Movie Everlasting: An Analytical Model for the Continuity of Consumer Preferences in Film

Team 30: Zhiqi Fu, Benji Kan, Yizheng Wang, Bo Zhang

Citadel Data Open East Coast Regional Fall 2020 (September 14-20)

Abstract

In this report, we describe a model for predicting the perpetuation of movie interest over the long term. Whereas movie trends are reflective of trends in popular culture at large, the film industry has the difficult task of predicting future trends in consumer preference. Given the unpredictability of culture and society in the modern age, we present a model that focuses on answering the question: What features influence the lifespan of movies, determining whether a movie has just a brief burst of popularity before fading out or surviving as an enduring classic? After labeling movies by category through their long-term sustainability and their consumer approval rating, we associate content tags to each movie with a relevance score. Using these genome tags, we then train a model via logistic regression to extract trends in these features, and we then test the accuracy of this model on predicting trends in a test data set. Overall, our model has a 92% precision score in accurately classifying movies as highly sustainable and highly rated across time.

Contents

1 Executive Summary

1.1 Introduction and Background

Movies are the essence of modern culture; from iconic lines to glamorous stars, the film industry continues to have an outsize cultural impact worldwide. Beyond its role as a recreational pasttime, entertainment has produced lasting social and political change as well, by inducing prominent national conversations and elevating voices and figures that influence international discourse. In a myriad of ways, cinema serves both to reflect and to shape culture in a universal manner, across continents and time.

But just as movies shape social and political progress, so must movies be shaped by the technological zeitgeist. From their humble beginnings as black-and-white silent films to modern IMAX motion pictures with vibrant graphics, movies have reflected technological progress, steadily improving in quality throughout the decades. Yet, the traditional film industry faces dual threats from the increasing popularity of streaming services and a Second Golden Age in TV shows. In order to adapt to both the opportunities and challenges of present day competition, the entertainment industry must continue to innovate with state-of-the-art technology and capitalize on preeminent trends in consumer preferences.

1.2 Motivation and Question

Consumer preferences can change nearly instantaneously, especially in the rapidly-evolving modern landscape. Moreover, consumers can afford to be picky in today's widening entertainment markets, from unlimited streaming services with indulgent algorithms to increased access to content of comparable quality. With fickle consumers preferences and high-caliber competitors, the entertainment industry bears the unenviable burden of delivering precisely to audiences' sweet spots.

In a perfect world, film companies would find data on consumer preferences and create movies that cater to all possible tastes, but the movie production process is extremely labor-

and time-intensive. Thus, movie companies must be selective with the movies they decide to film and fully produce. However, any input from consumers can only reflect the sentiment of the time of survey, and as a result of consumers' rapidly changing tastes compared to the lengthy process of movie production, it is especially challenging to predict specific future trends.

To address the question of consumer preferences without data from the future, we focus on extracting information about movies that have universal appeal across time. In essence, we ask the question:

What features influence the lifespan of movies, determining whether a movie has just a brief burst of popularity before fading out or surviving as an enduring classic?

After all, many commercial movies, even highly successful ones, have a large short-term impact, but are quickly forgotten and consigned to oblivion. On the other hand, there are certain qualities that allow a movie to retain some sort of popularity years into the future, even if its initial impact was smaller than a classic Hollywood tent-pole blockbuster. There must be some aspect of these longer-lasting movies, whether that be a specific factor in the content, or some quality that provides it with a loyal and sustainable fanbase across generations. At the same time, in addition to highly-praised movies, there exist notorious movies that are infamously remembered and rewatched. One example of a movie disaster that remains well-known despite crushing reviews is *The Room*, a classic movie—for the wrong reasons.

In answering this question, we hope to characterize the common qualities of movies that remain well-known regardless of the cultural zeitgeist, determining the factors that allow a movie to stand the test of time.

1.3 Data, Labels, and Features

In order to determine consumer preferences, we consider trends from rating data given directly by users themselves. We divide the data into four categories: 1) low rating and low sustainability; 2) low rating and high sustainability; 3) high rating and low sustainability; and 4) high rating and high sustainability.

Rather than considering transient factors that change over time, such as movie companies, directors, writers, or stars that may lose their appeal over time, we isolate the features that are more universal: the content metadata and relevance from `genome_score`. We then distill the features that distinguish movies in the four labeled categories; in particular, the tags that separate a well-liked, long-surviving category 4 movie from the rest.

1.4 Key Findings and Significance

After our analysis, described throughout Section 2, we create a model that classifies category 4 movies, those with high rating and high sustainability, with 92% accuracy. Moreover, we can similarly classify category 2 movies, those with a long lifetime but poor ratings, with 71% accuracy. Our model is discussed in more detail in Section 2.3.

These results suggest that there are indeed qualities that characterize the continuity of movies over time. We have exhibited a logistic regression model that can accurately predict movies perpetuation beyond its first year of release, and we can distinguish between movies that are well-received and poorly-received by consumers, even into the future. By extracting the direct categories of our model outputs, the film industry can better predict the success of certain movies given a couple specific keywords associated with the movie.

