ID Cover Page

Summary of WP Student Team

# Big Data meets the Big Screen: A Comprehensive Machine Learning Study about the Movie Industry

**Group constitution:**

| Student Name | Program | Individual Title |
|---|---|---|
| Benedikt Tremmel | Business Analytics | Predictions from Movie Reviews Leveraging Bert: Exploring the Impact of Sentiment Polarity and Review Length |
| Justin Bruce Sams | Business Analytics | The Differential Impact of Critic and User Ratings on Box Office Success: Exploring the Roles of Seasonality, Star Power, and Studio Size |
| Zeno Eule | Business Analytics | Quantitative Analysis of Factors Driving Movie Sequel Success: A Predictive Modeling Approach |
| Tomás Shoemaker Simão Gonçalves | Business Analytics | Understanding Box-Office Dynamics: The Role of Timing and Audience in Movie Revenue Prediction |
| Luc Marcel Pellinger | Business Analytics | Temporal Modeling for Movie Success Prediction: A Comparative Study on the Applicability of different Deep Learning Models in Entertainment Analytics |

**Work project carried out under the supervision of:**

**Advisor**: Sreyaa Guha

**Reading observations for jury:**
If applicable

Note: this template is a support document for the jury members before the defense, only.

A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

# Big Data meets the Big Screen:

# A Comprehensive Machine Learning Study about the Movie Industry


Benedikt Tremmel, 60253

Justin Bruce Sams, 59279

Zeno Eule, 59093

Tomás Shoemaker Simão Gonçalves, 47016

Luc Marcel Pellinger, 58611


Work project carried out under the supervision of:

Sreyaa Guha

17/12/2024

# Abstract

This thesis investigates factors influencing movie success using machine learning and deep learning techniques. Traditional machine learning methods analyze key determinants of box office performance, such as audience and critic ratings, release timing, and sequel dynamics. Deep learning approaches, including Natural Language Processing and Time Series Classification, examine patterns in sequential movie data. By integrating diverse data sources and predictive modeling techniques across the movie lifecycle, this study provides a comprehensive perspective on audience engagement and market dynamics. The findings advance academic understanding of the topic and offer actionable recommendations for industry stakeholders aiming to optimize performance.

**Keywords:** Box Office Success, Movie Success Factors, Traditional Machine Learning, Natural Language Processing, Time Series Analysis

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

# 1. Introduction

The film industry uniquely blends cultural importance with economic impact (Dellarocas, Zhang, & Awad, 2007). For instance, Steven Spielberg's 2004 film *The Terminal* exemplifies how audience reception can significantly affect box office performance. Despite a substantial $110 million budget and a star-studded cast featuring Tom Hanks, the movie earned only $77 million, falling well short of expectations. Analysts attributed this underperformance largely to negative word-of-mouth from consumers, underscoring the critical role of audience-driven perceptions in determining financial outcomes.

Online reviews and ratings have further amplified the importance of consumer perceptions in the movie industry (Pang, Liu, & Golder, 2022). Online reviews are pivotal in shaping how films are evaluated and received by the mass audience. Studies show that 88% of Americans consider online reviews to be trustworthy and beneficial, associating them with improved reliability (Watson & Wu, 2022). For movie enthusiasts, these critiques serve as an essential decision-making tool, guiding their viewing choices and shaping their confidence in the movie they select (Loupos, Peng, Li, & Hao, 2023). The influence of reviews extends beyond audiences; exhibitors are also affected, as evidenced by a positive correlation between review ratings and the number of screens allocated to a film (Wang & Guo, 2017). This highlights the broader impact of consumer feedback on the industry's supply chain dynamics.

Economically, the sector is a major driver of employment and revenue, with over 400,000 individuals working in the U.S. motion picture and video game industries as of 2023 (Bureau of Labor, 2023). This highlights the movie industry's role as both an economic driver and a cultural influence in modern society.

The movie industry provides a rich data environment encompassing diverse data structures. While movie metadata is predominantly presented in a structured tabular format, it highlights key attributes such as cast, crew, genre, and release details, which can be readily analyzed using traditional data processing techniques. Furthermore, unstructured data, such as textual movie reviews, and sequential rating data over time offer significant opportunities for advanced analytics. These diverse data types allow for the application of sophisticated methods, such as natural language processing and time-series analysis, to extract valuable, data-driven insights into audience preferences, sentiment trends, and market dynamics.

This study employs advanced machine learning (ML) techniques to investigate the factors driving movie success across all stages of the film lifecycle: preparation, production, publicity, and release. The role of critic and user ratings are evaluated, considering seasonality, star power, and studio size in shaping box office outcomes. Release timing is analyzed for its strategic importance, particularly during high-demand periods, and the unique challenges of sequels are addressed by assessing factors like runtime, engagement, and reviews. Furthermore, natural language processing further complements the research to examine how textual reviews predict ratings, exploring the moderating role sentiment polarity and review length. Finally, multivariate time-series models are applied to predict post-release performance, emphasizing the ability of temporal data and machine learning to optimize decision-making and enhance financial and audience outcomes across the movie industry.

# 2. Literature Review

## 2.1. Defining Movie Success

Movie success has a twofold definition depending on the scope of stakeholder. While consumers are more drawn public perception in terms of ratings, movie producers and investors additionally focus on profitability and maximizing financial indicators (Dellarocas et al., 2007; Divakaran, Palmer, Sendergaard, & Matkovskyy, 2017). Therefore, measuring a movie's success involves various factors, including ratings and box office performance emerging as two central success metrics. Ratings, provided by both users and critics, serve as proxies for word-of-mouth, shaping audience perceptions and influencing their viewing decisions. Historically, capturing consumer word-of-mouth was challenging due to its transient nature in offline communities. However, internet-mediated platforms have revolutionized this dynamic by providing a persistent and publicly accessible record of consumer opinions. While these platforms offer valuable insights into audience sentiment, they also highlight the inherent subjectivity of ratings.

Moon, Bergey, & Iacobucci (2010) highlight the complexities of movie ratings, which are influenced by individual preferences, viewing histories, community opinions, and movie characteristics. Another study by Tsao (2014) further underscores the subjective nature of movie ratings, driven by individual biases and diverse preferences. Consumer reviews significantly impact both movie selection and post-viewing evaluations, but the variability in how audiences interpret and respond to positive or negative reviews reveals the inconsistencies in ratings. Factors such as individual taste, mood, and expectations influence ratings, making them a valuable yet subjective measure of success.

In contrast, box office performance provides a more objective and financial measure of success (Divakaran et al., 2017). Box office revenues not only reflect a movie’s profitability but also

influence strategic decisions, such as marketing and distribution. Accurately predicting box office success is critical for optimizing marketing efforts, such as determining the release date, setting the number of screens, and increasing audience awareness through targeted advertising.

Despite their differences, both ratings and box office performance face challenges as success metrics. Movies, as short lifecycle products, have limited windows for revenue generation, making accurate forecasting of early success critical for mitigating financial risks (Chung, Niu, & Sriskandarajah, 2012; C. Zhang, Tian, & Fan, 2022).

These dynamics highlight the importance of strategic planning and predictive modeling in the film industry. By addressing the challenges inherent in both ratings and box office performance, stakeholders can better navigate the risks and uncertainties of the highly competitive movie market, ultimately improving the chances of a film's success.

## 2.2. Factors Driving Movie Success

### 2.2.1. User and Critic Ratings in Movie Success

Critic ratings are evaluations provided by individuals or entities with expertise in a specific field (Pang et al., 2022). In cultural industries like movies, books, wine, and restaurants, critics act as intermediaries connecting producers with consumers. They are distinguished by their explicit expertise and authority in evaluating quality and taste within their respective fields.

User ratings are crowd-sourced evaluations shared by general audiences, often through online platforms (Divakaran et al., 2017). In the movie industry, these ratings allow individuals to express their opinions, preferences, and experiences regarding films, including those yet to be released. Movie-based online communities play a vital role in facilitating the exchange of these crowd-sourced opinions.

User ratings are crowd-sourced evaluations shared by general audiences, often through online platforms (Divakaran et al., 2017). In the movie industry, these ratings allow individuals to express

their opinions, preferences, and experiences regarding films, including those yet to be released. Movie-based online communities play a vital role in facilitating the exchange of these crowd-sourced opinions. The movie industry faces significant information asymmetry before a film's release, as consumers have limited knowledge about its quality (Pang et al., 2022). This gap often prompts audiences to consult external evaluations, such as user and critic ratings, to reduce uncertainty. Critics, regarded as experts in their field, connect producers with consumers by providing informed opinions based on their expertise. Similarly, online communities serve as platforms for users to share their opinions and experiences, offering valuable crowd-sourced information (Divakaran et al., 2017). The persistent and publicly accessible nature of online ratings allows potential viewers to form opinions based on the experiences of both peers and experts (Dellarocas et al., 2007).

Studies have explored the relationship between critic and user ratings, revealing a moderate positive correlation between the two (Dellarocas et al., 2007). While there is alignment in some respects, the relatively low correlation highlights the distinctiveness of these sources of information. Critics' evaluations are rooted in professional expertise, yet they may also reflect personal biases or genre-specific preferences (D'astous, Montreal, Touil, & Tunisia, 1999). On the other hand, user ratings are dynamic, changing over time as new viewers contribute their opinions, making them more reflective of audience sentiment throughout a movie's lifecycle (Divakaran & Nørskov, 2016). These differences justify treating critic and user ratings separately, as they provide complementary perspectives (Dellarocas et al., 2007).

Reliance on critic or user ratings depends on the timing and availability of information. When information is limited, such as before or during a movie's release, consumers often turn to critics for their expertise and structured evaluations (Flanagin & Metzger, 2013). However, as more user-

generated content becomes available, consumers shift their reliance toward peer reviews, which are considered more relatable (Divakaran & Nørskov, 2016; Flanagin & Metzger, 2013).

### 2.2.2. Effect of Electronic Word-of-Mouth (eWOM) in the Movie Industry

Word of Mouth (WOM) and the subcategory of Electronic Word-of-Mouth (eWOM) are common topics of interest in marketing and consumer behavior studies (H. Zhang, Yuan, & Song, 2020). Both address the importance of consumer resumes and feedback for a product's success as it affects product preferences and purchase decisions made by other consumers. Particularly, volume and valence are KPIs of interest within the field of WOM research. They have been demonstrated to impact the success of products and services. For movies, researchers have shown that eWOM can significantly impact a movie's success. Duan, Gu, & Whinston (2008) identified WOM volume (total number of reviews and ratings) as the key driver of the box office revenue of a movie as increased WOM volume. In contrast, WOM's valence (the sentiment of reviews) showed a rather indirect impact on the box office. They also identified a dynamic interaction and positive feedback mechanism between WOM and box office revenue, meaning that the box office revenue can influence the WOM volume, which subsequently can influence the box office revenue again. While their study focused on discussing and investigating the dynamics of WOM and box office revenue, they have yet to delve into how specific patterns can be used to inform detailed marketing strategies of decisions in real time. However, the study of H. Zhang et al. (2020) focused on the role of marketing activity and eWOM in movie diffusion in both pre- and post-release phases. They investigated the role of advertising, WOM, and eWOM on movie diffusion (H. Zhang et al., 2020) by segregating advertising, eWOM, and viewers into two different groups. Users were divided into innovators and imitators, while advertising and eWOM were divided into pre-release and post-release. Their research indicated that pre-release advertising was more effective in attracting the innovator's group, while post-release advertising and positive post-release eWOM were more

influential for imitators. What stood out in their study was that the innovator's group is sensitive to negative pre-release eWOM. In contrast, negative post-release eWOM could increase their interest due to heightened awareness. While their study highlighted the role and effect of eWOM on the movie's performance, it did not focus on the dynamic effect by incorporating temporal data into their model. Moreover, they did not conduct their study model's performance across multiple channels, leaving the difference between different platforms unexplored. Recently, Delre & Luffarelli (2023) found that eWOM significantly impacts the box office revenue, peaking in the early stage of a movie's release and then declining over the cycle of the movies. Mainly, they investigated interaction effects of eWOM volume and valence on the box office sales, indicating that the features show a positive effect during the first 4 weeks after a movie's box office release. They focused on analyzing data from IMDb and Rotten Tomatoes to build a regression analysis that focuses on trends and effects rather than the overall task of modeling success prediction using machine learning based on instance-level temporal and textual features. This indicates that their insights could be used to formulate a predictive model for multivariate TSMC.

### 2.2.3. Influence of Major Studios on Film Performance

Major studios often act as a quality signal for consumers, as their established reputation and resources suggest a higher likelihood of delivering high-quality productions (Kraft & Rao, 2024). The uncertainty surrounding quality is highest for movies produced by non-major studios. Regression analysis confirms this, showing that returns to signaling are most significant for non-major studio productions compared to major studios and foreign productions.

Major studios like Disney and Warner Bros. are likely to dominate the market through their superior resources, extensive marketing budgets, and robust distribution networks (Pang et al., 2022). According to Basuroy (2006) these elements enhance their ability to signal quality effectively, influencing consumer perceptions and box office performance. If a studio's signals, such as high

upfront investments, are credible, they can positively shape perceived quality and, consequently, box office revenues. Building on these insights, we propose that non-major productions face greater quality uncertainty and are likely to have weaker distribution networks, resulting in less user-generated content about their movies.

### 2.2.4. Seasonality Effects on Box Office Revenues

Seasonality significantly influences box office performance, with consumer demand varying throughout the year (Einav, 2007; Simonton, 2009). Big-budget films are strategically released during high-demand periods, such as early summer and the Christmas season, to maximize audience reach and revenue. Analysis revealed that movies released during peak seasons often benefit from higher audience turnout, driven by factors such as school holidays, increased leisure time, and reduced work commitments (Ahmad, Duraisamy, Yousef, & Buckles, 2017). In contrast, box office revenues tend to decline during the off-season, particularly from mid-summer to early September . To account for these variations, researchers can include seasonality as a control factor, recognizing that differences in release timing can have substantial effects on a movie's performance (Karniouchina, Carson, Theokary, Rice, & Reilly, 2023).

### 2.2.5. Strategic Release Timing and Its Impact on Revenue

Numerous studies have identified release timing as a critical determinant of box-office performance. Ryoo, Wang, & Lu (2021) analyzed the impact of release timing across various periods, finding that holiday releases consistently outperformed those during non-holiday windows. Their study attributed this phenomenon to increased audience availability and leisure time during holidays, which boost theater attendance. The authors also highlighted that weekend releases – particularly those combined with holidays – maximize revenues, as they capture both opening-day buzz and sustained attendance over an extended period. Despite these findings, the authors primarily focused on family-oriented films and did not explore the broader implications for

other audience demographics. Future research could build on these findings by systematically comparing timing effects across different timeframes.

### 2.2.6. Impact of Star Power on Audience Engagement

In this study, star power is measured by the number of followers an actor has on the social media platform Instagram. Actors with significant followings often attract audiences due to their established popularity, creating a strong draw for potential viewers (Divakaran & Nørskov, 2016). This form of social influence contributes to shaping audience expectations and quality perception. User-generated content, such as crowd-sourced ratings, captures a broader spectrum of audience characteristics, including preferences, past experiences, and social influences, compared to critic evaluations(Einav, 2007; Simonton, 2009). While critics emphasize technical and artistic elements, ratings from users reflect mass-market sentiment, making them more representative of general audience opinions. Social influences, such as the popularity of major stars, further enhance the relevance of consumer evaluations, as these ratings often strongly correlate with box office performance.

### 2.2.7. Brand Extension and Success in Movie Sequels

Sequels play a special role in the context of determining factors influencing movie success. Building on the understanding of general success factors, sequels represent a distinctive case where elements such as brand equity take on heightened importance. As a form of brand extension, sequels not only capitalize on the value of a successful original work but also set expectations for audiences, solidifying their position as reliable box office performers. This dual role underscores their strategic significance in the motion picture industry, where leveraging a well-established brand often translates into consistent audience engagement and profitability (Moon et al., 2010).

The reputation of the original film shapes audience expectations, influencing decisions to engage with the sequel (Belvaux & Mencarelli, 2021) . This reputation is often reinforced through creative

continuity, such as retaining directors, writers, and key cast members. Consistency in the creative team preserves the narrative and tonal coherence of the franchise, strengthening audience trust and delivering on promises established by the original (Moon et al., 2010). Similarly, maintaining genre consistency helps balance audience expectations for familiarity and innovation. Successful franchises frequently reward audiences with storytelling that evolves while adhering to established thematic and stylistic boundaries (Belvaux & Mencarelli, 2021).

Conversely, deviations in creative leadership or genre can disrupt established audience expectations. Changes in directors, writers, or cast members risk breaking the emotional connection audiences feel with the franchise, creating uncertainty about the sequel's quality. Similarly, significant shifts in tone or genre may alienate loyal fans while failing to attract new audiences, leading to what Hennig-Thurau, Houston, & Heitjans (2009) describe as "brand fragmentation." Managing these risks requires a careful balance between continuity and novelty to sustain long-term audience engagement and franchise momentum.

Despite the recognized importance of creative and genre continuity, gaps remain in understanding how these elements interact with broader market forces. Factors such as release timing and audience engagement strategies may amplify or mitigate the impact of brand extension dynamics, yet these interdependencies are underexplored (Lehrer & Xie, 2021). By addressing these gaps, this study seeks to provide a more comprehensive framework for understanding the mechanisms underpinning sequel success within the context of brand extension.

### 2.2.8. Key Factors Influencing Sequel Performance

Sequel success is shaped by measurable factors such as runtime, release timing, and audience engagement metrics, each contributing uniquely to performance outcomes (Belvaux & Mencarelli, 2021; Moon et al., 2010).

Runtime is a significant predictor of absolute box office performance, often seen as a marker of production investment and narrative depth, which attract audiences and drive revenue (Moon et al., 2010). Empirical studies found a positive relationship between runtime and revenue, with longer films typically generating higher earnings (Lehrer & Xie, 2021). However, excessively long runtimes can alienate some audience segments due to pacing issues or time constraints, underscoring the need for balance (Navarathna, Carr, Lucey, & Matthews, 2019). Despite these challenges, runtime remains a critical factor in shaping audience perceptions and marketability (Lash & Zhao, 2016).

Release timing is another crucial determinant. Shorter intervals between an original film and its sequel help maintain audience interest and preserve emotional connections with the original (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009). Conversely, longer gaps risk diminishing engagement, especially in competitive markets with an abundance of new content. On the other hand, excessively short intervals may lead to franchise fatigue, where audiences perceive sequels as rushed or redundant (Moon et al., 2010). Striking the right balance between sustaining interest and avoiding overexposure is crucial for maximizing sequel success.

Audience engagement metrics, such as review volume and sentiment, are also critical predictors. Research shows that engagement volume—reflected in reviews, comments, and ratings—correlates strongly with box office performance, as it signifies heightened audience awareness and interest (Navarathna et al., 2019). While sentiment provides additional context, it typically exhibits weaker correlations with revenue compared to engagement volume (Song, Huang, Tan, & Yu, 2019). These findings underline the importance of audience participation as a driver of success.

Despite the importance of these factors, gaps remain in understanding their interdependencies. For example, the impact of audience engagement may vary depending on release timing, yet such relationships are rarely explored (Lehrer & Xie, 2021). By examining runtime, release timing, and

audience engagement collectively, this study helps to provide a better understanding of the drivers of sequel success and their interrelationships.

## 2.3. Machine Learning in the Movie Industry

### 2.3.1. Importance of Accurate Predictions in the Movie Industrie

The prediction of box office performance is a critical challenge in the film industry, as the financial stakes of high-budget productions are significant (Lash & Zhao, 2016). Accurate forecasting allows stakeholders - including studios, investors, and distributors - to allocate resources efficiently, reduce uncertainty, and optimize returns (Lehrer & Xie, 2021). As films continue to grow in production and marketing costs, understanding the drivers of success has become an essential focus for both theoretical research and practical decision-making.

Box office predictions are particularly influenced by a variety of factors, including production budgets, marketing strategies, release timing, star power, and audience engagement (Song et al., 2019). These determinants interact in complex ways, influenced by dynamic market conditions and evolving consumer preferences. Advances in data analytics have significantly improved forecasting models, integrating structured variables (e.g., runtime, genre, and distribution strategy) with unstructured data, such as social media activity and audience sentiment, to provide real-time, actionable insights.

Among the wide array of movie types, sequels occupy a unique position within the industry. Building on the brand equity of their predecessors, sequels often benefit from pre-existing audience loyalty and recognition, making them a popular strategy to mitigate financial risk. This has contributed to sequels consistently ranking among the highest-grossing films globally (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009). However, sequels also present distinctive challenges for prediction. Unlike standalone films, they must balance continuity with novelty - preserving the narrative and thematic elements that resonated in the original while offering fresh

appeal to attract new viewers. Audience expectations, shaped by the first installment, create additional pressure for sequels to perform both critically and commercially, complicating forecasting efforts (Hennig-Thurau et al., 2009).

While many models focus on absolute metrics such as total box office revenue, these fail to capture the nuanced performance measures critical for sequels. Relative success metrics, which assess a sequel's performance compared to its predecessor, provide deeper insights into franchise sustainability and audience retention. For instance, a sequel that generates substantial revenue might still underperform if it fails to exceed the original's success, potentially signaling declining interest in the franchise (Hennig-Thurau et al., 2009).

This comprehensive perspective on predicting box office performance aims to enhance theoretical understanding and inform practical applications in an increasingly globalized and competitive market.

### 2.3.2. Traditional Machine Learning in the Movie Industry

The least squares method is a foundational tool for analyzing data trends, often visualized as a scatterplot where data points exhibit a roughly linear pattern (Burton, 2021; Fox, 2016). This method identifies the "line of best fit," minimizing the sum of squared deviations between each data point and the fitted line, ensuring the smallest total error and providing a precise linear approximation of the dataset. Building on this principle, Ordinary Least Squares (OLS) regression fits a regression plane to data points with a linear trend, capturing partial relationships between regression coefficients and the dependent variable while considering the influence of other variables (Burton, 2021; Fox, 2016). Although widely used, OLS regression relies on several key assumptions, such as linearity, independence of errors, homoscedasticity, normality of errors, and low multicollinearity among predictors. Violating assumptions can lead to biases, invalid

significance tests, and unreliable predictions, underscoring the importance of verifying these criteria.

OLS regression remains a cornerstone of statistical modeling, offering a robust framework for understanding linear relationships and deriving interpretable results. Its applicability and clarity make it a valuable tool, particularly for initial analyses and straightforward relationships. However, in the context of box office forecasting, where nonlinear interactions and complex predictor relationships frequently arise, complementary methods can provide additional insights. For example, factors such as production budgets, cast popularity, and release timing often exhibit interactions or nonlinear effects that may not be fully captured by linear models (Lash & Zhao, 2016; Somlo, Rajaram, & Ahmadi, 2011). Incorporating advanced techniques, such as machine learning algorithms or hybrid approaches, can help address these challenges while building on the solid foundation provided by OLS (Lash & Zhao, 2016). This combination of methods ensures both interpretability and adaptability to complex datasets, as shown by studies that explore the integration of econometrics and predictive analytics to capture richer patterns in data (Lash & Zhao, 2016; Lehrer & Xie, 2021).

The growing availability of granular data from digital platforms and social media further highlighted the need for more sophisticated methods. Machine learning techniques, such as decision trees, support vector machines, and ensemble methods, have proven effective in addressing these limitations. These methods handle high-dimensional data adeptly, uncovering complex patterns among predictors such as runtime, reviews, release timing, and audience engagement (Lehrer & Xie, 2021; L. Liao & Huang, 2021). Additionally, their ability to incorporate unstructured data, like social media interactions, provides real-time insights into audience behavior—capabilities that traditional econometric models lack (Song et al., 2019).

Hybrid approaches that integrate econometric models with machine learning have emerged as a powerful solution, combining structured and unstructured data to enhance forecasting accuracy. By addressing multicollinearity and leveraging machine learning's flexibility, these models produce predictions that are both precise and theoretically grounded (Lehrer & Xie, 2021). However, many studies remain U.S.-centric, focusing primarily on absolute success metrics and neglecting relative measures that provide deeper insights into franchise sustainability and audience retention (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009).

This study builds on these advancements by employing machine learning and hybrid approaches to predict both absolute and relative success metrics. Incorporating predictors like engagement metrics, runtime, and release timing alongside international data, it seeks to uncover the interdependencies that shape sequel performance. By addressing these gaps, it contributes to a more comprehensive predictive framework, enhancing theoretical understanding and supporting practical decision-making.

### 2.3.3. Deep Learning Models in the Movie Industry

Deep learning (DL), a specialized subfield of machine learning (ML), marks a major milestone in the advancement of artificial intelligence (AI) technologies (Darwish, Hassanien, & Das, 2020). By utilizing multiple processing layers to model complex and abstract patterns in data, DL has become a foundational component of modern AI applications. Its rapid progress has been fueled by the availability of large datasets, the development of powerful computational architectures, and advancements in algorithms that enable efficient learning at scale (Sastry et al., 2024). These factors have allowed DL to address highly complex problems across a wide range of industries. In the movie industry, DL has made significant contributions, transforming processes such as visual effects creation, audience analytics, and personalized content recommendation (Mühling et al., 2017; Tahmasebi, Ravanmehr, & Mohamadrezaei, 2021; Zhou, Zhang, & Yi, 2019). DL algorithms

enhance professional media production by automating tasks like labeling video content, recognizing faces, and identifying similar visuals, making video analysis and retrieval more efficient and streamlined. DL has also advanced recommender systems that incorporate user behaviors and social influence to enhance the personalization of content. Furthermore, in the film industry, deep learning models that analyze movie posters and other data have shown great success in predicting box-office revenues, helping to reduce financial risks.

Moreover, DL significantly influences related fields like Natural Language Processing (NLP) (Banbhrani, Xu, Soomro, Jain, & Lin, 2022) and Time Series Prediction , which face challenges in processing sequential data due to its complexity and temporal dependencies. By effectively capturing sequential patterns, DL has enabled breakthroughs in these fields, with notable applications extending to the movie industry as well (Y. Liao et al., 2022).

## 2.4. Natural Language Processing

The rise of digital platforms has led to an overwhelming amount of user-generated content, such as textual movie review. These reviews are a rich source of insights into audience opinions, but the sheer volume makes it challenging to analyze them manually. Natural Language Processing (NLP), a subfield of AI, has become an essential tool for addressing this challenge by enabling computers to understand and interpret human language (Banbhrani et al., 2022). NLP focuses on breaking down and analyzing text, transforming it from unstructured data into a format that machines can process (M. T. Khan et al., 2016).

### 2.4.1. Sentiment Analysis

One particularly useful application of NLP is sentiment analysis, which identifies the emotional tone of text, whether it's positive, negative, or neutral (Wadawadagi & Pagi, 2020). Sentiment analysis uses machine learning algorithms to process reviews and classify them based on the opinions expressed. This approach is widely explored in the movie industry with many different

applications (Berger, Kim, & Meyer, 2021). Danyal et al. (2024) conducted research focusing on extracting subjective information from film critiques, such as the reviewer's overall sentiment, the film's strengths and weaknesses, and viewing recommendations. Their study utilized sophisticated language models to analyze datasets from IMDD and Rotten Tomatoes. Their results demonstrated that sentiment analysis can effectively discern emotions and attitudes expressed in film critiques, providing valuable insights into subjective opinions and improving the ability to summarize movie reviews. By identifying trends in sentiment, this method helps filmmakers and marketers understand how a movie resonates with its audience. What makes sentiment analysis particularly valuable is its ability to handle large-scale data and uncover insights that might not be immediately obvious. For example, the results from J. H. Lee, Jung, & Park (2017) highlighted the feature's importance pinpointing that sentiment is helpful in predicting movie success metrics like ratings or box office success. Overall, sentiment analysis bridges the gap between vast amounts of audience feedback in an unstructured data format. By combining NLP and ML techniques, it offers a scalable, efficient way to understand audience opinions and improve decision-making in the movie industry.

### 2.4.2. Rating Prediction

Sentiment analysis, while widely useful, tends to oversimplify the nuanced opinions found in complex movie reviews (M. T. Khan et al., 2016). In contrast, rating prediction represents a more sophisticated approach within NLP, as it translates textual feedback into precise numerical ratings, providing a detailed and accurate quantification of user evaluations (Ahmed & Ghabayen, 2022). Star ratings in online reviews provide a standardized measure of perceived quality, with higher scores reflecting greater satisfaction. Converting textual reviews into quantifiable ratings enhances accessibility and interpretability for industry stakeholder, enabling more structured analysis of feedback (Archak, Ghose, & Ipeirotis, 2011). Rating prediction methods are used to predict the

score someone might give to a movie by analyzing the sentiment and content of their written review (Z. Y. Khan, Niu, Sandiwarno, & Prince, 2021). By converting subjective language into standardized satisfaction metrics, these methodologies play a crucial role in quantifying opinions within online reviews.

Recent advancements in machine and deep learning have significantly improved rating prediction by using complex architectures capable of recognizing intricate textual patterns. Deep learning utilizes artificial neural networks with multiple layers to analyze complex patterns in data, often exceeding the capabilities of traditional machine learning techniques (Janiesch, Zschech, & Heinrich, 2021). Each layer plays a specific role, with input layers receiving raw data, hidden layers extracting and transforming features, and output layers delivering the final prediction or result. Generally, the more layers a network has, the deeper it is, which often enhances its ability to learn and achieve better results, especially for complex tasks.

Chambua & Niu (2021) demonstrated the effectiveness of deep learning models in rating prediction, highlighting their ability to detect subtle emotional cues in text. Similarly, Ahmed & Ghabayen (2022) found that a Bidirectional Gated Recurrent Unit (Bi-GRU) model outperformed traditional methods, showcasing the strength of neural architectures in linguistic feature extraction. A Bi-GRU, an enhanced Recurrent Neural Network (RNN), processes data in both forward and backward directions, allowing it to capture contextual information from both past and future words, which improves pattern recognition in NLP tasks (He et al., 2020). Despite advancements in deep learning, challenges in handling domain-specific complexities and limited data have driven interest in innovative approaches like transfer learning.

