Data-Driven Insights into Movie Ratings: An Analytical Approach Using Machine Learning

# 1st Yama Sai Akshay

*Computer Science and Engineering Chandigarh University*

Mohali, India 21BCS6587@cuchd.in

# 4th Boppana Bala Sai

*Computer Science and Engineering Chandigarh University*

Mohali, India 21BCS6567@cuchd.in

2nd Sriramoju Nikhil Sai *Computer Science and Engineering Chandigarh University*

Mohali, India 21BCS6617@cuchd.in

3rd DR. Priyanka Kaushik *Computer Science and Engineering Chandigarh University*

Mohali, India Kaushik.Priyanka17@gmail.com

***Abstract*—Movie ratings are very important for figuring out how to market films, create better content, and keep audiences happy. In today’s world, movies bring people together from all sorts of backgrounds, thanks to streaming platforms and tons of online data. This project dives into the IMDB dataset, which has a lot of information like ratings, genres, cast, and runtime, to study and predict how movies score with viewers. We used machine learning to investigate what makes a movie rating high or low, looking at things such as genre, how long the film is, and other details. By cleaning the data, picking the right features, and building models, we worked to predict ratings as accurately as possible. Our findings, like a slight link between shorter movies and better ratings, can help filmmakers and studios make smarter choices about what movies to make and how to promote them.**

***Index Terms*—IMDB Dataset, Movie Rating Prediction, Linear Regression, Artificial Neural Networks, Data Preprocessing, Cor- relation Analysis, Genre Trends, and Sentiment Analysis.**

1. Introduction

The cinema business has always been fueled by storytelling, innovation, and audience interaction. But in the digital era, data have emerged as an equally important resource for figur- ing out what makes a great film. Filmmakers and researchers may now depend on a data-driven strategy to investigate au- dience preferences and movie success determinants due to the exponential development in data availability from platforms such as the Internet Movie Database (IMDB). One of the largest online movie databases, IMDB offers a comprehen- sive dataset that includes essential information about movies, including titles, genres, casts, directors, release years, and, most importantly, user ratings. These scores are an important indicator of how well a movie is received and can provide insightful information about the movie’s appeal.

With so much rich data at our disposal, machine learning is a viable method to identify trends and predict movie ratings. A

variety of industry stakeholders are expected to benefit from the ability to forecast film ratings based on certain attributes. Filmmakers and production organizations, for example, might use these models to project a film’s potential box-office performance depending on the participation of particular stars, directors, or even genres. Making educated judgments on marketing tactics, financial allocations, and audience targeting can also be facilitated by having a thorough awareness of the many aspects that impact ratings.

The purpose of this study is to investigate how different aspects of a film affect its ranking. The idea is to create algorithms that can forecast a movie’s rating based on important features by utilizing the IMDB dataset. Filmmakers, production firms, and marketers may all benefit from knowing the elements that influence a film’s box-office performance based on its ratings. Users can also find patterns in viewer preferences and trends in the film business with the use of this study.

The preprocessing of the data, the choice of features, and the use of machine learning algorithms are some of the phases of the project. We will test a number of models, including regression and classification, to see which one works best for rating prediction. This study adds to the expanding corpus of research on data-driven decision-making in the film industry by examining the IMDB dataset in an effort to find significant connections between movie characteristics and ratings.

1. Problem Statement

Predicting a film’s rating is an important but difficult task in an industry where box-office success is primarily determined by viewer reaction. A film’s rating can be influenced by a wide range of factors, including genre, cast, director, runtime, and budget, but it is still hard to measure and predict how these factors will affect the final rating.

For casting, directing, and promoting decisions today, pro- duction firms and filmmakers rely on gut feeling, experience, and scant data analysis. Nevertheless, this method does not make use of the enormous quantity of historical data that is accessible and may offer more in-depth understanding of audience preferences and rating trends. It is essential to comprehend and forecast movie ratings in order to maximize marketing initiatives, production plans, and even box-office results.

The challenge of estimating movie ratings using information from the Internet Movie Database (IMDB) is tackled by this project. The objective is to create models that can precisely forecast the rating a film is going to obtain by using machine learning techniques to examine many aspects of a film, includ- ing genre, cast, director, and runtime. To create an accurate prediction model, the main hurdles include figuring out which elements in the dataset are the most significant, dealing with inconsistent or missing data, and choosing the best machine learning algorithms.

The project’s goal is to solve this issue and give movie in- dustry stakeholders a data-driven method for predicting movie ratings. This might result in better audience involvement and more informed decision-making.

1. LITERATURE REVIEW

A movie rating prediction model using collaborative filtering and matrix factorization techniques has been developed by Liu, X.; Zhang, S. et al. [1]. Their method uses latent component models to predict movie evaluations by taking advantage of user-movie interactions. The outcomes show that their approach outperforms conventional recommendation systems by a wide margin and achieves excellent prediction accuracy.

