LetterBoxd Film Analysis

Genre/Theme/Production

Clayton Brittan

CSPB-4502

CU Boulder

clbr8512@colorado.edu

Abstract:

The film industry is constantly evolving and with it the viewing public's tastes. Because of this it is becoming more important to understand these trends both as filmmakers and viewers. By understanding the underlying factors that influence a film’s success we can determine if any particular factor plays a greater role in determining a film's success. This project aims to do just that, by studying the data available from LetterBoxd, a crowd sourced film review site, we can make predictions of certain films.

To understand the underlying factors that influence a film’s success we must ask the following questions:

What are the key factors that contribute to a film's success in terms of audience reception and ratings?

How do different factors such as release year, duration, genre themes, and other features influence a film's ratings on a platform like LetterBoxd?

Are there any discernible trends or patterns in audience preferences over time, and how do these trends correlate with changes in the film industry?

Can we accurately predict the success or ratings of a film based on its various attributes such as release year, duration, and genre themes?

What insights can be gained from analyzing crowd-sourced film reviews and ratings on platforms like LetterBoxd, and how can these insights inform filmmakers and viewers about evolving trends in the film industry?

In my exploration, I found that relying solely on features like release year, duration, and movie themes aren’t enough to succeed in accurately predicting movie ratings. It seems like we need a bit more depth in our features to make better predictions. Including more features that affect a film’s performance such as studios, directors, crew, equipment, sound design among others would allow for a more versatile model which likely would be able to have more accurate predictions.

In analyzing films using traditional classification models was not as effective as anticipated, graphical analysis of the film data served to reveal underlying trends

within the data itself. By analyzing the distribution of themes we were able to spot more trends then through the model analysis.

One notable discovery was the clear performance gap between different movie genres. Drama films tended to score higher than horror movies on average, shedding light on the varying tastes and preferences of moviegoers. This is likely due to the intent of film makers when making Drama movies vs Horror films, that of the former being to tell a compelling story while the latter aims to get a reaction out of the audience.

Another challenge faced was in dealing with outliers. The data seemed to follow a normal distribution, which made it tricky to predict those rare instances where a movie scores a 4 or higher. This caused our models to tend to predict films were between 2.5 - 3.5 despite their features.

So, while traditional model-based approaches have their merits, the findings suggest that a blend of graphical analysis and a broader range of features is required to better understand audience preferences and movie success in the ever-evolving film industry.

Introduction:

This project seeks to address several key questions pertaining to the factors influencing a film's quality and success. Primarily, it aims to investigate how various features such as actors, directors, studios, themes, genres, and duration contribute to predicting a film's quality. Additionally, the project aims to explore whether certain trends, such as the impact of actors' performances or the combination of themes and genres, are discernible in the model's predictions. One specific question of interest is how the viewing public's perception of film runtime affects a film's score.

These questions hold significant importance for filmmakers, producers, and audiences alike. Understanding the underlying factors that contribute to a film's quality and success can inform decision-making processes in the film industry, from casting choices to thematic elements. By uncovering trends and patterns in audience preferences, stakeholders can tailor their productions to better resonate with viewers and maximize their chances of success. Additionally, the ability to predict a film's quality based on its features and identify similar movies can enhance the movie-watching experience for audiences, providing personalized recommendations and insights into popular trends.

Related Work:

The dataset at hand has piqued the interest of various data scientists, some of whom have already begun to conduct interesting analyses to delve into the complex dynamics of film consumption and genre preferences in different countries. One notable effort is a project called "What the World Is Watching." In this work, data scientists chose to explore the landscape of a movie's popularity through a geographic lens, using datasets to characterize films based on the country in which they were made, and juxtaposing them with data on the countries they were watched. The project's methodology revolves around mapping the popularity of genres in different geographic regions, gaining insight into the different preferences of global audiences. By categorizing films according to the country in which they were produced and analyzing their reception in different viewing countries, the project aims to uncover subtle patterns and trends in genre prevalence on a global scale. Through sophisticated data visualization techniques and geographic maps, the project uncovers the cultural and regional factors that influence viewers' preferences and viewing habits. It's worth noting, however, that while this analysis provides valuable insights into the broader genre pop landscape, it doesn't delve into the underlying complexities within individual films. Unlike the current project, which aims to reveal the multifaceted determinants of film quality and success, the "What the World is Watching" project focuses primarily on macro trends and does not explore factors such as the impact of actors or thematic elements on audience reception.