#### 2.4.2.1. Transformative Approaches: Transfer Learning

The emergence of transfer learning has further advanced NLP by allowing pre-trained models to adapt effectively to specialized tasks with minimal labeled data (Galal, Abdel-Gawad, & Farouk,

2024). Transfer learning models leverage extensive general knowledge gained from large datasets, reducing reliance on large task-specific datasets and improving performance in tasks like rating prediction (Mosin, Samenko, Kozlovskii, Tikhonov, & Yamshchikov, 2023). Fine-tuning is a critical step in transfer learning, where pre-trained models are adapted to the specific requirements of a target task. By leveraging large-scale pre-trained models, which would often be impractical or infeasible to train independently, fine-tuning aligns these models with the nuances of the target task. This approach not only enhances performance but also demonstrates the efficiency of transfer learning in addressing specialized challenges in NLP, enabling the application of powerful models without requiring extensive computational resources or massive task-specific datasets. The foundation for many transfer learning models can be traced back to the introduction of the transformer architecture in the groundbreaking paper *Attention Is All You Need* (Vaswani et al., 2017). This architecture, based on a self-attention mechanism, revolutionized NLP by allowing models to better capture contextual relationships within text. Self-attention helps the model focus on the most important parts of a sequence, such as specific words in a sentence, enabling it to understand context more effectively and make more accurate predictions compared to traditional DL models. Following this breakthrough Bidirectional Encoder Representations from Transformers (BERT) emerged as a state-of-the-art NLP model (Devlin, Chang, Lee, & Toutanova, 2019). BERT's transformer architecture uses bidirectional self-attention, which evaluates each word in a sentence in relation to all others, providing a more comprehensive understanding compared to traditional unidirectional models. BERT's extensive pretraining on datasets like BookCorpus (over 11,000 books) and Wikipedia (2.5 billion words) equips it with deep language knowledge (Rajapaksha, Farahbakhsh, & Crespi, 2021). BERT's strength lies in its ability to understand complex semantic relationships within text, and this can be effectively leveraged by fine-tuning the model for task-specific challenges (Aljrees et al., 2024). Fine-tuning enables BERT

to adapt its extensive pre-trained knowledge to specialized domains, aligning its capabilities with the unique requirements of tasks like movie review rating prediction. By tailoring the model to the specific language patterns and context of the target dataset, fine-tuning maximizes BERT's potential, making it particularly effective for addressing domain-specific applications with high performance (Mosin et al., 2023).

#### 2.4.2.2. Refining Rating Prediction in the Movie Industry

The analyzed literature offers an extensive examination of rating prediction in multiple domains, highlighting the utilization of textual reviews to enhance precision. Ahmed & Ghabayen (2022) examine product and service evaluations from platforms like Amazon and Yelp, illustrating the efficacy of textual data in forecasting ratings for consumer products. Chambua & Niu (2021) additionally provide a comprehensive analysis of rating prediction methods for various consumer products and services. Aljrees et al. (2024) redirect attention to mobile application evaluations from the Google Play Store, highlighting the discrepancies between written reviews and numerical ratings. Although rating prediction in the movie review domain is relatively underexplored, sentiment analysis has been thoroughly examined. Rating prediction has been well explored for products, services, and mobile apps, but remains under-researched in the movie review sector. A key challenge is the high subjectivity of online reviews (Park, Song, & Sela, 2023). Unlike sentiment analysis, which captures general emotions, rating prediction provides a precise, quantitative assessment of nuanced evaluations. Addressing this gap can offer deeper insights into a film's reception for industry stakeholders.

## 2.5. Time Series Analysis

Similar to NLP, Time series prediction builds upon the sequential nature of data by introducing a temporal dimension, enabling the examination of changes over time (Ismail Fawaz, Forestier, Weber, Idoumghar, & Muller, 2019). Time series can be categorized as univariate or multivariate

sequences. Univariate sequences focus on a single feature, while multivariate sequences incorporate multiple features, which are often referred to as multiple channels (X. Chen et al., 2024). The scientific domain of time series analysis encompasses two primary tasks: time series forecasting (TSF) (Haben, Voss, & Holderbaum, 2023) and time series classification (TSC) (Ismail Fawaz et al., 2019; H. Liu et al., 2024). TSF involves predicting the future behavior of variables over specified time intervals, whereas TSC aims to assign labels to samples based on detected temporal patterns. TSF and TSC often involve high-dimensional data that challenges traditional ML methods, highlighting the advantages of deep learning for complex tasks in the movie industry (Y. Liao et al., 2022). For example, DL enables multimodal approaches to predict movie opening weekend revenue or success (Y. Liao et al., 2022; Madongo & Zhongjun, 2023). Similarly, Rhee & Zulkerine (2016) demonstrated how combining movie features like metadata, ratings, and user reviews in neural networks can predict box office profitability. Modeling approaches for time series often rely on foundational neural network architectures, such as recurrent neural networks (RNNs), including long short-term memory (LSTM) models (Ghanbari & Borna, 2021), or graph neural networks (GNNs) (Liu et al., 2024), both of which are typically implemented within a supervised learning framework. However, recent advancements have expanded time series analysis to accommodate big data scenarios where labeled samples are sparse or irregular (X. Chen et al., 2024). These developments leverage weakly-supervised learning paradigms, such as multiple instance learning (MIL) (Bing & Wang, 2017; Chen et al., 2024; Early et al., 2024; Tibo et al., 2020). Weakly-supervised learning is a machine learning approach where models are trained with incomplete, inexact, or noisy labels, contrasting with the fully labeled datasets typical of supervised learning (Ren, Wang, & Zhang, 2023).

### 2.5.1. Multiple Instance Learning for Movie Success Prediction

MIL is a deep learning approach that has been successfully applied in various domains. Carbonneau, Cheplygina, Granger, & Gagnon (2018) defined Multiple Instance Learning (MIL) as a form of weakly-supervised learning approach that is structured as a binary problem where the training set is composed of labeled bags, each containing multiple instances, but where labels are assigned to the bags rather than the individual instances. Based on the standard MIL assumption, a bag is considered positive if at least one instance within the bag is positive. For instance, Bing & Wang (2017) applied it for such as Breast Ultrasound Image Classification. Bakdi, Kristensen, & Stakkeland (2022) applied it for an intelligent predictive maintenance system in ship electric propulsion systems. Frameworks such as Early et al. (2024) proposed MIL for time series classification to make temporal dependences more distinguishable. Their approach addressed the issue that supervised TSC methods, often considered as a black-box, lack interpretability, which is why they proposed a novel framework to overcome this limitation and showcased the effectiveness of their weakly-supervised modeling approach via application across multiple domains. According to their writing, the framework offers an inherent mechanism to provide interpretable insights, indicating that the MIL framework can produce interpretable insights for temporal features. Picking up on this work, X. Chen et al. (2024) proposed a novel MIL model for multivariate time series data, demonstrating significant improvements in model interpretability due to the model's attention layer. This is particularly useful when predicting movie success, as it helps to understand the model's decision-making process. They pointed out how their proposed model could be extended to integrate multi-channel information, possibly by integrating cross-channel temporal attention or positional encoding.

### 2.5.2. The Role of Sentiment Analysis in MIL

As outlined before, WOM features are useful when it comes to predicting financial KPIs as well as ratings. Review sentiment specifically are crucial drivers to predict ratings and corresponding movie success, increasing sales in the box office. Considering the versatility of MIL approaches across domains, recent works within the field of MIL indicate that the weakly-supervised approach is extendable to fields other than medical appliances or predictive maintenance, such as sentiment analysis of movie reviews. Tibo, Jaeger, Frasconi, & Wrobel (2020) approached the application of movie-related review data in a predictive setting by formulating the sentiment analysis problem based on IMDb movie reviews to construct a nested bag MMIL (multi-multiple instance learning) Classifier. Picking up on this work, Deng & Yiu (2022) proposed a novel pipeline indicating that it is generally possible to integrate textual features extracted from news into a MIL pipeline to forecast stock trends.

### 2.5.3. Enhancing Model Interpretability in Time Series Prediction

Works in the MIL field, such as Early et al. (2024), indicate a need to explore more sophisticated integration of interpretability techniques in deep learning frameworks. Following Turbé, Bjelogrlic, Lovis, & Mengaldo (2023), interpretability is a critical aspect of machine learning that focuses on making models' decision-making processes understandable to humans. It is especially important in deep learning, where complex architectures often act as black boxes. Assis, Dantas, & Andrade (2024) argue that interpretability enables validation of models, builds trust with stakeholders, and provides insights that can inform practical decision-making when it comes to evaluating the robustness and transparency of predictions made by AI systems. Their results indicate that there is a trade-off between predictor's performance and interpretability (Assis et al., 2024). Looking at the common approaches, existing machine learning benchmarks such as the one for the TodyNet model by H. Liu et al. (2024) or TimeMIL by X. Chen et al. (2024) neglect interpretability in their

evaluation when benchmarking their models and aim to provide the highest performance in terms of predictive accuracy. Although X. Chen et al. (2024) point out their model's ability to deliver interpretable results, they do not extend their study to compare it with other models in terms of interpretability either by visually comparing models using common attribution techniques such as SHAP or Integrated Gradients which are more universally applicable across different architecture types nor by quantifying the results (Turbé et al., 2023).Moreover, X. Chen et al. (2024) point out that their model is inherently interpretable due to the attention maps produced by the transformer components within the model, the value of attention to produce interpretable outputs is contested in other papers (Bibal et al., 2022; Jain & Wallace, 2019).

Additionally, they do not consider accessing their model's interpretability by applying interpretability metrics such as Infidelity proposed in the Captum framework for interpretability by Kokhlikyan et al. (2020) or other metrics and methods (Namdari & Li, 2019; Yeh et al., 2019).

# 3. Data

Existing works from Lash & Zhao (2016) showcase the integration of different data sources, suggesting that various movie-related data sources can be combined to build sophisticated predictive solutions. Their framework integrates IMDb and Box Office Mojo data through API calls and web scraping, ensuring comprehensive data acquisition. Madongo & Zhongjun (2023) used different types of movie data to capture deeper insights. They combined this data in a single model to predict how much revenue a movie would make during its opening weekend. Both research papers indicate that movie-related data offers various potentials for analytics and can be processed to leverage various data sources in movie success prediction. Following these insights, chapter three comprehensively outlines the wide range of different datasets and sources used for this thesis. Due to the strong limitation regarding the available resources, the central datasets for Rotten Tomatoes (Leone, 2020), IMDb (Banik, 2017), Twitter (Dooms, 2021), The Numbers (The Numbers, 2024)were extracted from the data science community platform Kaggle and enhanced using additional open-source APIs such as Box Office Mojo or TMDB (Mohamadi, 2024; TMDB, 2023). Furthermore, the existing database was enhanced by social media information from Twitter and Instagram. This chapter briefly introduces each dataset, showcasing their potential for use in movie-related scientific frameworks. The movie industry benefits from a wealth of data that can be acquired in diverse forms and from various sources, offering opportunities for deeper insights and predictive analytics. Structured data, such as tabular formats containing movie metadata (e.g., genres, actors, budgets) and movie ratings, provide well-organized, easily interpretable datasets that form the foundation for traditional machine learning framework. In contrast, semi-structured data, such as multivariate time series (e.g. box office trends over time), captures dynamic patterns and temporal relationships essential for trend forecasting and operational decision-making. These

data types are complemented by unstructured textual data like movie reviews that can be processed using advanced NLP methodologies to valuable qualitative insights into audience feedback and preferences. Given this diverse data environment in the movie industry, this study leverages movie metadata, box office performance data as well as movie ratings and reviews.

## 3.1. Movie Metadata

Movie metadata appears in a wide variety of different formats ranging from structured to unstructured data. Data was sourced from well-established online movie platforms Rotten Tomatoes and IMDb. Metadata is displayed in a tabular fashion and is considered structured data. It includes various features that can be used to better understand different aspects of a movie. For instance, it contains title, production studio, languages, runtime and information about the cast.

An open-source movie data collection from Rotten Tomatoes offers a detailed repository of metadata for over 140,000 movies, making it a useful resource for exploring the characteristics and performance metrics of films. This dataset provides structured and diverse information that captures both the qualitative and quantitative aspects of movies.

The dataset contains identifying attributes like the movie's title, release year and rotten tomatoes id. The genre column lists thematic categories, often multiple per movie, enabling analysis of genre-specific trends and audience reception. The runtime column details the duration of each movie in minutes. Key creative contributors to a film are also documented. The director field identifies the individual(s) responsible for the movie's overall vision, while the cast column highlights the main actors and actresses, providing a basis for exploring the influence of star power on a movie's success. The data further includes the MPAA rating, which specifies the audience suitability of each film (e.g., G, PG, R). This classification can be examined in relation to the movie's target demographics and reception.

Quantitative metrics in the dataset are particularly valuable for performance analysis. The critics score and audience score columns provide aggregated ratings as percentages, reflecting the reception of the film among professional reviewers and general viewers, respectively. The box office revenue column, expressed in USD, represents the financial success of the movie, enabling research into the economic aspects of the film industry. Additionally, the synopsis column offers a textual summary of the movie, which can serve as a basis for natural language processing applications to analyze thematic content.

The dataset is robust and versatile, but it is not without challenges. Missing values, particularly in fields like box office revenue necessitate careful preprocessing or the use of additional data sources. Similarly, the multi-label nature of the genre field requires appropriate handling to extract meaningful insights.

Additional movie metadata was collected from IMDb. The dataset obtained provides a comprehensive view of cinematic metadata for 45,000 movies released on or before July 2017. This dataset integrates diverse information sourced from the TMDB Open API and GroupLens platforms, offering rich contextual data for movie analytics. It captures a wide range of features, including metadata such as movie titles, production details, budget, revenue, and release dates, as well as artistic elements like language, production companies, and production countries. Additional layers of data include cast and crew information, plot keywords, and user engagement metrics like ratings and vote counts. These features create a robust framework for exploring the interplay between qualitative and quantitative dimensions of cinema, allowing researchers to analyze trends over time, predict performance, and explore audience preferences.

## 3.2. Box-Office Performance

The box office data provides detailed information on the financial performance of films across domestic and international markets for both their whole box office lifecycle and individual days for a subset of movies. It includes key attributes such as the movie title, worldwide gross revenue, domestic revenue, and foreign revenue. Additionally, the data breaks down the percentage of domestic and foreign markets contributing to the total gross revenue.

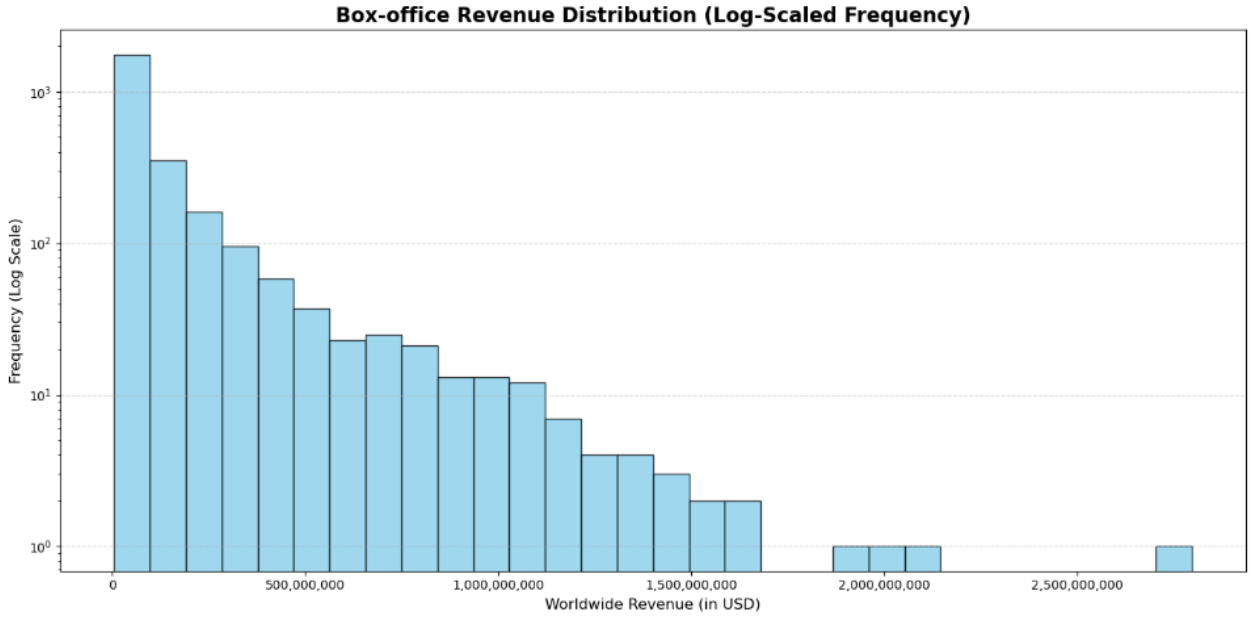



*Figure 1: Distribution of Movies Lifetime Box Office Revenues Worldwide*

The histogram in Figure 1 displays the distribution of worldwide box-office revenues, with the y-axis scaled logarithmically to better represent the varying frequencies across the available data. The x-axis represents revenue in USD, showing a concentration of films with lower box-office earnings, as indicated by the taller bars near the left. As revenues increase, the frequency decreases exponentially, as seen by the shorter bars toward the right. The logarithmic y-axis highlights the steep decline in frequency, with most films earning significantly less than $500 million, while only a few achieve revenues exceeding $1 billion. This distribution reflects the skewed nature of box-office earnings, where a small number of blockbuster films dominate total revenue.

The data is valuable for examining box-office trends. It allows for cross-referencing with other variables such as release dates, seasonal patterns, and audience behavior to investigate the impact of release timing on financial success. Moreover, the dataset's granularity supports the construction of predictive models to identify the factors that most significantly drive box-office performance.

To analyze the relationship between the original movies' metadata and their impact on sequel box office performance, we required a dataset explicitly identifying sequels. Detecting sequels is not always straightforward, as titles do not consistently include indicators like "Part 2" or similar phrasing. For this purpose, we sourced detailed data from "The Numbers", which specializes in worldwide box office performance for over 1,700 sequels. This dataset provides insights into the financial outcomes of sequel films, a key segment of the movie industry with notable commercial and audience interest.

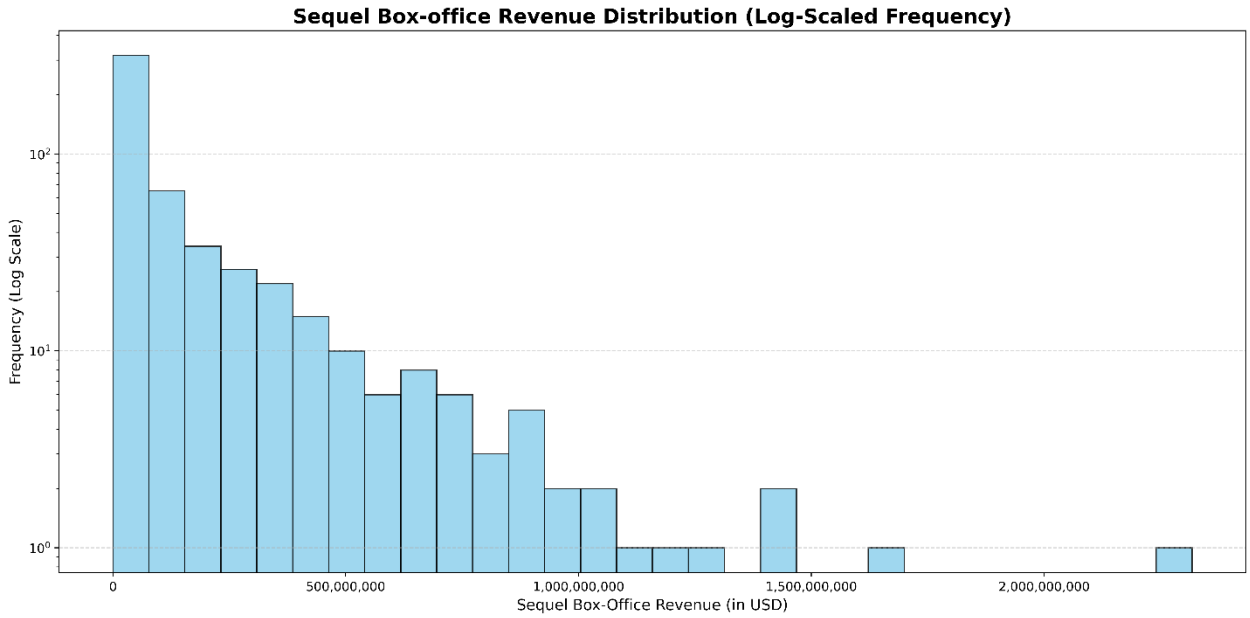



*Figure 2: Distribution of Movies Sequels Lifetime Box Office Revenues Worldwide.*

The dataset includes essential fields such as the sequel title, release year, and cumulative worldwide gross (in USD). These structured metrics enable comparative analyses, helping to identify factors influencing sequel success. For instance, Figure 2 illustrates the distribution of lifetime box office

revenues for movie sequels, showcasing the significant variation in financial performance within this category. Interestingly, when compared to the distribution of general movie box office revenues Figure 1, we observe notable similarities in their patterns. This resemblance suggests that sequels tend to follow the same general economic trends as movies overall. However, this also makes it particularly compelling to investigate what sets sequels apart - specifically, how unique factors like their connection to the original movie or audience expectations influence their performance.

While valuable, this dataset posed integration challenges when combined with other sources like Rotten Tomatoes. The absence of a shared unique identifier (e.g., a universal movie ID) necessitated reliance on title and release year for matching. Variations in formatting and language across datasets further complicated this process. For example, a sequel might be labeled "Part 2" in one dataset and "Part II" in another or titles in different languages (e.g., French or German) often appear inconsistently translated into English.

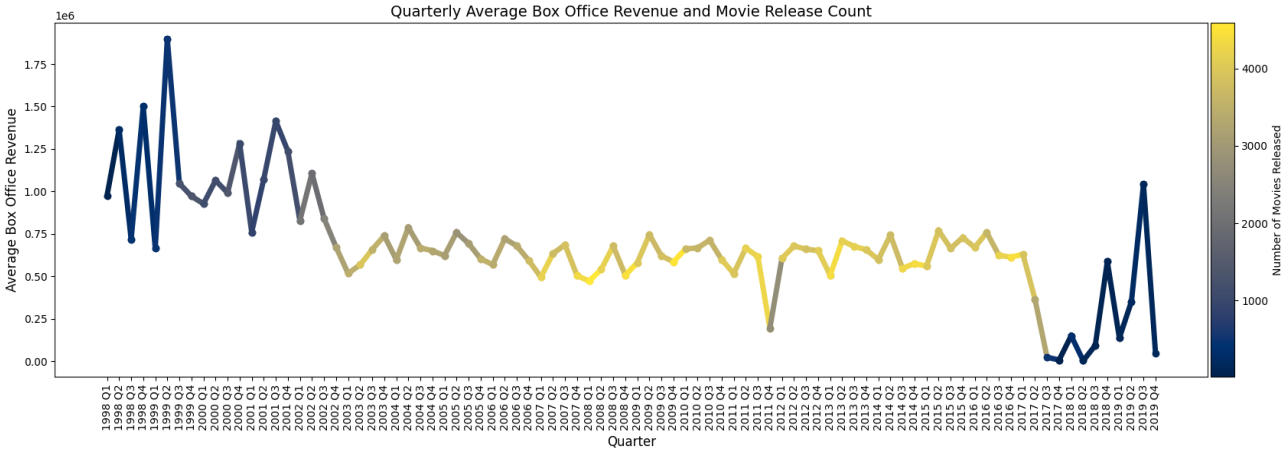



*Figure 3: Average quarterly Box Office Revenue recorded across samples in the daily Box Office Revenue dataset*

Further, box office data can also be represented in a contextualized time series format, representing the box office more granularly with a timestamp. The time series contains detailed information on daily box office revenues, theater counts, and movie performance metrics as time series, focusing on the USA and Canada markets. Each entry represents a daily record for a specific movie. An

outline of the trends for average quarterly Box Office revenue recorded over the years as well as the number of movies released per quarter of any year is illustrated in Figure 3.

The *daily* column captures the revenue generated by a movie on a particular day, expressed in US Dollars. *Theaters* refers to the number of theaters where the movie was screened on that day, providing insights into its market reach. The *BoxOffice* column aggregates a movie's total revenue. In addition to these key features, the dataset includes various metadata such as release dates, distributors, genres, ratings, and other performance metrics. These ancillary features help contextualize the daily and cumulative box office data. The dataset spans from the years 1999 to 2019 and provides a granular look at movie performance in the targeted regions.

## 3.3. Movie Reviews and Ratings

Reviews and ratings are very different compared to metadata. Following up on the data analysis, they often appear in a temporal, spatial format and are rather unstructured. Reviews and ratings capture critics and public perception about movies. This feedback-based data provides key-features for examining movie success and therefore provide valuable contributions to predictive analytics within the industry. Several data platforms were conducted to obtain relevant datasets. These platforms capture public perception combining review content, ratings and temporal information.

### 3.3.1. Movie Reviews: Rotten Tomatoes

Open-source movie review data was obtained from Rotten Tomatoes covering a wide range of critics and user feedback in an unstructured textual format. This data was accompanied by complementary structured features like corresponding ratings and timestamps. With over a million written reviews, the data obtained provides a granular perspective on how diverse audiences react to the movies documented in the accompanying metadata, enabling a detailed understanding of viewing reception and engagement. An exploratory data analysis was conducted to understand the dataset’s structure and prepare it for further analysis. Figure 4 displays the distribution of review

length defined as word count in KDE as well as boxplot. The visualizations show that the review content varied significantly, with an average length of approximately 21 words and a standard deviation of 9.42. The shortest reviews contained only a single word, while the longest extended to 55 words. The boxplot further supports highest distribution of word count around the average with a few outliers above the 50 words. The distribution revealed that 75% of reviews were 28 words or fewer, while half of the reviews contained 21 words or less, indicating a predominance of shorter, more concise reviews providing a realistic scenario.

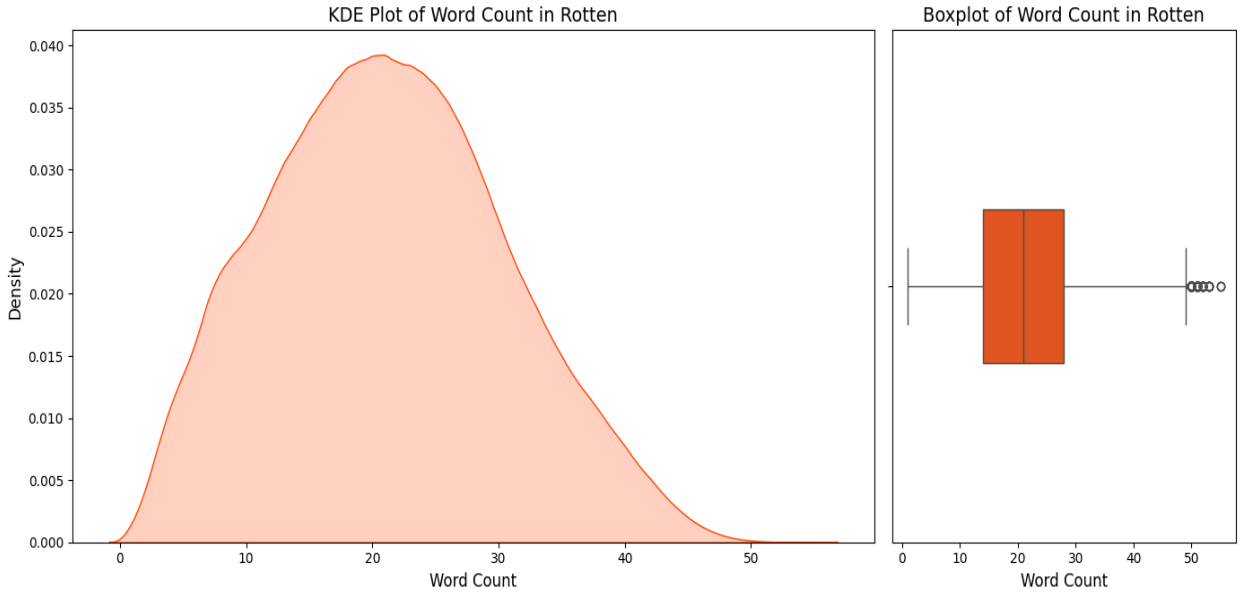



*Figure 4: Distribution of Word Counts for Rotten Tomatoes Movie Reviews*

### 3.3.2. Movie Ratings: Perspective across Platforms

Rating data was obtained from various platforms like Rotten Tomatoes as well as IMDb and Twitter. In Rotten Tomatoes data ratings are a critical component of review datasets, providing structured numerical insights to complement the textual content. The dataset obtained included over 1 million reviews on over 17,000 distinct movies. Rating scores were represented in various formats, including numeric fractions (e.g., 3/5), integers on diverse scales (e.g., 7, 100), and letter

grades (e.g., A, B+). Such variability in formats necessitates careful preprocessing to ensure consistency. To enable comparability, all ratings were normalized to a uniform 1–5 scale. Ambiguous data entries, such as "4/5.5," were flagged as errors and excluded from the data. This rigorous normalization process ensured consistency and interpretability, aligning with established best practices in predictive modeling.

Similarly, as the Rotten Tomatoes dataset, the Movie Tweetings dataset created by Dooms (2021) offers a dynamic and contemporary collection of movie ratings sourced from structured tweets on Twitter, encompassing around 39,000 unique movies with around 930,000 ratings recorded between the years 2013 and 2020. Designed to overcome the limitations of static datasets, it leverages real-time social media data, with ratings scaled from 0 to 10 and linked to IMDb identifiers for consistency and metadata enrichment. Twitter ratings are also recorded along with users. This enhances the potential of deriving potential features not only by focusing on the specific rating scores but also on the individual user's perceptions of the specific platform.

While Rotten Tomatoes gives an outline towards professional and semi-professional critic reviews, the IMDb dataset created by (Banik, 2017), specifically captures public opinion for movies as an entertainment product in a more holistic matter in terms of consumer ratings, in a large scale across streaming platforms. The platform captures ratings of viewers at a large scale for millions of different user ratings and reviews across different social media and streaming platforms such as Amazon, Disney+, Netflix, Twitter and many more. As a Big Data service, it captures a more holistic view of public perception and thus, the underlying data distributions can vary in their characteristic. The dataset used within this study captured around 26 million rating samples across 45,000 movies. Similar to the Twitter dataset, IMDb ratings appeared in a more standardized format with rating values ranging between 0 and 10. Moreover, due to the temporal characteristic of the recorded rating, the dataset also lends itself to time-series analyses by providing temporal

information on release dates and user interactions, which can be used to explore historical trends in cinema. The combination of metadata and audience-driven insights offers an opportunity to contextualize movie success and cultural impact within a broader temporal and industrial framework, making it a versatile resource for studying the evolution of the film industry and its audience dynamics.

#### 3.3.2.1. Rating Distribution across Platforms

Looking at the distribution of ratings across platforms displayed in Figure 5, various insights can be overserved. Ratings for Rotten and IMDb both exhibit a slightly left-skewed pattern, with Rotten centering around a score of 3 and IMDb around 3 to 4, reflecting a balance of positive and negative feedback. In contrast, Twitter is heavily right skewed, with the majority of ratings clustered at 4 and 5, indicating a strong positive sentiment bias. The disparity underscores the distinct purposes of these platforms: Rotten Tomatoes and IMDb are established as critical platforms trusted in the entertainment industry, incorporating a combination of user and critic feedback, which results in more balanced distributions. Meanwhile, Twitter, as a multipurpose social platform, encourages informal and spontaneous sharing of opinions, often leading to an emphasis on enthusiastic and positive sentiment at least in the available samples in the dataset used for this word.

