Analyzing Influential Factors of Movie Ratings

Balint Mucsanyi
Matrikelnummer 6013671

Daniel Dauner
Matrikelnummer 4182694

{balint.mucsanyi, daniel-wilfried.dauner}@student.uni-tuebingen.de

Abstract

We use the official IMDb datasets extended with scraped content from the imdb.com website. Our goal is to predict the IMDb rating based on features, such as the length, budget, or genre of a movie. We use logistic regression for this purpose. As a sub-study, we monitor the effect of COVID-19 on film ratings and revenues in the form of a hypothesis test.

1 Introduction

The central question of our project was quite general: what factors affect the success of a movie? To be able to quantify the level of appreciation of the public towards a particular movie, we considered two indicators: the *star rating* and the *box-office gross* of the film in question. To gather a representative dataset of films and their features, we used an extended version of the official IMDb dataset.

The IMDb website provides general information about movies, such as the runtime, the genre, or partially their budget and box-office gross, making it an ideal starting point for our purposes. The ratings of movies are aggregated by a weighted average and displayed on the title's main page. The exact method is not disclosed. According to IMDb [1], the method detects and rates fraudulent votes differently.

In the first part of our study, we studied the predictability of the IMDb rating of movies from various features with logistic regression. More advanced techniques were applied to investigate the capability of logistic regression in comparison. Using a non-linear transformation on our labels, we also fit a linear regression model and compared it to the other two approaches.

In our second experiment, we statistically examined the effect of the COVID-19 pandemic on movies that failed at the box office. We consider a movie as a box-office flop (often called a box-office bomb) if its budget exceeds its worldwide gross. Using multiple hypothesis tests, we show statistically significant trends of the "Covid-year" 2020.

2 Data Collection

IMDb offers a variety of datasets on their website[1] that are refreshed daily, which formed the basis of our experiments. The datasets contained 598,851 data points labeled as movies, of which only 273,557 movies had a user rating. The following features were used for logistic regression and were directly provided by IMDb: release year, runtime, number of rating votes, and genre. Some features were not provided by the official data but were available on the titles' main page. We additionally collected the estimated budget, the worldwide box-office gross, the number of user reviews, and the number of critic reviews of all movies with a user rating using a web-scraper over the course of 200 hours. During scraping, we

[1] www.imdb.com/interfaces (accessed on 14th January)

Table 1: Final training and test losses of a logistic regression model trained with steepest descent, using three different loss functions and three deep ReLU networks. The MAE variant performs best in terms of both training and test error between the logistic regression models. The ReLU networks were trained with L2 regularization and early stopping. The regularization parameter λ was tuned separately for each architecture. Test results for ReLU networks are not marginally better than using MAE logistic regression (not even a 0.1-star improvement could be achieved).

	Logistic Regression			ReLU-Net		
Variant	MAE	MSE	BCE	Depth-2	Depth-4	Depth-6
Final Training Loss	0.5803	0.6337	0.6721	0.4418	0.4289	0.4137
Test Loss	0.6107	0.6140	0.7978	0.5776	0.5678	0.5674

only accessed data that were not part of IMDb's `robots.txt` list and selected the delay between requests to closely model human activity.

The dataset included financial features from various decades in several currencies. Consequently, we had to account for inflation and convert the values into a uniform currency. For that, we used the inflation data of 16 currencies [2]. Movies were discarded if no inflation rates were available for the given currency. Each financial feature was converted from the release year to the year 2022. After that, all values were converted into USD. This was done using a CurrencyConverter library in Python, which takes the exchange rates from the European Central Bank.

The categorical features required additional preprocessing. IMDb considers 23 different genres. We decided to summarize these into 18 genres due to a large amount of sparsely represented categories. Another categorical feature is the age restriction rating. Merging these or creating useful meta-categories proved to be challenging due to non-overlapping categories and varying systems for nearly every country. Therefore, we only consider whether a movie has been rated or not.

Finally, we dropped any movies that did not have a complete set of features for the given experiment. The completed dataset consists of 10714 and 14674 movies for the movie rating prediction and the hypothesis test, respectively. As the hypothesis testing required fewer features, more movies were preserved.

3 Prediction of Movie Rating

In our first experiment, we chose the IMDb rating as the indicator of a film being cherished. We used 26 selected features, out of which 18 correspond to the *genre* and one to whether the film was *rated* or not. The remaining 7 were the *release year*, the *runtime*, the average *rating*, the number of *votes*, the *budget*, the worldwide *gross*, and the number of critic and user *reviews*.