Nonetheless, the project strives to place cinema's popularity within a global framework, providing a valuable perspective on the interplay of cultural influences, geographical factors, and audience preferences in shaping cinematic landscapes. Going forward, the insights gleaned from these analyses can provide a more complete picture of the multifaceted dynamics of film consumption and audience engagement across the globe.

(https://www.kaggle.com/code/suntzunami/what-the-world-be-watchin).

Data set:

The dataset for this project is titled, "Letterboxd Movie Data", curated by user gsimonx37. This dataset provides a comprehensive collection of information about films cataloged on the popular film review platform Letterboxd. The dataset covers a wide range of movie titles, including data points such as movie ratings, themes, genres, production details, and more. The dataset is divided into CSV files, each containing unique features and movie titles. In addition, the dataset can be structured to accommodate different areas of analysis, with separate files focusing on aspects such as geographic production locations, production crews, studios, genres, themes, and settings.

The dataset under consideration provides a comprehensive library of film-related information, with rich data on more than 900,000 films. Notably, the dataset's extensive content includes movie posters for each movie, with a total file size of 24GB. However, given our specific analytical focus of centering on features such as movie ratings, themes, and genres, we will abandon the use of these movie posters. The datasets are carefully organized into CSV files, and we focus on three key files: the movie, the theme, and the genre.

These CSV files contain a wealth of relevant details for each movie entry, from basic attributes such as movie titles and ratings to more complex aspects such as actors, directors, producers, release dates, languages, and country of origin. As a result, the dataset provides a comprehensive snapshot of the global cinema landscape, encompassing films from all corners of the world. Each CSV file in the dataset serves a different purpose to meet a specific aspect of film analysis. For example, one file might contain geographic information related to production locations, production crews, and studios, while another file might provide insights into genre classifications, theme elements, and settings. This division allows for a focused exploration of the different dimensions of the film industry, allowing researchers to delve into specific aspects of interest. However, it's important to note that while the dataset provides rich movie-related data, accessing the overall view requires merging multiple CSV files.

This integration process allows researchers to unlock the full potential of the dataset, leveraging its multifaceted information to fully understand different aspects of the film landscape.

(https://www.kaggle.com/datasets/gsimonx37/letterboxd/data)

Main Techniques Applied:

Before starting the data analysis, the dataset was divided into several CSV files, each containing relevant information related to the film under consideration. According to the most convenient and efficient method, a meticulous approach was taken to consolidate these disparate files into a unified and accessible environment, both through database integration or CSV consolidation. The use of the movie IDs provided in each file facilitates a seamless merging process, ensuring the integrity and integrity of the dataset. At this merge stage, any film that lacks the necessary entries is wisely removed from consideration, preemptively mitigating potential issues arising from incomplete or erroneous data entries. Once the dataset has been consolidated into a cohesive source, work is done on segmenting the data into smaller, more manageable subsections. There was a preliminary stratification according to the country of production, which made it possible to classify the films into a single target geographical cluster. This was done by systematically removing films whose country of production did not include “USA” as there is little to be learned from foreign films when assessing domestic film trends. Subsequent segments in these groupings further depict films based on genre, theme, producer, studio, and other relevant attributes, facilitating granular analysis of the dataset.

After segmenting the data, attention turns to exploring the correlations between various characteristics and identifying the most significant predictors of movie quality. A rigorous analysis was performed to determine the strength and importance of these correlations, with a focus on selecting the features with the highest relevance for subsequent modeling efforts. Through meticulous feature selection and optimization, the accuracy and efficiency of the subsequent prediction model are maximized, ensuring the robust and reliable performance of movie score prediction. After identifying the appropriate features, the model was constructed using advanced statistical and machine learning techniques. Both K-Nearest Neighbors and a Random Forest model were constructed to determine feature importance and make film predictions. The score of the film was used as the target variable. I Adopted a comprehensive training protocol to train and validate the models, leveraging curated validation datasets to evaluate your model's accuracy and performance. The principal measures of the model's performance were MSE and R-Squared MSE, as these were the most reliable metrics for continuous distributive models. Hyperparameter tuning further fine-tuned the model, optimizing its predictive power and improving overall accuracy. The resulting model, fine-tuned and optimized through iterative refinement, was expected to achieve its predictive capabilities, providing valuable insights into the expected performance of the film. In addition, a comprehensive data storage mechanism was implemented to preserve the model output and facilitate future analysis and visualization efforts. With these optimized models, it is possible to make informed predictions about the future success of films, providing valuable insights to stakeholders and informing the film industry's strategic decision-making process.