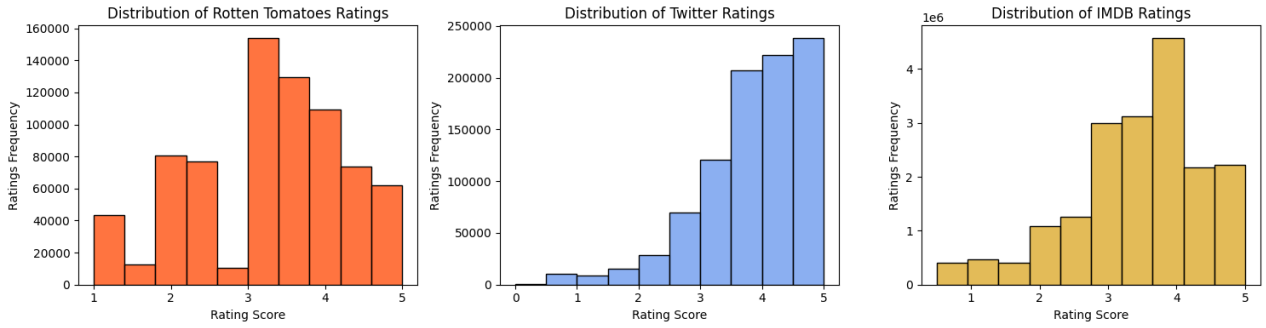



*Figure 5: Rating Distribution across Rotten Tomatoes, Twitter and IMDb*

### 3.3.2.2. Ratings in a Temporal Sequential Perspective

As mentioned, ratings and reviews are mostly attributed to a timestamp, extending them to time series. Following this, each movie's ratings form a sequence of values evolving in time, enabling spatial-temporal analysis to identify timely patterns that can be attributed to certain features within the data. Typical patterns are trends or seasonality, as outlined in section 2.2.5. An important characteristic of the data across different platforms is the varying length of the time sequences across features. Figure 6 visualizes the different distributions of time sequence lengths per movies across the three datasets. A filter was applied to include movies recorded for equal to or more than 28 days. However, it is to be mentioned that, due to movies differing in terms of time spent at the box office, samples can range from single days to years. The sequences for each platform show unique characteristics as each platform captures different perspectives within the market. For instance, the data derived from Rotten Tomatoes captures professional and non-professional critic reviews that often occur before and after the official release of a movie. IMDB and Twitter, on the other hand, capture the opinion of the mass audience in the post-release market stage, resulting in more significant amounts of available samples and, thus, longer time sequences. The broad span of unique characteristics within the data and features across platforms offers the opportunity for multifaceted analysis, combining different methods as well as data engineering techniques to form unique feature-sets for advanced data analysis within the movie industry.

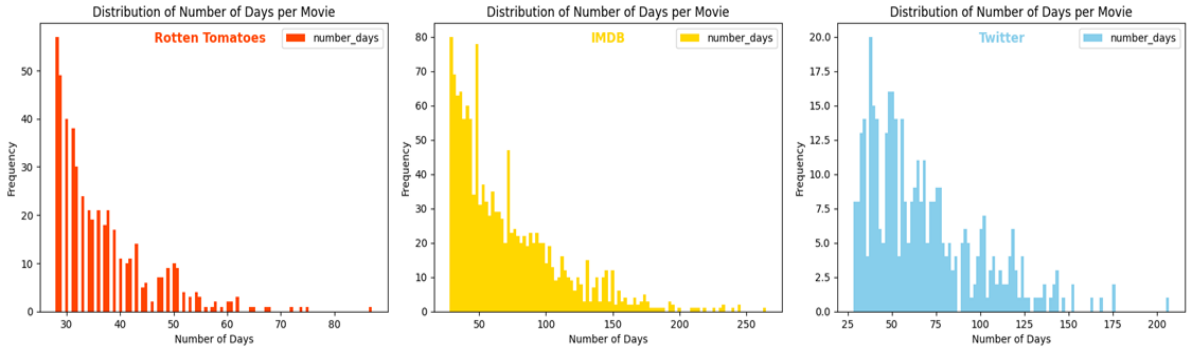



*Figure 6: Distribution of Sequence Lengths per dataset.*

# 4. Introduction to Individual Parts

Based on various data types and sources introduced in the previous section of this research paper, a comprehensive machine-learning framework exploring various factors that drive movie success is proposed.

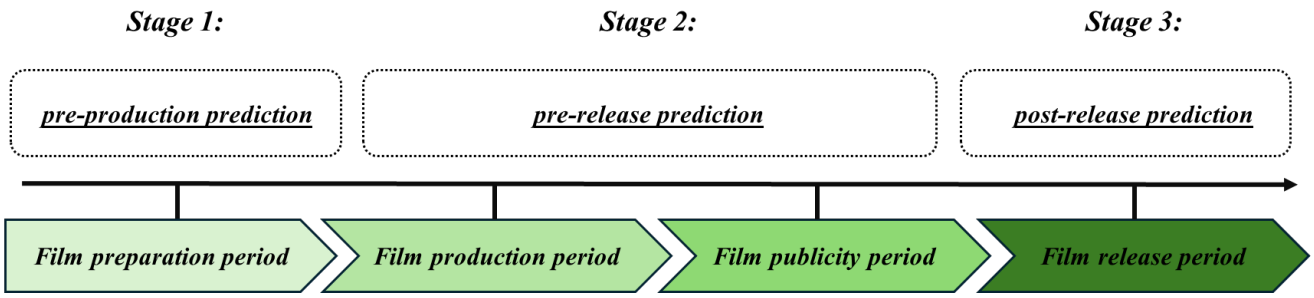



*Figure 7: Movie Lifecycle showcasing three stages for Machine Learning opportunities*

Figure 7 summarizes the main stages of a movie production lifecycle, defining characteristics within this diverse domain (Y. Liao et al., 2022). The cycle is divided into four sequential periods: Film Preparation Period, Film Production Period, Film Publicity Period, and Film Release Period. Each phase represents a critical stage in the filmmaking and distribution process. The preparation period involves pre-production activities, such as script development, casting, and budgeting. The production period focuses on filming and post-production tasks, including editing and visual effects. The publicity period encompasses promotional activities designed to generate awareness and excitement about the film. Finally, the release period marks the film's debut to audiences through theatrical or digital distribution channels, targeting revenue generation and audience engagement. This timeline highlights the structured progression of activities essential to a film's success. It also enables the directed deployment of various machine learning frameworks and predictive modeling approaches to enable informed decision making at various stages of a movie lifecycle.

This study examines the application of comprehensive machine learning frameworks across key stages of the film lifecycle, focusing on preproduction, prerelease, and post-release predictions. The post-release phase emphasizes feedback from ratings, reviews, and box office performance to evaluate success. The prerelease phase explores factors such as release timing and post-production elements to optimize audience engagement. Additionally, pre-production prediction is addressed by exploring actor cast as well as sequel production. By addressing these phases, the study highlights how predictive modeling supports data-driven decision-making throughout the film lifecycle. The following individual parts are introduced:

### The Differential Impact of Critic and User Ratings on Box Office Success: Exploring the Roles of Seasonality, Star Power, and Studio Size

The second study examines the role of critic and user ratings in shaping box office performance, considering how these early evaluations interact with contextual factors such as seasonality, star power, and studio size. Drawing on a longitudinal dataset spanning two decades, this research highlights the intricate relationships between professional and user-generated ratings and their capacity to influence consumer decision-making and revenue generation.

### Understanding Box-Office Dynamics: The Role of Timing and Audience in Movie Revenue Prediction

In the third study, attention shifts to the strategic implications of release timing, particularly during holidays, for box office revenues. This investigation distinguishes between family-friendly and adult-oriented films, analyzing the ways in which audience demographics and preferences interact with temporal contexts. Employing a combination of machine learning models and econometric techniques, this section provides a nuanced perspective on the role of timing in audience behavior and financial performance.

## Quantitative Analysis of Factors Driving Movie Sequel Success: A Predictive Modeling Approach

The fourth study focuses on the commercial success of movie sequels, exploring factors such as runtime, release timing, audience engagement, and reviews. By integrating predictive modeling techniques with feature importance analyses, this research provides a data-driven understanding of the challenges and opportunities associated with sequels, offering insights into the drivers of financial and critical success.

## Rating Predictions from Movie Reviews Leveraging Bert: Exploring the Impact of Sentiment Polarity and Review Length

This study explores the application of advanced natural language processing techniques, with a particular focus on transformer-based language models, to predict numeric movie ratings from textual reviews. By employing a large-scale dataset, this section investigates the potential of sentiment analysis and textual features to enhance predictive frameworks, utilizing both classification and regression methodologies to navigate the complexities of unstructured textual data in rating prediction.

## Temporal Modeling for Movie Success Prediction: A Comparative Study on the Applicability of different Deep Learning Models in Entertainment Analytics

Lastly, with focus on multivariate time series classification, the last study explores the applicability of Multiple Instance Learning (MIL) models for predicting box office performance based on time series data in the post-release stage of a movie. In a comparative approach, an MIL model was tested against supervised models for predictive power and interpretability. The research sheds light on the strengths and limitations of weakly-supervised approaches, particularly in handling movie-related features such as time series data and gives implications on how future model benchmarks could be conducted in a more use case-ready way.

# 5. The Differential Impact of Critic and User Ratings on Box Office Success: Exploring the Roles of Seasonality, Star Power, and Studio Size

## 5.1. Introduction

Audience and critical perception can be key drivers of financial success in the competitive landscape of the film industry (Dellarocas et al., 2007). This paper explores the relationship between user and critic ratings and their impact on box office performance, with a specific focus on how these ratings influence movies during the early stages of their release. The study aims to provide strategic insights into decision-making for promoting films and maximizing box office revenue under diverse conditions.

The research seeks to answer the question: *What are the key factors influencing movie revenue, focusing on the dynamics of critic and user ratings? How do non-major studio productions, a star cast, and off-season releases affect these ratings and box office performance?*

Utilizing a comprehensive dataset that includes movie information, daily box office revenues, and early-stage user and critic ratings, this research employs the ordinary least squares (OLS) regression methodology to analyze these relationships and predict the box office success. While this research focuses on post-release prediction, the findings provide valuable insights for understanding audience behavior and optimizing strategies for movie production and release planning.

## 5.2. Hypotheses Development

### 5.2.1. Gap Analysis

The existing literature on box office performance explores critic and user ratings often using statistical and machine learning methods like regression models to identify patterns and predict outcomes (Y. J. Lee, Hosanagar, & Tan, 2015; Pang et al., 2022). However, significant gaps remain.

While ratings are well-studied, factors such as studio size, actor popularity, and seasonality have received comparatively little attention (D'astous et al., 1999; Prosser, 2002; Wu, Liao, & Tang, 2023), leaving an incomplete understanding of the dynamics driving box office success.

A key gap lies in how critic and user ratings influence performance under varying conditions, such as studio size, release seasons, and a star cast. This study bridges gaps in understanding the contextual importance of ratings by integrating these factors into a unified framework. Unlike existing research that treats them as control variables; it offers a holistic analysis to uncover their combined impact on box office performance.

### 5.2.2. Hypotheses

Building upon the literature review, this research constructs three hypotheses to address critical gaps in understanding the dynamics between ratings, contextual factors, and box office success. Ratings, both user-generated and critic-provided, are recognized as predictors of a movie's financial performance.

**H1**: *The impact of critic ratings in the early stages of a movie release will be more significant on a movie's box-office success compared to user ratings for movies not produced by major studios.*

This hypothesis is based on evidence suggesting that major studios provide a strong quality signal to consumers, a factor often absent for smaller producers (Kraft & Rao, 2024). Additionally, critic ratings are widely considered as a reliable source, as they are authored by experts in the field (Pang et al., 2022).

**H2**: *The impact of user ratings in the early stages of a movie release will be more significant on a movie's box-office success compared to critic ratings for movies with actors in leading roles that are very popular on social media.*

Social media has become a significant channel for user engagement, with the popularity of actors often driving immediate interest and positive word-of-mouth (Divakaran & Nørskov, 2016). This

hypothesis examines how user ratings, amplified by the social media presence of leading actors, might outweigh traditional critic ratings for such movies.

**H3**: *Critic ratings have a higher impact on a movie's box-office success compared to user ratings in the first stage of a movie's release for movies released during off-seasons.*

Movies released during off-seasons typically experience lower initial engagement and overall visibility (Ahmad et al., 2017). In situations where user ratings might be sparse, potential viewers could tend to rely more heavily on critic opinions (Flanagin & Metzger, 2013).

By investigating small studio size, off-season releases and a star cast as primary influences, the study aims to deepen the understanding of how important user and critic ratings are in these scenarios and how they shape box office performance.

## 5.3. Research Methodology

### 5.3.1. Data Cleaning, Engineering and Preprocessing

The dataset for this study was compiled from multiple sources, including IMDb, Rotten Tomatoes, Box Office Mojo, and Instagram. Key movie details, such as titles, runtimes, release dates, production studios, genres, MPAA ratings, the amount of theater the movie was shown in, and cast information, were sourced from IMDb, Rotten Tomatoes, and Box Office Mojo. The critic ratings had different scales and therefore user and critic ratings were transformed to a standardized scale of 0-20. The dataset includes movies released between 1999 and 2019, ensuring comparability across features such as production budgets, social media influence, and other relevant factors, thereby providing a robust foundation for the analysis.

Daily revenue data and the number of theaters showing each movie were collected using the Box Office Mojo API. This data was aggregated to calculate the total box office revenues for the first two weeks following a movie's release, a critical period capturing peak audience engagement and

revenue. The first two weeks were chosen as they align with the average theatrical runtime of approximately four weeks (Fahey, 2015), providing insights into early performance trends.

To assess the influence of actor popularity, a list of 2,995 actors and actresses featured in the dataset was compiled. Instagram follower counts were manually collected to gauge the cast member's public appeal. Actors' follower counts were divided into quartiles: Q1 included cast with 11,325 followers or fewer, Q2 included those with up to 115,000 followers, Q3 included those with up to 892,000 followers, and Q4 included those with up to 42,300,000 followers. The use of dummy variables ensured that all data on actors' follower counts was retained, preventing the loss of information due to outliers. Furthermore, based on the follower data the feature *avg_following* was created. It represents the sum of all followers, divided by the cast size.

Movies can belong to multiple genres, necessitating the construction of dummy variables for each genre (Karniouchina et al., 2023). This binary representation ensured that the multi-genre nature of movies was accurately reflected in the analysis. Similarly, production studios were classified as major studios if they were part of companies such as Twentieth Century Fox, Warner Brothers, Universal Pictures, Paramount, Metro-Goldwyn-Mayer, Sony Pictures, Lionsgate, or Walt Disney (Pang et al., 2022).

To measure release-timing effects on the box office revenue, release dates were categorized into four seasons: *release_season_Winter* (December– February), *release_season_Spring* (March–May), *release_season_Summer* (June– August), and *release_season_Fall* (September–November). The capturing of the temporal effect of *year* was inspired by Pang et al. (2022). To account for the influence of different time periods, the year feature (1999–2019) was divided into three categories: *year_1999_2005*, *year_2006_2011*, and *year_2012_2019*. Creating dummy variables for each individual year was impractical due to data availability issues, such as sparse representation for certain years. This categorical grouping provided a more evenly distributed

representation of years. Furthermore, Facebook bought Instagram in 2012 and market the beginning of accelerated growth of the platform (Pew Research Center, 2024; Rusli, 2012).

To understand the interplay of smaller studio size, off-season releases and star power on ratings and box office success, the combined dataset was filtered into three subsets accordingly. Dataset 1 consists solely of movies released by non-major studios, Dataset 2 focuses on titles premiered in spring and fall, and Dataset 3 includes only those films featuring at least one actor in the Q4 quartile. The merging and filtering process is visualized in Figure 8. The first dataset comprises of 4,698 rows, the second one of 6,005 rows and the third totals 9,272 rows.

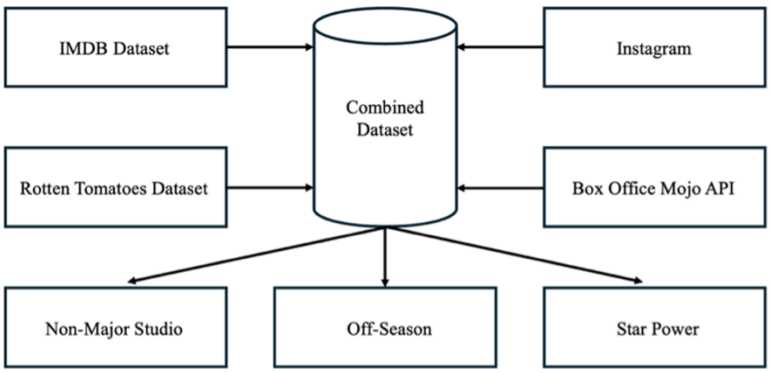



*Figure 8: Overview Data Structure*

Removing outliers from datasets is a common practice and to counter potential selection biases, outliers were identified and dropped after the filtering process (Fox, 2016). Furthermore, variables are often transformed to reduce their skewness (Karniouchina et al., 2023; Pang et al., 2022). The choice of square root or logarithmic transformation for independent features was guided by an assessment of skewness, kurtosis, and each variable's distribution. Variables were adjusted to achieve skewness values close to 0, kurtosis values near 3, and a distribution close to normal (Bai & Ng, 2005). After transformation, all variables were standardized using standard scaling to enhance model quality and ensure comparability (Fox, 2016).

### 5.3.2. Model Development

#### 5.3.2.1. Ordinary Least Squares Regression

The OLS method is a type of linear regression that minimizes the sum of squared deviations (Burton, 2021; Fox, 2016). It is widely used in academic research to statistically calculate

influences of features on a target variable (Divakaran & Nørskov, 2016; Karniouchina et al., 2023; Prosser, 2002). It allows for straightforward coefficient interpretation and was therefore chosen for this research.

#### 5.3.2.2. Feature Selection

Consistent with previous studies, multicollinearity was assessed by calculating the variance inflation factor (VIF) for the independent variables (Akdeniz & Talay, 2013; Bharadwaj, Noble, Tower, Smith, & Dong, 2017; Y. J. Lee et al., 2015). Features with a VIF greater than 5 were excluded from the dataset to prevent multicollinearity from affecting the coefficients.

Removing multicollinear features does not remove all correlation in the datasets. Therefore, and in accordance with other studies, the correlation between features was calculated for each dataset (Akdeniz & Talay, 2013; Prosser, 2002).

Lastly, in an iterative process, overly skewed features and those that violated the linearity assumption were either transformed using a different method or dropped from the dataset. This was necessary to conform to the assumptions for proper OLS regression as postulated by Fox (2016).

#### 5.3.2.3. Transformations

Following the example of Bharadwaj et al. (2017), the dependent variable *boxoffice* was transformed using the Box-Cox methodology (Cox, 1964). For the independent variables, logarithmic transformations were applied to *runtime* and *days_until_streaming*, while square root transformations were used for *budget, theaters*, *avg_following*, and *vote_count* to increase model performance and reduce skewness and extreme values (Fox, 2016).

#### 5.3.2.4. Model 1

Some variables were excluded during the process. Variables such as *release_season_Fall*, *year_2005_2010*, *country_count_2more*, *PG-13*, and *Q4* were removed due to multicollinearity issues. Additionally, *days_until_streaming* was dropped because it had a strong negative correlation

of -0.81 with *year_2012_2019*, and certain variables, including *romance*, *adventure*, *mystery*, wins, and seasonal release categories (Spring, Winter, and Summer), were excluded due to outliers and imbalanced distributions causing heteroskedasticity. The highest remaining correlation in the dataset was 0.44, observed between *drama* and *runtime*.

Furthermore, since the data was filtered to contain only smaller scale production companies, both features *non_major_studio* and *major_studio* were dropped from the dataset.

The equation for Model 1 is as follows:

$$\begin{aligned} boxoffice_1 = {} & \beta_0 + \beta_1 \times user\_rating + \beta_2 \times critic\_rating + \beta_3 \times runtime + \beta_4 \times R \\ & + \beta_5 \times drama + \beta_6 \times comedy + \beta_7 \times action + \beta_8 \times crime + \beta_9 \times thriller \\ & + \beta_{10} \times Q1 + \beta_{11} \times Q3 + \beta_{12} \times sqrt\_budget + \beta_{13} \times country\_count\_1 \\ & + \beta_{14} \times year\_2006\_2011 \end{aligned}$$

**5.3.2.5. Model 2**

Certain variables were excluded during the preparation process. *Q1*, *Q2*, *Q3*, and *Q4* were filtered out due to the star power filtering. Variables such as *release_season_Fall*, *not_major_studio*, *year_2006_2011*, *country_count2more*, and *PG-13* were removed because of multicollinearity. Additionally, *days_until_streaming* was excluded due to its high correlation of -0.78 with *year_2012_2019*, and the highest remaining correlation in the dataset was 0.64 between *sqrt_theaters* and *sqrt_budget*. Outliers and imbalanced distributions were addressed by excluding *wins*, while no other features showed significant heteroskedasticity or extreme outliers.

The equation for Model 2 incorporates one constant and 23 features.

$$\begin{aligned} boxoffice_2 = {} & \beta_0 + \beta_1 \times user\_rating + \beta_2 \times critic\_rating + \beta_3 \times runtime + \beta_4 \times R \\ & + \beta_5 \times drama + \beta_6 \times comedy + \beta_7 \times action + \beta_8 \times romance \\ & + \beta_9 \times crime + \beta_{10} \times thriller + \beta_{11} \times adventure + \beta_{12} \times mystery \\ & + \beta_{13} \times sqrt\_avg\_following + \beta_{14} \times major\_studio + \beta_{15} \times sqrt\_budget \\ & + \beta_{16} \times country\_count\_1 + \beta_{17} \times sqrt\_vote\_count + \beta_{18} \times sqrt\_theaters \\ & + \beta_{19} \times year\_1999\_2005 \\ & + \beta_{20} \times year\_2012\_2019 + \beta_{21} \times release\_season\_Winter + \beta_{22} \\ & \times release\_season\_Spring + \beta_{23} \times release\_season\_Summer \end{aligned}$$

#### 5.3.2.6. Model 3

Several features were excluded during the model-building process. Variables such as *release_season_Fall*, *release_season_Winter*, *release_season_Spring*, and *release_season_Summer* were removed during model filtering. Multicollinearity led to the exclusion of *major_studio*, *year_1999_2005*, and *country_count2more*, while *days_until_streaming* and *R* were dropped due to high correlations with *year_2012_2019* (-0.78) and *PG-13* (-0.76), respectively. The highest remaining correlation in the model was 0.64 between *sqrt_theaters* and *sqrt_budget*. Features such as *wins*, *romance*, *crime*, and *mystery* were also removed due to outliers and overly imbalanced distributions.

$$\begin{aligned} boxoffice_3 = {} & \beta_0 + \beta_1 \times user\_rating + \beta_2 \times critic\_rating + \beta_3 \times runtime + \beta_4 \times R \\ & + \beta_5 \times PG-13 + \beta_6 \times drama + \beta_7 \times comedy + \beta_8 \times action \\ & + \beta_9 \times thriller + + \beta_{10} \times adventure + \beta_{11} \times Q1 + \beta_{12} \times Q2 + \beta_{13} \times Q3 \\ & + \beta_{14} \times sqrt\_avg\_following + \beta_{15} \times not\_major\_movie \\ & + \beta_{16} \times sqrt\_budget + \beta_{17} \times country\_count\_1 + \beta_{18} \times sqrt\_vote\_count \\ & + \beta_{19} \times sqrt\_theaters + \beta_{20} \times year\_2012\_2019 \end{aligned}$$

## 5.4. Results and Evaluation

### 5.4.1. Regression Results

The following section presents the results of the regression analyses for Model 1, Model 2, and Model 3. These models are independent of one another and are not intended to represent iterative improvements. The results include coefficients, standard errors, z-values, p-values, and confidence intervals for all variables in each model. The significance levels for each variable are denoted by stars as follows: ($p < 0.001$): ***, ($p < 0.01$): **, and ($p < 0.05$): *, indicating highly robust, strong, and moderate evidence, respectively (Divakaran et al., 2017). The results can be seen in Tables 1, 2, and 3 and will be further discussed in the next section.

*Table 1: Model 1 Coefficient Overview*

| **Model 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **coefficient** | **std error** | **z** | **P>\|z\|** | **Significance Level** | **[0.025** | **0.975]** |
| **const** | 0.2736 | 0.052 | 5.268 | 0.000 | *** | 0.172 | 0.375 |
| **user_rating** | 0.0381 | 0.017 | 2.279 | 0.023 | * | 0.005 | 0.071 |
| **critic_rating** | 0.0756 | 0.016 | 4.761 | 0.000 | *** | 0.044 | 0.107 |
| **runtime** | -0.0934 | 0.019 | -4.919 | 0.000 | *** | -0.131 | -0.056 |
| **R** | -0.0997 | 0.036 | -2.788 | 0.005 | ** | -0.17 | -0.03 |
| **drama** | -0.5495 | 0.04 | -13.599 | 0.000 | *** | -0.629 | -0.47 |
| **comedy** | -0.2578 | 0.043 | -5.945 | 0.000 | *** | -0.343 | -0.173 |
| **action** | 0.0467 | 0.039 | 1.183 | 0.237 | | -0.031 | 0.124 |
| **crime** | -0.264 | 0.035 | -7.498 | 0.000 | *** | -0.333 | -0.195 |
| **thriller** | 0.2966 | 0.038 | 7.883 | 0.000 | *** | 0.223 | 0.37 |
| **Q1** | 0.0604 | 0.031 | 1.968 | 0.049 | * | 0.00 | 0.121 |
| **Q3** | 0.1471 | 0.021 | 6.984 | 0.000 | *** | 0.106 | 0.188 |
| **sqrt_budget** | 0.509 | 0.018 | 28.441 | 0.000 | *** | 0.474 | 0.544 |
| **country_count_1** | 0.017 | 0.031 | 0.541 | 0.589 | | -0.045 | 0.079 |
| **year_2006_2011** | -0.3003 | 0.033 | -9.095 | 0.000 | *** | -0.365 | -0.236 |

*Table 2: Model 2 Coefficient Overview*

| **Model 2** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **coefficient** | **std error** | **z** | **P>\|z\|** | **Significance Level** | **[0.025** | **0.975]** |
| **const** | 0.0286 | 0.047 | 0.606 | 0.544 | | -0.064 | 0.121 |
| **user_rating** | 0.011 | 0.01 | 1.147 | 0.251 | | -0.008 | 0.03 |
| **critic_rating** | 0.0607 | 0.01 | 5.81 | 0.000 | *** | 0.04 | 0.081 |
| **runtime** | 0.0687 | 0.013 | 5.199 | 0.000 | *** | 0.043 | 0.095 |
| **R** | -0.0952 | 0.027 | -3.544 | 0.000 | *** | -0.148 | -0.043 |
| **drama** | 0.0213 | 0.036 | 0.586 | 0.558 | | -0.05 | 0.093 |
| **comedy** | -0.0634 | 0.033 | -1.949 | 0.05 | | -0.127 | 0.00 |
| **action** | -0.3164 | 0.035 | -8.922 | 0.000 | *** | -0.386 | -0.247 |
| **romance** | 0.0005 | 0.044 | 0.01 | 0.99 | | -0.085 | 0.086 |
| **crime** | -0.1046 | 0.028 | -3.737 | 0.000 | *** | -0.159 | -0.05 |
| **thriller** | -0.0446 | 0.028 | -1.587 | 0.112 | | -0.1 | 0.01 |
| **adventure** | 0.1717 | 0.034 | 5.052 | 0.000 | *** | 0.105 | 0.238 |
| **mystery** | 0.2887 | 0.034 | 8.403 | 0.000 | *** | 0.221 | 0.356 |
| **sqrt_avg_following** | -0.0418 | 0.013 | -3.338 | 0.000 | *** | -0.066 | -0.017 |
| **major_studio** | 0.1754 | 0.027 | 6.612 | 0.000 | *** | 0.123 | 0.227 |
| **sqrt_budget** | 0.0235 | 0.021 | 1.124 | 0.261 | | -0.017 | 0.065 |
| **country_count_1** | 0.1496 | 0.023 | 6.622 | 0.000 | *** | 0.105 | 0.194 |
| **sqrt_vote_count** | 0.4544 | 0.018 | 25.812 | 0.000 | *** | 0.42 | 0.489 |
| **sqrt_theaters** | 0.509 | 0.017 | 30.582 | 0.000 | *** | 0.476 | 0.542 |
| **release_season_Winter** | 0.3619 | 0.031 | 11.829 | 0.000 | *** | 0.302 | 0.422 |
| **release_season_Spring** | -0.1619 | 0.031 | -5.227 | 0.000 | *** | -0.223 | -0.101 |
| **release_season_Summer** | 0.1581 | 0.03 | 5.21 | 0.000 | *** | 0.099 | 0.218 |
| **year_1999_2005** | 0.0857 | 0.03 | 2.857 | 0.000 | *** | 0.027 | 0.145 |
| **year_2012_2019** | -0.4621 | 0.031 | -14.995 | 0.000 | *** | -0.523 | -0.402 |

*Table 3: Model 3 Coefficient Overview*

| | ***Model* 3** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **coefficient** | **std error** | **z** | **P>\|z\|** | **Significance Level** | **[0.025** | **0.975]** |
| ***const*** | 0.3294 | 0.028 | 11.675 | 0.000 | *** | 0.274 | 0.385 |
| ***user_rating*** | 0.0041 | 0.006 | 0.657 | 0.511 | | -0.008 | 0.016 |
| ***critic_rating*** | 0.0206 | 0.006 | 3.195 | 0.001 | ** | 0.008 | 0.033 |
| ***runtime*** | 0.0327 | 0.008 | 4.201 | 0.000 | *** | 0.017 | 0.048 |
| ***PG-13*** | 0.1349 | 0.013 | 10.182 | 0.000 | *** | 0.109 | 0.161 |
| ***drama*** | -0.1631 | 0.017 | -9.412 | 0.000 | *** | -0.197 | -0.129 |
| ***comedy*** | -0.0994 | 0.018 | -5.394 | 0.000 | *** | -0.135 | -0.063 |
| ***action*** | -0.3073 | 0.016 | -19.427 | 0.000 | *** | -0.338 | -0.276 |
| ***thriller*** | -0.1285 | 0.017 | -7.346 | 0.000 | *** | -0.163 | -0.094 |
| ***adventure*** | -0.1419 | 0.019 | -7.496 | 0.000 | *** | -0.179 | -0.105 |
| ***Q1*** | -0.0123 | 0.012 | -0.989 | 0.323 | | -0.037 | 0.012 |
| ***Q2*** | 0.0347 | 0.011 | 3.148 | 0.002 | ** | 0.013 | 0.056 |
| ***Q3*** | 0.0033 | 0.011 | 0.296 | 0.767 | | -0.019 | 0.025 |
| ***sqrt_avg_following*** | -0.0279 | 0.008 | -3.457 | 0.001 | ** | -0.044 | -0.012 |
| ***not_major_studio*** | -0.0296 | 0.016 | -1.894 | 0.058 | | -0.06 | 0.001 |
| ***sqrt_budget*** | 0.2615 | 0.011 | 22.832 | 0.000 | *** | 0.239 | 0.284 |
| ***country_count_1*** | 0.0764 | 0.013 | 5.78 | 0.000 | *** | 0.05 | 0.102 |
| ***sqrt_vote_count*** | 0.459 | 0.009 | 50.1 | 0.000 | *** | 0.441 | 0.477 |
| ***sqrt_theaters*** | 0.3978 | 0.009 | 44.146 | 0.000 | *** | 0.38 | 0.416 |
| ***year_2012_2019*** | -0.47 | 0.016 | -30.169 | 0.000 | *** | -0.501 | -0.439 |

*Table 4: Models 1,2 and 3 Overview*

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| R-Squared: | 0.479 | 0.733 | 0.805 |
| Adjusted R-Squared: | 0.476 | 0.731 | 0.805 |
| F-Statistic: | 203.8 | 533.0 | 1060 |
| Prob. (F-Statistic): | 0.00 | 0.00 | 0.00 |
| user_rating \| P> t: | 0.0381 \| 0.023 | 0.0110 \| 0.251 | 0.0041 \| 0.511 |
| critic_rating \| P>t: | 0.0756 \| 0.000 | 0.0607 \| 0.000 | 0.0206 \| 0.001 |
| Transformation: | | | |
| user_rating % \| std | 15.86% \| 4.63 | 3.14% \| 4.86 | 1.6% \| 4.71 |
| user_rating (total) \| std | $4.5M \| 4.63 | $1.87M \| 4.86 | $1.1M \| 4.71 |
| critic_rating % \| std | 35.13% \| 4.94 | 17% \| 4.60 | 8.08% \| 4.61 |
| critic_rating (total) \| std | $9.9M \| 4.94 | $10M \| 4.60 | $5.3M \| 4.61 |
| Omnibus: | 4.417 | 0.751 | 5.096 |
| Prob. (Omnibus): | 0.110 | 0.687 | 0.078 |
| Jarque-Bera: | 3.833 | 0.798 | 5.107 |
| Prob. (JB): | 0.147 | 0.671 | 0.0778 |
| Durbin Watson: | 1.996 | 1.994 | 2.020 |
| Condition Number: | 8.36 | 11.5 | 11.0 |
| Skew: | 0.015 | 0.005 | -0.061 |
| Kurtosis: | 2.804 | 2.918 | 3.090 |
| Overfitting: | | | |
| RMSE training set: | 0.7222 | 0.5143 | 0.4462 |
| RMSE testing set: | 0.7357 | 0.5875 | 0.4366 |

To achieve the results shown in Tables 1,2,3 and, 4 the datasets were randomly split into 70% training and 30% testing sets prior to running the OLS regression. The adjusted R-squared denotes the accounted for proportion of variance in the dependent variable (Chicco, Warrens, & Jurman, 2021). The values for Model 1, Model 2, and Model 3 are 0.476, 0.731, and 0.805, indicating varying explanatory power. The F-statistics for Model 1 (203.8), Model 2 (533), and Model 3 (1060) are all associated with p-values of 0.00, signifying that each model is highly statistically significant in its own context (Fox, 2016). A Box-Cox transformation was applied to the dependent variable in all three models, after standardization, the coefficients reflect the change in box office performance in standard deviations, associated with a one-unit change in the predictor variable (Cox, 1964). To improve interpretation, the dependent variable, *boxoffice*, was back-transformed using the Box-Cox scaling coefficient. First, the predicted change is computed in the Box-Cox-transformed and standardized scale using the regression coefficient of the standardized user and critic ratings, and the standard deviation of the transformed box office (G. Chen, Lockhart, & Stephens, 2002; Cox, 1964). This change is then converted back to the original box office scale using the inverse Box-Cox formula. Finally, the absolute and percentage changes in box office revenue are calculated to provide an interpretable result.

