International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predicting the Success of A Movie Using Machine Learning Algorithms: An Analysis

Ahaan Anand

Modern School Vasant Vihar

Abstract:
Machine learning has been an integral part of reshaping the movie industry, and this paper delves into its transformative role. We examine the utilization of machine learning models, including Linear Regression, Decision Trees, and Random Forests, in analyzing data from IMDB's top-rated movies. The results of our analysis demonstrate that Random Forests achieve a remarkable 74% accuracy rate in predicting IMDB ratings. This paper serves as a comprehensive exploration of machine learning's profound impact on the movie industry, emphasizing its transformative role in content creation, production, and audience engagement. It underscores the enormous potential of machine learning in revolutionizing the movie industry, while also paving the way for the integration of Natural Language Processing (NLP) to enhance movie success predictions, propelling innovation within AI-driven cinematic endeavors. In conclusion, this research highlights the pivotal role of machine learning in the film industry, underscoring its ability to elevate the quality of content and engage audiences more effectively, while anticipating a future where technology and creativity harmonize to create unique and captivating cinematic experiences.

Keyword: Machine Learning, Movies, Predictions, Data-Driven Decision Making

1. Introduction

Machine Learning (ML) is an application of artificial intelligence which can learn, predict and also improve on its own responses. As the name suggests, ML is the area of study showing that machines learn just like humans. With the help of a powerful set of methods designed to identify intricate patterns from vast datasets and utilize these patterns to predict outcomes or make decisions when faced with new data, we arrive at the approach known as Machine Learning (ML) or Statistical Learning, allows computational systems to "learn" these underlying structures from the provided data and apply that knowledge to generalize and adapt to new data instances. Essentially, it empowers the system to grasp the essence of the data and make informed judgments based on that understanding, even when encountering novel and unfamiliar information.

A lot like a human baby, a machine slowly learns to process information leading it to create patterns within the data it has been fed. Any amount of statistical knowledge can be fed to a machine. In fact, the more data it is fed, the more accurate it becomes. Slowly, learning from the data, a machine can learn to differentiate between two or more things. For example, after feeding a machine 10000 pictures of cats and another 10000 pictures of dogs, the machine itself starts finding ways to differentiate between the two. After creating different patterns between the two, the computer will be able to segregate or classify these animals into categories of cats or dogs. Identifying the accuracy of the system, the statistical approach in

which the machine is trained may be modified to give better results. A machine learning algorithm can even be taught how to play a game of chess. It can even learn what to recommend to you for your watchlist on netflix.
The increasing presence of AI in the media and creative sectors is unmistakably evident. Throughout history, creative professionals have continuously sought novel tools to enhance their work, and AI has emerged as a notable addition to their repertoire. This technology aligns seamlessly with the unique demands of creative industries and is actively reshaping established paradigms.

The term "consumption" pertains to the interaction that end-users have with media content. This encompasses experiences like receiving music recommendations on audio streaming platforms or accessing news feeds via smartphone applications. On the other hand, "digital storytelling" entails the creation of videos that artfully combine audio, images, and video clips to narrate a compelling tale. The process of converting a creative narrative into a usable digital format requires a multidisciplinary skill set and an array of specialized tools, often involving electronic components, sensors, actuators, and software.

As we focus on examining the challenges encountered in the creation, consumption, and production of movies, several industries confront the daunting task of generating large, highly intricate numerical models of shapes and environments. This encompasses both the geometric aspects of objects and their surface appearances, including roughness, texture, and color. In the movie industry, similar challenges arise when designing virtual sets for actors, a scenario prevalent in both animated films and motion pictures incorporating pervasive special effects. The successful creation of such environments necessitates precise adherence to requirements concerning aesthetics, navigability, and plausibility, ultimately aiming for immersion.

To facilitate the work of movie and video game designers, the field of Computer Graphics has evolved a substantial body of research in content synthesis. These methods strive to automate certain facets of the content creation process, providing invaluable assistance to designers. The scope of automation includes tasks like automatically filling regions with textures or objects, generating detailed landscapes, flora, and urban landscapes, as well as creating building floor prints and environment layouts. AI methods play a crucial role in simplifying content creation, particularly in handling ill-posed problems and complex optimization objectives that may exhibit conflicting criteria. Rather than seeking a single, optimal solution, the focus lies in producing a diverse array of choices from which designers and producers can select.

Collaborating with users, these algorithms generate a broad spectrum of valid solutions (in a technical sense) while users explore and contribute aesthetic preferences. For the video game industry, algorithms that produce playable levels on-demand enhance replayability and substantially reduce content creation time, storage, and network bandwidth costs. Nevertheless, aesthetic aspects inherently remain subjective and are subject to the director's artistic taste.