Complementing model analysis, graphical visual analysis was a key tool used for a broader understanding of the intrinsic trends in a data set. Through visual exploration, we gain insight into the subtle relationships and patterns embedded in the data, revealing the themes and genres that are most closely associated with successful films. In addition, this visual analysis reveals key insights into actors who often appear in "good" films, providing valuable guidance for casting decisions and talent acquisition strategies in the film industry. By harnessing the power of graphical visualization techniques such as scatter plots, bar charts, and heat maps, a holistic view of data set dynamics is achieved, enabling stakeholders to make informed decisions and strategic interventions based on empirical evidence and actionable insights.

Our primary forms of validations are MSE and RMSE, as well as cross-validation. The MSE and RMSE are great measurements for determining the accuracy of predicted models. By subtracting the difference between what we expect vs our actual data we can get a numerical measure of our inaccuracy and try to minimize that measure. The RMSE provides a normalized version of the measurements as RMSE instead measures the standard deviation difference between the expected and actual. By training the data on a partitioned version of the dataset we can utilize cross validation to analyze our predictive performance. CV (cross validation) is sometimes a better evaluator as it can prove how good a model is at predicting new data. As that is one of the intended uses of the model it is valuable to utilize CV in our evaluation. Based on the efficacy of both of these methods, by implementing them both we can ensure our model is robust and accurate.

The primary tool used is Python, as it is the driving programming language behind the project, it is also very versatile and has great data analyzing and modeling capabilities. We will also be utilizing python libraries such as Pandas which is a very prominent data manipulation and analysis tool in Python; as well as NumPy for our mathematical calculations. Sciki-Learn includes many of the statistical models that we will be using so it is an essential tool for this project, as without it we don't have a model. We will also be using a variety of plotting tools such as MatplotLib, Seaborn, and PyPlot. Each excel in certain aspects and fall short in others so they will be used when necessary. SQL databases could be used depending on how the data processing step goes and what is required. The whole project is hosted using Git and we will be using it for project tracking and branch management. Finally we will be using Jupyter notebooks as a presentation tool to store a lot of the visualizations and modeling. By utilizing all these tools I efficiently performed all the tasks laid out and created a predictive model that is highly efficient and accurate. However with these tools, we can also allow for reproducibility, scalability and all around success.

Key Results:

In the field of predictive modeling for film score prediction, a key revelation has emerged about the indispensability of comprehensive feature selection. The effectiveness of such a model depends on the breadth and depth of features included, emphasizing the need to curate the feature set wisely. Notably, empirical evidence highlights the critical role of visual analysis as a powerful tool for revealing nuanced patterns and identifying underlying trends in film datasets. This shift from the traditional reliance on quantitative metrics to a more intuitive analytical paradigm reflects a paradigm shift in contemporary predictive modeling approaches. A notable observation gleaned from the analysis relates to the intrinsic skewness of the distribution of film scores. Despite concerted efforts to calibrate predictive models to produce results that reflect real-world movie values, a recurring trend persists that the prognosis tends to be in a narrow range, typically between 2.5 and 3.5 on the scoring continuum. This phenomenon is rooted in the complexity of the dynamics of data distribution, highlighting the complexity inherent in modeling the multifaceted landscapes of film evaluation.

Further review sheds light on the salienence of genre dynamics in shaping film reception and evaluation. There is a clear difference in scoring patterns, with genres such as drama and documentaries performing significantly better than horror productions. This stark difference highlights the subtle interplay between thematic content, narrative structure, and audience acceptance, shedding light on the different effects of genre categorization on film evaluation. It is speculated that the inherent propensity of genres such as drama and documentary to improve the quality of production has led to a consequent increase in critical acclaim, thus emphasizing the symbiotic relationship between craftsmanship and critical acceptance in the film environment.

One interesting avenue to explore is the potential impact of film budget allocations on the dynamics of film scores. Unfortunately, the lack of relevant budget data in the dataset does not allow for a comprehensive account of this key determining factor. Nonetheless, it is hypothesized that a comprehensive analysis combined with budgetary considerations will yield valuable insights into the subtle interplay between fiscal investment, production value, and film acceptance. Therefore, future efforts aimed at enhancing the comprehensiveness of the dataset to cover budget metrics will necessitate unraveling unknown dynamic contours of film evaluation.