The results are shown in Table 4 and detailed in the following description. In Model 1, both user and critic ratings are significant, with critic ratings at the 0.1% significance level and user ratings at the 5% significance level. A 4.94-point increase in critic ratings corresponds to a 35.13% increase in box office success, while a 4.63-point increase in user ratings leads to a 15.86% increase. In contrast, Model 2 finds user ratings insignificant, while critic ratings are highly significant at the 0.1% level, with a 4.6-point increase in critic ratings driving a 17% increase in box office performance. Model 3 similarly identifies critic ratings as significant at the 1% level, contributing

an 8.08% increase in box office success for a 4.61-point change, while user ratings remain insignificant. For reference, the total amount of change in U.S. Dollars is also recorded in Table 4. Normality of residuals was tested in all models using the Omnibus and Jarque-Bera tests, which assess the normality of residuals (Jarque & Bera, 1980; Urzúa, 1996). The null hypothesis of normality could not be rejected in any model, indicating that residuals closely follow a normal distribution. Supporting this conclusion, skewness values are near the target of 0, and kurtosis values approximate the target of 3 across all models. Residual independence was confirmed by the Durbin-Watson test, with all values close to the target value of 2 (Rutledge & Barros, 2002).

Residual-vs-fitted plots for each model further indicate homoscedasticity, as suggested by the alignment of residuals to the regression plane (Fox, 2016). The plots can be found in the appendixunder Figure 13. To test for overfitting, root mean squared error (RMSE) values were calculated for training and testing datasets in all models (Babyak, 2004). The roughly equal RMSE values indicate no overfitting.

Moreover, all models have a condition number significantly below 30, confirming the absence of multicollinearity.

## 5.5. Discussion

### 5.5.1. Hypotheses Testing

Hypothesis 1 predicts that critic ratings have a more significant impact on box office success than user ratings for movies not produced by major studios. The results support this hypothesis. Across all models, critic ratings are significant at the 0.1% or 1% level, while user ratings are either insignificant (Models 2 and 3) or less significant (Model 1). This suggests that critic ratings play a crucial role in driving box office success, particularly for movies without the quality signals associated with major studios (Kraft & Rao, 2024).

Hypothesis 2 posits that user ratings have a more significant impact than critic ratings for movies featuring actors popular on social media. The results do not support this hypothesis. In Model 2, user ratings are insignificant. This finding indicates that critic ratings remain the primary ratings driver of box office performance, even when accounting for social media popularity of the cast.

This could be explained by audience demographics and decision-making behaviors. According to the Board of Governors of the Federal Reserve System (2024), US citizens above 40 have the highest income and wealth, and therefore have the highest disposable income to spend in cinemas. Pew Research Institute (2024) reports that only about half of all adults in the USA use Instagram. If these individuals prioritize professional reviews over social media fueled hype, this could be an explanation towards the findings. Potential skepticism toward social media-driven promotion could lead audiences to seek reliable critiques from experts. Consumers critically analyze expert opinions, valuing unbiased insights (D'astous et al., 1999; Pang et al., 2022).

Hypothesis 3 suggests that critic ratings have a higher impact on box office success than user ratings for movies released during off-seasons. The results strongly support this hypothesis. Critic ratings are (highly) significant across all models, while user ratings remain insignificant in Models 2 and 3. This reinforces the critical role of professional reviews in influencing box office performance for movies released during periods of lower audience engagement.

Overall, the findings validate H1 and H3, emphasizing the consistent influence of critic ratings on box office success, while H2 is rejected due to user ratings being insignificant in Model 2.

### 5.5.2. Comparative Analysis of Results

This study aligns with the findings discussed in the literature review, detailing the information asymmetry in the movie industry and the trustworthiness of critic reviews (Pang et al., 2022).

Previous literature reveals no clear consensus on the extent or consistency of user and critic rating effects on box office performance. For instance, studies have found significant influences of both

user and critic ratings on box office success, with significance levels reported at the 0.1%, 5%, and 10% significance levels (Carrillat, Legoux, & Hadida, 2018; Divakaran et al., 2017; Pang et al., 2022). While most of the research supports the positive impact of these ratings and monetary success found in this research, some studies deviate. For example, Divakaran et al. (2017) and Carrillat et al. (2018) identify a slight but negative influence of critic ratings on box office performance.

Dellarocas et al. (2007) demonstrate that both user and critic ratings significantly boost box office revenue, particularly emphasizing a stronger relationship between user ratings and films with limited marketing efforts or fewer theater releases. While this contrasts with the findings of this study—where non-major studio films and off-season releases do not appear to be influenced by user ratings—it highlights the necessity of accounting for market conditions and promotional strategies when analyzing these relationships.

### 5.5.3. Theoretical Contributions & Managerial Implications

This thesis contributes to the literature on box office performance by demonstrating the dominant role of critic ratings in driving movie success, particularly for films released by smaller studios or released during off-seasons. It highlights the enduring influence of professional reviews as a quality signal, even in the age of digital and crowd-sourced opinions (Pang et al., 2022).

The limited impact of user ratings, even for movies featuring socially popular actors, challenges assumptions about the democratization of opinion and emphasizes the credibility gap between expert critiques and user-generated content. Managers should therefore prioritize building strong relationships with influential critics and investing in targeted press screenings, especially for smaller studios or off-season releases. Additionally, while social media star power alone may not guarantee higher user-driven revenue, leveraging celebrity influence alongside critic-focused marketing could create a more robust promotional strategy.

Furthermore, the study's adherence to OLS assumptions, along with a detailed data processing and methodology section, provides a well-structured and replicable framework. This ensures a template for future research to build upon in analyzing similar relationships.

### 5.5.4. Study Limitations

This study offers valuable insights into the interplay between critic and user ratings on box-office success but has limitations. Potential biases in the datasets include skewed user ratings and critics' genre preferences, which may affect generalizability. Advertising budgets, a potential key factor influencing audience perceptions and outcomes, are not considered. Moreover, actors' Instagram follower counts were collected in 2024 and may not accurately reflect their fame at the time of each movie's release. Additionally, the monetary-based features such as *boxoffice* and *budget* are not inflation adjusted and could insert a certain bias into the dataset. Finally, focusing on U.S. box-office data restricts the findings to a single market, overlooking regional variations in audience behavior and cultural preferences.

Future research could address these issues by incorporating global data and exploring genre-specific or regional biases for a broader understanding of audience decision-making. Additionally, this research could be expanded upon by integrating non-linear models like Poisson regression, to catch potential non-linear effects.

# 6. Understanding Box-Office Dynamics: The Role of Timing and Audience in Movie Revenue Prediction

## 6.1. Introduction

The film industry stands as one of the most prominent global entertainment sectors, generating billions of dollars annually. Amid intensifying competition, studios and distributors are under increasing pressure to maximize box-office revenues. A key determinant of box-office success lies in the strategic timing of movie releases. Holidays, characterized by increased leisure time and elevated audience availability, have long been viewed as prime opportunities for maximizing attendance (Ryoo et al., 2021). However, the extent to which holiday releases influence box-office performance remains a subject of debate, especially when considering variations across different types of audiences. Adult-oriented films, which target niche or mature demographics, and family-friendly (non-adult) movies, often catering to broader audiences, may experience these effects differently.

This thesis addresses a critical problem within this setting: how release timing during holidays impacts box-office revenues and whether these effects vary by audience demographic. Despite widespread industry practices that favor holiday releases, existing research offers conflicting conclusions about their efficacy. While some studies highlight the potential revenue boosts associated with holiday periods due to increased audience availability (Ryoo et al., 2021), others argue that such effects may be diluted by factors such as competition and audience segmentation (Huang, Boh, & Goh, 2017). Notably, previous research often treats timing as an isolated factor, failing to integrate its nuanced effects across different audience groups (Lash & Zhao, 2016).

This thesis builds on existing studies by exploring the differential impact of holiday timing on adult and non-adult movies. By employing advanced predictive modeling techniques and leveraging a

comprehensive dataset that includes audience metrics, critical reception, genres, and runtime, this study addresses two central questions:

1. Does releasing a movie during a holiday significantly impact its box-office performance?
2. How does this impact differ between adult-oriented and non-adult movies?

To investigate these questions, this research combines traditional econometric techniques, such as Ordinary Least Squares (OLS) regression, with advanced machine learning models, including CatBoost, Random Forest, and Gradient Boosting. The inclusion of interpretability methods, such as feature importance and SHAP analysis, further elucidates the roles of holiday timing and other predictors in driving box-office performance.

The findings reveal that the effects of holiday timing are highly dependent on audience segmentation. For non-adult movies, releasing during holidays consistently yields positive revenue outcomes, aligning with the notion that family-oriented content benefits from increased leisure time and collective viewing behaviors. In contrast, adult movies show marginal or even negative effects during holidays, possibly due to competition with family-friendly films dominating theaters. The overall significance of holiday timing is limited when considering all movie types collectively, underscoring the importance of targeted strategies based on audience demographics.

By providing actionable insights for film studios and distributors, this study contributes to the broader understanding of box-office performance drivers and offers nuanced guidance for optimizing release schedules. It fills critical gaps in the literature by systematically comparing the effects of holiday timing across audience types, presenting a more integrated perspective on the interplay between release timing, audience demographics, and box-office success.

## 6.2. Gaps in Literature and Hypothesis Formulation

Release timing has long been recognized as a critical determinant of box-office success, yet its nuanced effects remain underexplored (Elberse, 2007). This study examines the differential impacts of holiday periods on revenues, focusing on understanding how timing interacts with audience demographics.

Holidays are characterized by increased leisure time, higher audience availability, and a social inclination toward group activities, making them particularly appealing for movie releases. Prior research by Ryoo et al. (2021) demonstrated that films released during holiday periods consistently outperform those released at other times, attributing this to peak audience attendance, observing seasonal trends that align with school breaks and reduced work commitments. However, these studies often fail to address the heterogeneity of holiday impacts across different audience types (Moon et al., 2010). Building on these insights, this research hypothesizes:

**$H_1$:** *Releasing a movie during a holiday period significantly impacts box-office revenues.*

Family-friendly movies, in particular, are often timed to coincide with holidays to capitalize on increased family engagement and availability. Ryoo et al. (2021) highlighted the synergistic effect of holidays and family-oriented films, as these periods create optimal conditions for group attendance, identifying a strong seasonal pattern favoring kid's movies during school vacations. While these findings suggest that holiday timing amplifies the success of kid's movies, they lack a focused analysis of this relationship. This study hypothesizes:

**$H_{2A}$:** *Releasing a kid's movie during a holiday period positively impacts box-office revenues.*

Conversely, the dynamics of holiday timing may not favor adult-oriented films Lash & Zhao (2016). Mature-themed movies often target niche audiences, whose viewing habits may not align with the family-centric atmosphere of holidays. Ryoo et al. (2021) noted that holiday periods are dominated by broad-appeal films, potentially diminishing the visibility and attendance of adult

movies. It was suggested that holidays encourage social viewing experiences, which may be less relevant to the individualistic consumption patterns typical of mature audiences. These insights underpin the hypothesis:

**$H_{2B}$:** *Releasing an adult movie during a holiday period negatively impacts box-office revenues.*

In addressing these hypotheses, this study seeks to answer the overarching question: *How does release timing impact box-office revenues?* By examining the nuanced effects of holiday periods on different audience segments, this research contributes to a deeper understanding of timing as a strategic tool for optimizing box-office performance.

## 6.3. Methodology

### 6.3.1. Research Design

This study systematically investigates how holiday releases influence box-office success, focusing on adult and non-adult audience segmentation. The research was structured into stages, beginning with defining the research question and three hypothesis to explore the general and specific effects of holiday releases on revenues.

Data acquisition involved collecting comprehensive datasets on movie performance, audience metrics, and release dates from reputable sources. The data was cleaned to address missing values and standardize formats. Feature engineering created variables, including a binary holiday indicator, audience segmentation, and temporal factors.

Analytical techniques included OLS regression and more advanced machine learning models (e.g., CatBoost, Random Forest) to uncover relationships between features and revenues. Results were analyzed to provide actionable insights and suggest directions for future research.

### 6.3.2. Data Preparation

#### 6.3.2.1. Data Acquisition

For this research, two datasets were sourced from Kaggle, a platform providing public datasets for analysis. The first, *rotten_tomatoes_movies.csv*, was obtained from the "Clapper Massive Rotten Tomatoes Movies and Reviews" dataset (source). It includes key details about movies reviewed on Rotten Tomatoes, such as critic ratings, audience scores, and genres. The second, *box office.csv*, came from the "Box Office" dataset (source), providing financial data like budgets and gross revenue. These datasets, downloaded in CSV format, were preprocessed to address missing values and ensure compatibility, forming the basis for this analysis.

#### 6.3.2.2. Data Merging

Two datasets from Kaggle, *rotten_tomatoes_movies.csv* and *box_office.csv*, were merged to create a unified dataset for analysis. The first provided details on titles, genres, and ratings, while the second included financial data regarding revenue. Preliminary cleaning harmonized movie titles, used as key fields for an inner join. This join ensured only movies present in both datasets were retained, preserving data relevance. The merging resulted in a dataset of 1,221 rows, representing unique movies and combining critical review metrics with box-office performance. This comprehensive dataset served as the foundation for the study’s analyses.

#### 6.3.2.3. Data Cleaning

A structured cleaning process was applied to ensure data quality, addressing duplicates, missing values, and irrelevant columns. Duplicate rows were removed from both the Rotten Tomatoes and Box Office datasets to avoid redundancy. Missing values in essential columns, such as *genres*, *directors*, *original_release_date*, and *runtime*, were addressed by dropping rows where data was incomplete to maintain dataset integrity.

Irrelevant columns, including *rotten_tomatoes_link*, *movie_info*, and *streaming_release_date*, were removed to streamline the datasets. Similarly, in the Box Office dataset, columns like *Domestic*, *Domestic_percent*, and *Foreign_percent* were dropped due to redundancy. Rows labeled "NR" in the *content_rating* column were excluded for consistency.

This cleaning process produced datasets free of duplicates and irrelevant data, forming the foundation for further analysis.

### 6.3.3. Variables and Feature Engineering

#### 6.3.3.1. Variables Created

Feature engineering played a key role in transforming raw data into meaningful variables, capturing critical aspects of the dataset. These variables enhanced the analysis of relationships between movie characteristics, audience reception, and financial performance.

The *adult* variable classified movies as adult (1) or non-adult (0) based on their content rating, with "R"-rated movies categorized as adult. This binary classification enabled focused comparisons between demographic groups.

The *holiday* variable captured whether a movie's release date aligned with key holiday periods, which often coincide with higher audience availability (Ryoo et al., 2021). Using a custom function, three holiday windows were identified: Christmas and New Year (December 21–January 1), summer vacations (June 21–September 1), and Easter (the week before and after Easter Sunday, calculated using the Anonymous Gregorian algorithm). Movies released during these periods were assigned a value of 1, while others received a 0. This variable was pivotal for exploring seasonal patterns in box-office performance and audience engagement, offering valuable insights into the strategic importance of release timing.

Temporal variables, including *year* and *recent*, added context to the dataset. The year variable provided a reference for each movie's release, while recent measured a movie's age in years

(difference between release year and 2024). These features helped analyze trends over time and their correlation with audience preferences or financial performance.

Finally, ratio-based variables – *top_ratio*, *fresh_ratio*, and *rotten_ratio* – quantified the proportions of top critic reviews, positive ("fresh") reviews, and negative ("rotten") reviews, respectively. These metrics deepened the understanding of critical reception and its influence on movie success.

#### 6.3.3.2. Genre Encoding

The *genres* column, containing comma-separated strings, was transformed into binary features for analysis. The column was split into lists of individual genres, and the top 10 most common genres were identified. For each top genre, a binary column was created (e.g., *is_action*), indicating 1 if a movie belonged to the genre and 0 otherwise. Genre names were standardized by converting them to lowercase and removing special characters. The temporary split-genres column was then removed. This transformation converted categorical genre data into a machine-readable format, enabling integration into the analysis while preserving information on multiple genre assignments.

#### 6.3.3.3. Final Dataset

After merging, cleaning, and feature engineering, the final dataset comprised 1,221 rows, each representing a unique movie. This comprehensive dataset integrated critical, audience, and financial data, along with engineered features, providing a robust foundation for analysis.

For this research, the dataset was split into two subsets: Adult Movies (475 rows) and Non-Adult Movies (746 rows), based on the adult variable derived from the *content_rating* column. Movies rated "R" were classified as adult, while others were non-adult. This split enabled focused analysis of audience-specific dynamics and content influences on performance.

#### 6.3.3.4. Relevance to the Research Question

The central research question of this thesis is: How does releasing a movie during a holiday period influence its box-office success? To address this question, the feature engineering process focused on creating variables that capture the timing of a movie's release, audience segmentation, and critical reception, ensuring the dataset was well-aligned with the research objectives.

The *holiday* variable plays a pivotal role in this analysis. It was designed to identify whether a movie was released during key holiday periods, such as Christmas, summer vacations, or Easter, times when audience availability and movie-going activity are typically heightened (Ryoo et al., 2021). By incorporating this variable, the dataset captures a critical dimension of the research question, enabling an analysis of how these periods affect box-office performance.

Complementing this, the *adult* variable segments the dataset into adult and non-adult movies, facilitating a deeper exploration of audience-specific dynamics. This differentiation is vital for testing hypotheses related to demographic preferences, such as whether kid's movies perform better during holidays or whether adult movies face challenges competing in these periods.

Additionally, the *recent* variable accounts for the age of a movie, allowing for a nuanced understanding of how newer releases, especially during holidays, compare to older films (Moon et al., 2010). Variables such as *top_ratio*, *fresh_ratio*, and *rotten_ratio* further enrich the analysis by quantifying critical reception, providing insights into whether audience and critic respons(Lash & Zhao, 2016)ay releases (Lash & Zhao, 2016).

Together, these features align closely with the research question, enabling a robust examination of the interplay between holiday timing, audience segmentation, and critical reception, and their collective impact on box-office success.

### 6.3.4. Analytical Techniques

#### 6.3.4.1. Machine Learning Predictive Models

This research employed predictive models to explore relationships between movie features and worldwide box office revenues, using Ordinary Least Squares (OLS) regression and ensemble-based machine learning models.

OLS Regression served as a baseline, offering a straightforward approach to understanding linear relationships. The *worldwide* target variable (which represents the worldwide box-office revenue) was log-transformed to address skewness, and predictors were scaled and standardized. Comparisons were made with and without the holiday variable to assess its impact. OLS's simplicity and interpretability highlighted key predictors and their influence on revenues.

Advanced machine learning models, including Random Forest, Gradient Boosting, XGBoost, and CatBoost, were used to capture complex, non-linear relationships. A five-fold cross-validation technique ensured robust evaluation and reduced overfitting. During each fold, features were standardized, and model performance was measured using metrics such as Weighted Mean Absolute Percentage Error (WMAPE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE).

#### 6.3.4.2. Hypothesis Testing

This research investigates the impact of holiday periods on box office revenues, focusing on differences between kid's and adult movies.

The holiday variable, created during feature engineering, serves as the key independent variable in predictive models such as Ordinary Least Squares (OLS) regression and ensemble regressors. Statistical significance is evaluated using p-values and confidence intervals in the OLS model, assessing whether holiday timing significantly impacts revenues.

Stratified analysis for kid's and adult movies ensures hypotheses are rigorously tested, with metrics like Weighted Mean Absolute Percentage Error (WMAPE) used to evaluate model accuracy. These methods systematically test the hypotheses, providing insights into the influence of holiday timing on movie performance across demographics.

#### 6.3.4.3. Performance Evaluation Metrics

To evaluate the accuracy of the predictive models in this study, different performance metrics were employed based on the type of model. For Ordinary Least Squares (OLS), the Mean Squared Error (MSE) was used to assess the fit by measuring the average squared differences between predicted and actual values. For more advanced machine learning models, Weighted Mean Absolute Percentage Error (WMAPE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE) were applied. WMAPE accounts for variations in scale, making it well-suited for datasets like box-office revenues, where values vary significantly. MAPE provides an average percentage error, offering an intuitive measure for relative accuracy, while MAE, expressed in real-world units, indicates the typical deviation of predictions from actual values.

## 6.4. Results

The analysis of model performance and its implications provides crucial insights into the relationships between holiday periods, demographic segmentation, and box-office revenues. Both Ordinary Least Squares (OLS) regression and more advanced machine learning models were evaluated across three datasets: adult movies, non-adult movies, and the complete dataset, with and without the inclusion of the holiday variable.

### 6.4.1. Model Performance Overview

OLS Regression results indicated varying levels of significance for the *holiday* variable across datasets. In the adult movie dataset, the *holiday* variable had a negative coefficient (-0.1294) with

a marginally significant *p-value of 0.096 (*Table 13*)*. This suggests a weak but potential negative impact on box-office revenues, aligning with the third hypothesis ($H_{2B}$). For non-adult movies, the holiday coefficient was positive (0.1776) and statistically significant ($p = 0.014$) (Table 14), supporting the second hypothesis ($H_{2A}$) that releasing kid's movies during holidays positively influences revenues. In the complete dataset, the *holiday* variable was insignificant ($p = 0.249$) (Table 15), indicating no overarching impact of holidays on revenues ($H_1$). However, the significant positive effect for non-adult movies ($p = 0.014$) and the marginal negative impact for adult movies ($p = 0.096$) suggest audience-specific influences, partially supporting $H_1$ depending on segmentation. The following equation describes the linear relationship between the variables:

$$box_{office} = \beta_0 + \beta_1 audience_{count} + \beta_2 fresh_{ratio} + \beta_3 holiday + \beta_4 is_{actionandadventure} + \beta_5 is_{arthouseandinternational} + \beta_6 is_{comedy} + \beta_7 is_{drama} + \beta_8 is_{horror} + \beta_9 is_{mysteryandsuspense} + \beta_{10} is_{romance} + \beta_{11} is_{sciencefictionandfantasy} + \beta_{12} recent + \beta_{13} rotten_{ratio} + \beta_{14} runtime + \beta_{15} tomatometer_{rating}$$

The other machine learning models, evaluated using Weighted Mean Absolute Percentage Error (WMAPE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE), revealed additional patterns (Table 16). For adult movies, models including the *holiday* variable slightly underperformed compared to those without, with CatBoost achieving the best average *WMAPE of 0.5948* with the variable and *0.5984* without. This marginal difference suggests that the *holiday* variable is not a strong predictor for adult movies and aligns with the weak support for $H_{2B}$.

In contrast, non-adult movies exhibited a more substantial benefit from including the *holiday* variable, with Random Forest yielding the best performance (*WMAPE = 0.5499* with the *holiday* variable, *0.5489* without). This aligns with the OLS results, underscoring the positive influence of holiday releases on non-adult movies, as hypothesized in $H_{2A}$. For the complete dataset, the inclusion of the *holiday* variable had negligible negative impact, with CatBoost consistently

performing best (*WMAPE = 0.5491* with the *holiday* variable, *0.5465* without). These findings emphasize the varying significance of holidays depending on the audience demographic.

### 6.4.2. Implications of the Results

The results have clear implications for the film industry, particularly in strategic release scheduling. For adult movies, the evidence suggests that holiday periods may not offer a substantial advantage and could even detract from revenues, potentially due to competition with family-oriented content. Studios producing adult movies might benefit from avoiding holiday releases to avoid such competition, aligning with $H_{2B}$.

For non-adult movies, the positive association between holiday releases and revenues validates the industry practice of releasing family-friendly content during school breaks and festive seasons. This approach capitalizes on increased audience availability, as families are more likely to attend theaters together during these times. These findings support $H_{2A}$ and highlight the strategic importance of aligning release schedules with audience behavior.

The overall insignificance of the *holiday* variable in the complete dataset suggests that while holidays matter for specific demographics, their broader influence is limited. This underscores the importance of audience segmentation in understanding the drivers of box-office success.

In conclusion, the combined use of OLS and machine learning models provided a nuanced understanding of the relationship between holiday releases and box-office performance. The results validate $H_{2A}$, provide weak support for $H_{2B}$, and suggest limited overall significance for $H_1$. These insights offer valuable guidance for producers and distributors in optimizing release strategies to maximize revenue.

### 6.4.3. Feature Importance Analysis

#### 6.4.3.1. Key Features Across Datasets

In the non-adult movies dataset, both Random Forest and SHAP analyses consistently highlighted *tomatometer_count* as the most critical predictor, with an importance score of 0.3458 and SHAP importance of 40.7 million, respectively. This indicates that the volume of critical reviews significantly impacts box-office revenues for family-oriented films. Audience engagement metrics like *audience_count* also ranked high, reflecting the importance of broad audience appeal for non-adult movies. Interestingly, the *holiday* variable had minimal importance (0.0032 in Random Forest and 744,576 in SHAP), suggesting that while holidays boost revenues (as supported by $H_{2A}$), their predictive power is overshadowed by other factors like audience size and critical reception (Figure 14 and Table 18).

For the adult movies dataset, CatBoost analysis ranked *audience_count* as the most influential feature (21.02), followed by *top_ratio* (12.23). These results underscore the critical role of audience size and the proportion of top critic reviews in determining the success of mature-themed content. The SHAP analysis further corroborated these findings, with *top_ratio* and *audience_count* showing the highest contributions. The *holiday* variable held moderate importance (3.38 in CatBoost, 3.5 million in SHAP), suggesting a nuanced relationship between holiday timing and adult movie revenues, consistent with the weak support for $H_{2B}$ (Figure 15 and Table 17).

In the complete dataset, *audience_count* emerged as the dominant predictor across all analyses, with CatBoost ranking it at 21.02 and SHAP assigning it a massive importance of 77.2 million. This reinforces the centrality of audience engagement in box-office success across all movie types. Features like *top_ratio* and *recent* also played significant roles, indicating that movies with a higher proportion of top critic reviews and recent releases tend to perform better. The *holiday* variable had modest importance (3.38 in CatBoost and 6.8 million in SHAP), supporting $H_1$ that holiday timing

has a limited overarching effect on revenues when all movie types are considered together (Figure 16 and Table 19).

#### 6.4.3.2. Implications for Hypotheses and Research Question

The results from feature importance analyses have critical implications for the research question: How does releasing a movie during a holiday period influence its box-office success?

The modest importance of the *holiday* variable across all datasets underscores that while holidays can influence revenues, they are not the sole determinant of success. This supports rejecting the null hypothesis for $H_{2A}$, which posits that holiday releases positively impact non-adult movie revenues, but provides weak evidence for rejecting the null in $H_1$, regarding the general holiday impact, and $H_{2B}$, concerning negative impacts on adult movies. Audience-related features like *audience_count* dominated across all datasets, highlighting the universal importance of engaging a large audience. For non-adult movies, this suggests that holidays may indirectly boost revenues by increasing family attendance, aligning with $H_{2A}$. Features such as *top_ratio*, *fresh_ratio*, and *tomatometer_count* consistently ranked high, emphasizing the value of favorable critical reception, particularly from top critics, in driving revenues. For adult movies, this suggests that mature audiences prioritize quality, potentially mitigating negative effects of competition with family-friendly content during holidays, as suggested by $H_{2B}$. Additionally, the *recent* variable showed moderate importance across datasets, indicating that newer releases tend to attract more revenue. This trend offers context for how release timing, including holidays, interacts with audience preferences and market dynamics.

The feature importance analysis reveals that while holidays influence box-office performance, their impact is secondary to audience engagement and critical reception. This nuanced understanding aligns with the hypotheses and provides actionable insights for movie studios. For non-adult movies, leveraging holidays can amplify revenues by tapping into increased family attendance. In

contrast, adult movies may benefit from focusing on quality and avoiding direct competition during peak holiday periods. These findings contribute to a more refined strategy for optimizing release schedules and targeting audience segments effectively.