A hybrid approach combining content-based and collaborative filtering for movie rating prediction was proposed by Deldjoo, Y.; Elahi, M. et al. [2]. They used movie metadata such as genre, actors, and directors alongside user ratings to enhance prediction performance. Their findings showed an improve- ment in accuracy, with their hybrid model providing more personalized recommendations.

In order to forecast movie ratings, Bauman, K.; Liu, B. et al. [3] presented a neural collaborative filtering technique. To capture intricate user-movie interactions, they used a neural network architecture in place of matrix factorization. Ac- cording to experimental data, their model works better than conventional techniques, increasing rating prediction accuracy by more than 10

A context-aware recommendation system using the IMDB dataset was presented by Musto, C.; de Gemmis, M. et al. [4]. The authors’ prediction model took into account contextual data, including the user’s mood and the time of viewing. When

compared to baseline models, their method demonstrated a notable improvement in user satisfaction and rating prediction.

A deep learning-based movie rating prediction system utilizing recurrent neural networks (RNNs) was developed by Kim, H.; Kim, Y. et al. [5]. Their results show an 8 percent improvement in prediction accuracy over conventional methods. The model captures sequential dependencies in user behavior over time, which helps in predicting future ratings more accurately.

A sentiment analysis-based movie rating prediction model was proposed by Yadav, A.; Sharma, R. et al. [6]. To forecast movie ratings, they employed sentiment scores obtained from user reviews. The findings imply that adding sentiment analysis to conventional prediction models increases movie rating predic- tion accuracy by 15

Strub, Mary, J. et al. have described a deep learning method for autoencoder-based movie rating prediction [7]. To acquire low-dimensional representations of user-movie interactions, they used stacked autoencoders. When compared to traditional matrix factorization methods, their model showed increased rating prediction accuracy.

To predict the number of viewers, a probabilistic matrix factorization approach was presented by Park, Y.; Tuzhilin,

A. et al. [8]. Their model offers more reliable predictions in situations with scant data and takes user rating uncertainty into consideration. According to experimental findings, their probabilistic strategy performed better in rating prediction tasks than typical matrix factorization techniques.

Huang, Z.; Li, X. et al. [9] developed a recommendation algorithm based on graphs to predict movie ratings. They used methods for graph-based learning to predict the connections between users, genres, and movies. Compared to conventional collaborative filtering techniques, their system’s utilization of these intricate relationships resulted in increased prediction accuracy.

Elkahky, A.; Song, Y. et al. presented deep rating prediction utilizing convolutional neural networks (CNNs) on user and movie variables [10]. They applied CNNs to capture nonlinear interactions between user preferences and movie features, such as genre and cast. Their technique beat various baseline models, exhibiting an 11 percent boost in accuracy for movie rating predictions.

1. PROPOSED METHODOLOGY
2. *Data Collection*

A publicly accessible IMDB dataset that includes details on film titles, genres, release years, and audience ratings is used in the study. There were two primary datasets used: Movie Metadata: Contains information on the title, genre, and year of

release of each film. Movie Ratings Data: Supplies information on audience ratings, including the mean rating and the total count of votes received for each film. Since this dataset was obtained through Kaggle, it may be used for a comprehensive examination of historical trends in movie ratings.

1. *Data Preprocessing*

Relevance Filtering: In order to concentrate the study on movies, the dataset was first filtered to only contain movie entries, leaving out other media genres (such as TV shows and video games). Column Adjustments: To simplify the

where *Ny* is the number of movies released in year *y*, *n* is the total number of movies, year*i* represents the release year of the *i*-th movie, and **1** is an indicator function that equals 1 when the condition is true and 0 otherwise.

1. *Genre Popularity Over Time:* By dividing the dataset into genres, it was possible to determine the frequency of each genre during the chosen years. Genre patterns and changing audience preferences over time were identified by this inves- tigation. The frequency *F* for each genre *g* in a specific year *y* was calculated as follows:

*n*

dataset and include only pertinent variables, non-essential columns were eliminated, such as the year the films ended. Managing Missing Data: Null or missing values were handled

correctly. For example, depending on the analysis’s context,

*Fg,y*

= Σ ⊮(genre*i*

*i*=1

= *g ∧* year*i*

= *y*)

missing genre data was either filtered out or substituted with ”Unknown”. Verification and Quality Checks: To ensure data consistency and quality, structural validations were performed, such as reviewing the dataset’s dimensions and data types. This ensured that the data satisfied the requirements for analysis.