In summary, the empirical insights gleaned from this survey highlight the multifaceted nuances of predictive modeling in the field of film evaluation. From the necessity of a wide selection of features to the saliency of visual analysis and the different influences of genre dynamics, a nuanced understanding of the complexities that shape cinematic reception emerges. In addition, gaps related to budgetary considerations highlight the urgency of improving the comprehensiveness of the dataset to unravel the underlying determinants that control the dynamics of movie ratings. This convergence of empirical findings and methodological requirements creates fertile ground for future inquiry, at the intersection of predictive modeling, film evaluation, and data-driven inquiry.

Applications:

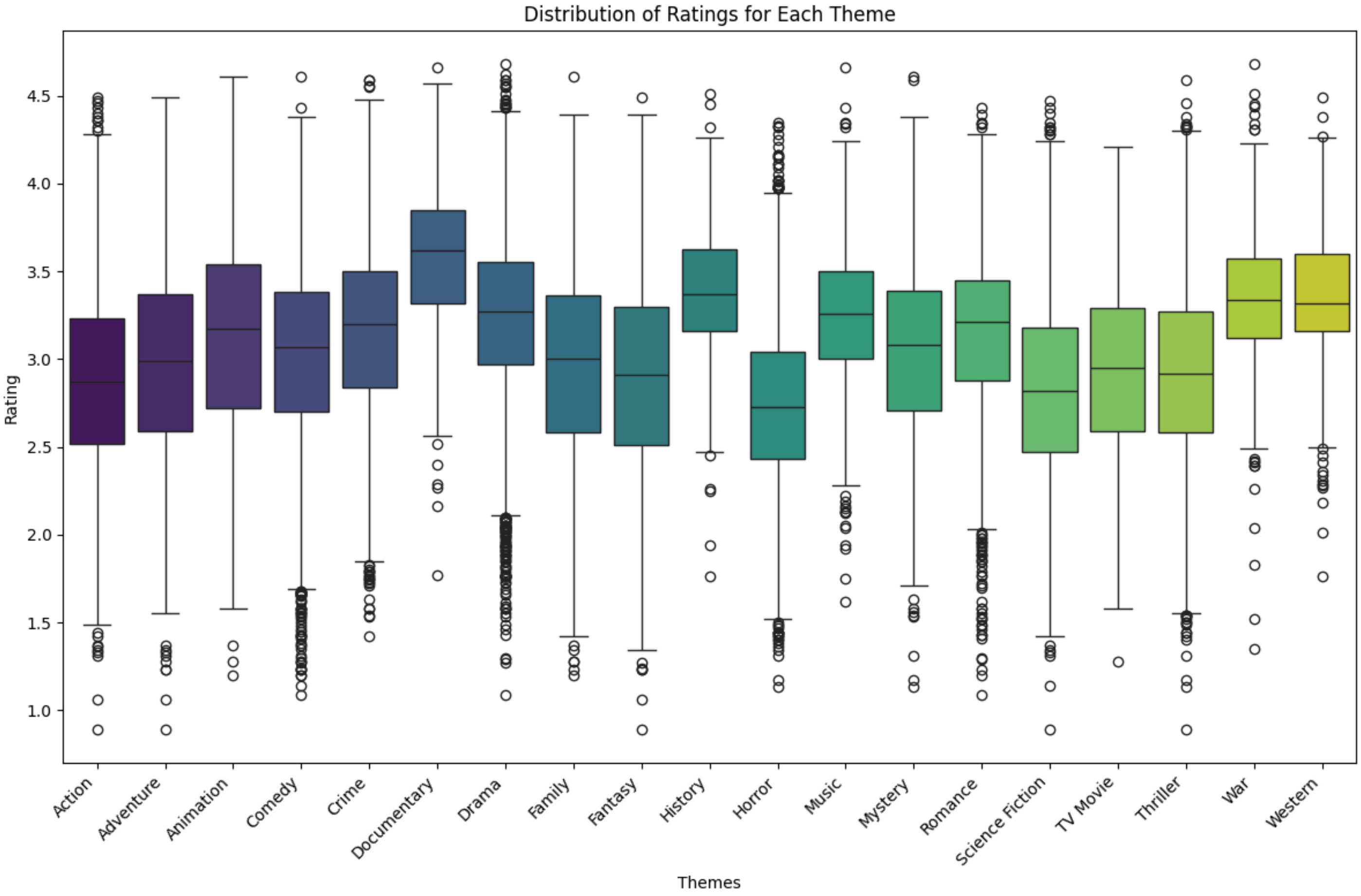
Potential applications of the results garnered from this project include:

Production companies and investors can leverage predictive models for integrated functional selection to strategically allocate resources and investments. By identifying types that are more inclined to receive favorable reviews, stakeholders can optimize production processes and portfolios to maximize returns and reduce risk.