## 6.5. Limitations & Future Research

This study offers valuable insights into the impact of holiday periods on box-office performance but has limitations. The data, while comprehensive, does not fully account for regional variations in holidays and cultural preferences, potentially overlooking localized trends. The holiday variable also simplifies complex phenomena, grouping diverse holiday periods into a single feature, which may dilute specific effects. Additionally, confounding factors like marketing budgets, star power, and competition were not explicitly controlled, complicating the interpretation of the holiday variable's impact. While machine learning models identified correlations, they do not establish causation. A more detailed segmentation by genre for instance could also enhance the analysis.

Future research could include region-specific data to explore cultural influences on holiday impacts and incorporate additional variables like marketing expenditure and competition to improve model comprehensiveness. Causal inference methods, such as difference-in-differences, could better isolate holiday effects, and dynamic modeling techniques could capture temporal trends. Further segmentation and alternative success metrics, such as streaming performance or critical acclaim, could provide a more holistic view of movie success, building on this study’s findings.

# 7. Quantitative Analysis of Factors Driving Movie Sequel Success: A Predictive Modeling Approach

## 7.1. Introduction

Within the movie landscape, movie sequels have become a strategic focus, functioning as brand extensions that build on the successes of their predecessors. However, the performance of sequels varies widely - some become box office triumphs, while others fail to meet expectations (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009). Understanding the factors driving sequel success is crucial for investors and studios seeking to mitigate risks and maximize returns. Existing research highlights the importance of incorporating diverse factors into predictive models for box office performance. Factors such as genre, runtime, and continuity in creative teams are known to influence sequel outcomes (Belvaux & Mencarelli, 2021; Moon et al., 2010). Additionally, the volume of audience reviews, have been shown to significantly impact sequel performance, often outweighing the influence of sentiment (Lehrer & Xie, 2021; Navarathna et al., 2019). The role of release timing, both in terms of seasonal strategies and the interval between original and sequel releases, also emerges as a critical determinant of success (Hennig-Thurau et al., 2009; Somlo et al., 2011). This study develops two ML models to predict sequel success. The first forecasts absolute box office revenue, while the second classifies whether a sequel outperforms its predecessor. By integrating variables like release timing, audience reviews, and film attributes, these models provide insights into both absolute and relative sequel performance, addressing gaps in existing research. The findings of this research have practical implications for decision-making in the film industry. Enhanced prediction accuracy can assist studios and investors in making informed decisions about resource allocation, marketing strategies, and production timeline.

## 7.2. Research Questions and Hypothesis Development

Despite advancements in understanding sequel success, key gaps remain particularly regarding relative success metrics, the integration of diverse predictors, and the dominance of U.S.-centric data. Most studies prioritize absolute box office revenue, overlooking how sequels perform relative to their predecessors. This relational metric offers deeper insights into audience expectations and brand extension dynamics (Hennig-Thurau et al., 2009). For example, a sequel might achieve high revenues but still underperform compared to its original if it fails to meet expectations. This underscores the importance of relative metrics as they capture franchise sustainability and the effectiveness of brand extensions, which absolute revenue figures alone cannot reveal, providing a nuanced understanding of how audience expectations and market dynamics evolve across sequels(Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009). Another limitation is the fragmented analysis of predictors like runtime, reviews, and release timing, often neglecting their interactions and relationship towards each other (Lehrer & Xie, 2021). Similarly, while review volume is known to predict absolute success, its role in relative success warrants further investigation (Navarathna et al., 2019). Geographic diversity is also underrepresented, with most studies focusing on U.S. markets. Cultural and economic differences significantly influence sequel dynamics, yet international variations remain poorly understood (Belvaux & Mencarelli, 2021). By incorporating global data, this study provides a broader perspective on how sequel success varies across markets. To address these gaps, this study examines runtime, release timing, and audience engagement metrics, integrating both absolute and relative success measures on an international collection of films. The effect of each metric will be explored on both of our success metrics, meaning the absolute sequel box-office and the relative sequel box-office compared to the original box-office. Building on the identified gaps, the following hypotheses are proposed:

**H1:** *Sequel-runtime has a significant positive effect on sequel success (absolute or relative).*

**H2:** *Time interval between the sequel and original movie has a significant negative effect on sequel success (absolute or relative).*

**H3:** *Review volume of the original movie has a larger significant positive effect on sequel success (absolute or relative) than review sentiment.*

By testing these hypotheses, the study seeks to advance theoretical understanding and provide actionable insights for optimizing sequel performance across diverse contexts.

## 7.3. Methodology

### 7.3.1. Research Design

The aim of this study is to examine the determinants of movie sequel success by employing predictive models and conducting hypothesis-driven analysis. Two distinct dimensions of success are analysed: absolute success, defined by box office revenue as a measure of standalone financial performance, and relative success, which evaluates a sequel's revenue performance in comparison to its predecessor. This dual approach provides a comprehensive understanding of sequel dynamics, incorporating both financial achievement and franchise sustainability, consistent with prior research (Hennig-Thurau et al., 2009; Moon et al., 2010). This is achieved by integrating hypothesis testing with ML modeling. Hypothesis testing is used to explore relationships between key factors and the success measures while ML techniques are employed to develop predictive models capable of quantifying the influence of these variables on sequel success.

### 7.3.2. Data Preparation

#### 7.3.2.1. Data Acquisition

The initial dataset was obtained from *The Numbers* on September 16, 2024. This source provided a list of sequels alongside their worldwide box office revenues. To complement this information, the corresponding original movies and their release years were manually researched and added to the dataset. Such manual enrichment ensured the inclusion of accurate and complete data on

original-sequel pairs, a crucial step given the lack of standardized sequel information across public datasets. To enhance the dataset, metadata and review information were integrated from the *Massive Rotten Tomatoes Movies and Reviews* collection on *Kaggle*, accessed on August 18, 2024. These datasets included key variables such as movie titles, release dates, critic and audience review scores, review sentiments, and audience engagement metrics.

#### 7.3.2.2. Combining and Matching Datasets

First, a string-similarity algorithm was used for automated comparing of movie titles and release years to identify potential matches. While this approach provided a strong initial framework, all matches were manually reviewed to ensure accuracy. Only high-confidence matches were included, and movies that could not be reliably linked to *Rotten Tomatoes* datasets were excluded. This rigorous process reduced the dataset to 969 rows, emphasizing precision in linking sequels to their originals. Additionally, the dataset was filtered to include only direct sequels, removing higher-order franchise instalments (e.g., parts three or four). This step ensured the analysis focused solely on comparisons between original movies and their immediate sequels, aligning with the study's objectives. This hybrid approach balanced efficiency with precision, adhering to best practices for handling large, non-standardized datasets (Lehrer & Xie, 2021).

#### 7.3.2.3. Data Imputation and Preparation

Missing original movies' box office revenues data was imputed using group averages based on the language and time period of the films. For instance, films were grouped by primary language (e.g., English or non-English) and categorized by decade (e.g., 1990s, 2000s). Imputation values were calculated within these groups to reflect typical earnings. Language correction factors were further applied to account for structural differences in box office performance between English and non-English films. Such imputation techniques align with prior studies emphasizing the importance of robust handling of missing data in predictive models (Navarathna et al., 2019).

To facilitate direct comparisons between sequels and their originals, the revenues for the original movies were adjusted for inflation. Yearly inflation rates were mapped to the release years of the films and then used to scale the original box office revenues year by year, converting them to the equivalent purchasing power of the sequel's release year. This adjustment ensured that all revenue figures were expressed in comparable terms, eliminating the distorting effects of inflation. The use of inflation-adjusted revenues is consistent with established methodologies in media economics and box office prediction (Hennig-Thurau et al., 2009; Lehrer & Xie, 2021).

#### 7.3.2.4. Final Dataset

The final dataset consisted of 603 rows and 24 columns. Each row represented a unique sequel-original pair, meaning the dataset contained information on 1,206 individual movies in total. The dataset integrated key variables such as inflation-adjusted revenues, review sentiments, audience engagement metrics, runtime differences, and content continuity indicators (e.g., shared directors, writers, and genres between originals and sequels). These features were designed to capture key drivers of sequel success as identified in the literature (Moon et al., 2010; Navarathna et al., 2019). The dataset includes movies with more than 20 different original languages to ensure global relevance of the findings. By focusing on direct sequel-original pairs with reliable and complete data, the dataset provided a robust foundation for subsequent modeling and hypothesis testing.

### 7.3.3. Variables and Feature Engineering

Runtime (*sequel_runtimeMinutes*) and year intervals were included as key variables, reflecting their established relevance in predicting box office outcomes. Studies suggest that longer runtimes are often perceived as indicative of higher production values or narrative complexity, which can positively influence revenue (Moon et al., 2010). To investigate this dynamic, the study incorporated the runtime of both the original and sequel films, along with their difference. Additionally, a variable for the time interval between the original and sequel's release was

calculated (*year_difference*). This variable is significant in understanding residual audience interest and the potential for franchise fatigue (Hennig-Thurau et al., 2009). Audience reviews provide insights into public perceptions and engagement, both of which are critical to box office success. The study utilized review volume or review count (*review_count*), sentiment (*scoreSentiment*), and mentions of the sequel in reviews of the original film. Sentiment, measured as the proportion of positive reviews, was included to capture audience favourability and to allow for analyses between volume and sentiment (Navarathna et al., 2019). To ensure consistency in ratings, raw review scores were normalized into a standard scale. Creative continuity between sequels and their originals is another critical factor, as retaining key creative personnel such as directors, writers, and distributors often ensures narrative consistency, which can reinforce brand identity and audience loyalty (Moon et al., 2010). For this study, binary variables were created to indicate whether the same director, writer, and distributor were involved in both films. Additionally, a variable reflecting whether the sequel's genre matched that of the original was introduced, recognizing the importance of genre consistency in maintaining audience expectations and interest. Categorical variables such as language and genre were encoded to capture their influence on sequel success. Language-specific differences, for instance, have been shown to affect global box office performance, with English-language films often dominating international markets (Belvaux & Mencarelli, 2021). To reflect these dynamics, the top five languages in the dataset were encoded as binary indicators. Similarly, the study identified the ten most common genres and encoded them to account for genre-specific patterns in audience preferences and revenue performance.

### 7.3.4. Analytical Techniques

#### 7.3.4.1. Machine Learning Models for Success Prediction

In this study two types of ML models were developed. Regression models (Model I) were used to predict absolute success, measured by box office revenue, while relative success was framed as a

binary classification task (Model II), predicting whether a sequel outperformed its original in revenue. Here ML techniques are particularly suited due to their ability to capture non-linear relationships and features interactions (Lehrer & Xie, 2021). Various algorithms were tested, including random forests and gradient boosting machines, chosen for their robust predictive performance and established use in entertainment analytics (Lash & Zhao, 2016). Linear regression (for absolute success) and logistic regression (for relative success) served as baseline models. Emphasis was placed on predictive accuracy rather than coefficient interpretability, as multicollinearity does not impact model performance. An 80/20 train-test split was applied for model evaluation, with k-fold cross-validation ($k = 5$) during training to prevent overfitting and provide robust performance estimates (L. Liao & Huang, 2021). Hyperparameter tuning was omitted to prioritize generalization over training optimization. Initial experiments showed that default hyperparameters in the algorithms provided reliable results without overfitting, aligning with prior findings (Lash & Zhao, 2016; Lehrer & Xie, 2021).

#### 7.3.5. Performance Evaluation Metrics

The regression models' predictive performance was assessed using *Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE)*, and *Weighted Mean Absolute Percentage Error (WMAPE)*. These metrics were chosen for their ability to capture accuracy amid high revenue variability. *MAE* measures average error in absolute terms, *MAPE* emphasizes relative error, and *WMAPE* accounts for revenue magnitude, mitigating disproportionate errors in low-revenue cases (Lash & Zhao, 2016; Lehrer & Xie, 2021; Moon et al., 2010). *WMAPE* was prioritized for its robustness and balanced assessment. For classification models predicting relative success, *accuracy, precision, recall, F1-score*, and *AUC-ROC* were used, with a primary focus on *accuracy* and *AUC-ROC* to evaluate model discrimination (Lash & Zhao, 2016). Ablation studies tested the

influence of key features by sequentially removing them and analysing performance changes, following Lehrer & Xie (2021), to deepen insights into variable contributions to predictions.

#### 7.3.5.1. Hypothesis Testing

Statistical hypothesis testing complemented the ML analyses by evaluating the significance of relationships between selected features and sequel success. The null hypothesis posited no significant relationship, while the alternative hypothesis posited a significant relationship. Hypotheses were tested for both absolute and relative success to ensure consistency across dimensions. For predicting absolute box-office success, *Spearman's rank correlation coefficient* was used to assess monotonic relationships between predictors, such as year differences and sequel revenues. This non-parametric method accounted for potential non-linearities, as recommended in similar studies (Navarathna et al., 2019). Linear regression models provided further insights, with metrics like *regression slope, intercept*, and $R^2$ used to quantify relationships. The following Linear Regression Formula was used:

$$absolute\_box\_office\_success = \beta_0 + \beta_1 \cdot independent\_variable$$

For relative success, logistic regression estimated the probability of success based on the selected predictors. The following Logistic Regression Formula was used:

$$relative\_box\_office\_success = \frac{1}{1 + exp\big(-(\beta_0 + \beta_1 \cdot independent\_variable)\big)}$$

Predictor significance was assessed using *p-values* from maximum likelihood estimation, while *McFadden's pseudo* $R^2$ and the *AUC-ROC* score evaluated model fit and discrimination (L. Liao & Huang, 2021). By integrating *Spearman's rank correlation* for continuous outcomes and logistic regression for binary outcomes, this study evaluated the significance and predictive power of key features. A significance threshold of $p < 0.05$ ensured that findings were statistically robust.

## 7.4. Results

### 7.4.1. Model Performance

#### 7.4.1.1. Model I: Absolute Success

Among the evaluated models, CatBoost demonstrated the highest predictive performance, with a *WMAPE* of *61.8%* and an *MAE* of approximately *$89M*. This success is caused by its ability to effectively model non-linear relationships and complex feature interactions. In contrast, Linear Regression was the least effective, with *WMAPE* exceeding *83%* and the highest *MAE* of *$118M*, highlighting its limitations in capturing the intricate dynamics of box office revenues. The results for the intermediate models, including Random Forest, XGBoost, and Gradient Boosting, are summarized in 13.3.1.1, which provides a detailed comparison of all metrics. These findings highlight the importance of using advanced tree-based models like CatBoost and Random Forest for predicting absolute success, particularly when working with datasets characterized by non-linear interactions and feature complexities.

#### 7.4.1.2. Model II: Relative Success

Among the models, CatBoost was the best, achieving the highest *accuracy (77.7%)* and *AUC-ROC (86.3%),* demonstrating its superior ability to model complex, non-linear relationships. Its *F1-score* of *77.1%* reflects a well-balanced *precision* and *recall*, confirming its reliability across metrics. Logistic Regression, while simpler, delivered competitive results with an *accuracy* of *77.1%* and an *AUC-ROC* of *84.4%.* These outcomes underscore its viability as a baseline model, despite its inherent linear assumptions. In contrast, Random Forest had the weakest performance, with an *accuracy* of *74.1%* and inconsistent metric results. Although its *AUC-ROC* of *84.6%* indicated reasonable discriminatory power, its lower *accuracy* and tendency to overfit made it less effective. Complete performance details for all models can be found in 0.

#### 7.4.1.3. Comparative Analysis and Implications

CatBoost consistently outperformed other models across both tasks, emphasizing its ability to handle high-dimensional data and non-linear interactions effectively. Its gradient-boosting framework integrates categorical and continuous features without extensive preprocessing, which made it particularly well-suited for the complex dynamics of sequel success prediction.

Conversely, the limitations of simpler models, such as Linear Regression, highlight the importance of choosing algorithms aligned with the complexity of the data and prediction task. While linear models offer interpretability, their inability to capture intricate feature interactions diminishes their effectiveness. Similarly, Random Forest, while competitive, demonstrated issues with scalability and generalization compared to gradient-boosting approaches.

### 7.4.2. Feature Importance

#### 7.4.2.1. Impurity-Based Rankings

The impurity-based rankings, derived from the best-performing models for each predicting, provide insights into the relative influence of key features associated with the hypotheses. The full rankings are visualized in the appendix (13.3.1.1 and 13.3.1.2). In Model I, review count emerged as a dominant feature, ranking second overall with a contribution of *12.30%*, while sequel runtime also proved influential, contributing *8.09%* to the model's predictions. Meanwhile, year difference *(4.10%)* and review sentiment *(2.99%)* were less prominent but still meaningful. In Model II sequel runtime took on a more pronounced role, with an importance of *10.91%,* suggesting that runtime might be particularly relevant when distinguishing a sequel's performance relative to its original. Review count, while still significant *(7.02%)*, played a slightly less dominant role than in the Model I, while review sentiment *(5.14%)* gained relative importance. Year difference *(4.56%)* maintained a consistent level, reinforcing the idea that timing impacts audience engagement across both success measures similarly. These observations suggest that engagement metrics, particularly the

sheer volume of audience interactions, are consistently vital for box office success. However, nuances emerge depending on the type of success being predicted: Sequel runtime seems to carry greater weight for comparisons within a franchise, while review volume adds more value when evaluating absolute performance. This highlights the multifaceted nature of audience preferences and their impact on sequel success.

#### 7.4.2.2. Ablation Studies

To assess the practical importance of individual features, ablation studies were conducted by systematically removing predictors and observing the resulting impact on model performance. We focused on the best-performing model for both tasks, CatBoost. Details for other models are provided in 13.3.2.1 and 13.3.2.2. In the absolute success model, CatBoost achieved the best performance with a *WMAPE* of *61.8%* and a *MAE* of *$89.31M*. Excluding sequel runtime caused a large performance drop, with *WMAPE* increasing to *65.4%* and *MAE* rising to *$94.65M*. This underscores its critical role as influential predictor of absolute box office revenue. Removing year difference slightly improved *WMAPE* to *60.8%,* but this may indicate compensatory effects from other features. The *MAE* reduced marginally to *$87.99M*, suggesting that its contribution is less pronounced compared to sequel runtime. The exclusion of the review sentiment had minimal impact, with *WMAPE* changing only slightly to *61.6%* and *MAE* to *$88.98M*. This highlights its relatively lower practical importance. Omitting review volume resulted in an increased *WMAPE* of *65.4%* and *MAE* of *$94.80M* reinforcing the idea that sentiment plays a secondary role compared to volume. For the relative success task, CatBoost achieved a *ROC AUC* of *0.863* and an *accuracy* of *77.7%* with the full feature set. Removing sequel runtime caused a substantial decrease in *ROC AUC* to *0.846* and *accuracy* to *75.4%,* confirming its importance for distinguishing successful sequels. Excluding year difference reduced *ROC AUC* to *0.853*, with accuracy holding steady at *77.7%.* While year difference remains relevant, it is less critical than sequel runtime for this task.

Omitting review volume had negligible effects, with *ROC AUC* decreasing only slightly to *0.862* and *accuracy* decreasing to *77.1%.* Interestingly, removing the review sentiment improved *ROC AUC* slightly to *0.865*, suggesting that the model may reallocate predictive weight effectively in its absence and confirming that volume is more important than sentiment.

#### 7.4.2.3. Implications of Feature Importance

The feature importance analyses emphasize the value of quantitative engagement metrics, particularly review volume and sequel runtime, as key drivers of sequel success,while year difference and sentiment showed limited influence. These findings are consistent with prior research advocating for the prioritization of audience awareness and engagement over qualitative measures in modeling (Lash & Zhao, 2016; Lehrer & Xie, 2021).

The ablation study results align closely with the tendencies observed in the impurity-based rankings. Across both models, review volume and sequel runtime consistently emerged as the most impactful features, driving predictive accuracy. While year difference showed some relevance, review sentiment had minimal impact in both contexts, reinforcing the dominance of engagement volume as a predictor. Full metrics for all models are detailed in 13.3.1 and 13.3.2.

The insights from these analyses provide actionable guidance for industry stakeholders. Studios and marketers can focus on amplifying audience engagement metrics, such as increasing review visibility and volume, as these are the most predictive indicators of box office performance.

### 7.4.3. Hypothesis Testing

#### 7.4.3.1. Sequel Runtime

The analysis (13.3.3.1) revealed a significant positive relationship between sequel runtime and absolute box office revenue. The *Spearman correlation coefficient* of *0.340* and a *p-value* of *0.000* confirm that longer runtimes are associated with higher revenue. This supports the rejection of the null hypothesis and validates the alternative hypothesis (H1). Linear regression further supports

these results, indicating that each additional minute of runtime increases revenue by *$4.32M*. However, the $R^2$ value of *0.131* suggests that runtime alone explains a modest proportion of the variability in revenue, highlighting the need to consider other factors. For relative success, sequel runtime also showed a significant positive relationship with the likelihood of outperforming the original (13.3.3.2). The *p-value* of *0.000* and a *coefficient* of *0.0425* suggest that longer runtimes increase the log-odds of relative success. The *Pseudo* $R^2$ of *0.105* indicates moderate explanatory power. These findings provide evidence towards H1, showing that runtime plays an important role in a sequel's ability to outperform its predecessor. This leads to the rejection of the null hypothesis.

#### 7.4.3.2. Time Interval Between Original and Sequel

The time interval between the movies was found to have a weak but statistically significant negative relationship with absolute revenue (13.3.3.1). The *Spearman correlation coefficient* of *-0.127* and a *p-value* of *0.004* support the rejection of the null hypothesis and acceptance of the alternative hypothesis (H2). However, the regression analysis revealed a minimal slope of *-$0.92M* per year and an $R^2$ of *0.001*, indicating that the time interval has a negligible practical impact while still being statistically significant. For relative success, the time interval exhibited a significant negative relationship with the likelihood of a sequel earning more than its predecessor (13.3.3.2). The *p-value* of *0.000* and a *coefficient* of *-0.0518* confirm this finding, suggesting that longer gaps between releases reduce the chances of relative success. However, the *Pseudo* $R^2$ of *0.028* indicates limited explanatory power. This leads to the rejection of the null hypothesis and acceptance of the alternative hypothesis (H2), although the practical significance of the time interval remains low.

#### 7.4.3.3. Number and Sentiment of Reviews

The analysis (13.3.3.1) confirmed that the absolute number of reviews is far more strongly correlated with absolute revenue than the sentiment of those reviews. The *Spearman correlation coefficient* for volume was *0.580*, compared to *0.127* for sentiment. Both were statistically

significant (*p-values* of *0.000* and *0.004,* respectively). Linear regression showed that review count explains *27.8%* of the variance in revenue, compared to just *2.6%* for sentiment. These results strongly support rejecting the null hypothesis and accepting the alternative hypothesis (H3), affirming that engagement volume is a much stronger predictor than sentiment for absolute success. Similarly, for relative success, the absolute number of reviews was a much stronger predictor than sentiment (13.3.3.2). The *p-value* of *0.000* for review count was highly significant, with a *Pseudo* $R^2$ of *0.068*. In contrast, sentiment had a weaker *p-value* of *0.029* and a *Pseudo* $R^2$ of *0.007*. The coefficient values in this context are not directly comparable, as review count represents absolute numbers while sentiment is measured on a normalized scale (0 to 1). These findings confirm that review volume is significantly more impactful than sentiment in predicting relative success. As a result, the null hypothesis is rejected, and the alternative hypothesis (H3) is accepted.

## 7.5. Discussion

The positive link between sequel runtime and success (absolute $R^2 = 0.131$, relative *Pseudo* $R^2 = 0.105$) supports prior studies indicating longer runtimes signal higher production value and narrative depth (Lehrer & Xie, 2021; Moon et al., 2010). However, its modest impact shows runtime is secondary to engagement metrics like review volume. These findings extend Lehrer & Xie (2021) work, affirming runtime's relevance globally. Similarly, the weak negative correlation between sequel success and year difference (absolute $R^2 = 0.001$, relative *Pseudo* $R^2 = 0.028$) supports Hennig-Thurau et al. (2009) argument that shorter gaps sustain interest. Yet, its minimal impact suggests marketing and franchise reputation might be more crucial. This global validation broadens the relevance of timing in diverse markets. Audience reviews emerged as the strongest predictor ($R^2 = 0.278$ for volume vs. *0.026* for sentiment), echoing Navarathna et al. (2019) and broadening their U.S.-centric findings. Global data strengthens these insights and the application of relative success metrics support addressing a literature gap. By comparing absolute and relative

success predictors, this research highlights the importance of examining franchise dynamics through relational metrics, which are underexplored in global contexts (Hennig-Thurau et al., 2009). Additionally, the integration of ML not only replicates but strengthens prior models by applying them to a diverse dataset, making the findings robust across cultural and economic variations. Ultimately, this research emphasizes the centrality of audience engagement and demonstrates the scalability of established predictors beyond U.S.-centric datasets, offering both theoretical and practical advancements in understanding sequel success.

## 7.6. Limitations and Future Research

This study has several limitations. The dataset, while extensive, may be biased by its geographic, temporal, or genre-specific scope, reducing the generalizability of findings. Additionally, key external factors such as marketing campaigns, concurrent releases, and shifting audience preferences were not captured. The predictive models, though robust, face challenges like overfitting and limited applicability to new data. Moreover, constraints in data availability restricted deeper exploration of features like screenplay characteristics and audience behaviour. It is also important to note that the correlation analyses performed in this study do not imply causation, and further experimental or longitudinal research would be needed to establish causal relationships. Future research should expand datasets to include external data, incorporate data sources like real-time audience sentiment and production-level details, and combine ML with econometric methods to better model nonlinear relationships. Additionally, future studies could focus on longer-running sequels, such as parts three, four, and beyond, to examine how franchise longevity and evolving audience dynamics impact success metrics.

# 8. Rating Predictions from Movie Reviews Leveraging Bert: Exploring the Impact of Sentiment Polarity and Review Length

## 8.1. Introduction

In the digital age, the movie industry vast amounts of textual reviews annually, offering rich but unstructured data that profoundly influence audience perceptions, box office outcomes, and industry strategies (Pang et al., 2022). Transforming this valuable feedback into actionable insights is crucial for post-release evaluation and data driven decision making. However, translating textual reviews into precise numeric ratings remains an underexplored challenge in the domain of movie reviews.

Natural Language Processing (NLP) has emerged as a subfield of Artificial Intelligence (AI) and enables computers to process large quantities of human-written text (Banbhrani et al., 2022). The introduction of the transformer architecture in 2017 marked a significant milestone, driving advancements in AI and NLP (Vaswani et al., 2017). Bidirectional Encoder Representations from Transformers (BERT) leverage this architecture to provide powerful tools for various NLP tasks like sentiment analysis and rating prediction from textual data (Devlin et al., 2019; Galal et al., 2024).

This study aims to leverage BERT's capabilities to fine-tune the model on textual movie reviews to predict corresponding rating scores. Sentiment polarity and review length are integrated into rating predictions, using both regression and classification. Sentiment polarity and review length are explored as moderating factors in rating prediction performance (Ahmed & Ghabayen, 2022; Bilal & Almazroi, 2023). The findings contribute to a deeper understanding of rating prediction in the movie domain and provide practical guidance for enhancing analytical approaches in post-release evaluation.

## 8.2. Literature Review & Hypothesis Development

The emergence of transfer learning has significantly advanced NLP by allowing pre-trained models to adapt effectively to specialized tasks (Mosin et al., 2023). Transfer learning models leverage extensive general knowledge gained from large datasets, reducing reliance on large task-specific datasets and improving performance in various NLP tasks. Fine-tuning is a critical step in transfer learning, where pre-trained models are adapted to the specific requirements of a target task. This technique has also been applied in the movie industry, with BERT being used for sentiment analysis of textual movie reviews (Danyal et al., 2024).

Sentiment analysis typically classifies reviews into discrete categories such as positive, neutral, or negative (M. T. Khan et al., 2016). While this approach is effective in many scenarios, it oversimplifies the nuanced opinions often expressed in complex movie reviews (F. H. Khan, Qamar, & Bashir, 2016). Rating prediction advances this process by estimating the numerical ratings that correspond to the textual feedback, offering a more precise quantification of user evaluations (Ahmed & Ghabayen, 2022). Although BERT has demonstrated exceptional performance across various NLP tasks, its application to movie rating prediction is still limited, highlighting opportunities for further research (Z. Liu, 2020; Chambua & Niu, 2021). Sentiment polarity, represented on a continuous scale, ranging from -1 to 1, effectively captures subtle emotional shifts and variations in sentiment analysis (Hemmatian & Sohrabi, 2019; Reagan, Danforth, Tivnan, Williams, & Dodds, 2017). Studies demonstrate that incorporating sentiment polarity into rating prediction frameworks significantly enhances model performance (Ahmed & Ghabayen, 2022; Aljrees et al., 2024; Chambua & Niu, 2021). Despite these advancements, a notable research gap persists regarding the role of sentiment polarity as a moderating factor in rating prediction tasks. Specifically, there is limited understanding of how variations in sentiment

intensity influence rating prediction performance leveraging transfer learning models like BERT. To address this gap the following hypothesis is proposed:

**H1**: *Movie reviews with strong sentiment polarity (greater than or equal to 0.75 or less than or equal to -0.75) improve rating prediction performance.*

It is hypothesized that BERT will more effectively identify extreme ratings (e.g., 1, 5) associated with strong polarity compared to ambiguous middle ratings (2, 3, 4). Validating this hypothesis will determine whether extreme sentiments enhance rating prediction performance while neutral or mixed sentiments introduce ambiguity.

Research further found that specific review attributes such as length, readability, and emotional expression significantly impact interpretability and depth (Lutz, Pröllochs, & Neumann, 2022). Short reviews often lack sufficient detail for effective NLP, whereas longer reviews provide richer context, enhancing predictive performance (Chambua & Niu, 2021). Bilal & Almazroi (2023) demonstrated that BERT models exhibit improved performance when processing longer input texts, as they effectively leverage the model's ability to capture complex semantic relationships (Galal et al., 2024). Despite these findings, there is a notable research gap regarding the direct influence of review length on rating prediction tasks, particularly within transfer learning frameworks like BERT. To fill this gap, the following hypothesis is proposed:

**H2**: *Longer movie reviews yield more precise rating predictions than shorter reviews.*

Testing this hypothesis will clarify how review length impacts model performance, enabling effective applications of BERT in capturing rich semantic nuances unique to movie reviews.

This study explores how sentiment polarity and review length affect model performance in rating prediction from movie reviews leveraging BERT. Answering this research question will provide insights into factors influencing model performance, enabling more tailored and effective applications within the movie industry.

### 8.2.1. Data & Research Methodology

### 8.2.2. Data

This study utilized a dataset of 1,130,017 critic reviews from Rotten Tomatoes, a widely recognized platform for aggregating movie reviews. Covering over 17,000 movies, it was collected on October 31, 2020 (Leone, 2020). Each entry includes a textual movie review with corresponding rating, providing a solid foundation for rating prediction as a supervised machine learning task. During preprocessing, 33% of entries with missing fields or duplicate content were removed to ensure data reliability (García, Ramírez-Gallego, Luengo, Benítez, & Herrera, 2016). Ratings were normalized to a 1–5 scale for consistency by adjusting various data entry formats like fractions, integers, and letter grades (Kim, Sung, Park, & Kim, 2016). Reviews with fewer than 10 words were excluded to enhance contextual richness and reduce bias (Helan & Sultani, 2023; Singh, Rajput, Sharma, Satakshi, & Kumar, 2024). The final dataset contained 757,062 complete reviews with a mean score of 3.24 (Standard Deviation = 1.03) and a slight positive skew, offering a robust foundation for a comprehensive research methodology (Mansoury, Burke, & Mobasher, 2021).