Within the domain of aesthetics, AI machine learning techniques prove especially well-suited. By-example content synthesis is one notable methodology gaining prominence, enabling the generation of new content that emulates input data. This process involves matching various features such as colors, sizes, curvatures, and geometric properties, thereby imbuing the produced content with the aesthetics of

the original input. The pursuit of advancements in AI, including generative adversarial networks, is an active area of research aimed at enriching this particular domain.

The importance of my paper on machine learning in the movie industry stems from its focus on the extensive utilization of this technology and its profound impact on the field. The movie industry holds a prominent position within the media and entertainment industry, exerting significant influence on societal culture and on people's lives. Embracing state-of-the-art innovations such as machine learning has proven to be transformative in this regard. Filmmakers have been empowered to efficiently design difficult virtual sets, create realistic special effects, and automate content synthesis, thereby elevating the visual experience for movie audiences. Moreover, machine learning has revolutionized movie marketing and distribution, with tailored recommendations and personalized advertisements catering to individual preferences. As a result, my paper aims to shed light on the pivotal role of machine learning in propelling the movie industry forward, fostering creativity, and delighting audiences on a global scale.

Machine learning algorithms are broadly categorized into three types:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Reinforcement Machine Learning

Supervised machine learning is a crucial component of modern AI systems, enabling accurate predictions and decision-making by learning from labeled data. It has found diverse applications in natural language processing, image recognition, and medicine. However, challenges include data quality and biases in training sets. Ongoing research focuses on improving data collection and algorithm robustness. Despite these challenges, supervised machine learning continues to drive innovation and has promising implications across industries, making machines more intelligent and shaping our future.

Unsupervised machine learning is an essential technique in artificial intelligence, where algorithms learn from data without any explicit guidance or known outcomes. Unlike supervised learning, it doesn't rely on labeled data for training. Instead, it autonomously discovers patterns and structures within the data, making it useful for tasks like clustering, dimensionality reduction, and anomaly detection. Its versatility finds applications in various fields, offering valuable insights and simplifying data analysis without the need for labeled data. This method fosters innovation and enriches our understanding across diverse domains. For example: if we were to provide the machine learning program with thousands of images of various animal species and then ask

Reinforcement machine learning is an AI approach based on rewards. It involves an agent interacting with an environment, taking actions to achieve a goal. The agent learns from rewards or penalties received for its actions and fine-tunes its policy over time. This method is useful for tasks where the optimal strategy requires experimentation. Reinforcement learning has found applications in games, robotics, autonomous vehicles, and recommendation systems. However, it faces challenges like balancing exploration and exploitation and computational complexity. Still, it holds promise for creating adaptive and intelligent AI systems.

1.1. Machine Learning Models

In this section we discuss three machine learning models used in this paper, namely logistic regression, decision trees and random forest.

Logistic regression is a statistical model used for binary classification tasks. It predicts the probability of an outcome (e.g., yes/no, 1/0) based on one or more predictor variables. It employs the logistic function to constrain the output between 0 and 1, making it suitable for modeling probabilities. The algorithm estimates coefficients for each predictor and these coefficients indicate the strength and direction of the relationships between predictors and the outcome.

Decision trees are widely used in machine learning for classification and regression tasks. They recursively split data based on input features, forming a tree-like structure with internal nodes representing decisions and leaf nodes providing predictions. They are easy to interpret, handle various data types, and can calculate non-linear relationships. However, the lack of integrity to decision trees can be addressed using methods like Random Forest. Decision trees remain valuable due to their transparent decision-making processes.

Random Forest creates multiple decision trees during training and combines their predictions to improve accuracy and reduce overfitting. Each tree is constructed using a random subset of the training data and a random subset of the features. The algorithm then aggregates the predictions from these trees, typically by a majority vote classification. Random Forest is known for its robustness and ability to handle complex data while providing insights into feature importance.

2. Method

The data was sourced from a Kaggle dataset comprising information on IMDB movies. It consists of 15 columns and 1000 rows. The columns include year of release, the director as well as star actors, the metascore as well as IMDB rating, the length of the movie, the genres that the movie may be defined under, the gross value which the movie was able to collect as well as the certificate that the movie has received.

