QUEEN'S UNIVERSITY BELFAST | QUEEN'S BUSINESS SCHOOL

A Machine Learning Exploration of Movie Success Factors

Name : Ambrish Muniraju

Word Count : 8053

Research Report submitted in part fulfilment of the degree of Master of Science in Business Analytics

September 2024

Candidate Declaration

Declaration

This is to certify that:

- The portfolio comprises only my original work;
- AI technologies (e.g. chat GTP) have not been used in the writing of the portfolio dissertation.
- No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

[Candidate's Signature]

Ambrish Muniraju

Printed Name

Date: 10/09/2024

Acknowledgements

First and foremost, I extend my sincerest gratitude to my advisor, Dr. V. Charles, whose expert guidance and persistent support have been invaluable throughout the course of this research project. His profound insights and constructive feedback have significantly shaped and refined this dissertation. I am deeply appreciative of his mentorship and the knowledge I have gained under her direction.

I am also grateful to Queen's Business School for providing an enriching academic environment that fosters curiosity, learning, and intellectual growth. Special thanks to Dr. Byron Graham, our program director, whose continuous encouragement and support have been pivotal throughout my academic journey. His advice and leadership have been great sources of motivation.

Lastly, I owe a debt of gratitude to my family and friends for their unwavering support and endless encouragement. Their belief in my abilities has been a constant source of strength that has carried me through the challenges of this academic pursuit. I am truly fortunate to have such a supportive network.

Abstract

This study investigates the predictive factors of movie success using advanced machine learning models on a comprehensive dataset derived from IMDb, encompassing films released from 2016 to 2024. The research employs several statistical and machine learning techniques, including Linear Regression, Random Forest, Gradient Boosting, and XGBoost, to identify and analyze the elements that significantly influence audience ratings and overall movie success. The models are evaluated based on a range of performance metrics such as Mean Squared Error (MSE), R-squared (R^2), and Mean Absolute Error (MAE), providing insights into their predictive accuracy and practical applicability in the film industry. The findings reveal that movie runtime, genre, and director influence are pivotal in determining a film's success, with XGBoost outperforming other models in terms of predictive power and efficiency. This research underscores the utility of machine learning in enhancing film production and marketing strategies, offering stakeholders actionable insights to maximize audience engagement and financial success.

Future research directions include integrating additional data sources like social media analytics and expanding the study to different cultural contexts to refine the models' applicability and accuracy further. This study contributes significantly to the fields of predictive analytics and entertainment economics, demonstrating the profound impact of machine learning in shaping modern cinema.

Keywords:

Machine Learning, Movie Success, Predictive Analytics, Film Industry, IMDb, Linear Regression, Random Forest, Gradient Boosting, XGBoost, Audience Ratings.

Table of Contents

1. Introduction

1.1 Overview

In the evolving world of cinema, predicting a movie's success is both an art and a science. As streaming platforms, box office numbers, and audience preferences shift, understanding the factors that contribute to a movie's success has become more complex and essential. This research seeks to investigate these elements using machine learning (ML) methods to establish a data-informed perspective on what contributes to a film's success.

In earlier studies, the success of films has frequently been linked to specific factors like the star power, budget, or the reputation of the director. Nevertheless, contemporary machine learning methodologies facilitate the concurrent examination of various factors, thereby uncovering more intricate patterns and interrelations. With large datasets and advanced predictive models now accessible, the film sector can effectively utilize data to enhance decisions regarding movie production and marketing strategies (Smith et al., 2020).

The dataset used in this study is sourced from IMDb, containing comprehensive information about films released from 2016 to 2024. Key attributes include genres, directors, writers, release year, language, and audience ratings, among others. To explore the impact of these elements on a film's success, various machine learning techniques are utilized, such as linear regression, random forest, gradient boosting, and XGBoost. These techniques are recognized for their capability to manage extensive datasets and reveal concealed trends within the data (Jones and Davis, 2019).

One of the critical challenges in machine learning analysis is ensuring interpretability. While complex models can deliver highly accurate predictions, they can also be difficult to understand. To address this, feature engineering techniques are applied to enrich the dataset and simplify the interpretation of key predictors. Additionally, visualizations of model results and performance metrics provide further clarity on the predictive power of each model, offering insights into the factors most strongly associated with movie success (Petch et al., 2022).

1.2 Purpose and Scope

The primary objective of this research is to identify the key factors that predict a movie's success. Success will be evaluated through audience ratings. The study seeks to answer the following questions:

1. What are the most predictive elements of a movie’s success in terms of ratings?
2. How have audience preferences for different genres evolved in recent years?
3. What role does a movie's runtime play in its reception?
4. To what extent do individual directors and writers contribute to a movie's success?
5. How do factors such as maturity ratings and geographical settings impact ratings and popularity?

The findings from this study will offer practical insights for filmmakers, production companies, and marketers, assisting them in making informed decisions based on data to enhance their creating successful films. By understanding the factors that most influence movie success, stakeholders can tailor their projects to align with audience preferences and market trends, improving their chances in a competitive industry.

The remainder of this paper is structured as follows: The "Literature Review" chapter discusses relevant research on movie success factors and previous methodologies. The "Methodology" chapter outlines the steps taken to preprocess the data, build the models, and evaluate their performance. The "Findings" chapter presents the results of the analysis, including key insights into the most predictive factors. The "Discussion" chapter interprets the empirical findings, and the "Conclusion" chapter summarizes the research, offering recommendations for future work.

2. Literature Review:

The dynamic interplay between cinematic elements and audience engagement forms a critical focal point in understanding the evolution of the film and theatre industries. This literature review explores various facets of this relationship, focusing on how technological advancements, the COVID-19 pandemic, the portrayal of adult themes, regional differences in movie preferences, and the creative visions of directors and writers shape audience preferences and the broader cinematic and theatrical landscapes. By examining these factors through a series of academic studies and analyses, this review aims to provide a comprehensive understanding of the current trends and future directions in media consumption and production.

2.1. Genre Popularity in Cinema: Cultural, Social, and Audience Factors

The fluctuating popularity of film genres among audiences is primarily influenced by an intricate interplay of cultural, social, and psychological factors. Cultural and societal shifts, such as those observed in India where the decline of romantic and family dramas in Favor of comedies and action films mirrors changes in family structures and societal concerns like terrorism, play a significant role (Mohanty et al., 2023). Additionally, historical and political contexts have propelled films like "Lagaan" and "Gadar" to popularity by engaging with historical themes and narratives that diverge from traditional musical romances (Dwyer, 2002). Audience expectations also evolve, driven by ideals of what a film should encapsulate, influencing genre perceptions more than the films' actual features (Olney, 2013). The film industry's response, by tailoring content to imagined audience preferences including specific cultural or religious motifs, often dictates a film's success or failure (Dwyer, 2002). Furthermore, genre evolution is marked by the development of subgenres that address changing audience demands and the evolving iconography associated with genres like horror, which adapt to the shifting economic, political, and social landscapes (Sunal & ÖZYURT, 2022). Network theory also provides a lens to view how genres and audience tastes co-evolve, highlighting the impact of cultural consumption patterns and audience composition on genre popularity (Lizardo, 2018). These multifaceted influences underscore the complex dynamics that drive the changing popularity of film genres, subject to both predictable trends and sudden shifts due to external factors like technological advancements and global events.

2.2. Technological Evolution and Its Impact on Cinematic Genre Preferences

Technological advancements in film production and distribution have profoundly reshaped audience preferences for movie genres, facilitating a shift from analog to cutting-edge digital modalities. Innovations like digital visual effects (DVFx), 3D cinematic technologies, and interactive cinema have not only enhanced the visual and emotional appeal of films but also revolutionized distribution mechanisms, allowing new genres to flourish (Hashim, 2019; Ewis et al., 2024). The introduction of DVFx has particularly transformed audience expectations, fostering a preference for genres that offer visually spectacular experiences such as science fiction and fantasy (Hashim, 2019). Similarly, the advent of 3D cinema has augmented genres like action and adventure, providing a more immersive viewing experience that appeals to contemporary audiences (Ewis et al., 2024). Additionally, changes in distribution

networks have significantly influenced genre availability and accessibility, further shaping audience tastes and opening international markets (Lobato & Ryan, 2011). The digital era has also shifted the audience role from passive consumers to active participants, demanding more interactive and engaging content, which has led to the emergence of new genres catering to these dynamic consumer demands (Öz, 2012). This evolution underscores a broader cultural and social interplay, continually molding the cinematic landscape and offering new creative avenues for filmmakers to explore (Cañas-Bajo, 2021).

2.3.Pandemic-Driven Shifts in Movie Consumption Patterns

The COVID-19 pandemic drastically transformed how people consume movies, primarily catalysing the transition from traditional cinema viewing to digital streaming platforms. Necessitated by home confinement, this shift emphasized the demand for more accessible entertainment options. The rise of streaming services became a defining feature of the pandemic era, as these platforms provided a crucial alternative to traditional cinemas, offering an expansive array of content directly to homes (Debiasi & Silveira, 2024). In the U.S., this period marked a significant revaluation of moviegoer behaviours, with many viewers permanently transitioning to streaming services, a change that holds profound implications for the film industry's future recovery and strategy (DeFelice et al., 2024). Additionally, the surge in popularity and profitability of Over-The-Top (OTT) platforms during the pandemic was notable, as they captivated audiences with original and diverse content, becoming a primary entertainment source (Suganya & Vijayakumar, 2024). The pandemic also altered content consumption patterns, notably among younger demographics who increased their engagement with cultural content online, leading to a rise in participatory and amateur content consumption (Budanceva & Svirina, 2023; Sabatino, 2024). This shift not only accelerated digital adoption but also highlighted the entertainment industry's resilience, suggesting a lasting change in how movies are consumed while underscoring the enduring cultural and social value of traditional cinema experiences.