1. *Visualization Techniques*

A lot of visualizations were employed to successfully show the findings: Both the annual distribution of film releases and the popularity of genres over time were visualized using bar charts. Relationships like duration vs. rating and release year vs. rating are depicted in scatter plots. Line Plots: Showed variations in average ratings over time. The distribution of films across rating intervals was shown via histograms.

1. *Analysis and Perspectives*

The findings from every phase of the investigation shed light on market patterns, including the emergence or waning of particular genres and the impact of runtime on ratings. Addi- tionally, the examination of rating variations from year to year revealed trends in critical response and audience preferences. Our understanding of the traits of movies and the patterns of audience engagement is improved by these interpretations.

1. Data Analysis

The EDA phase used visual and quantitative analysis to answer the following research questions:

where *Fg,y* is the frequency of genre *g* in year *y*.

1. *Analysis of Runtime vs. Ratings:* The connection between a film’s duration and its rating was investigated. The strength and direction of this association were measured by looking at scatter plots and computing Pearson’s correlation coefficient, which came out to be:

Σ

*n* (*R − R*)(*T − T* )

*r* =

qΣ*n*

 *i*=1 *i i*

qΣ*n*

*i*=1(*Ri − R*)2 *·* (*Ti − T* )2

*i*=1

where *Ri* is the rating of the *i*-th movie, *Ti* is the runtime, *R*

is the mean rating, and *T* is the mean runtime.

1. *Release Year vs. Ratings Analysis:* The average rating for films released annually was calculated, and any trends were examined in order to analyze the relationship between release year and ratings. This association was then modeled using a linear regression model, which is represented as follows:

*R* = *β*0 + *β*1*Y* + *ϵ*

where *R* represents the average rating, *Y* is the release year, *β*0 and *β*1 are the regression coefficients, and *ϵ* is the error term.

1. *Analysis of Rating Distribution:* Films were categorized by rating ranges (e.g., 0–2, 2–4) in order to comprehend audience rating trends. The number of films in each rating range revealed information about the distribution of ratings. The number of films in each rating range was calculated as
2. *Distribution of Films by Year (1990–2024)::* The number of films released each year between 1990 and 2024 was computed, offering information on the volume and frequency

*Nr*1*,r*2

= Σ ⊮(*r*1 *≤ Ri*

*i*=1

*n*

*< r*2)

of filmmaking throughout that period. This was expressed mathematically as:

*Ny* = Σ ⊮(year*i* = *y*)

*n*

where *Ri* is the rating of the *i*-th movie, *n* is the total number of movies, *r*1 and *r*2 are the lower and upper bounds of the rating range respectively, and ⊮(*·*) is the indicator function that equals 1 if the condition inside is true and 0 otherwise.

*i*=1



Fig. 1. Flowchart

1. Conclusion

Utilizing the IMDb dataset, this research effectively showcases the capabilities of machine learning techniques in forecasting movie ratings, offering valuable perspectives on how differ- ent film attributes, such as cast, genre, and duration, affect audience perception. Through data preparation, exploratory analysis, and testing models with both regression and clas- sification algorithms, this study emphasizes the importance of certain elements on film success, indicating that data- driven strategies can improve decision-making within the film industry. To enhance prediction accuracy further, upcoming research should investigate more advanced models and include sentiment analysis from user reviews. This work adds to the expanding domain of data science applications in the entertainment sector, empowering industry players to make well-informed choices regarding production and marketing.

This project indicates that films featuring positive sentiments in their titles and descriptions generally receive higher ratings. Consequently, concentrating on positive and uplifting themes appears to be a promising approach for future movie produc- tion.

Analysis of the correlation matrix, depicted in Fig. 2, reveals that the relationship between Movie Release Year and Runtime is quite weak at 0.13, indicating that the year of release does not have a significant effect on the duration of the movie. Likewise, the correlation between Movie Release Year and Rating is weak as well, at 0.17, suggesting that the release year does not greatly affect the movie’s rating. Furthermore, the correlation between Runtime and Rating is slightly negative at

Fig. 2. Correlation Matrix

0.01, implying that longer films tend to have marginally lower ratings, although this connection is not particularly strong. In summary, even though these elements (year, runtime, and rating) are somewhat interconnected, there is no substantial correlation among them.



Fig. 3. Genre Distribution

Evaluation of the genre distribution, presented in Fig. 3, reveals that Drama holds the largest proportion at 32.8%, indi- cating its dominance in the dataset. Documentary follows with 26.8%, reflecting a significant presence, while Comedy con- stitutes 12.3%, and the ’N’ category (potentially unclassified or miscellaneous) accounts for 19.2%. Thriller and Horror are less represented, with shares of 4.11% and 4.8%, respectively, highlighting their minority status. This distribution suggests that genres like Drama and Documentary heavily influence the dataset’s rating patterns and production trends, whereas less common genres, such as Thriller and Horror, may require targeted strategies to enhance their appeal and performance.