### 8.2.3. Feature Engineering: Sentiment Polarity

The initial step in the methodology involved feature engineering sentiment polarity scores to determine the optimal feature for subsequent analysis and hypothesis testing. Sentiment polarity captures subtle emotional shifts and variations in sentiment analysis, represented on a continuous scale from -1 to 1 (Hemmatian & Sohrabi, 2019; Reagan et al., 2017).

Three approaches for sentiment polarity scoring were employed, including VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto & Gilbert, 2014) and two pre-trained sentiment models, BERT and RoBERTa, which share a foundation in the transformer-based BERT architecture (Devlin et al., 2019; W. Liao, Zeng, Yin, & Wei, 2021). Together these models provide complementary insights into the sentiment expressed in movie reviews.

VADER is a rule-based tool designed for short texts and social media analysis, known for its effectiveness in these contexts (Hutto & Gilbert, 2014). It calculates sentiment polarity scores within a range of -1 to 1 by combining word valence with linguistic modifiers, such as negations and intensifiers, to determine sentiment strength.

While VADER employs a straightforward approach, BERT-models, on the other hand, leverage transformer-based architectures that provide advanced capabilities for sentiment analysis (Devlin et al., 2019). BERT captures bidirectional context by analyzing preceding and succeeding tokens, while RoBERTa enhances this with dynamic masking and larger batch sizes for improved sentiment detection (W. Liao, Zeng, Yin, & Wei, 2021). These models were extensively pretrained on large sets of data and can be directly leveraged to predict movie review sentiments by loading -pretrained parameters from platforms like Hugging Face. Since the models generate categorical outputs, the results were scaled to a continuous range (-1 to 1) using confidence scores. This approach facilitates effective feature engineering with advanced BERT models while ensuring compatibility with VADER and enabling cross-method comparisons Specifically, the 3-class *cardiffnlp/twitter-roberta-base-sentiment* model, fine-tuned on 124 million tweets (2018–2021), was used to classified sentiment into three categories: negative, neutral, and positive (Jahin, Shovon, Mridha, Islam, & Watanobe, 2024). Furthermore, a more granular five-class *nlptown/bert-base-multilingual-uncased-sentiment* model, fine-tuned on 150k multilingual reviews, was leveraged to predict sentiment across five-star rating classes (Miah et al., 2024). The three models were compared and selected based on their ability to produce sentiment polarity distributions aligned with real-world contexts. Performance was further evaluated using the Spearman correlation coefficient between the model's predicted sentiment scores and review ratings (Cavallo, 2020). This ensured the chosen models reliably captured sentiment trends.

### 8.2.4. Rating Prediction

This study employs transfer learning by fine-tuning pretrained BERT variants on unstructured movie review data to predict numeric ratings. The models are loaded with trainable parameters and are therefore optimized for task-specific performance using two distinct approaches: regression for continuous ratings and classification for discrete categories.

#### 8.2.4.1. Model Architecture

A comprehensive framework including baseline and advanced architectures is proposed. BERT-Small (29M parameters) is deployed for both models (Qiao, Zhu, & Gong, 2022). The advanced model is further upgraded to BERT-Base (110M parameters) (Devlin et al., 2019). BERT is based on the transformer architecture, and processes input text using self-attention mechanism and feedforward layers to capture contextual information from both directions of the text (Vaswani et al., 2017).

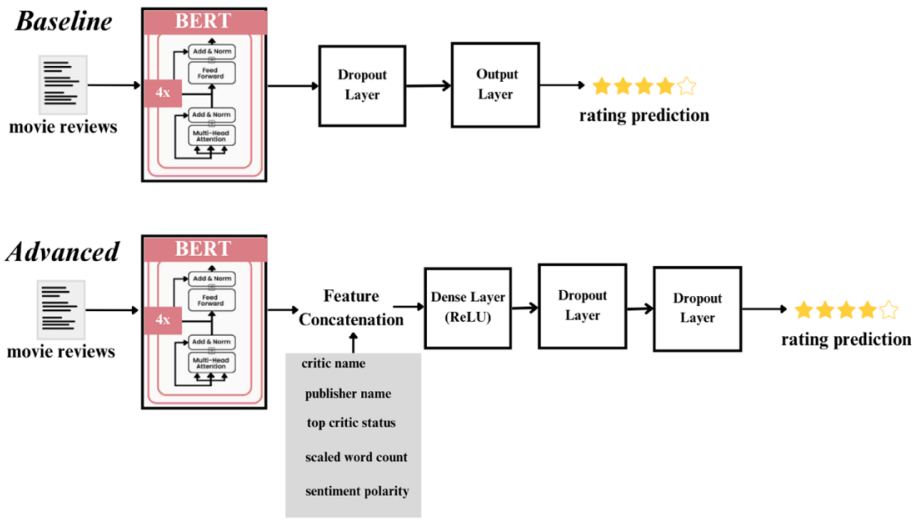



*Figure 9: Comparison of Baseline and Advanced Model Architectures*

Figure 9 visualizes the architectural differences in the baseline and advanced models. The baseline model processes textual movie reviews exclusively, passing the [CLS] token output through a dropout layer and an output layer to predict rating values (Lv, Liu, Zhao, Xu, & You, 2023). The

advanced model builds on this structure by integrating additional features, such as critic name, publisher name, top critic status, scaled word count, and sentiment polarity. These features are concatenated with the [CLS] token output, forming a hybrid representation (X. Liu, Tian, Niu, Li, & Han, 2023). The concatenated representation is then passed through a dense layer with ReLU activation, followed by a dropout layer, and finally through the output layer. The key architectural difference between regression and classification lies in the output layer design. Regression uses a single neuron with a linear activation function to predict continuous rating values (Dubey, Singh, & Chaudhuri, 2022). In contrast, classification employs a multi-neuron output layer with a SoftMax activation function, converting logits into probabilities for assigning inputs to distinct rating categories (Vasanthakumari, Nair, & Krishnappa, 2023).

#### 8.2.4.2. Preprocessing & Data Split

***Regression***

For regression, ratings were predicted on a continuous scale from 1-5 to capture subtle differences. (Kumar & Hanji, 2024). During splitting, stratified sampling was used to preserve the original dataset's distribution of the target variable (rating) across the training, validation, and test sets, with random selection applied within each stratum (Mohapatra, Bhoi, Mallick, Jena, & Mishra, 2022). This resulted in 603,007 training samples, 75,376 validation samples, and 75,376 test samples.

***Classification***

For classification, ratings were discretized into five categories (1–5 stars) using a floor-rounding method (Maulidi, Syahrini, Oktavia, Ihsan, & Emha, 2021). Stratified sampling with random selection within each stratum preserved proportional representation of the target variable, while undersampling addressed class imbalance (Thabtah, Hammoud, Kamalov, & Gonsalves, 2020). This produced in 197,430 training samples, 24,675 validation samples, and 24,680 test samples, ensuring sufficient data for effective model training and evaluation.

#### 8.2.4.3. Performance Metrics

***Regression***

For regression, Mean Squared Error (MSE) is used as the loss function, penalizing larger deviations by squaring differences (Köksoy, 2006). Root Mean Squared Error (RMSE) provides a complementary metric by taking the square root of MSE (Chicco et al., 2021). Mean Absolute Error (MAE), the primary evaluation metric, calculates average absolute differences, offering a straightforward and interpretable measure for rating predictions (Lai & Hsu, 2021).

***Classification***

For classification a SoftMax activation function with categorical cross-entropy loss ensures valid probability distributions across distinct rating classes (Darwish et al., 2020). Accuracy, calculated as the ratio of correct predictions to total instances, serves as the primary evaluation metric (Kotsiantis, Zaharakis, & Pintelas, 2006). To account for the ordinality of ratings, 1up1down Accuracy considers predictions within one class of the true label as correct, aligning with ordinal metrics like the Closeness Evaluation Measure, which penalizes errors based on their distance from the true label (Amigó, Gonzalo, Mizzaro, & Carrillo-de-Albornoz, 2020).

### 8.2.5. Hypotheses Testing

Hypotheses were evaluated on the test set by grouping predictions into sentiment polarity and review length categories, with MAE for regression and Accuracy for classification computed across groups. Shapiro-Wilk tests were deployed to confirm non-normal distributions in model performance metrics (Shatz, 2024). A one-tailed Mann-Whitney U test evaluated whether extreme sentiments enhanced performance (Karch, 2021). For review length, the Kruskal-Wallis test detected differences, followed by pairwise one-tailed Mann-Whitney U tests to assess directional trends between groups. (Guo, Zhong, & Zhang, 2013). This approach provided a comprehensive evaluation of the effects of sentiment polarity and review length on model performance.

## 8.3. Results

### 8.3.1. Sentiment Polarity Model Selection

Figure 10 compares the sentiment polarity distributions of the three models. VADER shows a sharp neutral peak with a close mean (0.20) and median (0.27), indicating a balanced but neutrality-heavy distribution. Twitter RoBERTa displays a polarized U-shape, overemphasizing extreme sentiments, as reflected in its mean (0.20) and higher median (0.32). In contrast, Multilingual BERT offers a balanced distribution with subtle sentiment shifts, evidenced by its nearly identical mean (0.11) and median (0.12). This alignment suggests a balanced and realistic sentiment polarity distribution with minimal skew or extreme biases.

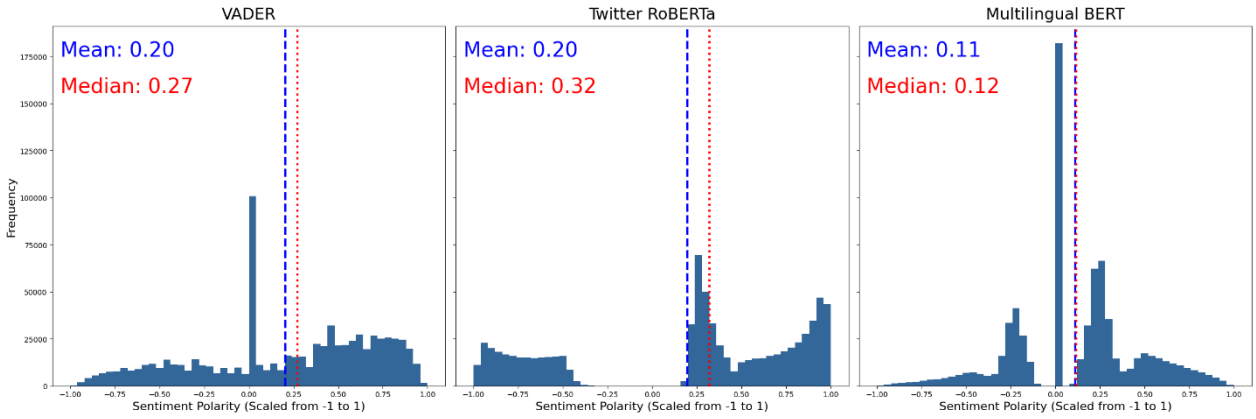



*Figure 10: Sentiment Polarity Distribution across Models*

Models are further evaluated based on their spearman correlation with review ratings showing statistical significance. VADER shows limited alignment with a correlation of $r_s = 0.253, p < 0.001$. Twitter RoBERTa performs best, achieving a correlation of $r_s = 0.515, p < 0.001$, effectively capturing both extreme and intermediate sentiments. Multilingual BERT follows closely with a correlation of $r_s = 0.507, p < 0.001$, demonstrating strong alignment with ratings, though slightly behind RoBERTa. Thus, it is the most suitable, offering a realistic and balanced distribution that better reflects real-world sentiment compared to Twitter RoBERTa. Therefore, the five-class sentiment polarity was selected as a feature for advanced model architectures and to evaluate **H1**.

### 8.3.2. Rating Prediction Performance Evaluation

#### *Regression Evaluation*

The evaluation of regression models, as summarized in Table 5, highlights progressive improvements with the adoption of more sophisticated architectures and additional features. The Advanced Regression Model Large, achieved the strongest performance, with an MSE of 0.3305, RMSE of 0.5749, and an MAE of 0.4277. On a 1–5 rating scale, the MAE indicates that the model's predicted ratings deviate by an average of only 0.43 points from the actual ratings, demonstrating a high level of precision. This low margin of error reflects the model's ability to capture nuanced variations in movie reviews. The results emphasize the advantages of leveraging a deeper architecture in combination with contextual features, validating the efficacy of BERT-based models for continuous rating prediction tasks.

*Table 5: Regression Evaluation Table of different Performance Metrics for Baseline and Advanced Models*

| | Model | Loss (MSE) | RMSE | MAE |
|---|---|---|---|---|
| **Regression** | **Baseline Model** (BERT Small) | 0.4678 | 0.684 | 0.5154 |
| | **Advanced Model** (BERT Small) | 0.3853 | 0.6207 | 0.4593 |
| | **Advanced Model Large** (BERT Base) | **0.3305** | **0.5749** | **0.4277** |

#### *Classification Evaluation*

The evaluation of classification models, summarized in Table 6, further highlights performance improvements with increasing model complexity. The most sophisticated model, Advanced Model Large, achieved an Accuracy of 54.33% and a 1up1down Accuracy of 91.07%. The high 1up1down Accuracy demonstrates the model's ability to predict ratings within one class above or below the true label in over 91% of cases, capturing the general direction of ratings effectively. However, while the Advanced Model Large demonstrates strength in handling complex classification tasks,

the lack of Accuracy improvement compared to Advanced Model highlights limitations in fully leveraging the larger architecture, suggesting challenges in learning and optimization.

*Table 6: Classification Evaluation Table of different Performance Metrics for Baseline and Advanced Models*

| | Model | Loss | Accuracy | 1up1down Accuracy |
|---|---|---|---|---|
| **Classification** | **Baseline Model** (BERT Small) | 1.2078 | 0.5083 | 0.8999 |
| | **Advanced Model** (BERT Small) | 1.3512 | **0.5472** | 0.898 |
| | **Advanced Model Large** (BERT Base) | **1.1289** | 0.5433 | **0.9107** |

### 8.3.3. Hypotheses Testing

To assess **H1** on the test set predictions were grouped into strong and neutral/moderate sentiment groups. Shapiro-Wilk tests were deployed to reveal significant deviations from normality ($p < .001$) across all models for both groups, justifying the use of the Mann-Whitney U test. Subsequently, one-tailed Mann-Whitney U tests were deployed across all models to evaluate the hypothesis. Test results for regression and classification models are presented in Table 7.

*Table 7: Sentiment Polarity Hypothesis (**H1**) Test Results across Models*

| | Model | Mann-Whitney U | Result |
|---|---|---|---|
| **Regression** | **Baseline Model** | p = 0.990 | *Non-significant* |
| | **Advanced Model** | p = 0.598 | *Non-significant* |
| | **Advanced Model Large** | p = 0.598 | *Non-significant* |
| **Classification** | **Baseline Model** | p < 0.001 | ***Significant*** |
| | **Advanced Model** | p = 0.011 | ***Significant*** |
| | **Advanced Model Large** | p = 0.026 | ***Significant*** |

Regression models did not support the hypothesis, with $p > 0.05$ in all cases, indicating that sentiment polarity strength has no significant impact on regression-based rating prediction performance. This result suggests that regression models are robust to sentiment variability, as they effectively capture subtle nuances across the entire spectrum of sentiment intensity. Consequently, extreme sentiment cues may not provide additional predictive value in continuous rating tasks. In

contrast, all classification models found significant support for **H1**. The Baseline Classification Model ($p < 0.001$), relying solely on textual reviews, demonstrates the strongest dependency on sentiment polarity. As indicated by slight increases in p-values with advanced model configurations, the dependency on sentiment polarity decreases by incorporating multi-feature inputs. This suggests that more complex architectures rely less on sentiment polarity and instead leverage a broader range of textual and contextual patterns for predictions. These findings demonstrate that sentiment polarity plays a critical role in classification tasks, while regression models were not significantly influenced by sentiment strength, indicating a fundamental difference in how these models process and utilize sentiment-related information.

Furthermore, **H2** was assessed by grouping test set predictions into short, medium, and long review length categories. Kruskal-Wallis tests were used to assess statistical significance, where the direction of differences was further explored through pairwise Mann-Whitney U tests. Table 8 reports the findings for both regression and classification models.

*Table 8: Review Length Hypothesis (H2) Test Results across Models*

| | Model | Kruskal-Wallis | Pairwise Mann-Whitney U | | Result |
|---|---|---|---|---|---|
| **Regression** | **Baseline Model** | $p = 0.911$ | - | | *Non-significant* |
| | **Advanced Model** | $p = 0.210$ | - | | *Non-significant* |
| | **Advanced Model Large** | $p = 0.373$ | - | | *Non-significant* |
| **Classification** | **Baseline Model** | $p < 0.001$ | Short vs. Medium | $p = 0.998$ | *Non-significant* |
| | | | Medium vs. Long | $p = 0.980$ | |
| | | | Short vs. Long | $p = 0.999$ | |
| | **Advanced Model** | $p = 0.226$ | - | | *Non-significant* |
| | **Advanced Model Large** | $p = 0.497$ | - | | *Non-significant* |

The Kruskal-Wallis test revealed no significant differences ($p > 0.05$) in all models except the Baseline Classification Model, where significant differences were observed across groups ($p < 0.001$). However, pairwise comparisons using Mann-Whitney U tests revealed no significant support for improved performance with longer reviews. These findings provide no significant

support for **H2**. While some variability was observed, the directionality toward longer reviews did not meaningfully influence outcomes. This highlights BERT's robustness in handling varying review lengths, showing that even shorter reviews can achieve strong rating prediction performance.

## 8.4. Insights & Discussion

Ahmed & Ghabayen (2022) proposed a two-phase framework that predicts sentiment polarity and then leverages it for rating prediction, achieving RMSEs of 0.56 on Yelp and 0.65 on Amazon Books, outperforming baseline models like Bi-GRU and Bi-LSTM. In comparison, this study's Advanced Regression Model Large achieved an RMSE of 0.5749 on a movie review dataset by Rotten Tomatoes, closely matching Yelp's performance and surpassing Amazon Books by 0.0751. These results underscore the importance of sentiment polarity in enhancing rating predictions and demonstrate BERT's ability to handle the nuanced complexity of movie reviews compared to product or service reviews. Regression model performance improved consistently across all metrics, demonstrating the model's capacity to effectively learn from multi-feature inputs and benefit from larger pretrained BERT variants. In contrast, classification performance followed a different trend. Z. Liu (2020) applied a BERT-based classification approach for rating prediction on Yelp data, achieving an Accuracy of 69.11%. In this study, the Advanced Classification Model Large reported an Accuracy of 54.33% on Rotten Tomatoes data, reflecting the greater subjectivity of movie reviews. Moreover, classification results stagnated, with Accuracy showing no improvement when upgrading from Advanced Classification Model to Advanced Classification Model Large. This was further supported by high 1up1down Accuracy results of up to 91.07%, compared to relatively lower overall Accuracy scores. These findings highlight the classification model's difficulty in differentiating between distinct categories, as it does not account for the natural order of star ratings. BERT's widespread use in text classification tasks has led to a

predominance of classification-based approaches in rating prediction research (Galal et al., 2024). Aljrees et al. (2024) utilized BERT to predict Google Play Store ratings, potentially reflecting a preference for distinct categories over continuous scales due to data availability. However, with Rotten Tomatoes movie reviews offering a nuanced, continuous scale, this study demonstrates the effectiveness of equipping the output layer with a linear neuron for regression-based rating prediction leveraging BERT. Furthermore, prior research has consistently emphasized the importance of sentiment polarity in rating prediction models (Ahmed & Ghabayen, 2022; Aljrees et al., 2024). This study extends that understanding by showing that the impact of sentiment polarity is positively moderated by the strength of sentiment expressed in reviews specifically in classification tasks. Classification BERT models perform significantly better with extreme sentiments, whereas regression models showed no significant support for the hypothesis. Finally, while Bilal & Almazroi (2023) found that longer sequences enhance BERT's predictive performance, this study's results challenge that view, specifically within the realm of the movie domain. These findings suggest that review length does not enhance predictive performance, demonstrating BERT's robustness in handling diverse textual data across varying lengths. This study highlights the importance of aligning modeling approaches with task characteristics and data availability in rating prediction. Regression models effectively capture nuanced insights on continuous scales, while classification models excel with extreme sentiments but struggle with ordinal structures. BERT's consistent performance across varying review lengths further underscores the model's potential in versatile deployment across a wide range of review platforms. This study advances review rating prediction by exploring sentiment polarity and review length as moderating factors, offering a novel perspective on leveraging transfer learning with BERT in a comprehensive framework to predict ratings from textual movie reviews.

## 8.5. Limitations & Future Research

This study has several limitations that offer opportunities for future research. One key challenge is the subjectivity of ratings, as textual reviews do not always align with numerical ratings, introducing bias that can limit the model's ability to learn and generalize. While previous research has addressed this issue (Aljrees et al., 2024), further investigation could enhance understanding of this topic. This study demonstrated promising performance, with regression models improving across all metrics when upgraded to more complex and larger architectures. However, computational constraints restricted the use of larger architectures, such as BERT-Large, and extended training cycles. Future research should investigate these factors to evaluate their impact on predictive performance. Additionally, the rejection of **H2** highlights BERT's robustness across varying review lengths, suggesting potential for social media rating prediction where reviews are often brief and variable. Moreover, this study focused solely on leveraging foundational BERT, neglecting other advanced transfer learning models such as RoBERTa, XLNet and GPT-based architectures (Shao, Basit, Karri, & Shafique, 2024). Benchmarking these models against BERT in could provide insights into their potential performance advantages.

In conclusion, this study demonstrates the effectiveness of leveraging pre-trained transformer-based models like BERT, fine-tuned for complex domain-specific task such as movie rating prediction. This work offers valuable insights into the strengths and limitations of regression and classification approaches, emphasizing the importance of tailoring problem framing to domain specific requirements and data availability. Going beyond simple performance evaluation, this study explored moderating factors in review rating prediction with BERT, providing insights for effective model application. These findings highlight the transformative potential of NLP, not only within the movie industry but across a wide range of domains that rely on textual feedback.

# 9. Temporal Modeling for Movie Success Prediction: A Comparative Study on the applicability of a MIL Model in Entertainment Analytics

## 9.1. Introduction

Existing research has extensively examined the characteristics and importance of various input features in identifying drivers of movie success, with a strong emphasis on model and feature interpretability (J. H. Lee et al., 2017). Features such as review sentiment and audience ratings have been consistently analyzed to understand their impact on box office revenues (Duan et al., 2008; J. H. Lee et al., 2017). Temporal patterns, including seasonality effects, have also been identified as critical factors influencing box office performance (Einav, 2007; Karniouchina et al., 2023; Simonton, 2009). Building on this body of work, this study shifts focus to emphasize the temporal characteristics of features within the available data. It explores the application of advanced deep learning frameworks for predicting box office success in the post-release stage. By integrating temporal and post-release features as multivariate time series, this research seeks to explore the accuracy and interpretability capabilities of the weakly-supervised TimeMIL (X. Chen et al., 2024) compared to a long short term recurrent neural network (LSTM) (Ghanbari & Borna, 2021) as baseline and TodyNet (H. Liu et al., 2024) representing an advanced supervised architecture.

## 9.2. Research Question & Hypothesis Development

Deep learning (DL) models for movie success predictions have explored advanced techniques, including multimodal frameworks (Madongo & Zhongjun, 2023) and stacking fusion methods (Y. Liao et al., 2022). While Y. Liao et al. (2022) argue that post-release success prediction has limited value for initial investment decisions, other studies highlight the benefits of incorporating external word-of-mouth (eWOM), user sentiment, and ratings from social media and consumer platforms, which can significantly improve predictive accuracy (Rhee & Zulkerine, 2016;

Subramaniyaswamy, Viginesh, Vishnu, & Logesh, 2017). Although post-release forecasts may not guide initial funding strategies, they offer actionable insights for marketing and distribution planning in the post-release stage (C. Zhang, Tian, & Fan, 2022). This underscores the need for interpretable methods that identify key performance drivers. Multiple-instance learning (MIL), a DL framework that assigns labels to sets of instances (bags) rather than individual samples (Carbonneau et al., 2018), has shown promise in improving interpretability, as evidenced by studies in other domains (Early et al., 2024). While MIL has been used in other fields to integrate textual and temporal features, including sentiment data, for financial forecasting or classification (Deng & Yiu, 2022; Tibo et al., 2020), its application to movie performance prediction remains underexplored. X. Chen et al. (2024) introduced TimeMIL, a foundational weakly-supervised MIL model for multivariate time series classification (MTSC) and demonstrated competitive performance against supervised models on UEA benchmark datasets (Bagnall et al., 2018; Middlehurst et al., 2024). However, their focus on standard benchmarks does not address real-world use cases and characteristics of the movie domain. Whether MIL models can outperform supervised methods in predicting movie success and delivering interpretable outputs has yet to be established. While X. Chen et al. (2024) highlighted the inherent interpretability of MIL due to its attention mechanism, concerns raised by Jain & Wallace (2019) and Bibal et al. (2022) about the reliability of attention-based explanations suggest the need for further validation. This study addresses these gaps through the following research question (RQ): TO which extend can a weakly-supervised MIL model integrating early movie performance data and eWOM sentiment analysis accurately classify movies into success categories (flop, average, hit) and provide superior performance and interpretability compared to traditional time-series classification methods? To explore this question, several hypotheses are proposed:

***H1****: Weakly-supervised MIL models for time series, utilizing early movie performance data (e.g., critic reviews, audience ratings, and box office revenue), will outperform supervised time-series classification methods in predicting movie success.* The hypothesis is supported by prior works demonstrating MIL's efficacy in leveraging textual and temporal features (Deng & Yiu, 2022; Tibo et al., 2020) and X. Chen et al. (2024)`s findings on MIL's competitive accuracy to other supervised models such as TodyNet (H. Liu et al., 2024).

***H2:*** *Weakly-supervised MIL models for time series will offer more interpretable insights into factors contributing to a movie's success compared to supervised methods, as evaluated through attribution techniques.* H2 addresses the need for pipelines interpretability, building on research on feature-contributions for predictive models in the movie industry (Y. Liao et al., 2022; Yang, Xu, & Tu, 2023). Prior studies further underscore the importance of identifying key factors influencing movie performance, validating the need for model-specific feature interpretations (Rhee & Zulkerine, 2016; Subramaniyaswamy, Viginesh, et al., 2017). It is supported by recent work on the advantageous value of the MIL framework for interpretability (Early et al., 2024). TimeMIL represent a novel framework to perform multivariate time series classification in the MIL framework. However, its applicability is not yet widely evaluated for specific use cases of movie success prediction. Consequently, this study proposes to explore the advantages of the MIL framework by comparing TimeMIL to supervised models for the use case of movie success prediction. It will incorporate ratings, sentiment analysis, and box office revenue into time-series data, to address accuracy and interpretability in movie success predictions.

## 9.3. Methodology

This section outlines the data sources, preprocessing steps for creating the final time series datasets, and the framework for comparing the models in terms of interpretability and performance. All datasets were curated, processed, and combined to create comprehensive time series datasets based

on daily ratings and critical sentiments of movies. This enables the reimplementation and exploration of predictive models for movie success classification based on temporal input features. Figure 11 visualizes the experiment pipeline to outline the different steps conducted in the study.

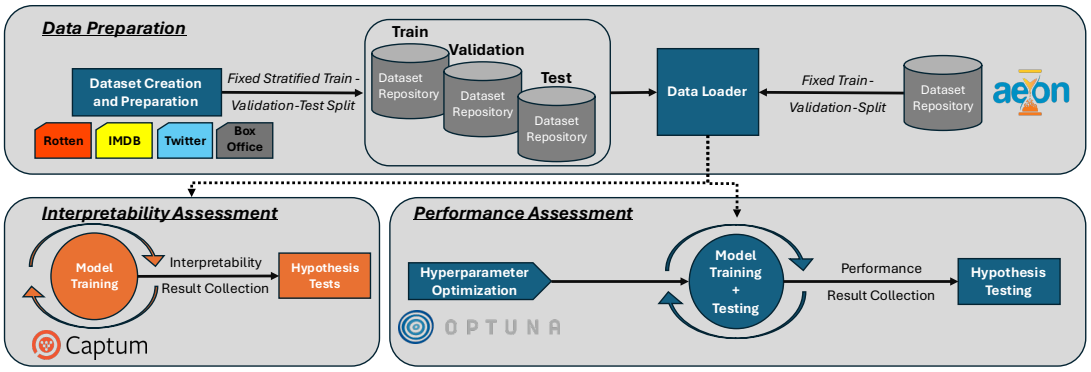



*Figure 11: Evaluation Framework to assess the Performance and Interpretability of Models*

### 9.3.1. Datasets and Data Preparation

The central datasets for this study consisted of rating data retrieved from Rotten Tomatoes (Leone, 2020), IMDb (Banik, 2017), as well as Twitter (Dooms, 2021). Additionally, the time series was enhanced by adding movies lifetime revenues, budgets as well as daily Box Office revenues for each movie using data from Box Office Mojo (Mohamadi, 2024). The data, as outlined in section 3.3, specifically the ratings from Rotten Tomatoes, were standardized and cleaned. Further, missing values for budgets and revenues were imputed by integrating APIs such as TMDB during preprocessing (Lash & Zhao, 2016). To form the final time series, the features were aggregated for each platform by calculating mean, min, max, and variance scores for the platforms' rating features to create a time series over daily observations for each movie. As critic reviews were available for the Rotten Tomatoes dataset, the critics' sentiment was extracted and combined with daily statistics like the rating feature (Yang et al., 2023) using the most promising approach outlined in the preceding work focusing on sentiment polarity feature engineering in section 8.3.1. The data was then joined with daily box office performance data, including daily box office revenues, the daily number of theaters where the movie was shown, and daily mean box office gross per theater. Table

38 and Table 39 in the appendix summarize the features and label distributions over each dataset. Target labels were derived for each movie by estimating the movie's return-on-invest (RoI) using the movie's budget and lifetime revenue described in Equation 1 in the appendix. As outlined by Rhee & Zulkerine (2016), following a classification approach to label movies into three distinct classes can lead to accurate predictions. Moreover, by producing both class labels and associated probabilities, the model reveals how close a film is to flopping, rather than locking it into a single rigid category. Consequently, three classes offer a good trade-off for task complexity for the model while giving enough insight to derive a particular trend when looking at the output probabilities. Thus, the RoI was encoded with the label's "flop" (RoI > 1.0), "average" (1.0 < RoI < 3.0), and "hit" (3.0 < RoI). As described in section 3.3.2.2 movies had between 1 and 270 recorded samples, requiring sequence length adjustments during preprocessing. To ensure sufficient data, only movies with over 28 samples post-release were included, with sequences padded or truncated to a maximum of 60 days fitting them to the median sequence lengths across the different datasets. Additionally, the data was standardized for the training process due to varying ranges between features (Fox, 2016). Following the standard practice and to ensure comparability, each model was trained on the same stratified train-validation-test split (X. Chen et al., 2024; García, Herrera, & Es, 2008; H. Liu et al., 2024).