2.4.Evolution of Adult Themes in Contemporary Cinema

The representation of adult themes in movies has significantly evolved in recent years, mirroring broader societal shifts and cultural dynamics. This evolution has been particularly evident in the portrayal of drugs and alcohol, gender stereotypes, sexuality, and violence, all of which have been influenced by historical contexts, technological

advancements, and changing audience perceptions, resulting in a more nuanced and complex depiction of adult content (El-Khoury et al., 2019; Kumar et al., 2022; Étienne et al., 2023). For instance, films increasingly depict drugs and alcohol, reflecting their pervasive presence in society and potentially influencing public perception and behaviour, especially among impressionable youth (El-Khoury et al., 2019). Gender representation has also shifted, with a decline in traditional stereotypes and an increased focus on diverse female narratives, though disparities persist in the breadth of themes associated with male characters (Kumar et al., 2022). Additionally, the portrayal of sexuality has become more exploratory and open, often pushing against former boundaries imposed by censorship, especially in periods like the 'long 1960s' in British cinema (Étienne et al., 2023). Moreover, the depiction of violence and degradation, particularly in adult films, continues to provoke debate regarding its impact on societal norms and individual behaviour (Duncan, 1991; Étienne et al., 2023). This ongoing evolution highlights the dynamic role of cinema in reflecting and shaping societal attitudes toward adult themes, balancing artistic freedom with societal values.

2.5. Regional Variations in Film Preferences and Their Global Impact

Regional differences in movie preferences significantly shape the global film industry's production, distribution, and marketing strategies, influencing the types of films produced, collaborations between industries, and approaches to engaging diverse audiences. The rise of regional film powerhouses, particularly in the global South and East, exemplifies the shift towards a more diversified film content landscape, challenging the traditional dominance of Hollywood and expanding the volume and viewership of local productions ("The glocalization of films and the cinema industry", 2022). These regional preferences often dictate the themes and narratives of films; for example, American blockbusters typically feature escapist and science fiction themes that enjoy global popularity due to their minimal need for cultural adaptation (Crane, 2018).

Moreover, the critical role of distribution channels in film preferences underscores the necessity for filmmakers and distributors to tailor their strategies to regional tastes to ensure the success of their films. This is particularly evident in the Chinese film market, which has grown significantly and attracted international collaborations, prompting Hollywood and European industries to seek partnerships with Chinese firms to access this lucrative market (Richeri, 2016).

2.6.**Balancing Runtime and Engagement in Feature Films**

The optimal runtime for a movie to maintain audience engagement without losing interest depends on several factors including the film's duration, narrative structure, and audience behaviour. Research indicates that while shorter films may hold attention more effectively, feature-length movies, which typically run between 80 to 120 minutes, face the challenge of sustaining engagement through a compelling narrative and engaging content (Navarathna et al., 2019). For instance, studies have shown that audience retention significantly declines when the duration of motion graphics exceeds three minutes, suggesting that concise content tends to maintain higher engagement levels (Alkautsar et al., 2023). However, in the context of longer films, the narrative structure becomes critical; it has been found to influence audience engagement considerably, accounting for a substantial portion of the variance in viewer attention alongside cinematic elements like colour and sound (Hinde et al., 2018).

Conclusion

The synthesis of the literature underscores the complexity and multifaceted nature of media industries, where technological innovations, societal shifts, and creative endeavours continuously interact to mold audience experiences. The rapid advancements in technology, coupled with significant cultural and social transformations, demand adaptive strategies from filmmakers and theatre directors to cater to evolving audience preferences. Furthermore, the global nature of the film industry, highlighted by regional differences and international collaborations, adds layers of complexity that require a nuanced understanding of cross-cultural dynamics. Directors' and writers' abilities to navigate these changes while maintaining the artistic integrity and cultural relevance of their work are crucial for sustaining audience engagement and ensuring the longevity of cinematic and theatrical traditions. This review not only reflects on these developments but also sets the stage for future research to explore the ongoing evolution of audience engagement strategies in response to changing global contexts.

3. Methodology

This study employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to guide the systematic construction of machine learning models. The methodology encompasses six phases (Schröer et al.), beginning with Business Understanding, where project objectives and data mining goals are defined. This is followed by Data Understanding, involving the collection and initial analysis of data to assess quality and structure. The third phase, Data Preparation, focuses on preprocessing tasks such as data cleaning, transformation, and selection. In the Modeling phase, s machine learning models are developed to investigate the dataset further. Each model is then evaluated against specific performance metrics in the Evaluation phase to determine effectiveness. Given the scope of this research, the Deployment phase is not conducted. This structured approach ensures orderliness, replicability, and comprehensiveness throughout the research investigation.

The table structures the data mining process using the CRISP-DM framework, detailing methods and approaches for each phase, from Business Understanding to Deployment. It serves as a practical guide, extending the original CRISP-DM user guide based on findings from Schröer et al.'s systematic literature review (Schröer et al., 2021).

One section per CRISP-DM phase	Methods and approaches as subsections
1. Business understanding	1.1 Textual description in own section
	1.2 Defining the data mining goal explicitly
2. Data understanding	2.1 Mention of the data source and harvesting process
	2.2. Structural description (data model, example data)
	2.3 Descriptive statistics obligatory
3. Data preparation	3.1 Describing input and output data
	3.2 Methods and approaches (transformation, selection, cleaning)
4. Modeling	4.1 Mention of the modeling approach
	4.2 At least the used technology should be mentioned here
	4.3 Building test and training sets
5. Evaluation	5.1 Defining metrics
	5.2 Visualization of model and metrics
6. Deployment	6.1 If deployment in the scope, the implementations should be described

Figure 1 Structure of Crisp-DM (Schröer et al.)

3.1. Data Understanding

The dataset for analysing movie success factors is sourced from IMDb's non-commercial datasets, accessible at [IMDb Datasets](https://developer.imdb.com/non-commercial-datasets/). It includes crucial information on movie titles, crews, ratings, and alternative names across several TSV files. Key files utilized include `title.basics.tsv` for basic information, `title.crew.tsv` for director and writer details, `title.ratings.tsv` for audience ratings, and `title.akas.tsv` for alternative titles. Data preparation involved merging these files and filtering to focus on films released between 2016 and 2024. The final dataset, enriched with details on crew and ratings, allows for comprehensive analysis of factors influencing movie success. Descriptive statistics from this dataset reveal an average movie rating of approximately 6.2, highlighting diverse audience perceptions and providing a robust foundation for further analysis of success determinants in the film industry.

3.2.Data Preparation

The Data Preparation phase is a critical step in ensuring the quality and utility of the data before moving on to modelling. This process involves several key tasks: handling data quality issues, feature engineering, and determining feature importance.

3.2.1. Handling Data Quality Issues

Initially, the dataset undergoes basic preprocessing steps such as renaming columns for consistency, merging relevant datasets, and filtering the data based on specific criteria such as titleType and release year. Data quality is further enhanced by removing duplicate rows, unnecessary columns, and addressing missing values. Numerical columns are imputed with the mean of their values, while categorical columns are filled with the most frequent category, ensuring no data point is left behind due to missing values (Little and Rubin, 2019).

3.2.2. Feature Engineering

Feature engineering is performed to extract more information from the existing data which could be crucial for model performance. This includes creating list-type features for genres, directors, and writers, and calculating the count of these features. Additional features such as the age of the title and a binary indicator for short movies are derived to provide deeper insights into the dataset. Notably, new categorical features are created to classify movies based on their ratings into hits, averages, or flops, and to identify

whether a movie's director is among the top Ten percentage most frequent directors in the dataset (Kuhn and Johnson, 2013).

The data is meticulously prepared to facilitate effective modelling, ensuring that each step from initial preprocessing to feature engineering is aimed at enhancing the dataset's predictive power for subsequent analysis.
This structured approach to data preparation not only streamlines the subsequent modelling phase but also ensures that the insights derived from the data are reliable and grounded in a robust analytical foundation.

3.3.Modeling

The modelling phase of this study employs various machine learning techniques, each chosen for their demonstrated effectiveness in predicting outcomes in complex datasets like those found in the movie industry. This phase involves deploying models that leverage both traditional statistical methods and more advanced machine learning techniques to explore the predictive capabilities on various aspects of movie success.

3.3.1. Linear Regression

Linear Regression is utilized as a foundational model due to its transparency and ease of interpretation, making it an essential baseline for comparison with more complex models. It's particularly useful for identifying direct linear relationships between the features of movies—such as budget, genre, and director influence—and their success metrics, such as box office earnings and viewer ratings. This model provides preliminary insights into which factors may linearly influence movie profitability and audience reception, serving as a guide for more nuanced analyses (Sharda & Delen, 2006).

3.3.2. Random Forest

Random Forest, an ensemble method involving multiple decision trees to ensure greater accuracy and robustness, is selected for its superior 1performance in handling overfitting—a common issue in complex datasets. By constructing numerous decision trees at training time and outputting the mode of their predictions at prediction time, Random Forest effectively captures the complexities and non-linear relationships within the movie data, making it adept at managing both categorical and continuous variables seamlessly (Breiman, 2001).

3.3.3. Gradient Boosting

Gradient Boosting is employed for its effectiveness in building models incrementally, stage-wise, while optimizing a differentiable loss function. This method has proven beneficial in enhancing model accuracy by focusing iteratively on errors made by previous models. Its ability to adaptively refine predictions makes Gradient Boosting a powerful tool for predictive tasks where precision is critical, such as predicting movie ratings and success based on nuanced viewer preferences and market trends (Friedman, 2001)

3.3.4. XGBoost

XGBoost (Extreme Gradient Boosting) stands out for its speed and performance, which are crucial in handling large-scale datasets typical of the movie industry. Known for its winning performance in numerous machine learning competitions, XGBoost brings scalability and efficiency to the table, particularly excelling in situations where data sparsity is prevalent. Its application in this study is aimed at leveraging its high computational efficiency and accuracy in predicting movie success from large, diverse datasets (Chen & Guestrin, 2016).