Our results may be applied to predict the universality of movies, which provides a robust metric to predict performance regardless of temporal context. These conclusions about consumer preferences have far-reaching applications in movie recommendation systems for streaming services, film company executive decisions, and industry understanding the predictability of movie popularity regardless of contemporary trends.

2 Technical Findings

2.1 Data Transformation and Feature Engineering

We begin by aggregating the `ratings` dataset into statistics for each movie by year that the rating is assigned. Since we aim to describe overall consumer trends in the data, we do not distinguish between individual preferences. We group all ratings for a given movie by the year the rating is assigned, and compute the mean and standard deviation for each year the movie received ratings.

To ensure that our data is high-quality, we first sort out all movies for which there is only one year of ratings data, as we cannot compare its change in popularity over time. This dropping process also excludes movies that were released prior to the start of the `ratings` dataset. Then, we determine the average of all ratings across time for each movie, which we can use to divide all movies into a highly-rated and poorly-rated section. In our EDA in Section 2.2, we find that while there are minor fluctuations in average rating across time, users' average rating over time remains fairly tight within 0.1 of 3.5. We also compute a retention percentage metric for each movie, calculated by dividing the count of the number of ratings received by a movie in the first year versus the number of ratings received in the last year in the dataset. This retention percentage measures how popular a movie remains across time, long after its initial release.

We find that the median of the average rating for each movie is 3.27, and the median of the retention percentage is 20%. Thus, for average rating r retention percentage p_r, we assign movies to categories based on the following rules:

Category 1: $r < 3.27$. $p_r < 0.2$

Category 2: $r < 3.27$. $p_r > 0.2$

Category 3: $r > 3.27$. $p_r < 0.2$

Category 4: $r > 3.27$. $p_r > 0.2$

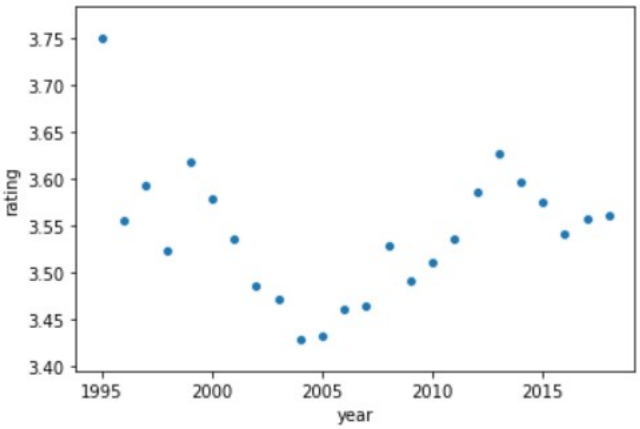


Figure 1: Average rating assigned to movies in each year over time.

Since we select the two-factor divisions based on the median values, each category contains approximately one quarter of the full available data.

We then split the movies into two populations: a train dataset, and a test dataset. Since we hope to train our model on movies with a longer range of data across time, we select the training dataset to be all movies with ratings in 2009 or earlier, and we reserve the test dataset to be movies released after 2010. An additional advantage of this split is that our test dataset is for more recent movies, which will be a better indicator of how predictive our model is. In addition, we are guaranteed that these two sets are disjoint, but each contains the entire range of movie genres released that year.

2.2 Exploratory Data Analysis and Insights

We begin by investigating average rating of movies over the years, shown in Figure 1.

We note that the y-axis is scaled, and so the actual variation in ratings over time is quite small—around 3.5 ± 0.1 out of a maximum score of 5. This gives us confidence to compare ratings across time.

We also visualize the distribution of retention percentages for movies throughout time, displaying several in Figure 2.

As can be seen, there is natural fluctuation in the number of ratings a movie receives.

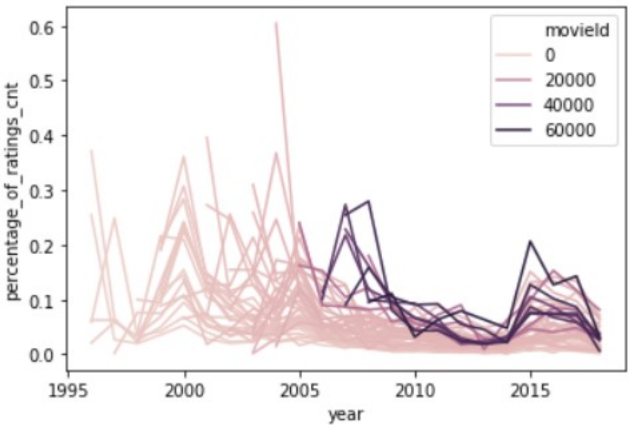


Figure 2: Movie retention of viewers across time, as a percentage of number of ratings received in its first year in the dataset.