### 9.3.2. Model Selection

During the experimentation, three models were compared from three different domains. Particularly, the weakly-supervised TimeMIL model proposed by (X. Chen et al., 2024) was reimplemented in order to compare its performance along with TodyNet, a supervised GNN model proposed by H. Liu et al. (2024) that showed competitive performance for at least 14 out of the 26 UEA datasets (Bagnall et al., 2018) in the benchmark paper of X. Chen et al. (2024). Each of the models captures distinct characteristics of different architectural concepts. On the one hand,

TimeMIL, an architecture based on the MIL framework, uses a tokenized transformer with wavelet-based positional encoding to extract sparse and temporal patterns. Wavelet positional encoding applies transformations to capture hierarchical, positional information from sequential inputs. The extracted patterns are then learned by employing a weakly-supervised learning strategy (X. Chen et al., 2024). TodyNet, on the other hand, employs the framework of dynamic graph neural networks to capture spatial-temporal dependencies through dynamic graph construction and hierarchical pooling (H. Liu et al., 2024). Additionally, to explore how TimeMIL and TodyNet compare in terms of performance, a baseline model was implemented to capture whether any of the models generally perform significantly differently compared to less sophisticated architectures. To establish a baseline, the LSTM architecture is deployed, as it is considered a well-established DL technique for time-series predictions. LSTMs focus on modeling long-term sequential dependencies with memory cells, differing from the other models' graph-based and weakly-supervised approaches (Ghanbari & Borna, 2021). TodyNet and TimeMIL had to be reimplemented using the available information from their corresponding GitHub repositories (X. Chen et al., 2024; H. Liu et al., 2024).

### 9.3.3. Assessment of Model Performance

To ensure the correct reimplementation of TimeMIL and TodyNet, they were tested on UEA datasets (Bagnall et al., 2018) via the Aeon package by Middlehurst et al. (2024). Additionally, to maintain a fair comparison between the three architectures, each model's hyperparameters were tuned for 30 trials using the Optuna package (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). This reduced randomness when choosing optimal hyperparameters for the comparative experiments. Finally, to evaluate model performance, each fine-tuned model was then trained 30 times for 200 epochs minimizing CrossEntropyLoss (Ho & Wookey, 2020) using the AdamW-optimizer (Loshchilov & Hutter, 2017) together with a ReduceLROnPlateau-scheduler (Al-Kababji, Bensaali,

& Dakua, 2022) with a patience level of 50. Regarding the selection of performance metrics, Accuracy, F1, Recall, Precision, and AUC were calculated during each training run. Due to class imbalances, and as Accuracy tends to hide classification errors while not accounting for the class distributions (Grandini, Bagli, & Visani, 2020), the focus of the performance comparison was the macro average F1 and macro average AUROC score. They reflect performance across all classes more fairly by integrating precision-recall trade-offs and overall discriminative capability, leading to a better estimate of the real performance of each model.

### 9.3.4. Assessment of Model Interpretability

When it comes to interpretability techniques, three different categories can be distinguished. Some techniques propagate sample and feature importance via model gradients, which is why they are referred to as gradient methods (Schlegel & Keim, 2021). Common gradient-based methods are Saliency (Simonyan, Vedaldi, & Zisserman, 2013), Integrated Gradients (IntGrad) (Sundararajan, Taly, & Yan, 2017), or Deeplift (Shrikumar, Greenside, & Kundaje, 2017). Other techniques use network weights and biases together with a per-model layer-defined set of weighing rules. Furthermore, more sophisticated techniques include surrogate methods such as SHAP, Shapley, or Lime, which generate interpretable attributions by sampling and modeling perturbed instances (Schlegel & Keim, 2021). To compare the models in terms of interpretability, the attributions were evaluated using the metric Infidelity. Infidelity measures how faithfully the applied techniques and attributions capture the model's internal reasoning and assesses how attributions to a model align with the changes in its output when the input is perturbed, reflecting the explanation's accuracy, which is why it was chosen as the primary metric. Perturbation is an important concept in attribution evaluation by applying small changes to the model's input data to see how the model adjusts. Low infidelity scores indicate that attributions align well with the model's behavior when the input is minimally (Kokhlikyan et al., 2020; Yeh et al., 2019). Following Schlegel & Keim

(2021), visually evaluating models' feature importance using attributions is still a difficult task even for domain experts. Thus, to further investigate the visual dispersion and certainty of attributions, the entropy was calculated for the model's attributions to estimate how spread out the assigned importance is across all input features. A low entropy value indicates that the model's attributions focus on several key features and time steps. Conversely, a high entropy indicates that a model's importance scores are more evenly distributed, making it harder to pinpoint which features drive its predictions (Namdari & Li, 2019; Perez, Skalski, Barns-Graham, Wong, & Sutton, 2021).

### 9.3.5. Hypothesis Assessment

To ensure scientifically correct hypothesis test application, the distribution of the metrics was estimated by applying the Shapiro-Wilk test on the pairwise differences of evaluation metrics between the MIL and the supervised models (García et al., 2008; Shatz, 2024). Depending on the distribution of the performance metrics, the MIL model was compared in a pairwise approach either by applying a paired t-test as the primary method or a Wilcoxon signed rank test as the secondary method (Rainio, Teuho, & Klén, 2024). Additionally, the Bonferroni correction was applied to standardize the p-values across tests. In case the results of these tests are ambiguous, a more novel significance test, referred to as the Almost Stochastic Order (ASO) test (del Barrio, Cuesta-Albertos, & Matrán, 2017; Dror, Shlomov, & Reichart, 2019), implemented by Ulmer, Hardmeier, & Frellsen (2022), was applied to investigate stochastic dominance and assess the hypothesis stated in this study. The test calculates epsilon $\epsilon$ so that $H_0$: $\epsilon_{min} \geq \tau$ can be tested. Epsilon can be interpreted as a confidence score similar to a p-value but does not represent the same. The lower it is, the certain one can be that the model is better than the model it is compared to. Similarly to standard statistical tests, the boundary tau ($\tau$) (commonly set to a value of 0.5 or 0.25) can be interpreted similarly to the significance boundary alpha ($\alpha$) in standard tests (Ulmer et al., 2022).

## 9.4. Results

### 9.4.1. Model Performance Evaluation

Table 9 summarizes the model's average F1 Scores and AUROC values over multiple runs for the different datasets. TodyNet consistently achieved the highest F1 Score for the Rotten Tomatoes and IMDb datasets, with scores of 0.510 and 0.553, respectively. The LSTM performed nearly as well as TodyNet for IMDb, achieving a close F1 Score of 0.552. However, for the Twitter dataset, TimeMIL outperformed the other models with an average F1 Score of 0.504, while TodyNet and LSTM trailed behind with scores of 0.491 and 0.424, respectively across experiments. These results indicate that while TodyNet generally excels in classification tasks across datasets, TimeMIL demonstrates a comparative advantage when predicting data from Twitter. Regarding the models' ability to distinguish classes, similar to the F1 Score, TodyNet consistently achieved the highest AUROC for the Rotten Tomatoes and IMDb datasets, with AUROC values of 0.749 and 0.764, respectively. Similarly, TimeMIL excelled, although by a very minor proportion, on the Twitter dataset, compared to TodyNet and LSTM. However, TimeMIL, while outperforming LSTM in the Twitter dataset for AUROC and F1 Score, generally had the lowest AUROC scores across the other datasets, indicating relatively weaker overall discriminability. Following the results of the Shapiro-Wilk test summarized in Table 45 in the appendix, all the pairwise differences were normally distributed, so the paired t-test was applicable to evaluate the model's significant performance.

*Table 9: Comparison of models' average performance on the F1 Score and AUROC*

| ***Models*** | ***TimeMIL*** | | ***TodyNet*** | | ***LSTM*** | | ***Rank TimeMIL*** | |
|---|---|---|---|---|---|---|---|---|
| ***Datasets\Metrics*** | ***F1*** | ***AUROC*** | ***F1*** | ***AUROC*** | ***F1*** | ***AUROC*** | ***Abs F1*** | ***Abs AUROC*** |
| ***Rotten Tomatoes*** | 0.456348 | 0.658596 | **0.510062** | **0.749291** | 0.503833 | 0.698412 | 3 | 3 |
| ***IMDb*** | 0.459950 | 0.651499 | **0.553272** | **0.763592** | 0.552268 | 0.742273 | 3 | 3 |
| ***Twitter*** | **0.503830** | **0.750273** | 0.491021 | 0.750155 | 0.423803 | 0.729751 | 1 | 1 |

Table 10 summarizes the hypothesis test results for pairwise comparisons with a significance level alpha (α) of 0.05. Based on the paired t-test results, significant differences were observed for most

pairwise comparisons over each dataset and evaluation metric. The hypothesis testing results validate the observations for the average performance of the models for the Rotten Tomatoes and IMDb datasets. They confirm that TodyNet excels for movie review data as time series and the GNNs competitiveness for the Twitter time series. Although TimeMIL is superior to the LSTM model regarding the Twitter dataset, its performance is not statistically superior to TodyNet. There was no significant difference between TimeMIl and TodyNet for both the F1 score and AUROC. This indicates that, although the global mean comparison indicated minor superiority of TimeMIL to TodyNet, the performance was not significant enough across experiments.

*Table 10: Paired t-test for statistical significance of models superiority across evaluation metrics*

| ***MIL-Model*** | ***Metric*** | ***Dataset*** | ***Test statistic*** | ***p-value*** | ***Test Result*** |
|---|---|---|---|---|---|
| *TimeMIL vs. LSTM* | *F1* | *Rotten* | *4.5676* | *0.000084* | *A significant diff. for LSTM* |
| | | *IMDb* | *14.3439* | *<0.001* | *A significant diff. for LSTM* |
| | | *Twitter* | *-6.6023* | *<0.001* | *A significant diff. for LSTM* |
| | *AUROC* | *Rotten* | *5.2239* | *0.000014* | *A significant diff. for LSTM* |
| | | *IMDb* | *15.3737* | *<0.001* | *A significant diff. for LSTM* |
| | | *Twitter* | *-3.2751* | *0.002737* | *A significant diff. for LSTM* |
| *TimeMIL vs. TodyNet* | *F1* | *Rotten* | *5.0527* | *0.000022* | *A significant diff. for TodyNet* |
| | | *IMDb* | *7.6377* | *<0.001* | *A significant diff. for TodyNet* |
| | | *Twitter* | *-1.03* | *0.311506* | *No significant diff. for TimeMIL* |
| | *AUROC* | *Rotten* | *11.7481* | *<0.001* | *A significant diff. for TodyNet* |
| | | *IMDb* | *16.6109* | *<0.001* | *A significant diff. for TodyNet* |
| | | *Twitter* | *-0.0185* | *0.98533* | *No significant diff. for TimeMIL* |

### 9.4.2. Interpretability Evaluation

The interpretability assessment involved attribution techniques such as such as IntGrad or DeepLift. According to the results presented by Nguyen & Martínez (2020), Integrated Gradients provide the most general explanations while offering low effective complexity, meaning that the attributions are efficient and simple to interpret. Additionally, Deeplift was used to ensure comparability over several interpretability techniques. IntGrad and Deeplift ranked in the top tier in the assessment of attribution techniques for the transformer model and bidirectional LSTM model in the comparison of attribution techniques of Turbé et al. (2023). SHAP seemed to be a more sophisticated technique but proved unfeasible due to the incompatibility of SHAP with certain

layers in the compared networks, such as LayerNorm or Identity from TimeMIL. Similarly, although ranked highest in the assessment of Turbé, Bjelogrlic, Lovis, & Mengaldo, Shapley values calculation was not applicable, which can be explained following the findings of Kolpaczki, Bengs, Muschalik, & Hüllermeier (2024), due to the number of features leading to extensive computational complexity of Shapley values resulting in *O(n*n!)* complexity with *n* as the number of features. Figure 12 visualizes an example of how attributions of methods such as DeepLift can be used to interpret how a sample is evaluated by the model when classifying the sample (movie) and which features and timestamps contributed negatively, neutrally, or positively to the decision. The red color signals that the timestep had a negative contribution, while the green signals positive attribution scores for the step. Although this can be rather abstract to interpret, such attributions can still be used to determine whether certain timesteps and features had a more substantial influence during the classification process of the sample movie than other time steps. The accuracy of model attributions can be evaluated by calculating the infidelity score for each model. Table 11 summarizes the average infidelity for each model, attribution technique, and dataset.

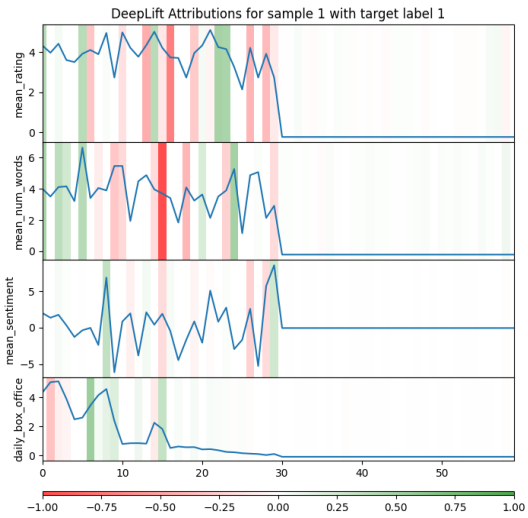


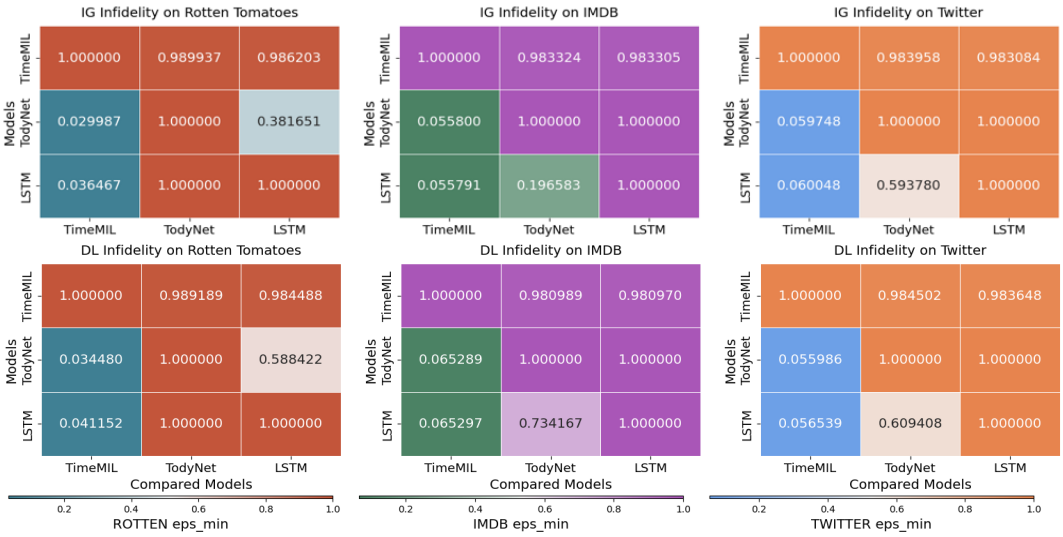

*Figure 12: Example of Deeplift attributions on a movie from the test sample labeled as 'average'*

*Table 11: Comparison of interpretability performance on the experiments average Infidelity across models*

| ***Models*** | ***TimeMIL*** | | ***TodyNet*** | | ***LSTM*** | | ***Rank TimeMIL*** | |
|---|---|---|---|---|---|---|---|---|
| ***Dataset\Infidelity*** | ***IntGrad*** | ***DeepLift*** | ***IntGrad*** | ***DeepLift*** | ***IntGrad*** | ***DeepLift*** | ***IntGrad*** | ***DeepLift*** |
| ***Rotten Tomatoes*** | 0.003376 | 0.003767 | **0.000529** | **0.000514** | 0.000781 | 0.000741 | 3 | 3 |
| ***IMDb*** | 0.573625 | 0.773938 | 0.001446 | 0.001152 | **0.001098** | **0.001143** | 3 | 3 |
| ***Twitter*** | 0.022442 | 0.025024 | 0.000584 | 0.000522 | **0.000384** | **0.000341** | 3 | 3 |

When comparing the average entropy scores in Table 46 in the appendix to the infidelity scores for TimeMIL, one can observe that the average Entropy scores for TimeMIL are larger compared to TodyNet and the LSTM, indicating that its attributions are more evenly distributed. During the assessment of the infidelity metric, the traditional hypothesis test resulted in ambiguous statistics and p-values, making a conventional hypothesis testing approach inapplicable to evaluating the infidelity metric. The statistical tests results for the Shapiro Wilk and Wilcoxon test are summarized in Table 47 and Table 48 in the appendix. Thus, the multiple ASO test with a confidence level of 0.95 was applied to estimate the significance of the model's interpretability performance. Table 12 summarizes epsilon $\epsilon$ from the ASO test for the infidelity metric across the attribution schemes and models. The model in the row is stochastically dominant over the model in the column if $H_0$ can be rejected. Considering $\tau = 0.2$, TodyNet and the LSTM showcase stochastic dominance over TimeMIL, and $H_0$ for both models can be rejected over all three datasets when compared to TimeMIL. Regarding Deeplift, similar results indicate that the models produce more faithful attributions when used with the methodology compared to TimeMIL. Noticeably, for the Rotten Tomatoes dataset, $\epsilon$ for IntGrad and Deeplift is lower for the supervised models, indicating that the technique is stochastically even more dominant compared IMDb and Twitter data.

*Table 12: ASO Epsilon Scores across Attributions Methods and Datasets for each architecture*



## 9.5. Discussion

Although Assis et al. (2024) argued that there is a trade-off between performance and interpretability, this characteristic was not reflected in this study's results when comparing weakly supervised to supervised models for predicting movie success via temporal modeling. The supervised models performed superior across evaluation dimensions for performance and interpretability, significantly outperforming the weakly-supervised TimeMIL model. Building on the work of on selecting the optimal evaluation metric, one can see why accuracy is not always the best metric to rely on when evaluating and comparing models (Grandini et al., 2020). This is undermarked by the results summarized in the box plots for the metrics Accuracy, F1, and AUROC in Figure 30 in the appendix, one can observe that, when just focusing on a single metric such as Accuracy when evaluating and comparing the classifier's performance, TimeMIL performs significantly better than the other models. However, when looking at the score distributions of the F1 and AUROC metrics, this seems less evident in the case of TodyNet. This visual example underscores the hypothesis test results, which state that TimeMIL is not significantly better than TodyNet for the Twitter dataset, emphasizing why a quantitative evaluation of models is strictly necessary when choosing the appropriate modeling approach. Regarding ***H1*** which stated superior predictive performance of MIL compared to supervised models, the results of the experiments provide partial support for the hypothesis that weakly-supervised MIL models, such as TimeMIL, outperform supervised time series classification methods in classifying movies into success categories. Specifically, TimeMIL demonstrated superiority over the LSTM model on the Twitter dataset, showcasing its ability to handle social media data sources such as tweets effectively. This aligns with the ground assumption of the weakly-supervised framework, as weakly-supervised approaches can leverage unlabeled data and capitalize on the diversity and noisiness of information. However, TimeMIL's performance on other datasets, such as Rotten Tomatoes and IMDb, was

consistently inferior to that of the supervised models, particularly TodyNet. TodyNet achieved significantly higher F1 scores and AUROC values for these datasets, demonstrating that supervised models may be advantageous when handling movie ratings and Box office success data as time series. Furthermore, while TimeMIL marginally outperformed TodyNet on the Twitter dataset, this difference was not statistically significant, suggesting that TimeMIL's strength on partially labeled data is less pronounced than more robust supervised architecture like TodyNet. These findings highlight the contextual strengths and limitations of weakly-supervised MIL models. While they excel in leveraging unlabeled, potentially noisy data, applying them to more other datasets and scenarios remains challenging. This suggests that supervised models, especially those optimized for specific data domains, may outperform weakly-supervised models in many real-world scenarios. Consequently, the hypothesis **(*H1*)** is not supported, as TimeMIL's advantage seems dataset-dependent rather than universal, underscoring the importance of selecting model architectures tailored to the nature of the data in time series classification tasks. Generally, comparing different deep learning architectures and the differences in the underlying deep learning assumptions for interpretability was a rather difficult task. Suitable and feasible to compute metrics were rather difficult to implement, picking up on existing works within the field of explainable AI (Nguyen & Martínez, 2020; Turbé et al., 2023). This is why a fusion of a set of concepts was applied, which considered concepts of corresponding works within the field of explainable AI and model interpretability. The interpretability results do not support ***H2***, which proposed that weakly-supervised MIL models like TimeMIL offer more interpretable insights into movie success factors than supervised models. TimeMIL failed to outperform traditional supervised models such as TodyNet and LSTM, showing higher infidelity scores and less focused attributions. Consistently, TimeMIL had worse infidelity across datasets indicating that it cannot compete with supervised techniques when it comes to time series classification for movie success prediction. These findings

indicate that supervised models, particularly TodyNet, are more effective for generating faithful and interpretable attributions in classifying movie success.

## 9.6. Limitations & Future Research

Although the datasets in this study provided diverse perspectives, constructing a well-defined dataset was challenging. The available data samples across rating platforms limited the final predictive performance of the models. Additionally, the models' predictive limitations can also be argued from a technical perspective. Regarding the input sequence preprocessing, the challenge of fixed sequence lengths could be addressed in the future by exploring variable lengths, incorporating adaptive mechanisms during training, and treating sequence length as a tunable hyperparameter for optimization. Moreover, the dataset-specific results highlight the importance of selecting appropriate DL architectures. Thus, future work could explore additional DL frameworks and architectures for movie success prediction using time series data. The interpretability assessment, constrained by a limited set of attribution methods evaluated through the Infidelity metric, faced several limitations. As noted by Turbé et al. (2023) certain models may appear less interpretable due to poor alignment with the chosen attribution techniques rather than their inherent properties. Aligning with the results presented in their paper, Shapley and SHAP appeared to be less applicable in the chosen comparison setup, as both techniques are less scalable than the attribution methods used. Further, only the Infidelity metric was evaluated neglecting other methods. However, it highly depends on the selected perturbation scheme, which raises questions about generalizability across different methods and models (Turbé et al., 2023). Consequently, important dimensions of interpretability, such as sensitivity, simplicity, and stability, remain unexplored in this context (Nguyen & Martínez, 2020; H. Zhang et al., 2020). Future work should test a broader range of attribution techniques, adopt more robust perturbation frameworks, and incorporate additional interpretability metrics to gain a more holistic view of model interpretability.

# 10. Conclusion & Implications for the Movie Industry

The movie industry presents a compelling landscape for the application of advanced analytics and machine learning frameworks due to its reliance on diverse data types and sources. This study explores the use of both structured and unstructured data, demonstrating how these distinct forms of information can be harnessed to address various challenges and opportunities within the domain. Traditional machine learning methods were employed on structured tabular data, such as box office revenues**,** to build predictive models and identify critical factors influencing movie success. For unstructured data, natural language processing (NLP) techniques were applied to textual information such as movie reviews to predict ratings, uncover sentiment patterns, and derive deeper insights into audience preferences. Additionally, the analysis of predictive models for multivariate time series data, encompassing features for movies ratings, review sentiment, box office revenues, provided a dynamic perspective how data around audience behavior and market fluctuations can be applied in novel frameworks. By capturing how these features evolve and interact across temporal dimensions, the study highlights the potential of applying regression analysis, NLP and fusion of sequential models with interpretability techniques to understanding trends, forecasting demand, and adapting strategies to optimize outcomes in the highly competitive movie industry. This integrative approach underscores the importance of leveraging multiple data modalities, each contributing unique insights that collectively enhance decision-making and operational efficiency. Building on the diverse data types, this study implemented a range of machine learning frameworks to extract actionable insights. Supervised learning methods, including regression analysis and classification models, were applied to structured tabular data to predict outcomes such as box office success or genre-specific performance. For text data, deep learning techniques leveraging pretrained transformer models, such as BERT, were employed. These models enabled the extraction

of contextual insights and enhanced the prediction of ratings, sentiment analysis, and audience preferences by capturing the nuanced semantics of textual content. In contrast, weakly-supervised learning frameworks, such as Multiple Instance Learning (MIL), were compared with supervised models for the case of multivariate post-release time series data, where complete labeling was challenging. Critic ratings emerged as a more significant driver of box office performance than user ratings, particularly for smaller studios and off-season releases. Positive critical reviews served as vital quality signals, especially for films with limited brand recognition. By contrast, user ratings showed limited predictive power, likely due to their variability and susceptibility to biases. Seasonal trends amplified the importance of critic ratings, with off-season films relying more heavily on professional reviews to attract audiences. These findings highlight the enduring influence of expert evaluations and the importance of timing in maximizing movie success. For studios, strategically fostering relationships with critics and aligning release schedules with periods of lower competition can amplify box office returns, particularly for niche or independent productions. The impact of releasing movies during holiday periods is highly audience dependent. For non-adult movies, holiday releases significantly boost box-office revenues, aligning with increased audience availability and family viewing habits. Conversely, adult movies experience insignificant marginal or negative effects during holidays, likely due to competition with family-oriented films. Including holidays as a predictive factor shows limited overall significance when considering all movie types collectively. These findings suggest that strategic release timing tailored to audience demographics is essential. For the film industry, leveraging holiday periods for family-friendly content while reserving less competitive windows for adult-oriented films could optimize revenues, refine marketing strategies, and enhance overall return on investment. Sequel modeling revealed that audience engagement, measured by review volume, was the most influential factor in predicting both absolute and relative success. Advanced machine learning models, such

as CatBoost, outperformed traditional approaches, highlighting the importance of capturing non-linear relationships and feature interactions. Metrics like runtime and sentiment played secondary roles, with engagement metrics proving far more predictive of financial outcomes. For the movie industry, these results emphasize the need to maintain sustained audience interaction through marketing campaigns and social media presence, particularly for franchise films. Studios can use these predictive models to assess the financial viability of sequels, optimize resource allocation, and design strategies that prioritize audience retention. Transfer learning transformer-based models, particularly BERT-based frameworks, proved effective in extracting nuanced insights from textual data, achieving high performance in sentiment analysis and rating prediction. Notably, classification models benefited from strong sentiment polarity, while regression models demonstrated resilience across varied sentiment distributions. These findings suggest that advanced NLP techniques can enable movie studios to better understand audience feedback and preferences, facilitating more precise targeting and informed creative decisions. For the industry, this underscores the potential of leveraging automated sentiment analysis tools to refine marketing strategies, improve audience engagement, and enhance content development based on data-driven insights. When applied to multivariate time series data, Multiple Instance Learning (MIL) did not perform as effectively as supervised models like TodyNet in terms of predictive accuracy. Although MIL sowcased better performance particularly in unstructured data applications like Twitter, it did not outperform supervised methods signaling that movie data might not be the perfect fit. These results suggest that while interpretability-focused techniques like MIL can be applied to data involved in decision-making processes for movies, supervised models remain the preferred choice both in terms of predictive accuracy and interpretability. However, for the movie industry, integrating such techniques into analytics pipelines in scenarios where labels are scarce could improve the understanding of temporal dynamics, such as post-release audience behavior or the

impact of streaming trends, enabling studios to make more data-informed distribution and marketing decisions. This comprehensive machine learning framework covering a wide range of datatypes and applications in the movie industry proves solid potential. This could be further enhanced by adopting a multimodal approach by integrating diverse data sources to develop more comprehensive and robust analytical frameworks that go beyond parallel model application but leverage various strengths across a sophisticated system (Madongo & Zhongjun, 2023; Miah et al., 2024). By combining structured data, such as box office performance, audience demographics, and release schedules, with unstructured data from movie reviews, sentiment, and visual or audio content, researchers can unlock synergies across data modalities that provide a more nuanced understanding of audience behavior and market dynamics. Further leveraging additional data types like image and video and integrating them into NLP techniques could enable more precise insights into the impact of soundscapes on viewer experiences. Such multimodal frameworks would allow for a broader analytical bandwidth, enhancing predictive performance and interpretability in areas (Mühling et al., 2017; Tahmasebi et al., 2021; Zhou et al., 2019). To advance the application of analytics in a multimodal fashion accounting for different data representations, academic research can contribute to more sophisticated decision-making processes and innovative strategies for addressing the dynamic challenges of the movie industry. In conclusion, by applying advanced machine learning techniques in the movie industry, this study presents a transformative opportunity for the domain to address evolving challenges and capitalize on emerging trends. Through combination of structured and unstructured data with predictive frameworks, industry stakeholders can develop strategies that enhance production efficiency, optimize release schedules, and better cater to audience preferences. This study highlights the critical role of machine learning in enabling studios to remain competitive in an increasingly data-driven landscape, urging the adoption of advanced analytics to navigate the complexities of modern cinema.

# 11. Limitations & Future Research

This study provides valuable insights into the factors influencing movie performance, yet several limitations need to be addressed to fully contextualize the findings and guide future research. These limitations arise from challenges in data acquisition, feature availability, methodological approaches, and computational constraints, which collectively influence the scope and depth of the analyses.

The studies either used a single or a combined selection of data sources, such as Rotten Tomatoes, IMDb, Twitter or BoxOfficeMojo, to create their input data, picking up on existing work (Subramaniyaswamy, Vaibhav, Prasad, & Logesh, 2017). As showcased by Y. Liao et al. (2022), one must point out that other sources exist to retrieve movie-related data that could potentially be used to conduct research, such as data from the Douban movie platform, search engine data such as Baidu, or data from other Microblogs. However, the scope of data collection was constrained by reliance on open-source platforms, which limited the comprehensiveness of the datasets. Specifically, box office data was predominantly focused on North American markets, excluding the broader global landscape and its associated cultural and regional variations. This narrow focus limits the generalizability of findings to diverse geographical contexts. Future research should aim to expand data acquisition efforts to incorporate global box office data and explore differences in audience behaviors across distinct geographical and cultural landscapes, providing a more holistic understanding of movie performance drivers.

Several important features that are likely to influence movie success were unavailable, notably marketing and advertising budgets. These variables are critical in shaping audience awareness and engagement and could significantly enhance predictive models. Additionally, the datasets exhibited heavy skews in certain features, such as genres, which may overrepresent popular categories while

underrepresenting niche ones. Ratings data, another key feature, is inherently subjective and susceptible to biases from both users and critics, potentially affecting the accuracy of performance predictions. Addressing these limitations through more diverse and balanced data sources would greatly improve future analyses. While the machine learning models employed in this study demonstrated strong predictive capabilities, they were inherently limited in their ability to establish causal relationships due to inherit irregularity within the movies data. Identifying correlations between variables is valuable for predictive analytics, but without causal inference, it remains difficult to determine which factors directly drive movie success. A notable limitation across the studies was the lack of detailed audience segmentation. While some analyses differentiated between adult and non-adult movies, there was limited exploration of granular divisions based on demographics, cultural preferences, or regional behaviors. This lack of segmentation restricts the ability to identify specific audience behaviors and preferences, which could be critical for tailoring marketing strategies and release decisions. Future research should prioritize deeper audience segmentation to uncover nuanced patterns influencing movie performance. Limited computational resources affected the ability to explore advanced techniques, such as dynamic modeling, causal inference methods, or the use of larger and more sophisticated transformer architectures. These constraints also restricted the granularity of fine-tuning and optimization processes, potentially limiting the models' predictive accuracy. Addressing these computational challenges in future work would enable more robust analyses, such as the inclusion of broader datasets, advanced model architectures, and more nuanced hyperparameter optimization. This research paper provides a detailed exploration of various factors influencing movie performance, with each study contributing particular insights into predictive analytics around movies, consumer behavior, and strategic decision-making processes within the movie industry.