We used logistic regression to predict the IMDb rating of films based on the aforementioned covariates. The label, in this case, is a continuous value between 1 and 10. Thus, the output of the sigmoid non-linearity was further transformed as $\hat{y} = 9 \cdot \sigma(\theta^\top \mathbf{x}) + 1$. As the choice of the loss function was not immediately obvious, we experimented with the regular binary cross-entropy, the mean squared error, and the mean absolute error as well. For evaluation purposes and easier interpretability, we used the mean absolute error as our test metric for all three approaches. We divided our dataset of 10714 samples with a 90%-10% split into training and test sets. As scikit-learn does not allow setting custom loss functions or having non-binary targets, we implemented the models in PyTorch. We trained the models using steepest descent with a learning rate of $1e-3$ and the following stopping criterion: $\frac{1}{|\theta|}\|\nabla J(\theta)\|_2^2 < 1e-4$. The results are summarized in Table 1. The MAE variant achieves the lowest training and test error.

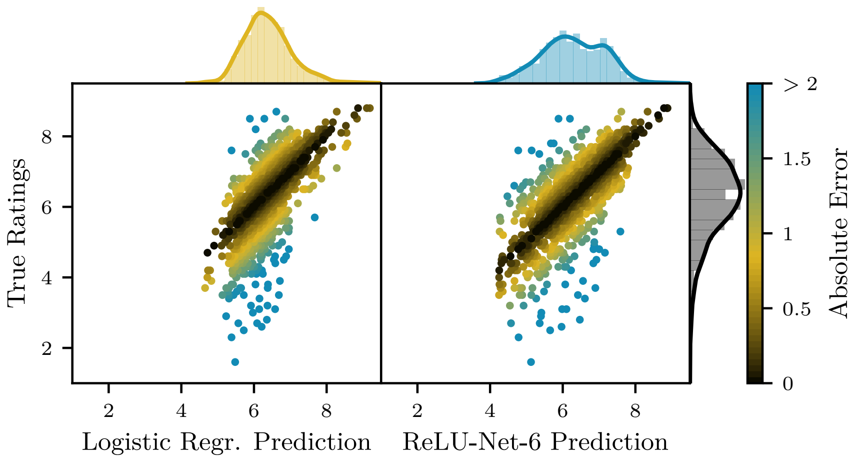

Figure 1: Prediction map of MAE logistic regression and the largest ReLU-Net. Although MAE logistic regression achieves a below-one-star MAE, it performs poorly outside of the $5-8$ rating range. The ReLU-Net has a larger range where its predictions are mostly correct.

We also tried transforming the IMDb ratings to real numbers by using the inverse of our scaled and shifted sigmoid: $y^{(\text{new})} = -\log\left(\frac{9}{y^{(\text{old})}-1} - 1\right)$. This made it possible to try a logistic regression model as well, which has a closed-form solution. The test loss, in this case, was 0.8216, the worst of these four methods.

Although all four approaches achieve a below-one-star mean absolute error, we wanted to investigate whether a model with a non-linear decision boundary could achieve an even better generalization error. For this, we considered ReLU networks with varying depth, trained with a learning rate of $1e-3$. We chose the MAE loss, as that performed best in the linear case. The hidden layers contained 32 neurons each. The results are summarized in Table 1. No architecture achieved a 0.1-star error improvement compared to MAE logistic regression.

We did not manage to significantly improve the accuracy of logistic regression by using deeper models. This could indicate that our covariates are not descriptive enough to go below the 0.5-0.6 test MAE regime. However, as our method and architecture selection were not by any means extensive, this remains to be a hypothesis. In the future, we would like to experiment with kernel machines [6] and random forests [4] as well.

4 Movie Industry in 2020

In the second part of our project, we turned to the second indicator that a film is generally enjoyed by the public: the box-office gross. According to the Motion Pictures 2020 Theme Report [3], the worldwide grosses were heavily negatively affected by COVID-19. We wanted to verify this claim with our available data. For this, we used 14674 movies (all movies in the IMDb database with available gross and budget) and considered the frequency of flops in years preceding 2020 and in 2020 separately.