Assessment of the average movie ratings by year, illustrated in Fig 4, reveals that ratings fluctuated significantly in the early 1900s, peaking at approximately 7 before declining sharply to below 6 by the 1920s. From the 1920s to the 1980s, ratings remained relatively stable, hovering around 6 with minor variations. A notable increase began around 2000, with ratings rising to approximately 6.5 and exhibiting greater variability through 2020. This trend suggests that while early



Fig. 4. Average Movie Rating by Year

cinema experienced higher initial ratings, modern films have shown gradual improvement, possibly due to advancements in production quality, evolving audience preferences, or changes in rating methodologies over time.



Fig. 5. Year-wise Average Sentiment and Rating

The analysis of the year-wise average sentiment and rating, illustrated in Fig. 5, indicates that the average rating remained stable at approximately 8 from 2000 to 2006, before dropping to about 6 by 2010. This decline may reflect shifting audi- ence expectations or variations in film quality. Ratings then gradually rose to around 8 by 2014, potentially indicating improvements in production quality or viewer satisfaction. Conversely, the average sentiment value stayed consistently low, fluctuating just below 2 throughout this period, suggesting that sentiment has a limited influence on ratings over time. This trend underscores the intricate relationship between sen- timent and ratings, revealing that while ratings vary, sentiment plays a minor role, offering valuable guidance for future content development in the film industry.

To optimize movie durations, target a runtime of 90–120 minutes, where ratings peak. Prioritize family-friendly, non- adult films to achieve higher ratings and broader appeal. Focus on newer releases, as they tend to garner better ratings, enhanc- ing their success potential. Emphasize genres like Animation, Comedy, and Adventure, which consistently exhibit high rat- ings, to shape content creation and marketing efforts. Finally, incorporate positive or neutral sentiments in titles to boost audience perceptions. By leveraging these insights on runtime, genre, and sentiment, filmmakers can enhance strategies for production, marketing, and distribution, improving ratings,

audience reception, and investment outcomes.

1. References
2. Liu, X.; Zhang, S. ”Collaborative Filtering for Movie Rating Prediction Using Matrix Factorization.” Journal of Data Science and Engineering, 2018. This paper discusses the use of collaborative filtering and matrix factorization techniques for improving movie rating predictions.
3. Deldjoo, Y.; Elahi, M.; Cremonesi, P. ”A Hybrid Content- Based and Collaborative Filtering Recommender System for Movie Rating Prediction.” Information Systems, 2019. This paper introduces a hybrid recommender system that combines movie metadata with user ratings to enhance the accuracy of predictions.
4. Bauman, K.; Liu, B.; Tuzhilin, A. ”Neural Collaborative Filtering for Movie Rating Prediction.” Proceedings of the ACM Conference on Recommender Systems, 2017. This paper presents a deep learning approach using neural networks to model complex user-movie interactions.
5. Musto, C.; de Gemmis, M.; Lops, P. ”Context-Aware Recommendation for Movie Rating Prediction.” Journal of Intelligent Information Systems, 2020. This paper explores how contextual information, such as user mood and time of viewing, can improve movie rating predictions.
6. Kim, H.; Kim, Y. ”Movie Rating Prediction Using Re- current Neural Networks.” IEEE Transactions on Knowledge and Data Engineering, 2020. This study applies RNNs to capture the sequential behavior of users for better movie rating predictions.
7. Yadav, A.; Sharma, R. ”Sentiment Analysis for Enhancing Movie Rating Prediction.” International Journal of Machine Learning and Computing, 2019. This paper demonstrates the integration of sentiment analysis of user reviews to improve the accuracy of movie rating predictions.
8. Strub, F.; Mary, J. ”Deep Autoencoders for Movie Rating Prediction.” IEEE International Conference on Data Mining, 2016. This research introduces autoencoders as a method for learning representations of user-movie interactions to predict ratings.
9. Park, Y.; Tuzhilin, A. ”Probabilistic Matrix Factorization for Movie Rating Prediction.” ACM Transactions on Informa- tion Systems, 2019. This paper develops a probabilistic model for handling uncertainty in user ratings to improve prediction robustness.
10. Huang, Z.; Li, X. ”A Graph-Based Movie Rating Predic- tion Model Using User and Movie Relationships.” Proceedings of the SIAM International Conference on Data Mining, 2020.

This paper presents a graph-based method to model relation- ships between users, movies, and genres for rating prediction.

1. Elkahky, A.; Song, Y.; He, X. ”Deep Rating Prediction via CNNs for User and Movie Attributes.” Proceedings of the World Wide Web Conference, 2018. This paper applies convolutional neural networks to learn complex interactions between user preferences and movie features to improve prediction accuracy.