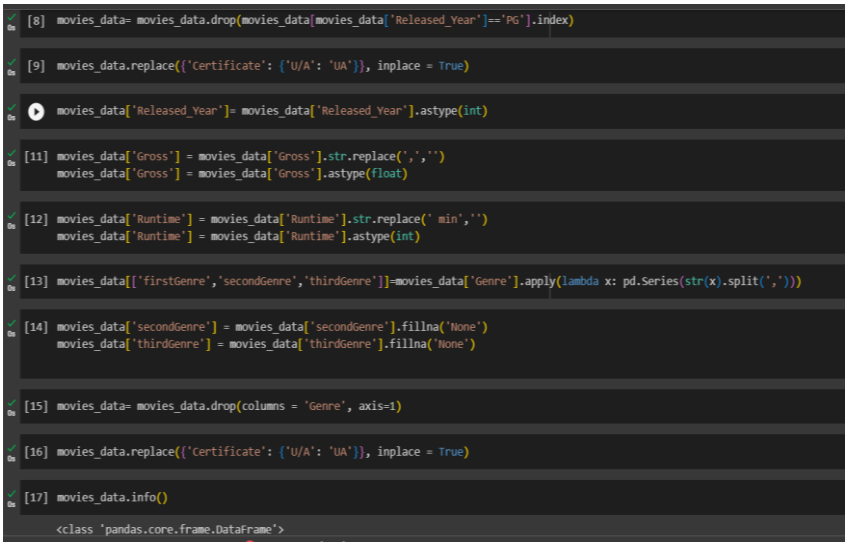

Figure 1. Cleaning of the dataset

The imported datatset includes 1000 best ranked movies and TV shows according to IMDB ratings which are largely based on public opinion of a motion picture. The dataset also included information such as name of the film, link to a poster of the film, and a basic overview of the plot of the movie. Keeping in mind that the aforementioned code must be able to predict the success of a movie, without Natural Language Processing, these columns hold no relevance in this context and have hence been removed. The "Gross" and "number of votes" columns initially had object data type due to the presence of commas between the digits of the number. Similarly, the runtime was taken up as object datatype due to "min" written next to the numbers (since min denotes the unit of the runtime column). We modified these columns to convert them into numerical values with either int or float data type. Figure 1 shows a snapshot of the code lines used to achieve the same.

The genre column was restructured into three distinct columns: firstGenre, secondGenre, and thirdGenre. Consequently, the original genre column, which contained amalgamated genre information, was omitted from the dataset. In instances where movies lacked a second or third genre, we designated these columns as 'None' and exclusively employed the firstGenre column. Furthermore, we undertook a comprehensive procedure to address missing values in both numerical and categorical columns. Missing values in numerical columns were imputed using the mean, while categorical columns like 'certificate' were imputed using the mode, ensuring the dataset's completeness and reliability for subsequent analysis. Columns such as 'Director','star1','star2','star3','star4','secondGenre', and 'thirdGenre' were not used because of the vast number of variables in these columns, which would make the data sparse after encoding.

3. Results and Discussion

The columns of the dataset were analyzed to understand the dataset better. This was done via Exploratory Data Analysis (EDA). As shown in Figure 2, we understand that the dataset covers a wide range of movies and TV series from the 1920s to 2020. Similarly, in figure 4, we observe that the runtime varies from 60 to 250 minutes, the maximum movies being approximately 100 minutes long.

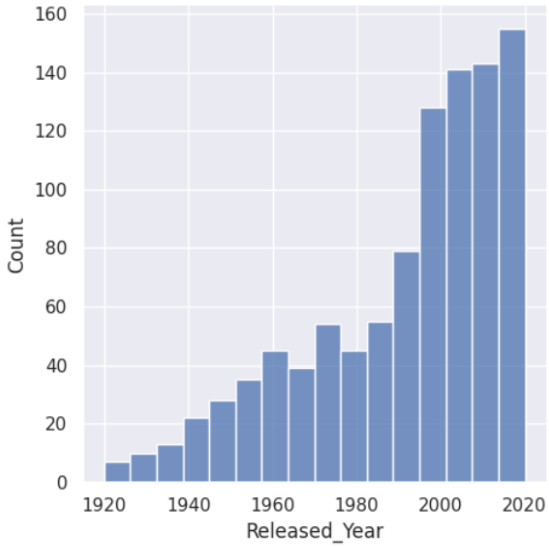

Figure 2. Number of movies released in each of the years between 1900s to 2020s.

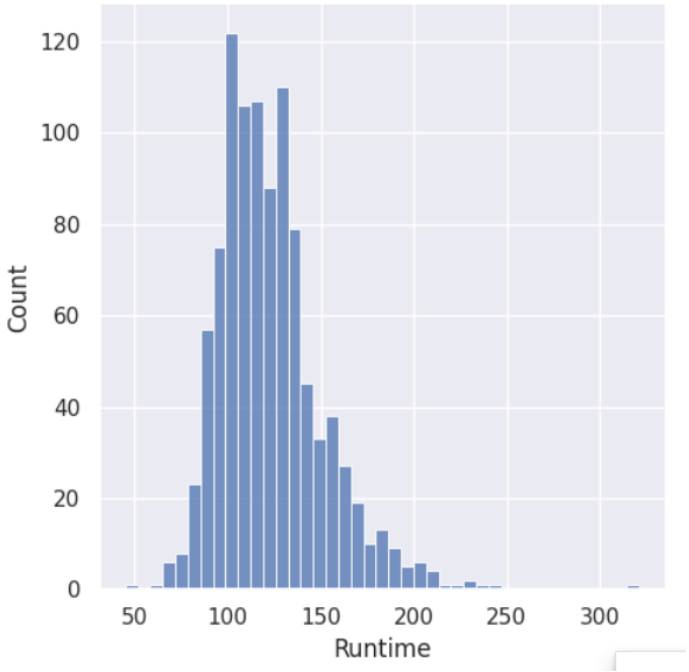

Figure 3. Number of movies displayed along with their respective runtimes.