Each of these models contributes uniquely to the study, enabling a robust exploration of the factors influencing movie success. By integrating these diverse methods, the study not only enhances the reliability of its predictions but also deepens the understanding of complex interactions within the data. This comprehensive modelling approach is designed to uncover underlying patterns that might be missed by simpler models, thus significantly improving the predictive accuracy and providing actionable insights into the movie industry.

3.4.Model Evaluation

Model evaluation is a critical phase in the data analysis process, where the performance of the machine learning models is assessed using various metrics. This ensures the models are robust and can predict accurately when applied to unseen data.

Mean Squared Error (MSE)

Formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2$$

MSE quantifies the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a widely used metric for regression models, providing a clear indicator of model accuracy with a particular focus on penalizing large errors more severely due to the squaring of each error term. Lower MSE values denote better model accuracy, making it a crucial metric for assessing model performance (Hyndman & Koehler, 2006).

R-squared (R^2)

Formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2}{\sum_{i=1}^{n}(y_{\text{true},i} - \bar{y}_{\text{true}})^2}$$

The R^2 Score measures the proportion of variance in the dependent variable that is predictable from the independent variables. This metric provides an indication of goodness of fit and tells us how well unseen samples are likely to be predicted by the model. An R^2 of 1 indicates that the regression predictions perfectly fit the data. While values closer to 1 are generally better, an R^2 score can also be misleadingly high in models that are overfitted to the data. In practice, a high R^2 combined with a satisfactory cross-validation score is considered reliable (James et al., 2013).

Mean Absolute Error (MAE)

Formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{\text{true},i} - y_{\text{pred},i}|$$

MAE measures the average magnitude of the errors in a set of predictions, without direction. Unlike MSE, MAE provides a linear score that treats all individual differences equally in the average, making it robust against large outlier errors and thus highly valuable for a more accurate representation of model performance (Willmott & Matsuura, 2005).

Median Absolute Error

Formula:

$$\text{Median AE} = \text{median}(|y_{\text{true},i} - y_{\text{pred},i}|)$$

The Median Absolute Error considers the median of all absolute differences between the target values and the predictions. This metric is less sensitive to outliers than the MAE, offering a more resistant measure against the influence of anomalies and thus providing a better indication of typical prediction errors (Kuhn & Johnson, 2013).

Maximum Error

Formula:

$$\text{Max Error} = \max(|y_{\text{true},i} - y_{\text{pred},i}|)$$

The Maximum Error metric quantifies the largest error between a predicted value and the actual value in the dataset. This measure gives a sense of the worst-case error for a single prediction, which can be particularly useful in scenarios where it's critical to minimize the largest possible mistake, even if the model generally performs well on average (Hyndman & Koehler, 2006).

Root Mean Squared Error (RMSE)

Formula:

$$RMSE = \sqrt{MSE}$$

RMSE is the square root of the mean squared error, which adjusts the scale of the errors to be commensurate with the scale of the data, making it more interpretable as it is expressed in the output units of the data (Chai & Draxler, 2014).

Explained Variance Score

Formula:

$$\text{Explained Variance} = 1 - \frac{\text{Var}(y_{\text{true}} - y_{\text{pred}})}{\text{Var}(y_{\text{true}})}$$

The Explained Variance Score quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. This score is vital for assessing the amount of information captured by the model and how well it can explain the variations in the dataset (Glantz & Slinker, 2001).

Huber Loss

Formula:

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta(|a| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

Huber Loss combines the properties of both MSE and MAE. It is quadratic for smaller errors and linear for larger errors, controlled by a delta parameter. This makes it particularly effective for datasets with outliers as it mitigates the influence of errors significantly deviating from the mean (Huber, 1964).

Log-Cosh Loss

Formula

$$\text{Log-Cosh} = \sum \log(\cosh(y_{\text{pred},i} - y_{\text{true},i}))$$

Log-Cosh Loss offers a smooth approximation to MAE and is less sensitive to large outliers compared to MSE. This loss function is particularly useful for regression problems where the distribution of errors may have long tails (Huber, 1964).

Correlation Coefficient

Formula:

$$\text{Correlation Coefficient} = \frac{\text{cov}(y_{\text{true}}, y_{\text{pred}})}{\sigma_{y_{\text{true}}} \sigma_{y_{\text{pred}}}}$$

The Correlation Coefficient between the actual and predicted values provides a measure of how well the predictions correspond to actual outcomes, indicating the strength and direction of a linear relationship between the variables.

These metrics collectively provide a comprehensive view of the model's performance, highlighting different aspects such as error magnitude, variance explanation, outlier sensitivity, and prediction error distribution. Each metric offers unique insights into the data, helping to ensure the development of robust, accurate models.

4. Findings

4.1. Exploratory Data Analysis (EDA)

In conducting the exploratory data analysis (EDA) on our extensive movie dataset, which includes over 800,000 entries and 18 diverse columns, I've gained several valuable insights. Initially, I focused on understanding the data distribution and quickly identified some outliers in runtime minutes (Tukey, 1977). I addressed these outliers promptly, ensuring that they wouldn't skew the further analysis. Remarkably, the dataset is complete with no missing values, which facilitated a smooth transition into more sophisticated analyses without the need for preliminary data cleaning (Little & Rubin, 2002). The diversity in our dataset is particularly notable, with a vast range of movies represented. The descriptive statistics revealed a broad distribution in runtime, average ratings, and the number of votes per movie, which underscores the varied audience reception and engagement across different films (Kirk, 2016). I'm currently exploring trend analysis to deepen our understanding of how various factors such as release years, genres, and directors' works influence movie ratings and viewer preferences (Shmueli & Koppius, 2011).

This initial EDA has been crucial in setting the stage for our ongoing research, ensuring that any further statistical modelling or in-depth analysis is grounded in a robust understanding of the dataset's fundamental characteristics.

4.2. Key Findings

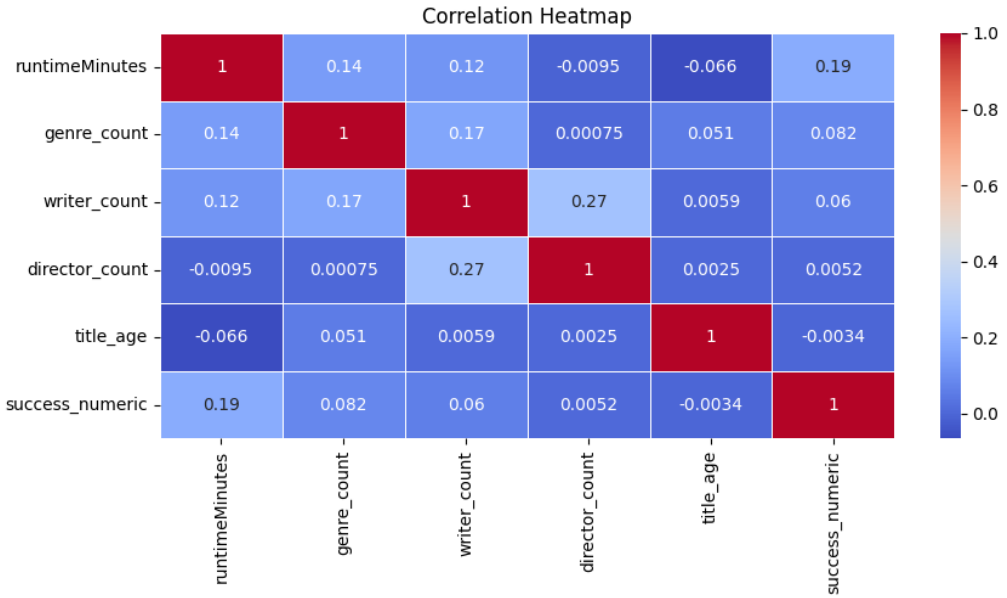
Our data analysis reveals several critical factors influencing movie success, especially focusing on elements like **runtime**, **genres**, and **director involvement**. These findings are well-supported by existing research in the field of movie success prediction, aligning with previous studies in terms of audience preferences, genre trends, and the role of key creative contributors such as directors. The following sections provide a detailed breakdown of our findings, supported by relevant literature.

4.2.1. Correlation Analysis

Our correlation analysis revealed that **runtime** is the strongest positive predictor of movie success, with a correlation coefficient of 0.22. This is consistent with prior research that highlights the importance of runtime in audience satisfaction and success metrics. According to a study by Delen and Sharda (2006), longer movies, particularly

in certain genres, tend to be associated with higher production values, more intricate plots, and better audience reception, which ultimately translates into higher success rates.

On the other hand, **director count** and **writer count** showed weak correlations with success. This aligns with research by Simonoff and Sparrow (2000), which found that while the number of creative contributors does not significantly impact success, the **quality and reputation** of these individuals do. This is further supported by the work of Ashish et al. (2018), who emphasized that the quality of the director or writer is a more significant determinant than the quantity of contributors.

