb\_news\_articles, 4)}_4 + -0.76653 \cdot \text{ns(movie\_meter\_IMDBpro, 5)}_1 - 1.09715 \cdot \text{ns(movie\_meter\_IMDBpro, 5)}_2 - 1.45149 \cdot \text{ns(movie\_meter\_IMDBpro, 5)}_3 - 1.65408 \cdot \text{ns(movie\_meter\_IMDBpro, 5)}_4 - 0.02986 \cdot \text{ns(movie\_meter\_IMDBpro, 5)}_5 + -0.97869 \cdot \text{ns(release\_year, 5)}_1 - 0.93934 \cdot \text{ns(release\_year, 5)}_2 - 1.07575 \cdot \text{ns(release\_year, 5)}_3 - 1.95785 \cdot \text{ns(release\_year, 5)}_4 - 0.90530 \cdot \text{ns(release\_year, 5)}_5 + 0.52113 \cdot \text{drama} - 0.36499 \cdot \text{horror} - 0.29161 \cdot \text{action} - 0.47229 \cdot \text{colour\_film\_dummy} + \epsilon$$

Final Model