Marketing agencies and distributors can leverage predictive modeling insights to tailor targeted advertising campaigns and promotional strategies. By discerning audience preferences and genre affinities, marketing efforts can be tailored to resonate with specific demographics, resulting in improved audience engagement and box office performance.

Content streaming platforms and digital publishers can adopt predictive models to optimize content management and acquisition strategies. By predicting the potential acceptance of different genres, platforms can curate catalogs that cater to the tastes of different audiences while mitigating the risk of underperforming titles

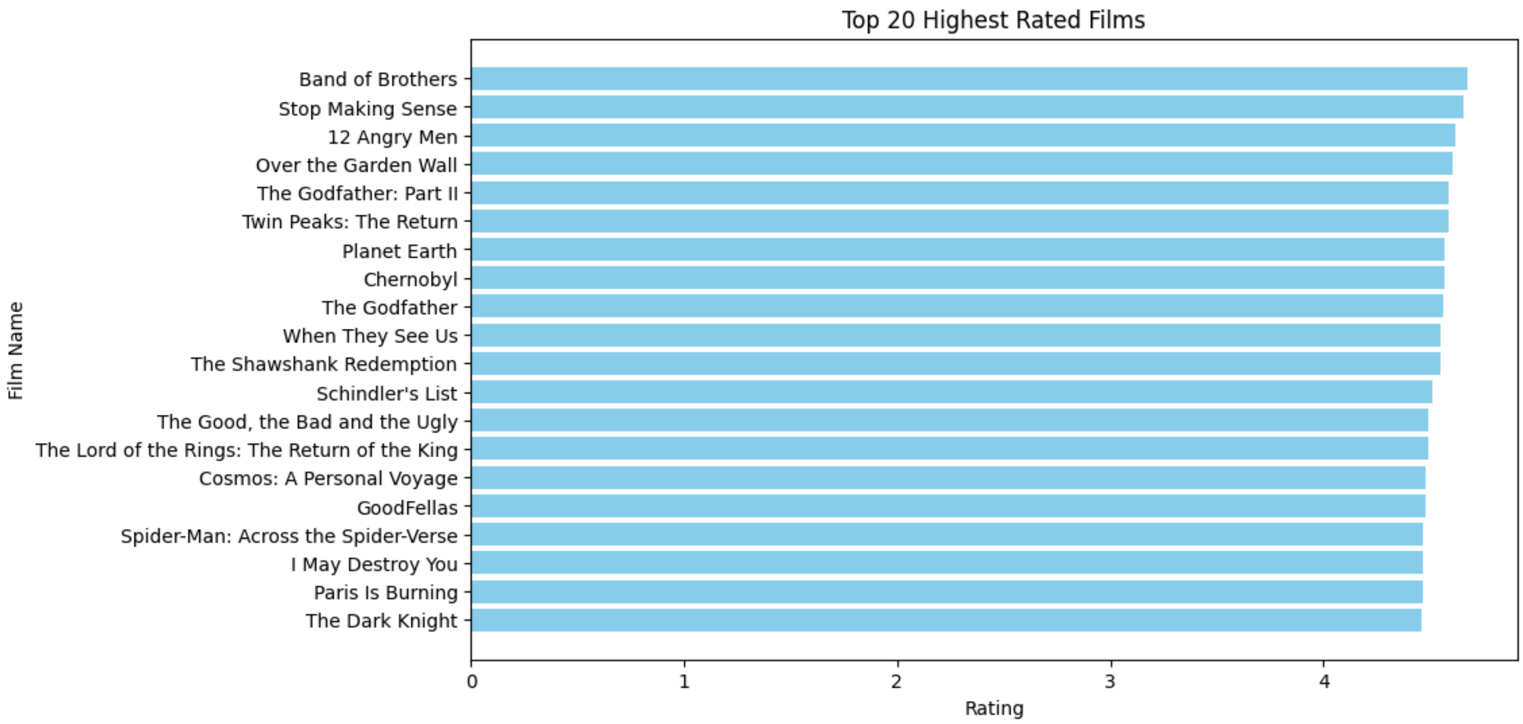
Visualization:

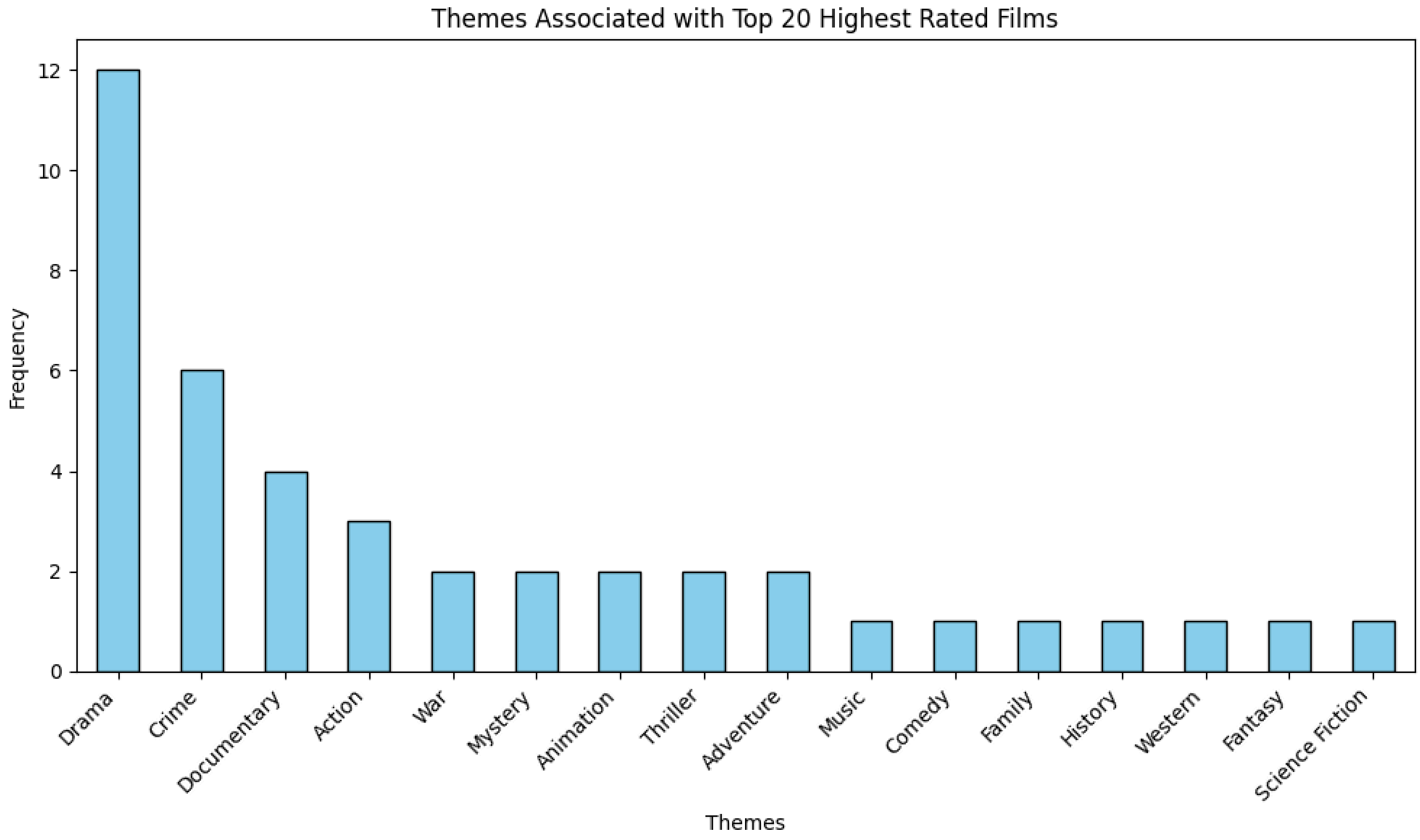


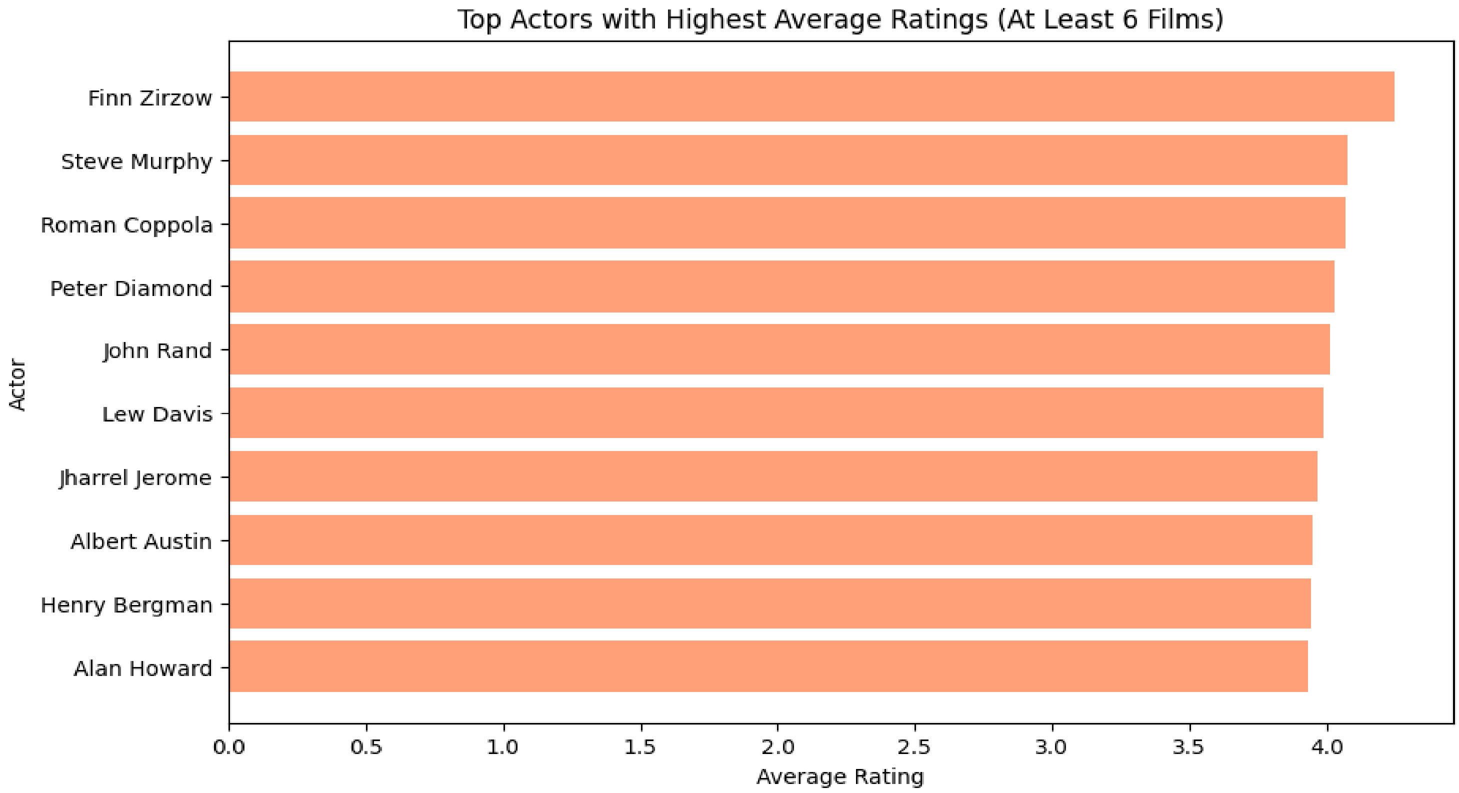
From the distribution of themes we can see that particular themes tend to outperform others. Notably documentary films on average perform the highest, However the highest rated films are not necessarily documentaries. The Drama genre contains the highest rated films and contains a significant amount as shown by the outliers. If we look at the distribution of top films and their associated themes this relationship becomes even clearer.