Our null hypothesis was that the number of flops in 2020 comes from the same distribution as in all preceding years. We considered a standard $\alpha = 0.05$ one-sided test, treating only more flops as extreme cases. The rationale behind this was that we wanted to study whether COVID-19 had a negative effect or not. The p-value in this case is

$$p\text{-value} = \sum_{m=m_{2020}}^{N_{2020}} p(m \mid H_0),$$

where m_{2020} is the number of flops in 2020 and N_{2020} is the number of all movies in 2020 with available gross and budget. As the true probability of a film being a flop before 2020 is

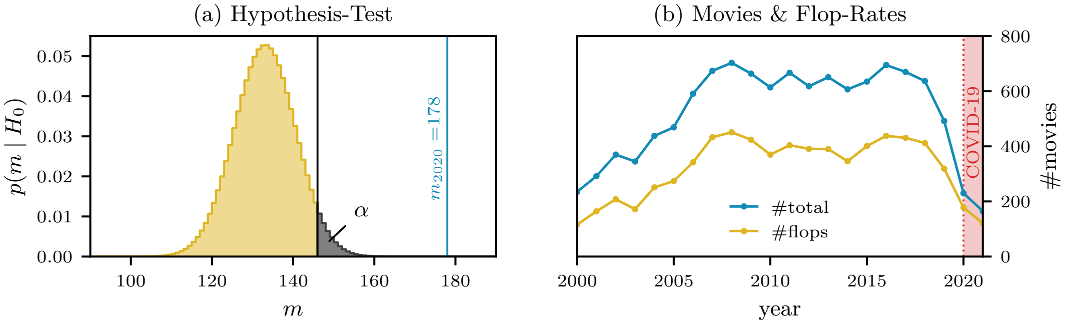

Figure 2: (1) One-sided hypothesis test. The results are significant (p-value = 7.234e−10). (2) Number of movies (blue) and the flop rate (yellow) plotted against years. In 2018-2021, a significant drop in the number of movies can be observed, with a jump in the flop rate in 2020.

unknown, we have to integrate over it as

$$p(m \mid H_0) = \int_0^1 p(m \mid f, H_0) p(f \mid H_0)\, df,$$

as seen in Lecture 5 in Data Literacy [5]. The posterior of f is a beta distribution, and the likelihood is binomial. The resulting distribution is thus a beta-binomial one, with parameters $N_{2020}, m_{<2020} + 1, N_{<2020} - m_{<2020} + 1$. Here $m_{<2020}$ is the number of flops before 2020, and $N_{<2020}$ is the number of movies before 2020. The results of the test can be seen in Figure 2.

Even though we got a significant result, we can easily verify that we would also get a significant result for the years preceding 2020 too. This is interesting in itself, as it indicates that in the previous years, it's quite unlikely that the number of flops follows the distribution of the flops in all of the years before. However, if we now calculate the p-value for a fixed year for all possible year intervals [lower_bound, year), we can see that 2020 was the first year since 2007 where the result is significant for *every possible interval*.[2] This highlights the year 2020 as an interesting outlier in the middle of an already significant trend of increasing flops in previous years.

Finally, we wanted to connect the two indicators of a good movie that we considered in our project: the IMDb rating and the box-office gross. We did this by further refining our hypothesis test, dividing our dataset into two parts: one with an IMDb rating of less than 5 and one with greater than or equal to 5. We followed the same method introduced previously on the two datasets separately. The result for the bad movies (IMDb rating less than 5) was not significant (p-value = 0.1962); thus, we cannot state that the rate of flopping for bad movies changed based on this test. Conversely, the results for the good movies (IMDb rating greater than or equal to 5) are significant (p-value = 4.4533e−9, meaning that it is very unlikely that the probability of flops for good movies stayed the same in 2020.

5 Conclusion

To answer the question "What movies do people like?", we considered two important metrics: the IMDb rating and the box-office gross. In our first experiment, we tried to predict the IMDb rating from several features of a movie. With MAE logistic regression, we achieved a below-one-star error rate. By experimenting with deeper models, we couldn't get this error significantly lower.

In the second experiment, we wanted to study the effects of COVID-19 on the worldwide gross of movies. The test showed significant results for 2020 compared to all previous years. To connect the two indicators, we divided our dataset based on the IMDb rating and conducted

[2] We support this claim in code.

the same test separately. We only got a significant result for good movies, meaning that there was most likely a greater change in the probability of flopping for good movies than for bad ones.

References

[1] Imdb ratings faq. https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV. Accessed: 2022-1-27.

[2] Inflation calculator. https://www.officialdata.org/. Accessed: 2022-1-27.

[3] Motion pictures 2020 theme report. https://www.motionpictures.org/wp-content/uploads/2021/03/MPA-2020-THEME-Report.pdf. Accessed: 2022-1-28.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Philipp Hennig. Data literacy, ws 2021/22 - lecture 05: Testing hypotheses. Accessed: 2022-1-28.

[6] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008.