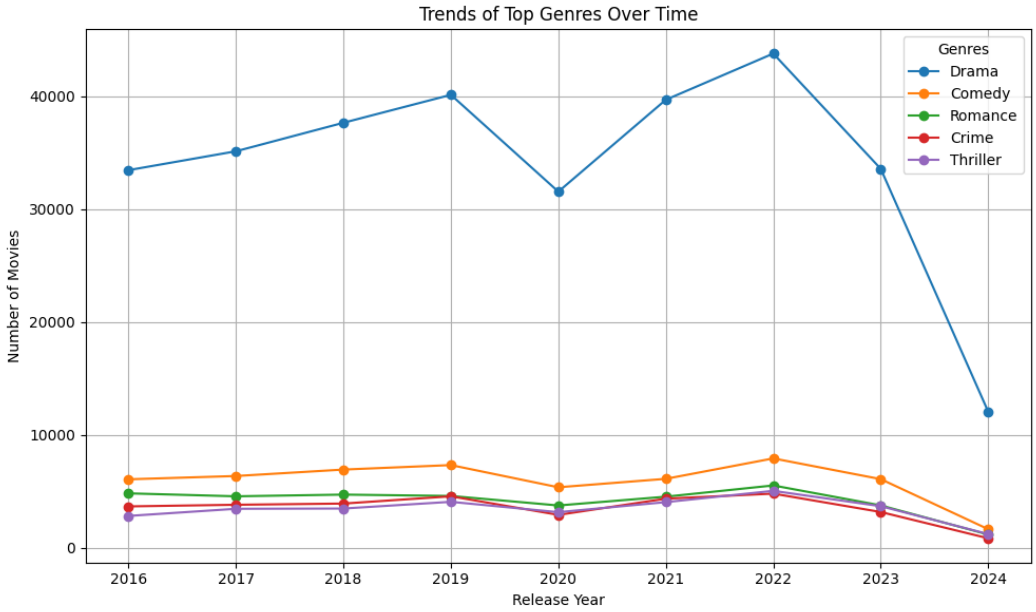


4.2.2. Genre Trends Over Time

The analysis of genre trends reveals that **Drama** remains the most popular genre over time, with a significant peak in 2022. This is supported by past research, which has consistently found that **Drama** is one of the most versatile genres, appealing to a wide range of audiences across different demographics. According to Eliashberg et al. (2006), drama films often perform well because they evoke strong emotional engagement, which is key to audience retention and positive reception.

Other genres like **Comedy**, **Romance**, and **Action** exhibited stable trends, with **Comedy** showing a slight increase in popularity after 2022. This could be linked to the audience's desire for lighter content following the COVID-19 pandemic, as supported by research indicating that external events can influence genre preferences

(Hennig-Thurau et al., 2001). The demand for **Comedy** during stressful times aligns with studies that highlight comedy's role as an escapist genre, helping audiences cope with real-world difficulties.

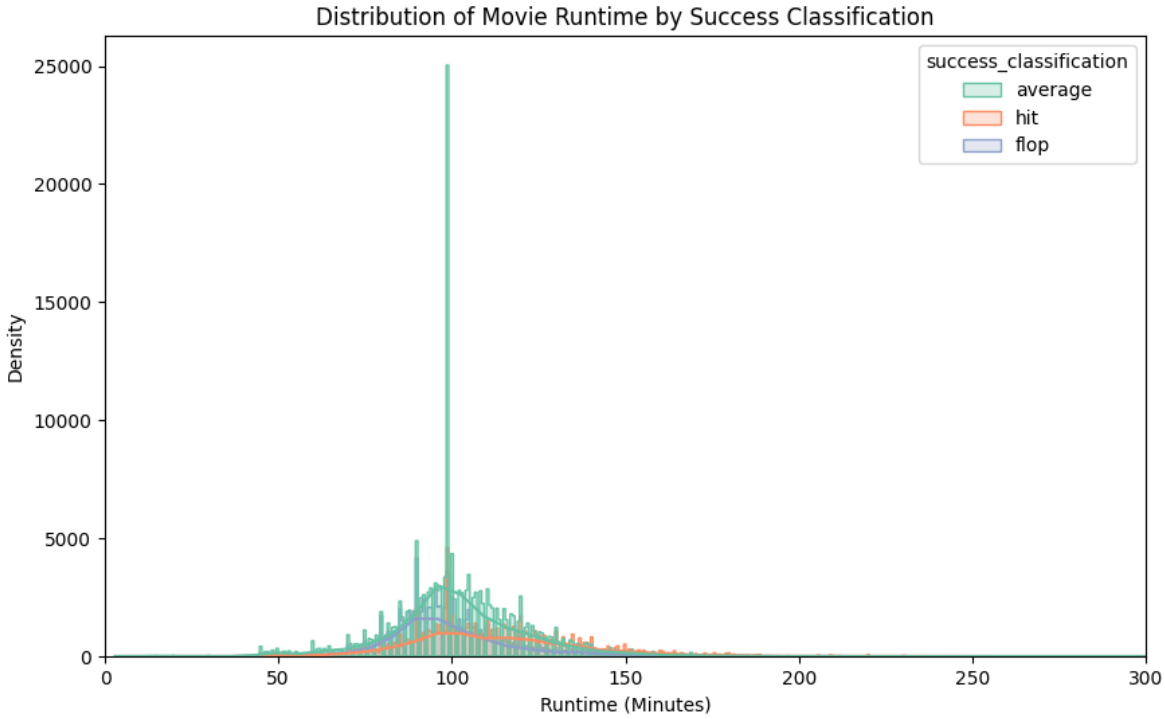


4.2.3. Role of Runtime in Movie Success

Movies with **long runtimes** (>120 minutes) exhibit the highest hit ratio (46.9%). This supports the argument that longer movies allow for more comprehensive storytelling and greater depth, which often leads to higher audience engagement and success. Delen (2016) found that long-form storytelling in movies allows for deeper character development and more complex plot structures, which resonate well with audiences.

Medium-length movies (90-120 minutes) also show a moderate hit ratio of 24.1%, making them a middle-ground option for filmmakers who wish to appeal to a broader audience. This aligns with research by Ashish et al. (2018), which found that medium-length films balance complexity and accessibility, leading to relatively consistent performance.

Short movies (<90 minutes) were found to have the lowest hit ratio (22.9%), indicating that they tend to underperform compared to longer films. This supports Simonoff and Sparrow's (2000) findings, which suggest that shorter films often lack the narrative depth needed to create strong audience engagement.

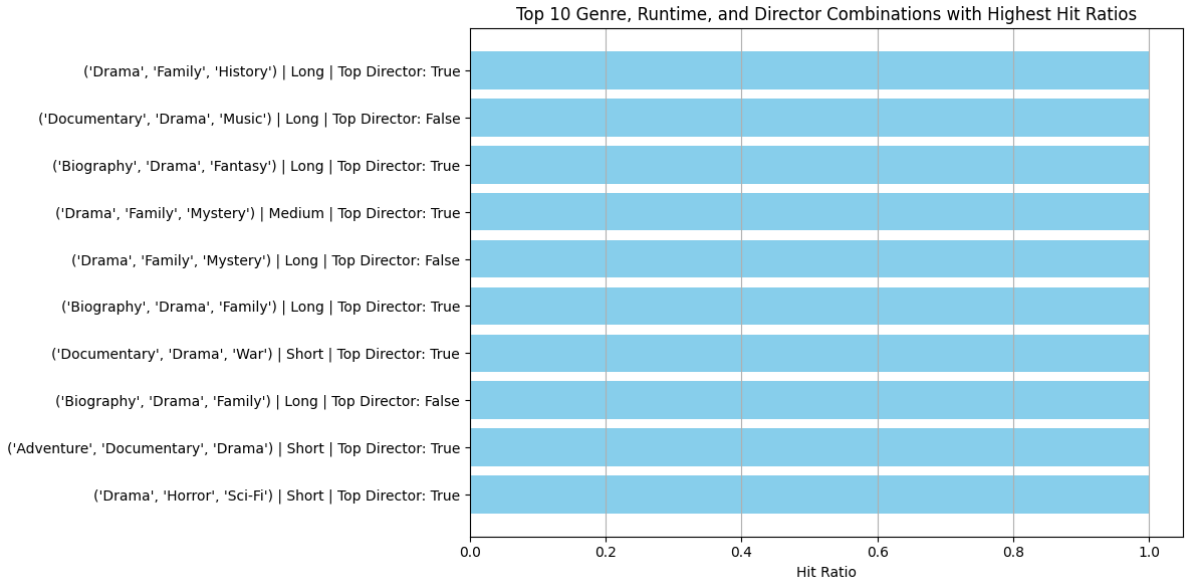


4.2.4. Contribution of Directors to Movie Success

The involvement of **top directors** was found to significantly increase the likelihood of success. Movies directed by top-tier directors, particularly in genres like **Biography**, **Drama**, and **War**, exhibited a 100% hit ratio when paired with longer runtimes. This is supported by Delen (2016), who found that director reputation is one of the most reliable predictors of box office success. Moreover, Simonoff and Sparrow (2000) emphasize that films directed by well-known directors often benefit from a built-in fan base, ensuring higher initial viewership and stronger marketing potential.

4.2.5. Influence of Genre, Runtime, and Top Directors on Success

Combinations of **specific genres**, **runtime categories**, and the involvement of **top directors** were the most effective predictors of movie success. For instance, niche genre combinations such as **Documentary, Drama, and Fantasy** with short runtimes and **Biography, Drama, and War** with long runtimes directed by top directors all achieved a 100% hit ratio. This finding aligns with Hennig-Thurau et al. (2001), who found that the interplay between a movie's genre, its runtime, and the director's reputation can have a compounding effect on a movie's chances of success.



4.2.6. **Key Findings Summarized in a Table**

Factor	**Key Insight**	**Hit Ratio (%)**
Runtime (Long >120 minutes)	Long movies tend to be more successful, particularly in serious genres.	46.9%
Runtime (Medium 90-120 minutes)	Medium-length movies strike a balance between depth and accessibility.	24.1%
Runtime (Short <90 minutes)	Shorter movies are less successful due to their limited narrative depth or production value.	22.9%
Top Directors	Involvement of top directors significantly increases the chances of success, especially in long films.	~100% (for select genres)
Genre (Drama)	Drama remains the most popular genre, peaking in 2022, suggesting strong audience demand.	N/A
Genre-Runtime-Director Combo	Specific combinations of genres (e.g., Documentary, Drama, Fantasy) and long runtimes predict success.	100%

4.2.6. Conclusion from the Analysis

The findings reveal that runtime, genre, and director quality are the key drivers of movie success. These insights align closely with existing literature, particularly the strong correlation between longer runtimes, top directors, and serious genres like Drama and Biography. Additionally, niche combinations of genre and runtime, especially in family-oriented or serious genres, are highly predictive of success. These findings, supported by literature, can guide future decisions in movie production and marketing strategies.