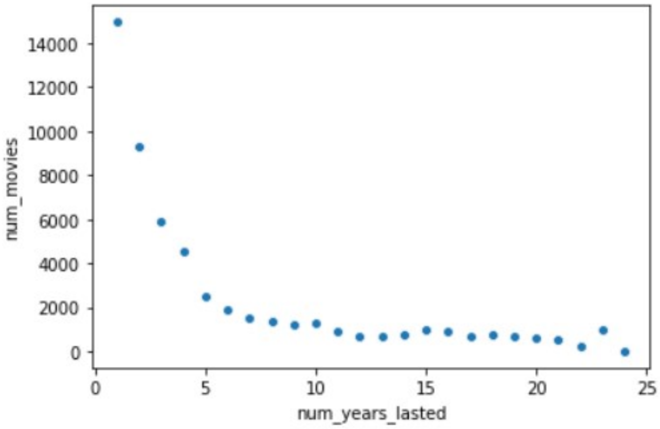
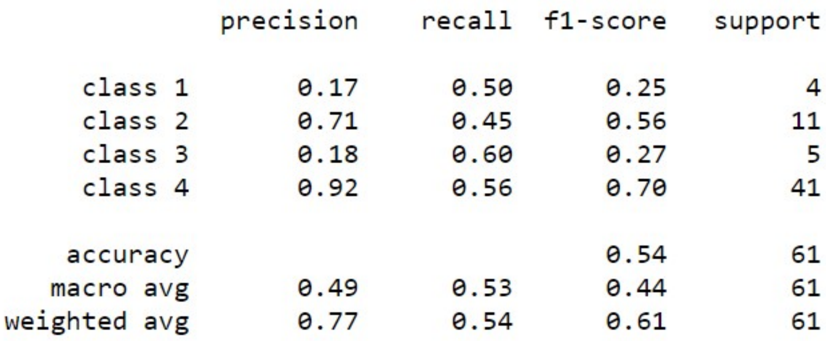


Figure 3: Number of movies with ratings for number of years in the dataset.

However, this fluctuation is quite small, and validates our decision to split the movies based on the final retention percentage, as a ratio compared to its initial number of ratings upon release.

Finally, we also visualize the distribution of movies as a function of the number of years it continues to receive ratings across time, shown in Figure 3.

We see a exponentially decaying distribution, as is broadly expected.



```
              precision    recall  f1-score   support

     class 1       0.17      0.50      0.25         4
     class 2       0.71      0.45      0.56        11
     class 3       0.18      0.60      0.27         5
     class 4       0.92      0.56      0.70        41

    accuracy                           0.54        61
   macro avg       0.49      0.53      0.44        61
weighted avg       0.77      0.54      0.61        61
```

Figure 4: Results of our model on the test data, described in Section 2.1.

2.3 Feature Selection and Models

We focus on features due to content as given by genome tags, as other factors cannot remain as constant the long run. Thus, we utilize the `genome_tags` and `genome_scores` datasets. Since the content tag metadata is inherently made up of categorical variables, we considered three potential modelling methods: logistic regression, random forest, and an SVM classifier. While SVM relies n Euclidean distances, for which we have no good representation of for the dense matrix of genome tag assignments, and our matrix is much denser than can be reasonably handled by a random forest model, we utilize a logistic regression to form our model.

We train our logistic regression model on the training data described in Section 2.1. Our model creates a classifier based on the association of various tags, and based on the relevance of those tags to the movie, to a movie, with our assigned category label. We display the results of our model on the test data in Figure 4.

We note that our model performs at 92% accuracy in classifying movies in category 4, or movies that have both high rating and high sustainability. moreover, our model performs at 71% accuracy in classifying movies in category 2, or movies with high sustainability across time but low ratings. This result indicates the overall quality of our model, as we can predict with relatively high accuracy if a movie will be well-regarded or poorly-regarded, given that it has a longer lifetime.

Comparatively, our model performs poorly in classifying movies in category 1 and 3, which are movies with lower sustainability, and thus contain less data from viewers. This result is not particularly surprising; many movies that are continually rewatched many years into the future, whether for good or ill reasons, share similar, universal qualities. However, movies that burn out all have their own distinctive qualities, and differ in unique but subtle ways. As a result, our model performs relatively poorly on these classes of movies.

3 Appendix A: Future Research Directions

While our work on consumer preferences over the long-term has significant applications of broad interest to consumers, film producers, writers, and industry executives at large, we suggest several future directions for further study.

While our ratings data provides a direct view of consumer impressions of movies, future study could include collecting a large sample of such data from a controlled set of viewers, sampled more broadly from the population. At the present moment, our data is limited by the movies that individual consumers watch, and the limited overlap between individuals. In addition, the ratings could be standardized for individual bias in higher or lower ratings, based on a common set of movies that all individuals in the sample rate.

Moreover, an interesting research direction would be to more carefully investigate the relevance of movie genre to the long-term sustainability of movies. Our model focused on the direct content of the movies, but applying a machine learning algorithm to provide scores of relevance to genres could provide a better weighted model for the overall predicatbility of movie continuity.