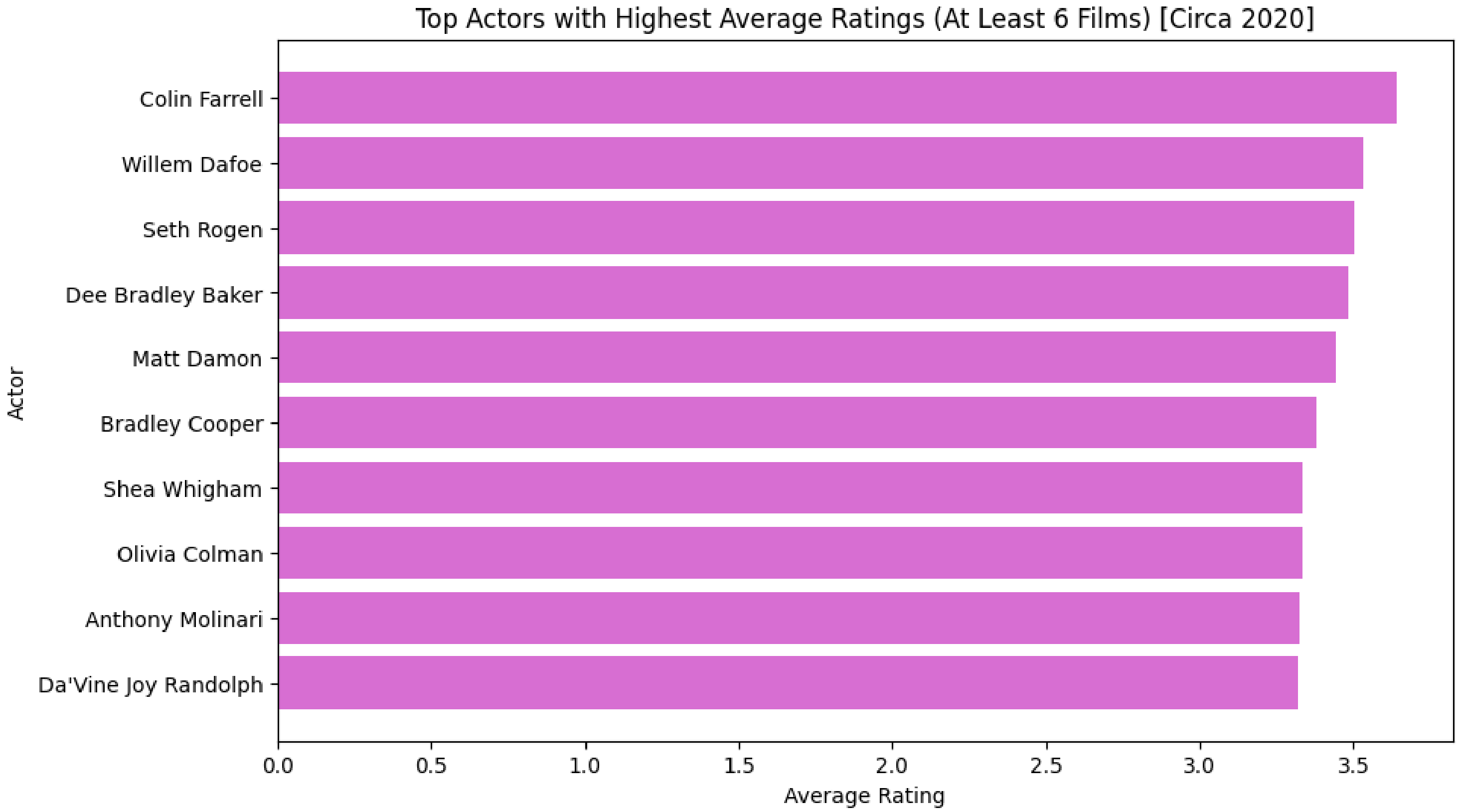
The feature importance of release date and runtime are also important however contribute less to the overall model predictions and are allocated as:

```
Feature: date, Importance: 0.29043848409355005
Feature: minute, Importance: 0.2778505425441498
```


Top 20 Highest Rated Films
Film Name
Band of Brothers
Stop Making Sense
12 Angry Men
Over the Garden Wall
The Godfather: Part II
Twin Peaks: The Return
Planet Earth
Chernobyl
The Godfather
When They See Us
The Shawshank Redemption
Schindler's List
The Good, the Bad and the Ugly
The Lord of the Rings: The Return of the King
Cosmos: A Personal Voyage
GoodFellas
Spider-Man: Across the Spider-Verse
I May Destroy You
Paris Is Burning
The Dark Knight
0
1
2
3
4
Rating


Themes Associated with Top 20 Highest Rated Films
Frequency
0
2
4
6
8
10
12
Drama
Crime
Documentary
Action
War
Mystery
Animation
Thriller
Adventure
Music
Comedy
Family
History
Western
Fantasy
Science Fiction
Themes


Top Actors with Highest Average Ratings (At Least 6 Films)
Actor
Finn Zirzow
Steve Murphy
Roman Coppola
Peter Diamond
John Rand
Lew Davis
Jharrel Jerome
Albert Austin
Henry Bergman
Alan Howard
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Average Rating


Top Actors with Highest Average Ratings (At Least 6 Films) [Circa 2020]
Actor
Colin Farrell
Willem Dafoe
Seth Rogen
Dee Bradley Baker
Matt Damon
Bradley Cooper
Shea Whigham
Olivia Colman
Anthony Molinari
Da'Vine Joy Randolph
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Average Rating