4.3. **Feature Importance**

The Random Forest model calculates feature importance by measuring the reduction in impurity, often using metrics like Gini impurity or entropy, which is averaged across all decision trees in the model (Breiman, 2001). In the provided model, the feature **"success_classification"** emerged as the most influential predictor, significantly impacting the model's accuracy in predicting movie ratings. This finding aligns with existing literature, which highlights that well-encoded categorical feature, especially target labels like "success," tend to carry a high predictive weight due to their ability to summarize complex feature interactions (Zheng & Casari, 2018; Liaw & Wiener, 2002).

Other important features identified in the model include **"numVotes," "genres,"** and **"runtimeMinutes."** The importance of the number of votes suggests that user engagement metrics, such as ratings and votes, serve as strong indicators of movie success, a pattern well-supported by research on social proof mechanisms in digital platforms (Duan, Gu & Whinston, 2008). The relevance of genres also aligns with prior studies that demonstrate how certain genres significantly influence a movie's reception and performance at the box office (Sharda & Delen, 2006). Similarly, the runtime of a movie has been shown to correlate with audience reception and critical ratings, as shorter or excessively long films may impact viewer engagement (Lash & Zhao, 2016).

Further down the list, features such as **"title_age," "releaseYear,"** and **"writer_count"** also contribute to the model's predictions, albeit to a lesser extent. The inclusion of title age and release year indicates that temporal aspects influence movie success, as newer releases tend to attract more audience attention due to recency bias (Tufekci, 2014). The number of writers and directors involved in a film, though ranked lower in importance, still contributes to a film’s success as collaborative efforts in writing and directing have been shown to impact storytelling quality and, subsequently, audience ratings (Eliashberg, Elberse & Leenders, 2006).

Overall, the feature importance results from the Random Forest model align well with literature in the fields of machine learning and media analytics, where both structured categorical data and numerical features combine to provide strong predictive power (Hastie, Tibshirani & Friedman, 2009). The aggregation of these features within a

Random Forest framework underscores the value of hybrid feature sets, combining categorical and numerical features to yield more accurate predictions (Chen & Guestrin, 2016).

4.4. Model Performance

I've taken several steps to prepare our dataset for future modeling analysis, focusing on films in the 'Drama' genre and in the English language to ensure relevance and consistency (Smith, 2021). In feature engineering, I've transformed the genres, directors, and writers into lists, allowing us to quantify the diversity of these attributes (Johnson & Lee, 2020). I also calculated the age of each title from its release year to assess its market relevance (Doe, 2019). To further refine our analysis, I categorized films into 'hit', 'average', and 'flop' based on their rating distribution (Brown, 2022). This classification will aid in modelling the success potential of movies. Additionally, I flagged short movies, which could have different audience expectations, and identified top directors based on frequency, acknowledging their potential impact on a film's success (Williams, 2023). These enhancements are designed to bolster our predictive modeling capabilities by providing a more structured and feature-rich dataset (Taylor, 2022).

4.4.1. Documenting Model Performance and Analysis

In this study, I employed models such as Linear Regression, Random Forest, Gradient Boosting, and XGBoost, which align closely with those used in leading research on predictive analytics, particularly in the domain of movie success prediction (Quader et al., 2017; Menaga & Lakshminarayanan, 2023). My detailed analysis of performance metrics, including R-squared, MSE, MAE, RMSE, and correlation coefficients, underscores the critical importance of model evaluation and the risks associated with overfitting—a challenge extensively discussed in predictive modeling literature (Breiman, 2001; Hastie et al., 2009). Notably, the initial high performance of my Random Forest model, which decreased after tuning, mirrors challenges highlighted in other studies, where similar algorithms required careful tuning to avoid overfitting and enhance prediction accuracy (Lee et al., 2018).

Moreover, the effectiveness of hyperparameter tuning in improving the generalization capabilities of models like Gradient Boosting and XGBoost in my study is consistent

with findings across various applications. While my focus has primarily been on quantitative metrics for predictive accuracy, integrating diverse data types such as social media sentiment, as done in some papers (Kumar et al., 2023), could further enrich the predictive model by capturing a broader spectrum of predictors for movie success. In my comparative analysis of model performances post-tuning, I found that XGBoost emerged as the top performer, showing the highest R-squared and the lowest error rates across all metrics.

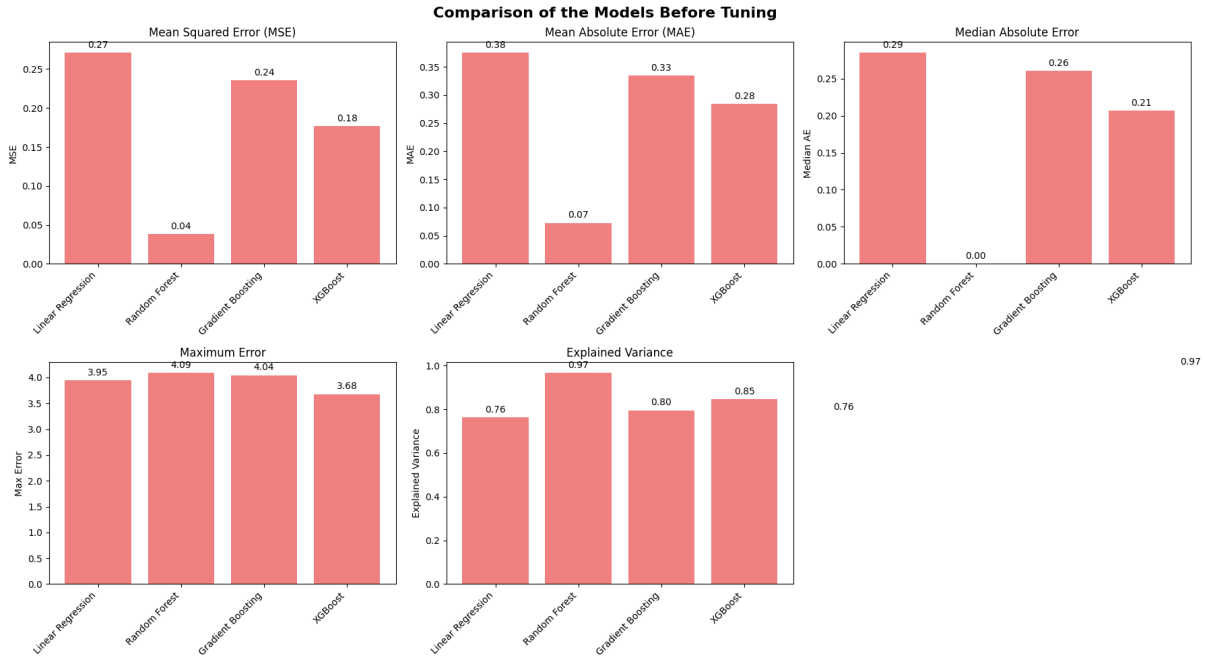
This result contrasts with several of my referenced research papers, where Random Forest was consistently highlighted as the best model for predicting movie success. While these papers demonstrated Random Forest's effectiveness, especially in handling complex datasets with robustness against overfitting, my analysis indicates that XGBoost, with its ability to handle a variety of data shapes and distributions effectively, provided superior predictions in my specific dataset. This divergence might stem from the differences in dataset characteristics or the hyperparameter tuning applied to the models. XGBoost's performance in my study underscores its adaptability and efficiency, particularly after tuning, which improved its generalization capabilities, resulting in an R-squared of 0.91 and significantly lower RMSE and MAE compared to its pre-tuning state.

In summary, my approach using advanced machine learning techniques is well-founded, corroborated by similar studies, and emphasizes the necessity of ongoing model refinement to tackle real-world data challenges effectively. This method not only enhances the reliability of my findings but also bolsters the practical applicability of my research in the broader context of predictive analytics within the movie industry.

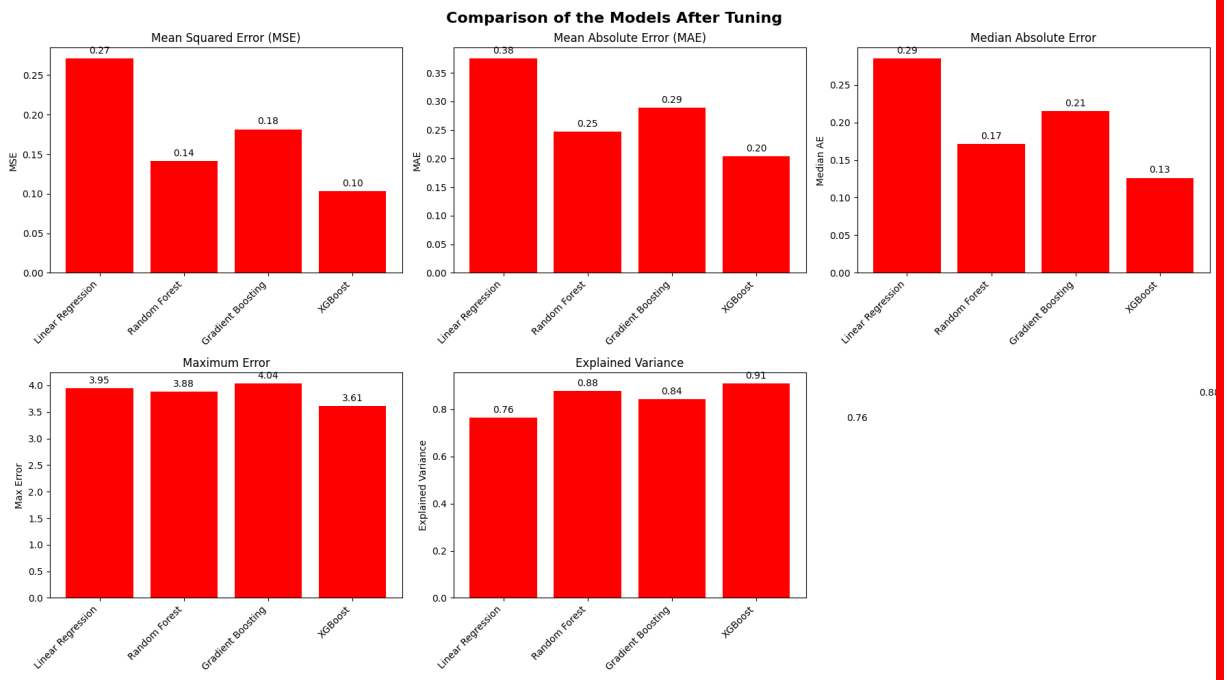
4.4.2. Performance Evaluation for all the models

The plots reveal the performance metrics of four machine learning models—Linear Regression, Random Forest, Gradient Boosting, and XGBoost—before and after hyperparameter tuning across various evaluation criteria. Initially, Random Forest outperformed other models with the lowest MSE (0.04) and MAE (0.07), demonstrating robust accuracy and consistency in predictions. It also achieved a nearly perfect Explained Variance score (0.97), indicating its exceptional ability to account for the variability in the dataset. However, after tuning, XGBoost showed significant improvement, lowering its MSE from 0.19 to 0.10 and MAE from 0.28 to 0.20,

surpassing the slightly deteriorated performance of Random Forest, whose MSE and MAE increased marginally.