In our analysis, we initially faced a perplexing situation where machine learning models, such as Linear Regression and XGBoost Regressor, yielded exceptionally high training accuracy but failed to deliver accurate predictions during testing. To rectify this, we adopted a pragmatic approach by categorizing the IMDb rating values. By grouping ratings into discrete intervals, such as rounding off values between 6.5 and 7.4 to '7' and those between 7.5 and 8.4 to '8,' we achieved a more robust predictive model. Figure 4 shows the modifications done in the “Certificates” column to reduce variables and sparseness of data.

```
[ ]  df4["Certificate"]=df4["Certificate"].replace("U","UA")
     df4["Certificate"]=df4["Certificate"].replace("PG-13","PG")
     df4["Certificate"]=df4["Certificate"].replace("G","PG")
     df4["Certificate"]=df4["Certificate"].replace("TV-PG","TV")
     df4["Certificate"]=df4["Certificate"].replace("TV-14","TV")
     df4["Certificate"]=df4["Certificate"].replace("16","TV")
     df4["Certificate"]=df4["Certificate"].replace("TV-MA","TV")
```

Figure 4. Grouping the classes of certificates

With this adjusted dataset, Random Forests exhibited a remarkable 74% accuracy, effectively providing a valuable range of IMDb rating values for movies. This transformation not only enhanced the accuracy of our predictions but also facilitated a more practical understanding of the movie ratings, allowing for a broader spectrum of insights in our analysis.

Model	Accuracy
Random Forests	74.00%
Decision Trees	71.00%
Logistic Regressions	53.56%

In categorizing IMDB ratings into groups of 7, 8, and 9, we utilized three different machine learning models: Random Forest Classifier, Decision Trees, and Logistic Regression. The Random Forest Classifier led the pack with a 74.00% accuracy, thanks to its ensemble approach that combines multiple decision trees to enhance performance and reduce overfitting. Decision Trees followed closely with a 71.00% accuracy, although it's simpler and quicker, it tends to overfit especially when dealing with numerous features. Logistic Regression lagged behind at 53.00% accuracy, as its linear nature made it less adept at capturing the complex patterns in the data. The results underline the robustness of Random Forest for complex classification tasks, even though there is room for improvement across all models through hyperparameter tuning and feature engineering.

Conclusion

This paper delves into the intricate use and application of Machine Learning as a pivotal AI module within the movie industry. It provides a comprehensive overview of machine learning and its various types, illuminating the vast spectrum and potential applications of machine learning algorithms in the cinematic domain. Machine learning, as a technological marvel, has ushered in remarkable advancements in content creation, production processes, and audience engagement. It empowers filmmakers to seamlessly craft intricate virtual sets, conjure lifelike special effects, and even automate content generation, delivering an immersive visual experience.

What's more, machine learning has revolutionized movie marketing and distribution, offering personalized recommendations and tailored advertisements that cater to individual preferences. Notably, the importance of this paper extends beyond the movie industry's extensive adoption of this technology. It underscores the profound influence the movie industry experts have within the broader media and entertainment sector, shaping societal culture and people's lives. Moreover, this paper spotlights the potential of incorporating Natural Language Processing (NLP) techniques, paving the way for enhanced analysis and more accurate predictions regarding a movie's success. By integrating NLP, a more comprehensive and refined analysis can be conducted, elevating the overall effectiveness of predictive models and ultimately redefining the dynamics of the industry.

References

1. AI in the media and creative industries - New European Media
2. Anantrasirichai, N., Bull, D. Artificial intelligence in the creative industries: a review. *Artif Intell Rev* **55**, 589–656 (2022). https://doi.org/10.1007/s10462-021-10039-7
3. Sylvia M. Chan-Olmsted (2019) A Review of Artificial Intelligence Adoptions in the Media Industry, International Journal on Media Management, 21:3-4, 193-215, DOI:10.1080/14241277.2019.1695619

4. N. Quader, M. O. Gani and D. Chaki, "Performance evaluation of seven machine learning classification techniques for movie box office success prediction," 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2017, pp. 1-6, doi: 10.1109/EICT.2017.8275242.
5. Lee, K., Park, J., Kim, I. *et al.* Predicting movie success with machine learning techniques: ways to improve accuracy. *Inf Syst Front* 20, 577–588 (2018). https://doi.org/10.1007/s10796-016-9689-z