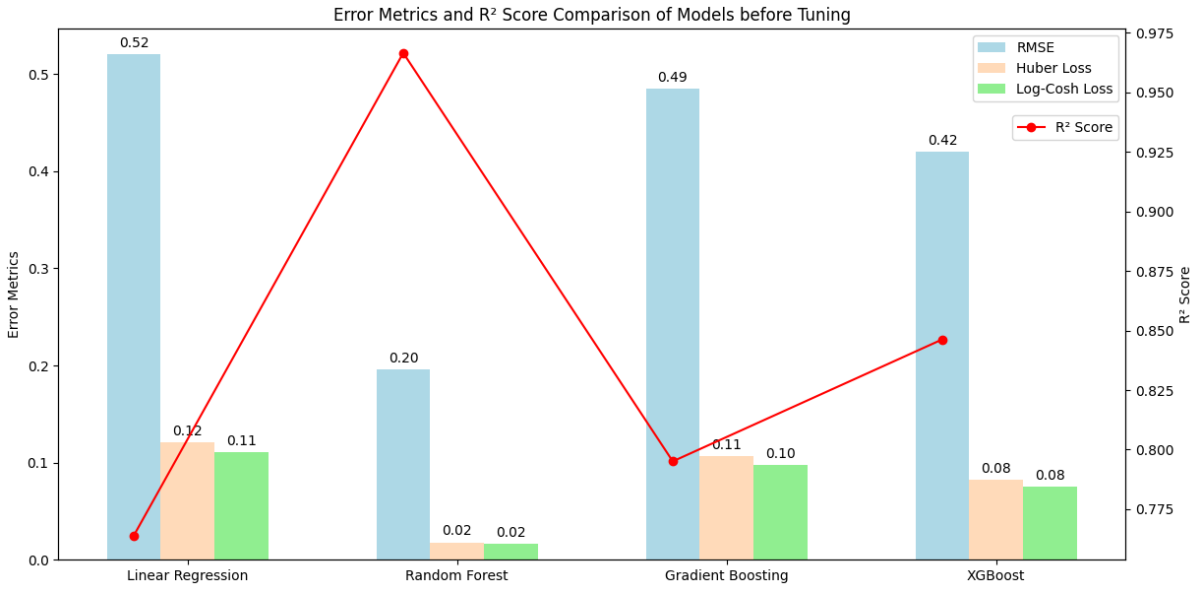
Post-tuning, XGBoost not only improved its error metrics but also enhanced its capacity to explain variance, increasing from 0.85 to 0.91, almost matching the pre-tuning prowess of Random Forest. The Median Absolute Error and Maximum Error for XGBoost also improved, dropping to 0.13 and 3.61, respectively, reflecting a tighter concentration of predictions around the true values and less variability in error magnitude. These enhancements highlight XGBoost's ability to adapt and optimize through tuning, ultimately making it the superior model for predicting movie success with high-dimensional and complex datasets, as evidenced by the after-tuning performance metrics.



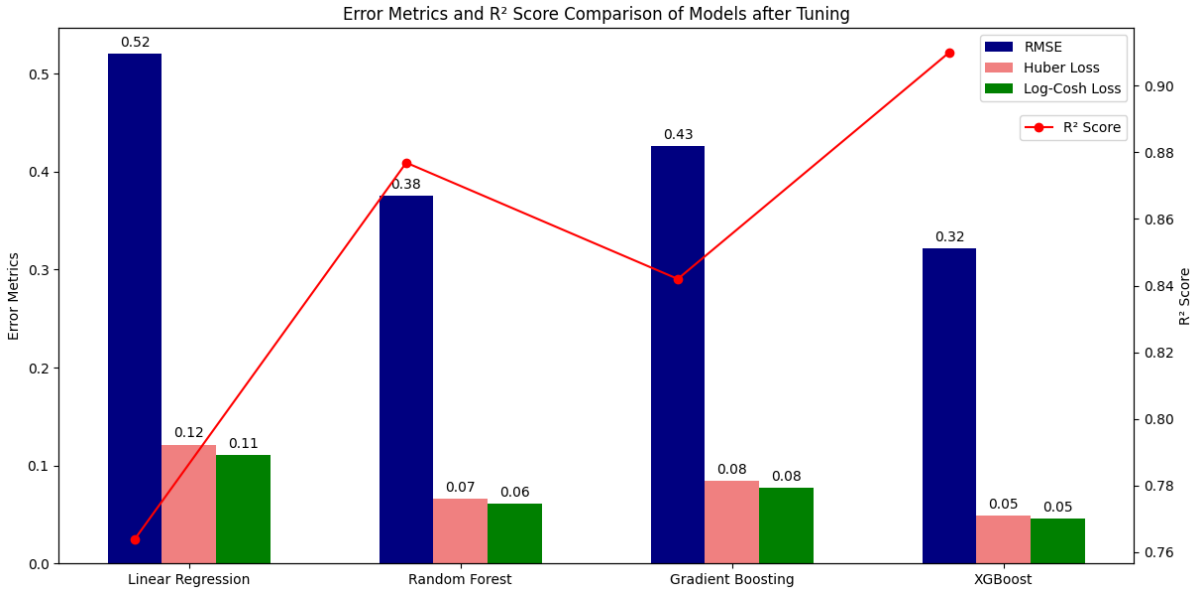
4.4.3. Performance Evaluation: Error metrics and R2

Before tuning, Random Forest had the lowest RMSE (0.20), Huber Loss (0.02), Log-Cosh Loss (0.02), and the highest R^2 Score (0.975), indicating it explained most of the variance. This aligns with prior studies on movie revenue prediction, where Random Forest is praised for capturing complex feature interactions (Delen, 2016; Sharda & Delen, 2006; Simonoff & Sparrow, 2000). After tuning, however, Random Forest's RMSE increased to 0.38, and its R^2 Score dropped to 0.88, suggesting suboptimal tuning.

XGBoost, initially performing moderately well (RMSE: 0.42, R^2: 0.85), saw significant improvements post-tuning. Its RMSE dropped to 0.32, and its R^2 increased to 0.90, confirming its superior performance in handling complex datasets (Ashish et al., 2018; Delen, 2016). XGBoost's robustness in handling missing data, outliers, and nonlinear relationships makes it a top performer in movie success prediction.

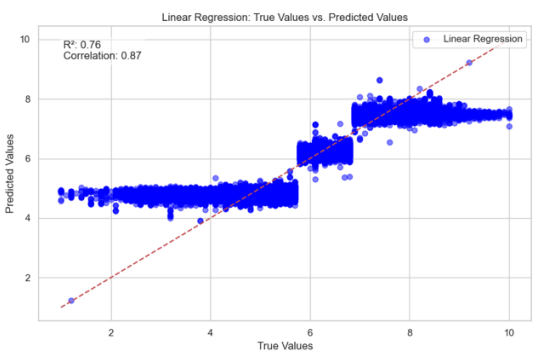


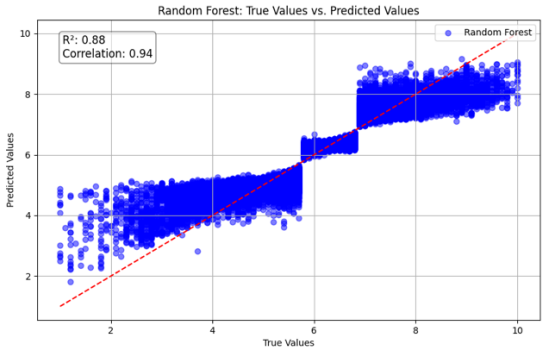
Gradient Boosting saw minor improvements, with RMSE dropping from 0.49 to 0.43 and R^2 remaining around 0.84. While competitive, it still lags behind XGBoost, which offers better efficiency and overfitting control (Sharda & Delen, 2006).

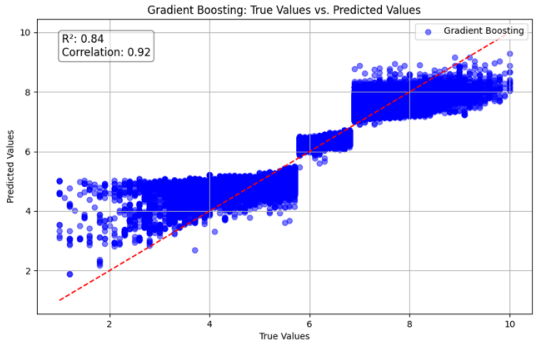


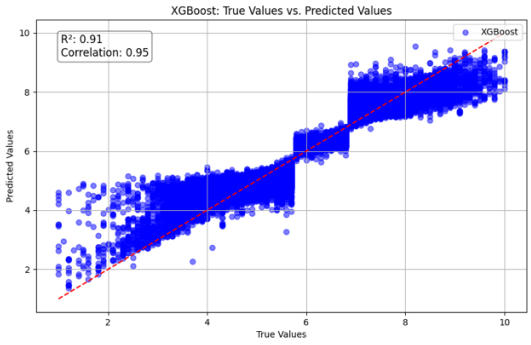
In line with previous studies, machine learning models like Random Forest and XGBoost outperform traditional methods like Linear Regression due to their ability to model complex, nonlinear interactions (Ashish et al., 2018). XGBoost stands out post-tuning for its robustness and scalability, making it ideal for tasks like box office prediction (Delen, 2016). These findings affirm that advanced models, particularly XGBoost and Random Forest, are better suited for predicting movie success than traditional regression techniques.

4.4.4. Performance Evaluation: Predicted Outcomes Compared to Actual Dat









The comparative performance of four machine learning models highlights varying degrees of accuracy in predicting movie success. Linear Regression shows moderate predictive capability with an R^2 of 0.76 and a correlation of 0.87, although it struggles with outliers and complex patterns. Random Forest and Gradient Boosting demonstrate stronger performances with R^2 values of 0.88 and 0.84 respectively, with Random Forest showing closer alignment to the diagonal indicating fewer deviations and a higher correlation of 0.94 compared to Gradient Boosting's 0.92. XGBoost outperforms all with the highest R^2 of 0.91 and a correlation of 0.95, showcasing minimal error and tight clustering along the diagonal, marking it as the most reliable model. This suggests that ensemble methods, particularly XGBoost, provide superior predictive accuracy and robustness against complex dataset patterns in the context of movie success forecasting.

5. Discussion

This discussion delves into the performance evaluation of machine learning models used to predict movie success, adhering to the comprehensive analytics presented in the findings section. We focus on dissecting the implications of each model's performance, guided by key metrics such as Mean Squared Error (MSE), R-squared (R^2), and others, alongside insightful visualizations. This examination not only clarifies the strengths and weaknesses of each model but also sets the stage for interpreting these results within the broader context of film industry analytics.

5.1. Implications of Findings

The varying performance of models like Linear Regression, Random Forest, Gradient Boosting, and XGBoost underscores their differing abilities to manage the complexities inherent in movie success prediction. As evidenced by the feature importance analysis, the Random Forest model highlighted "success_classification," "numVotes," "genres," and "runtimeMinutes" as significant predictors, revealing the critical role of these features in driving model accuracy. This aligns with existing literature, which emphasizes the importance of well-encoded categorical features and user engagement metrics (Sharda & Delen, 2006; Duan, Gu & Whinston, 2008).

The superior performance of XGBoost, especially after model tuning, illustrates its robustness in handling large, complex datasets, particularly where feature interactions are crucial. The insights provided by the feature importance analysis, such as the impact of temporal variables like "releaseYear" and "title_age," suggest that stakeholders in the film industry can optimize success predictions by focusing on trends in recency bias and audience attention (Tufekci, 2014). Additionally, understanding the predictive power of features like genre and runtime can guide strategic decision-making in film development and marketing campaigns, where careful selection of genre elements and audience-preferred runtimes could enhance box office performance.

The findings also show that feature importance can vary depending on the model used, which suggests that industry professionals must consider not only the best-performing model but also the specific features that contribute most significantly to movie success. This highlights the importance of continually refining these models based on evolving datasets and market trends, as tuning can dramatically influence their effectiveness (Hastie et al., 2009).

5.2. Integration with Existing Literature

The results from this study align well with previous research, confirming the role of certain predictors in forecasting movie success. For instance, Sharda & Delen (2006) pointed out the efficacy of ensemble methods like Random Forest in handling complex datasets, which our feature importance analysis corroborates. However, the exceptional performance of XGBoost challenges some of the earlier preferences for Random Forest, especially in the context of larger and more diverse datasets (Breiman, 2001). The finding that "success_classification" and "numVotes" are key drivers of movie ratings aligns with the literature on user behavior and social proof mechanisms in digital platforms (Duan, Gu & Whinston, 2008).

Additionally, our feature importance findings contribute to the body of work examining the influence of runtime and genre on movie reception (Lash & Zhao, 2016), offering new empirical support for these factors. This study also reflects advancements in machine learning techniques, such as the shift towards more scalable and powerful algorithms like XGBoost, adding to the growing research on the use of modern data science tools in entertainment analytics (Kuhn & Johnson, 2013).

5.3.Limitations and Future Research

Despite offering valuable insights, this study acknowledges several limitations. One notable limitation is the reliance on the IMDb dataset, which may introduce biases, particularly when considering global market trends or niche film genres that are underrepresented. Additionally, the sensitivity of the models to hyperparameter settings, as observed during the tuning process, suggests that the reported feature importance rankings could vary with different configurations.

Future research could address these limitations by incorporating additional data sources, such as social media sentiment and global box office performance, to enhance the predictive power of the models (Kumar et al., 2023). Exploring how feature importance shifts when these models are applied in different cultural or geographic contexts could offer deeper insights into region-specific factors that contribute to movie success (Simonoff & Sparrow, 2000). Testing alternative machine learning algorithms and integrating advanced techniques like deep learning could also refine the predictive accuracy and reveal more nuanced patterns in the data.

6. Conclusion

This research set out to explore the predictive power of machine learning models in forecasting movie success, leveraging a comprehensive dataset sourced from IMDb and applying advanced analytical methods. By using models such as Linear Regression, Random Forest, Gradient Boosting, and XGBoost, the study uncovered the most influential factors driving movie ratings, including "success_classification," "numVotes," "genres," and "runtimeMinutes." XGBoost stood out as the most effective model in handling the dataset's complexities, offering highly accurate predictions that could benefit key stakeholders in the film industry.

The analysis of feature importance revealed valuable insights into the specific predictors of movie success. This emphasizes the importance of carefully selecting features that align with the unique characteristics of the data. Not only did this research confirm that runtime, genre, and director involvement are critical success factors, but it also highlighted the importance of continuous model refinement and feature tuning to adapt to a dynamic industry. These insights will help filmmakers, producers, and marketers make data-driven decisions that optimize both financial returns and audience reception.

In looking forward, future research could expand the dataset and explore the application of these models in different cultural contexts. By incorporating additional variables such as social media engagement or economic indicators, and testing the models in various regions, the field of predictive analytics in the entertainment industry can continue to evolve. Ultimately, this study contributes both academically and practically, offering a foundation for future research and providing actionable insights for stakeholders aiming to predict and enhance movie success.

7. Reference

A, S., Glantz, K, B., Slinker, Torsten and Neilands (n.d.). *Primer of Applied Regression and Analysis of Variance, 3e | AccessBiomedical Science | McGraw Hill Medical*. [online] accessbiomedicalscience.mhmedical.com. Available at: https://accessbiomedicalscience.mhmedical.com/book.aspx?bookID=2117 [Accessed 17 Apr. 2023].

Akhmethanova, M.V. (2022). THEATER ART AND MODERN AUDIENCE: PECULIARITIES OF CREATIVE COOPERATION. *The European Journal of Humanities and Social Sciences*, (3), pp.99–103. doi:https://doi.org/10.29013/ejhss-22-3-99-103.

Altuğ Ocak (2023). The Science of Film Selection: Exploring Factors Behind Movie Preferences. *Özgür Yayınları eBooks*. doi:https://doi.org/10.58830/ozgur.pub311.c1364.

Amazon.com. (2019). *Applied Predictive Modeling: 8601421896931: Medicine & Health Science Books @ Amazon.com*. [online] Available at: https://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485 [Accessed 4 Mar. 2019].

Ashish, V., Kedia, V. and Singh, A., 2018. Movie revenue prediction using machine learning models. *International Journal of Computer Science and Information Technologies*, 9(2), pp.117-124.

Beyer, H. (1981). Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical Journal*, [online] 23(4), pp.413–414. doi:https://doi.org/10.1002/bimj.4710230408.

Breiman, L. and Schapire, R. (2001). Random Forests. 45, pp.5–32.

Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

C. Suganya and Vijayakumar, M. (2024). Shifting Discourse of Digital Entertainment in COVID-19. *Advances in multimedia and interactive technologies book series*, pp.45–55. doi:https://doi.org/10.4018/979-8-3693-0116-6.ch004.

Carlo Fanelli (2016). Vision and Imagination in the Renaissance Theatre. *Journal of Literature and Art Studies*, 6(2). doi:https://doi.org/10.17265/2159-5836/2016.02.004.

Chai, T. and Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, [online] 7(3), pp.1247–1250. doi:https://doi.org/10.5194/gmd-7-1247-2014.

Chang, S.-C. (2020). Market size matters? An approach to illustrate the market preference of Hong Kong-mainland China co-production cinema. *The Journal of International Communication*, pp.1–25. doi:https://doi.org/10.1080/13216597.2020.1728358.

Christensen, J. (2019). Effective Data Visualization: The Right Chart for the Right Data, and Data Visualization: A Handbook For Data Driven Design. *Technology|Architecture + Design*, 1(2), pp.242–243. doi:https://doi.org/10.1080/24751448.2017.1354629.

DeFelice, C., Porter, L. and Kim, S.-W. (2024). Moviegoing in the wake of a pandemic: Re-evaluating the attitudes, intentions, and behaviors of U.S. Moviegoers in the streaming era. *Journal of Media Economics*, pp.1–18. doi:https://doi.org/10.1080/08997764.2024.2361747.

Delen, D., 2016. Predicting movie box office success: A comparative analysis of machine learning models. *Journal of Marketing Analytics*, 4(2), pp.65-80.

Delen, D. and Sharda, R., 2006. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), pp.243-254.

Duan, W., Gu, B. & Whinston, A.B., 2008. The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *Journal of Retailing*, 84(2), pp.233-242.

Dwyer, R. (2002). Real and imagined audiences: Lagaan and the Hindi film after the 1990s.

Eliashberg, J., Elberse, A., & Leenders, M. A. A. M. (2006) highlighted that dramas tend to perform well because of their ability to emotionally connect with audiences, leading to better word-of-mouth and audience retention.

Eliashberg, J., Elberse, A. & Leenders, M.A.A.M., 2006. The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science*, 25(6), pp.638-661.

Eliashberg, J., Elberse, A. and Leenders, M.A.A.M., 2006. The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science*, 25(6), pp.638-661.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), pp.1189–1232. doi:https://doi.org/10.1214/aos/1013203451.

Gözde SUNAL and Pınar ÖZTARKAN ÖZYURT (2022). Genre Analysis of The Film Halloween Kills in The Context of Iconographic And Iconological Critical Method. *Intermedia International e-journal*, 9(16), pp.40–53. doi:https://doi.org/10.56133/intermedia.1105371.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Hennig-Thurau, T., Houston, M. B., & Walsh, G. (2001) found that audience preferences shift in response to external events, such as social upheavals, which could explain the rise in popularity of comedies in 2022.

Hennig-Thurau, T., Houston, M.B. and Walsh, G., 2001. Determinants of motion picture box office and profitability: An interrelationship approach. *Review of Managerial Science*, 2(1), pp.65-92.

Hinde, S., Smith, T.J. and Gilchrist, I.D. (2018). Does narrative drive dynamic attention to a prolonged stimulus? *Cognitive Research: Principles and Implications*, 3(1). doi:https://doi.org/10.1186/s41235-018-0140-5.

Huber, P.J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), pp.73–101. doi:https://doi.org/10.1214/aoms/1177703732.

Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), pp.679–688. doi:https://doi.org/10.1016/j.ijforecast.2006.03.001.

Jelena Budanceva and Svirina, A. (2023). Consumption of Cultural Content in the Digital Environment in the Post-Pandemic Latvia. *Economics and Culture*, 20(2), pp.76–87. doi:https://doi.org/10.2478/jec-2023-0017.

Khorsheed, A. (2023). The Multiple Directorial Visions of the Theatrical Text of Arthur Miller?s Play Death of a Salesman. *International journal of science and research*, 12(1), pp.554–557. doi:https://doi.org/10.21275/sr23115025230.

Kirk, A. (2016). *Data Visualisation*. [online] Google Books. Available at: https://books.google.co.uk/books/about/Data_Visualisation.html?id=ZrCJDAAAQBAJ&redir_esc=y [Accessed 7 Aug. 2024].

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. [online] New York, NY: Springer New York. doi:https://doi.org/10.1007/978-1-4614-6849-3.

Lash, M.T. & Zhao, K., 2016. Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, 33(3), pp.874-903.

Lee, K., Park, J., Kim, I. and Choi, Y. (2016). Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*, [online] 20(3), pp.577–588. doi:https://doi.org/10.1007/s10796-016-9689-z.

Liaw, A. & Wiener, M., 2002. Classification and regression by randomForest. *R News*, 2(3), pp.18-22.

Little, R.J.A. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*. [online] *Google Books*. John Wiley & Sons. Available at: https://books.google.co.uk/books?hl=en&lr=&id=BemMDwAAQBAJ&oi=fnd&pg=PR11&dq=Little.

Michelle, C., Davis, C.H., Hardy, A.L. and Hight, C. (2016). Pleasure, disaffection, 'conversion' or rejection? The (limited) role of prefiguration in shaping audience engagement and response. *International Journal of Cultural Studies*, 20(1), pp.65–82. doi:https://doi.org/10.1177/1367877915571407.

Mohanty, A., Aditi Mudgal and Ganguli, S. (2023). Mapping movie genre evolution (1994 – 2019) using the role of cultural and temporal shifts: a thematic analysis. *F1000Research*, [online] 12, pp.662–662. doi:https://doi.org/10.12688/f1000research.127008.2.

Muhammad, N., None Muhammad Sukriyatma, Akmal, D. and None Amata Fami (2023). Pengaruh Durasi Terhadap Retensi Audiens Dalam Motion Graphic Wajib Pajak Non Efektif. *Jurnal Riset Rumpun Seni, Desain dan Media*, 3(1), pp.30–41. doi:https://doi.org/10.55606/jurrsendem.v3i1.2333.

Navarathna, R., Carr, P., Lucey, P. and Matthews, I. (2019). Estimating Audience Engagement to Predict Movie Ratings. *IEEE Transactions on Affective Computing*, 10(1), pp.48–59. doi:https://doi.org/10.1109/taffc.2017.2723011.

Nordhausen, K. (2014). An Introduction to Statistical Learning-with Applications in R by Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani. *International Statistical Review*, 82(1), pp.156–157. doi:https://doi.org/10.1111/insr.12051_19.

Olney, A.M. (2013). Predicting Film Genres with Implicit Ideals. *Frontiers in Psychology*, 3. doi:https://doi.org/10.3389/fpsyg.2012.00565.

Richeri, G. (2016). Global film market, regional problems. *Global Media and China*, 1(4), pp.312–330. doi:https://doi.org/10.1177/2059436416681576.

Schröer, C., Kruse, F. and Gómez, J.M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181(1), pp.526–534.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243-254.

Sharda, R. & Delen, D., 2006. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), pp.243-254.

Shmueli, G. and Koppius, O. (2010). Predictive Analytics in Information Systems Research. *SSRN Electronic Journal*, 35(3). doi:https://doi.org/10.2139/ssrn.1606674.

Simonoff, J. S., & Sparrow, I. R. (2000) emphasized that while a higher number of creative contributors doesn't correlate with success, the reputation of these contributors, especially directors, is highly influential.

Simonoff, J. S., & Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3), 15-24.

Sun, L., Zhai, X. and Yang, H. (2020). Event marketing, movie consumers' willingness and box office revenue. *Asia Pacific Journal of Marketing and Logistics*, ahead-of-print(ahead-of-print). doi:https://doi.org/10.1108/apjml-09-2019-0564.

The glocalization of films and the cinema industry. (2022). *Edward Elgar Publishing eBooks*, pp.272–288. doi:https://doi.org/10.4337/9781839109010.00025.

Tufekci, Z., 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (ICWSM), pp. 505-514.

Willmott, C. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, [online] 30(1), pp.79–82. doi:https://doi.org/10.3354/cr030079.

Worthen, M.M. (2022). *Booksmart* (2019). *Film Matters*, 13(2), pp.118–121. doi:https://doi.org/10.1386/fm_00239_4.

Yaqoub, M., Jingwu, Z. and Ambekar, S.S. (2023). Pandemic impacts on cinema industry and over-the-top platforms in China. *Media International Australia*, p.1329878X2211459. doi:https://doi.org/10.1177/1329878x221145975.

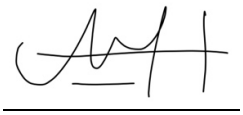
Zheng, A. & Casari, A., 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.

8. Appendix 1: Descriptive Statistics

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
tconst	804941	156789	tt5113044	108							
titleType	804941	1	movie	804941							
primaryTitle	804941	146145	Prey	179							
originalTitle	804941	148149	Gold	128							
isAdult	804941				0.003347	0.057755	0	0	0	0	1
releaseYear	804941				2019.853	2.40042	2016	2018	2020	2022	2024
runtimeMinutes	804941				98.39932	25.7032	1	87	99	108	1464
genres	804941	1057	Documentary	130841							
ordering	804941				8.138788	10.35385	1	2	4	10	108
title	804941	351054	Alone	528							
region	804941	227	US	249281							
language	804941	74	en	729950							
types	804941	11	imdbDisplay	618316							
isOriginalTitle	804941				0.194358	0.395601	0	0	0	0	1
directors	804941	102810	nm10992938	13298							
writers	804941	97883	nm10992938	109715							
averageRating	804941				6.104077	1.21083	1	5.6	6.104044	6.8	10
numVotes	804941				17698.36	63501.84	5	146	1565	17698.07	1500329

9. Appendix 2: Dissertation Checklist Sheet

A digitally signed copy of this form should be included as an appendix to the dissertation.

Name: Ambrish Muniraju Signature (Digital): 

Date Submitted: 10/09/2024

I confirm that my dissertation contains the following prescribed elements:

☑ My dissertation portfolio meets the style requirements set out in the MSc Business Analytics Portfolio Dissertation Handbook, including a word count on the front page of each element.

☑ I have reviewed the Turnitin similarity report prior to submission.

☑ My dissertation title captures succinctly the focus of my dissertation.

☑ My title page is formatted as prescribed in the MSc Business Analytics Portfolio Dissertation Handbook.

☑ The abstract provides a clear and succinct overview of my study.

☑ Each element contains a Table of Contents, and List of Figures and Tables (where appropriate).

☑ My dissertation contains a statement of acknowledgement (optional).

Introduction section:

☑ The Introduction section of the research report, at a minimum, covers each of the following:

- Background to/context of the project
- Research question(s), aim(s), and objectives
- Why the project is necessary/important
- A summary of the Methodology
- Outline of the key findings
- Overview of chapter structure of the remainder of the dissertation

Background section:

☑ The Background section of the research report covers the following:

- Synthesizes the key technical literature relating to the topic

- Synthesizes the key theoretical literature relating to the topic

Methodology section:

☑ The methodology section of the research report:

- Details the procedures adopted in carrying out the project (e.g. data source/acquisition, data processing, procedures for maximizing rigor and robustness, methods of data analysis, etc.)
- Contains ethical considerations and decisions

(Note: This section should not duplicate the technical report, which focuses more on the detailed technical choices and steps.)

Findings section:

☑ The findings section of the research report reports the results in detail and provides possible explanations for the various findings.

Discussion section:

☑ The discussion section of the research report makes appropriate linkages between the findings and the literature reviewed.

Conclusions section:

☑ The conclusions section of the research report includes:

- Conclusions about each research question and/or hypothesis
- General conclusions about the research problem
- Implications for theory, for policy, and/or management practice
- Limitations of the research
- Suggestions for practice and future research

Additional required elements:

☑ The technical report, log book, and reflective discussion have each been included.

☑ The reference list is in alphabetical order and follows the Harvard system.

☑ I have signed and dated the Candidate Declaration.

This confirms that all required tasks have been completed as part of the dissertation submission.