# 12. Bibliography


Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B. (2017). Movie success prediction using data mining. *8th International Conference on Computing, Communication and Networking Technologies*, 1–4. IEEE. https://doi.org/10.1109/ICCCNT.2017.8204173

Ahmed, B. H., & Ghabayen, A. S. (2022). Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, *13*(7), 3423–3432. https://doi.org/10.1007/s12652-020-01807-4

Akdeniz, M. B., & Talay, M. B. (2013). Cultural variations in the use of marketing signals: A multilevel analysis of the motion picture industry. *Journal of the Academy of Marketing Science*, *41*(5), 601–624. https://doi.org/10.1007/s11747-013-0338-5

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework* (pp. 1–10). Retrieved from http://arxiv.org/abs/1907.10902

Aljrees, T., Umer, M., Saidani, O., Almuqren, L., Ishaq, A., Alsubai, S., … Ashraf, I. (2024). Contradiction in text review and apps rating: Prediction using textual features and transfer learning. *PeerJ Computer Science*, *10*(7), 1–25. https://doi.org/10.7717/peerj-cs.1722

Al-Kababji, A., Bensaali, F., & Dakua, S. P. (2022). *Scheduling techniques for liver segmentation: ReduceLROnPlateau vs OneCycleLR* (pp. 1–8). Retrieved from http://arxiv.org/abs/2202.06373

Amigó, E., Gonzalo, J., Mizzaro, S., & Carrillo-de-Albornoz, J. (2020). *An Effectiveness metric for ordinal classification: Formal properties and experimental results* (pp. 1–12). https://doi.org/10.18653/v1/2020.acl-main.363

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485–1509. https://doi.org/10.1287/mnsc.1110.1370

Assis, A., Dantas, J., & Andrade, E. (2024). The performance-interpretability trade-off: A comparative study of machine learning models. *Journal of Reliable Intelligent Environments*, *11*(1), 1–15. https://doi.org/10.1007/s40860-024-00240-0

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Bagnall, A., Dau, H. A., Lines, J., Flynn, M., Large, J., Bostrom, A., … Keogh, E. (2018). *The UEA multivariate time series classification archive*. Retrieved from http://arxiv.org/abs/1811.00075

Bai, J., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business and Economic Statistics*, *23*(1), 49–60. https://doi.org/10.1198/073500104000000271

Bakdi, A., Kristensen, N. B., & Stakkeland, M. (2022). Multiple instance learning with random forest for event logs analysis and predictive maintenance in ship electric Propulsion System. *IEEE Transactions on Industrial Informatics*, *18*(11), 7718–7728. https://doi.org/10.1109/TII.2022.3144177

Banbhrani, S. K., Xu, B., Soomro, P. D., Jain, D. K., & Lin, H. (2022). TDO-Spider Taylor ChOA: An optimized deep-learning-based sentiment classification and review rating prediction. *Applied Sciences*, *12*(20), 1–26. https://doi.org/10.3390/app122010292

Banik, R. (2017). The Movies Dataset IMDb. Retrieved December 13, 2024, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download&select=links.csv

Basuroy, S. D. K. K. T. D. (2006). An empirical investigation of signaling in the motion picture industry. *Journal of Marketing Research*, *43*(2), 287–295. https://doi.org/https://doi.org/10.1509/jmkr.43.2.287

Belvaux, B., & Mencarelli, R. (2021). Prevision model and empirical test of box office results for sequels. *Journal of Business Research*, *130*, 38–48. https://doi.org/10.1016/j.jbusres.2021.03.008

Berger, J., Kim, Y. D., & Meyer, R. (2021). What makes content engaging? How emotional dynamics shape success. *Journal of Consumer Research*, *48*(2), 235–250. https://doi.org/10.1093/jcr/ucab010

Bharadwaj, N., Noble, C. H., Tower, A., Smith, L. M., & Dong, Y. (2017). Predicting innovation success in the motion picture industry: The influence of multiple quality Signals. *Journal of Product Innovation Management*, *34*(5), 659–680. https://doi.org/10.1111/jpim.12404

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., & Watrin, P. (2022). Is attention explanation? An introduction to the debate. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 3889–3900. Long Papers. https://doi.org/10.18653/v1/2022.acl-long.269

Bilal, M., & Almazroi, A. A. (2023). Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer Reviews. *Electronic Commerce Research*, *23*(4), 2737–2757. https://doi.org/10.1007/s10660-022-09560-w

Bing, L., & Wang, W. (2017). Sparse representation based multi-instance learning for breast ultrasound image classification. *Computational and Mathematical Methods in Medicine*, *2017*(1), 1–10. https://doi.org/10.1155/2017/7894705

Board of Governors of the Federal Reserve System. (2024). Distribution of household wealth in the U.S. since 1989. Retrieved December 11, 2024, from https://www.federalreserve.gov/releases/z1/dataviz/dfa/distribute/chart/

Bureau of Labor. (2023). Occupational employment and wage statistics. Retrieved November 12, 2024, from https://www.bls.gov/oes/2023/may/naics4_512100.htm

Burton, A. L. (2021). OLS (linear) regression. In *The Encyclopedia of Research Methods in Criminology and Criminal Justice* (pp. 1–11). Wiley. https://doi.org/DOI:10.1002/9781119111931

Carbonneau, M. A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, *77*(2), 329–353. https://doi.org/10.1016/j.patcog.2017.10.009

Carrillat, F. A., Legoux, R., & Hadida, A. L. (2018). Debates and assumptions about motion picture performance: A meta-analysis. *Journal of the Academy of Marketing Science*, *46*(2), 273–299. https://doi.org/10.1007/s11747-017-0561-6

Cavallo, B. (2020). Functional relations and spearman correlation between consistency indices. *Journal of the Operational Research Society*, *71*(2), 301–311. https://doi.org/10.1080/01605682.2018.1516178

Chambua, J., & Niu, Z. (2021). Review text based rating prediction approaches: Preference knowledge learning, representation and utilization. *Artificial Intelligence Review*, *54*(2), 1171–1200. https://doi.org/10.1007/s10462-020-09873-y

Chen, G., Lockhart, R. A., & Stephens, M. A. (2002). Box-Cox transformations in linear models: Large sample theory and tests of normality. *The Canadian Journal of Statistics*, *30*(2), 1–59. https://doi.org/https://doi.org/10.2307/3315946

Chen, X., Qiu, P., Zhu, W., Li, H., Wang, H., Sotiras, A., … Razi, A. (2024). *TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning* (pp. 1–17). pp. 1–17. Retrieved from http://arxiv.org/abs/2405.03140

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Cox, D. B. G. E. P. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological))*, *26*(2), 211–252. https://doi.org/https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Danyal, M. M., Khan, S. S., Khan, M., Ullah, S., Mehmood, F., & Ali, I. (2024). Proposing sentiment analysis model based on BERT and XLNet for movie reviews. *Multimedia Tools and Applications*, *83*(24), 64315–64339. https://doi.org/10.1007/s11042-024-18156-5

Darwish, A., Hassanien, A. E., & Das, S. (2020). A survey of swarm and evolutionary computing approaches for deep learning. *Artificial Intelligence Review*, *53*(3), 1767–1812. https://doi.org/10.1007/s10462-019-09719-2

D'astous, A., Montreal, H., Touil, N., & Tunisia, P. (1999). Consumer evaluations of movies on the basis of critics' judgments. *Psychology & Marketing*, *16*(8), 677–694. https://doi.org/https://doi.org/10.1002/(SICI)1520-6793(199912)16:8<677::AID-MAR4>3.0.CO;2-T

del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2017). An optimal transportation approach for assessing almost stochastic order. In *Studies in Systems, Decision and Control* (Vol. 142, pp. 33–44). Springer International. https://doi.org/https://doi.org/10.1007/978-3-319-73848-2_3

Dellarocas, C., Zhang, X., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45. https://doi.org/10.1002/dir.20087

Delre, S. A., & Luffarelli, J. (2023). Consumer reviews and product life cycle: On the temporal dynamics of electronic word of mouth on movie box office. *Journal of Business Research*, *156*, 1–26. https://doi.org/10.1016/j.jbusres.2022.113329

Deng, Y., & Yiu, S. M. (2022). Deep multiple instance learning for forecasting stock trends using financial NewsDeep multiple instance learning for forecasting stock trends using financial news. *Conference: 8th International Conference on Artificial Intelligence*, 95–111. Academy and Industry Research Collaboration Center (AIRCC). https://doi.org/10.5121/csit.2022.121008

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Retrieved from https://github.com/tensorflow/tensor2tensor

Divakaran, P. K. P., & Nørskov, S. (2016). Are online communities on par with experts in the evaluation of new movies? Evidence from the fandango community. *Information Technology and People*, *29*(1), 120–145. https://doi.org/10.1108/ITP-02-2014-0042

Divakaran, P. K. P., Palmer, A., Sendergaard, H. A., & Matkovskyy, R. (2017). Pre-launch prediction of market performance for short lifecycle products using online community data. *Journal of Interactive Marketing*, *38*(1), 12–28. https://doi.org/https://doi.org/10.1016/j.intmar.2016.10.004

Dooms, S. (2021). MovieTweetings. Retrieved December 13, 2024, from https://github.com/sidooms/MovieTweetings

Dror, R., Shlomov, S., & Reichart, R. (2019). Deep dominance - how to properly compare deep neural models. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2773–2785. Association for Computational Linguistics. https://doi.org/https://doi.org/10.18653/v1/p19-1266

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), 233–242. https://doi.org/10.1016/j.jretai.2008.04.005

Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92–108. https://doi.org/10.1016/j.neucom.2022.06.111

Early, J., Cheung, G. K., Cutajar, K., Xie, H., Kandola, J., & Twomey, N. (2024). *Inherently interpretable time series classification via multiple instance learning*. Retrieved from http://arxiv.org/abs/2311.10049

Einav, L. (2007). Seasonality in the U.S. motion picture industry. *RAND Journal of Economics*, *38*(1), 127–145. https://doi.org/10.1111/j.1756-2171.2007.tb00048.x

Elberse, A. (2007). The power of stars: Do star actors drive the success of movies? *Journal of Marketing*, *71*(4), 102–120. https://doi.org/10.1509/jmkg.71.4.102

Fahey, M. (2015). The big crunch. Retrieved November 13, 2024, from https://www.cnbc.com/2015/11/17/why-movies-are-sometimes-here-and-gone-in-theaters.html

Flanagin, A. J., & Metzger, M. J. (2013). Trusting expert versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior*, *29*(4), 1626–1634. https://doi.org/10.1016/j.chb.2013.02.001

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). SAGE Publications.

Galal, O., Abdel-Gawad, A. H., & Farouk, M. (2024). Rethinking of BERT sentence embedding for text classification. *Neural Computing and Applications*, *36*, 20245–20258. https://doi.org/10.1007/s00521-024-10212-3

García, S., Herrera, F., & Es, H. U. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, *9*(12), 2677–2694. https://doi.org/https://jmlr.org/papers/v9/garcia08a.html

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics*, *1*(1), 1–22. https://doi.org/10.1186/s41044-016-0014-0

Ghanbari, R., & Borna, K. (2021). Multivariate Time-Series Prediction Using LSTM Neural Networks. *26th International Computer Conference, Computer Society of Iran, CSICC 2021*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CSICC52343.2021.9420543

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: An overview*. Retrieved from http://arxiv.org/abs/2008.05756

Guo, S., Zhong, S., & Zhang, A. (2013). Privacy-preserving Kruskal-Wallis test. *Computer Methods and Programs in Biomedicine*, *112*(1), 135–145. https://doi.org/10.1016/j.cmpb.2013.05.023

Haben, S., Voss, M., & Holderbaum, W. (2023). Time series forecasting: Core concepts and definitions. In *Core concepts and methods in load forecasting* (pp. 55–66). Springer International. https://doi.org/https://doi.org/10.1007/978-3-031-27852-5_5

He, H., Wang, H., Ma, H., Liu, X., Jia, Y., & Gong, G. (2020). Research on short-term power load forecasting based on Bi-GRU. *Journal of Physics: Conference Series*, *1639*(1), 1–7. IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1639/1/012017

Helan, A., & Sultani, Z. N. (2023). Topic modeling methods for text data analysis: A review. *AIP Conference Proceedings*, *2457*, 1–14. American Institute of Physics Inc. https://doi.org/10.1063/5.0118679

Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, *52*(3), 1495–1545. https://doi.org/10.1007/s10462-017-9599-6

Hennig-Thurau, T., Houston, M. B., & Heitjans, T. (2009). Conceptualizing and measuring the monetary value of brand extensions: The case of motion Pictures. *Journal of Marketing*, *73*(6), 167–183. https://doi.org/10.1509/jmkg.73.6.167

Ho, Y., & Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, *99*(1), 4806–4813. https://doi.org/10.1109/ACCESS.2019.2962617

Huang, J., Boh, W. F., & Goh, K. H. (2017). A temporal study of the effects of online opinions: Information sources matter. *Journal of Management Information Systems*, *34*(4), 1169–1202. https://doi.org/10.1080/07421222.2017.1394079

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–226. https://doi.org/https://doi.org/10.1609/icwsm.v8i1.14550

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Jahin, M. A., Shovon, M. S. H., Mridha, M. F., Islam, M. R., & Watanobe, Y. (2024). A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets. *Scientific Reports*, *14*(1), 1–28. https://doi.org/10.1038/s41598-024-76079-5

Jain, S., & Wallace, B. C. (2019). *Attention is not Explanation*. https://doi.org/https://doi.org/10.18653/v1/D19-1002

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*, 685–695. https://doi.org/10.1007/s12525-021-00475-2/Published

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259. https://doi.org/https://doi.org/10.1016/0165-1765(80)90024-5

Karch, J. D. (2021). Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–14. https://doi.org/10.1177/2515245921999602

Karniouchina, E. V., Carson, S. J., Theokary, C., Rice, L., & Reilly, S. (2023). Women and minority film directors in hollywood: Performance implications of product development and distribution biases. *Journal of Marketing Research*, *60*(1), 25–51. https://doi.org/10.1177/00222437221100217

Khan, F. H., Qamar, U., & Bashir, S. (2016). eSAP: A decision support framework for enhanced sentiment analysis and polarity classification. *Information Sciences*, *367–368*, 862–873. https://doi.org/10.1016/j.ins.2016.07.028

Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*, *4*(2), 1–19. https://doi.org/10.1186/s40294-016-0016-9

Khan, Z. Y., Niu, Z., Sandiwarno, S., & Prince, R. (2021). Deep learning techniques for rating prediction: A survey of the state-of-the-art. *Artificial Intelligence Review*, *54*(1), 95–135. https://doi.org/10.1007/s10462-020-09892-9

Kim, S. C., Sung, K. J., Park, C. S., & Kim, S. K. (2016). Improvement of collaborative filtering using rating normalization. *Multimedia Tools and Applications*, *75*(9), 4957–4968. https://doi.org/10.1007/s11042-013-1814-0

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., … Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch* (pp. 1–11). pp. 1–11. https://doi.org/10.48550/arXiv.2009.07896

Köksoy, O. (2006). Multiresponse robust design: Mean square error (MSE) criterion. *Applied Mathematics and Computation*, *175*(2), 1716–1729. https://doi.org/10.1016/j.amc.2005.09.016

Kolpaczki, P., Bengs, V., Muschalik, M., & Hüllermeier, E. (2024). *Approximating the Shapley Value without Marginal Contributions* (pp. 1–37). https://doi.org/https://doi.org/10.48550/arXiv.2009.07896

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Kraft, A., & Rao, R. S. (2024). Signaling quality via demand lockout. *Quantitative Marketing and Economics*. https://doi.org/10.1007/s11129-024-09288-x

Kumar, N., & Hanji, B. R. (2024). Aspect-based sentiment score and star rating prediction for travel destination using Multinomial Logistic Regression with fuzzy domain ontology algorithm. *Expert Systems with Applications*, *240*(1), 1–16. https://doi.org/10.1016/j.eswa.2023.122493

Lai, C. H., & Hsu, C. Y. (2021). Rating prediction based on combination of review mining and user preference analysis. *Information Systems*, *99*, 1–14. https://doi.org/10.1016/j.is.2021.101742

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), 874–903. https://doi.org/10.1080/07421222.2016.1243969

Lee, J. H., Jung, S. H., & Park, J. H. (2017). The role of entropy of review text sentiments on online WOM and movie box office sales. *Electronic Commerce Research and Applications*, *22*(4), 42–52. https://doi.org/10.1016/j.elerap.2017.03.001

Lee, Y. J., Hosanagar, K., & Tan, Y. (2015). Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Science*, *61*(9), 2241–2258. https://doi.org/10.1287/mnsc.2014.2082

Lehrer, S. F., & Xie, T. (2021). *The bigger picture: Combining econometrics with analytics improve forecasts of movie success* (No. 24755). Retrieved from http://www.nber.org/papers/w24755

Leone, S. (2020). Rotten Tomatoes movies and critic reviews dataset. Retrieved December 13, 2024, from https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/

Liao, L., & Huang, T. (2021). The effect of different social media marketing channels and events on movie box office: An elaboration likelihood model perspective. *Information and Management*, *58*(7), 1–16. https://doi.org/10.1016/j.im.2021.103481

Liao, W., Zeng, B., Yin, X., & Wei, P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. *Applied Intelligence*, *51*(6), 3522–3533. https://doi.org/10.1007/s10489-020-01964-1

Liao, Y., Peng, Y., Shi, S., Shi, V., & Yu, X. (2022). Early box office prediction in china's film market based on a stacking fusion model. *Annals of Operations Research*, *308*(4), 321–338. https://doi.org/10.1007/s10479-020-03804-4

Liu, H., Yang, D., Liu, X., Chen, X., Liang, Z., Wang, H., … Gu, J. (2024). TodyNet: Temporal dynamic graph neural network for multivariate time series classification. *Information Sciences*, *677*, 1–16. https://doi.org/10.1016/j.ins.2024.120914

Liu, X., Tian, J., Niu, N., Li, J., & Han, J. (2023). "Standard Text" relational classification model based on concatenated word vector attention and feature concatenation. *Applied Sciences (Switzerland)*, *13*(12), 1–17. https://doi.org/10.3390/app13127119

Liu, Z. (2020, December 11). *Yelp review rating prediction: Machine learning and deep learning models* (pp. 1–8). pp. 1–8. Retrieved from http://arxiv.org/abs/2012.06690

Loshchilov, I., & Hutter, F. (2017). *Decoupled Weight Decay Regularization*. 1–19. https://doi.org/https://doi.org/10.48550/arXiv.1711.05101

Loupos, P., Peng, Y., Li, S., & Hao, H. (2023). What reviews foretell about opening weekend box office revenue: the harbinger of failure effect in the movie industry. *Marketing Letters*, *34*(3), 513–534. https://doi.org/10.1007/s11002-023-09665-8

Lutz, B., Pröllochs, N., & Neumann, D. (2022). Are longer reviews always more helpful? Disentangling the interplay between review length and line of argumentation. *Journal of Business Research*, *144*(11), 888–901. https://doi.org/10.1016/j.jbusres.2022.02.010

Lv, X., Liu, Z., Zhao, Y., Xu, G., & You, X. (2023). HBert: A Long text processing method based on BERT and hierarchical attention mechanisms. *International Journal on Semantic Web and Information Systems*, *19*(1), 1–14. https://doi.org/10.4018/IJSWIS.322769

Madongo, C. T., & Zhongjun, T. (2023). A movie box office revenue prediction model based on deep multimodal features. *Multimedia Tools and Applications*, *82*(21), 31981–32009. https://doi.org/10.1007/s11042-023-14456-4

Mansoury, M., Burke, R., & Mobasher, B. (2021). Flatter Is Better: Percentile Transformations for Recommender Systems. *ACM Transactions on Intelligent Systems and Technology*, *12*(2), 1–16. https://doi.org/10.1145/3437910

Maulidi, I., Syahrini, I., Oktavia, R., Ihsan, M., & Emha, R. (2021). An analysis on determination of land area claim prediction for rice farmers insurance business in Indonesia using nonparametric bayesian method. *IOP Conference Series: Earth and Environmental Science*, *1796*(1), 1–9. https://doi.org/10.1088/1742-6596/1796/1/012006

Miah, M. S. U., Kabir, M. M., Sarwar, T. Bin, Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, *14*(1), 1–18. https://doi.org/10.1038/s41598-024-60210-7

Middlehurst, M., Ismail-Fawaz, A., Guillaume, A., Holder, C., Guijo-Rubio, D., Bulatova, G., … Bagnall, A. (2024). Aeon: A python toolkit for learning from time series. *Journal of Machine Learning Research*, *25*, 1–10. Retrieved from http://jmlr.org/papers/v25/23-1444.html.

Mohamadi, P. (2023). *Unofficial Python API for Box Office Mojo*. Retrieved from https://github.com/Stink-Po/boxoffice_api

Mohamadi, P. (2024). Box Office Mojo API. Retrieved December 13, 2024, from https://github.com/Stink-Po/boxoffice_api

Mohapatra, D., Bhoi, S. K., Mallick, C., Jena, K. K., & Mishra, S. (2022). Distribution preserving train-test split directed ensemble classifier for heart disease prediction. *International Journal of Information Technology (Singapore)*, *14*(4), 1763–1769. https://doi.org/10.1007/s41870-022-00868-2

Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing*, *74*(1), 108–121. https://doi.org/10.1509/jmkg.74.1.108

Mosin, V., Samenko, I., Kozlovskii, B., Tikhonov, A., & Yamshchikov, I. P. (2023). Fine-tuning transformers: Vocabulary transfer. *Artificial Intelligence*, *317*, 1–11. https://doi.org/10.1016/j.artint.2023.103860

Mühling, M., Korfhage, N., Müller, E., Otto, C., Springstein, M., Langelage, T., … Freisleben, B. (2017). Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications*, *76*(21), 22169–22194. https://doi.org/10.1007/s11042-017-4962-9

Namdari, A., & Li, Z. (Steven). (2019). A review of entropy measures for uncertainty quantification of stochastic processes. *Advances in Mechanical Engineering*, *11*(6), 1–14. https://doi.org/10.1177/1687814019857350

Navarathna, R., Carr, P., Lucey, P., & Matthews, I. (2019). Estimating audience engagement to predict movie ratings. *IEEE Transactions on Affective Computing*, *10*(1), 48–59. https://doi.org/10.1109/TAFFC.2017.2723011

Nguyen, A., & Martínez, M. R. (2020). *On quantitative aspects of model interpretability* (pp. 1–14). pp. 1–14. https://doi.org/https://doi.org/10.48550/arXiv.2007.07584

Pang, J., Liu, A. X., & Golder, P. N. (2022). Critics' conformity to consumers in movie evaluation. *Journal of the Academy of Marketing Science*, *50*(4), 864–887. https://doi.org/10.1007/s11747-021-00816-9

Park, S. K., Song, T., & Sela, A. (2023). The effect of subjectivity and objectivity in online reviews: A convolutional neural network approach. *Journal of Consumer Psychology*, *33*(4), 701–713. https://doi.org/10.1002/jcpy.1382

Perez, I., Skalski, P., Barns-Graham, A., Wong, J., & Sutton, D. (2021). *Attribution of predictive uncertainties in classification models* (pp. 1–16). https://doi.org/https://doi.org/10.1109/ICCV.2019.00505

Pew Research Center. (2024). Social media fact sheet. Retrieved December 1, 2024, from https://www.pewresearch.org/internet/fact-sheet/social-media/

Prosser, E. K. (2002). How early can video revenue be accurately predicted? *Journal of Advertising Research*, *42*(2), 47–55. https://doi.org/https://doi.org/10.2501/JAR-42-2-47-55

Qiao, Y., Zhu, X., & Gong, H. (2022). BERT-Kcr: Prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics*, *38*(3), 648–654. https://doi.org/10.1093/bioinformatics/btab712

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. In *Scientific Reports* (Vol. 14). Nature Research. https://doi.org/10.1038/s41598-024-56706-x

Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2021). BERT, XLNet or RoBERTa: The best transfer learning model to detect clickbaits. *IEEE Access*, *9*, 154704–154716. https://doi.org/10.1109/ACCESS.2021.3128742

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., & Dodds, P. S. (2017). Sentiment analysis methods for understanding large-scale texts: A case for using continuum-scored words and word shift graphs. *EPJ Data Science*, *6*(1), 1–21. https://doi.org/10.1140/epjds/s13688-017-0121-9

Ren, Z., Wang, S., & Zhang, Y. (2023). Weakly supervised machine learning. *CAAI Transactions on Intelligence Technology*, *8*(3), 549–580. https://doi.org/10.1049/cit2.12216

Rhee, T., & Zulkerine, F. H. (2016). Predicting movie box office profitability: A neural network approach. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and*

*Applications, ICMLA 2016*, 665–670. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICMLA.2016.138

Rusli, E. M. (2012, April). Facebook buys instagram for $1 billion. Retrieved December 16, 2024, from https://archive.nytimes.com/dealbook.nytimes.com/2012/04/09/facebook-buys-instagram-for-1-billion/

Rutledge, D. N., & Barros, A. S. (2002). Durbin-Watson statistic as a morphological estimator of information content. *Analytica Chimica Acta*, *454*(2), 277–295. https://doi.org/https://doi.org/10.1016/S0003-2670(01)01555-0

Ryoo, J. H., Wang, X., & Lu, S. (2021). Do spoilers really spoil? Using topic modeling to measure the effect of spoiler reviews on box office revenue. *Journal of Marketing*, *85*(2), 70–88. https://doi.org/10.1177/0022242920937703

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., … Coyle, D. (2024). *Computing power and the Governance of artificial intelligence* (pp. 1–104). https://doi.org/https://doi.org/10.48550/arXiv.2402.08797

Schlegel, U., & Keim, D. A. (2021). *Time series model attribution visualizations as explanations* (pp. 1–5). pp. 1–5. https://doi.org/https://doi.org/10.48550/arXiv.2109.12935

Shao, M., Basit, A., Karri, R., & Shafique, M. (2024). Survey of different large language model architectures: Trends, benchmarks, and challenges. *IEEE Access*, 1–41. https://doi.org/10.1109/ACCESS.2024.3482107

Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods*, *56*(2), 826–845. https://doi.org/10.3758/s13428-023-02072-x

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). *Learning important features through propagating activation Differences* (pp. 1–9). https://doi.org/https://doi.org/10.48550/arXiv.1704.02685

Simonton, D. K. (2009). Cinematic success criteria and their predictors: The art and business of the film industry. *Psychology and Marketing*, *26*(5), 400–420. https://doi.org/10.1002/mar.20280

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency Maps* (pp. 1–8). https://doi.org/https://doi.org/10.48550/arXiv.1312.6034

Singh, T., Rajput, V., Sharma, N., Satakshi, & Kumar, M. (2024). Sentiment analysis based distributed recommendation system. *Multimedia Tools and Applications*, *83*(25), 66539–66563. https://doi.org/10.1007/s11042-023-18081-z

Somlo, B., Rajaram, K., & Ahmadi, R. (2011). Distribution planning to optimize profits in the motion picture industry. *Production and Operations Management*, *20*(4), 618–636. https://doi.org/10.1111/j.1937-5956.2010.01166.x

Song, T., Huang, J., Tan, Y., & Yu, Y. (2019). Using user- and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Information Systems Research*, *30*(1), 191–203. https://doi.org/10.1287/isre.2018.0797

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 182–186. Palladam: IEEE. https://doi.org/10.1109/ISS1.2017.8389394

Subramaniyaswamy, V., Viginesh, V. M., Vishnu, P. R., & Logesh, R. (2017). Predicting movie box office success using multiple regression and svm. *Proceedings of the International Conference*

*on Intelligent Sustainable Systems*, 526. Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ISS1.2017.8389394

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks* (pp. 1–11). https://doi.org/https://doi.org/10.48550/arXiv.1703.01365

Tahmasebi, H., Ravanmehr, R., & Mohamadrezaei, R. (2021). Social movie recommender system based on deep autoencoder network using Twitter data. *Neural Computing and Applications*, *33*(5), 1607–1623. https://doi.org/10.1007/s00521-020-05085-1

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*(3), 429–441. https://doi.org/10.1016/j.ins.2019.11.004

The Numbers. (2024). All time worldwide sequel box office. Retrieved December 13, 2024, from https://www.the-numbers.com/box-office-records/worldwide/all-movies/cumulative/sequel

Tibo, A., Jaeger, M., Frasconi, P., & Wrobel, S. (2020). Learning and interpreting multi-multi-instance learning networks. *Journal of Machine Learning Research*, *21*, 1–60. https://doi.org/https://www.jmlr.org/papers/v21/18-811.html

TMDB. (2023). TMDB API. Retrieved December 13, 2024, from https://developer.themoviedb.org/reference/intro/getting-started

*TMDB API* (Vol. 3). (2023). TiVo Platform Technologies LLC. Retrieved from https://developer.themoviedb.org/docs/getting-started

Tsao, W. C. (2014). Which type of online review is more persuasive? The influence of consumer reviews and critic ratings on moviegoers. *Electronic Commerce Research*, *14*(4), 559–583. https://doi.org/10.1007/s10660-014-9160-5

Turbé, H., Bjelogrlic, M., Lovis, C., & Mengaldo, G. (2023). Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, *5*(3), 250–260. https://doi.org/10.1038/s42256-023-00620-w

Ulmer, D., Hardmeier, C., & Frellsen, J. (2022). *Easy and meaningful statistical significance testing in the age of neural networks* (pp. 1–20). Retrieved from http://arxiv.org/abs/2204.06815

Urzúa, C. M. (1996). On the correct use of omnibus tests for normality. *Economics Letters*, *53*(3), 247–251. https://doi.org/10.1016/S0165-1765(96)00923-8

Vasanthakumari, R. K., Nair, R. V., & Krishnappa, V. G. (2023). Improved learning by using a modified activation function of a convolutional neural network in multi-spectral image classification. *Machine Learning with Applications*, *14*(2), 100502. https://doi.org/10.1016/j.mlwa.2023.100502

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., … Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems*, 1–11.

Wadawadagi, R., & Pagi, V. (2020). Sentiment analysis with deep neural networks: Comparative study and performance assessment. *Artificial Intelligence Review*, *53*(8), 6155–6195. https://doi.org/10.1007/s10462-020-09845-2

Wang, H., & Guo, K. (2017). The impact of online reviews on exhibitor behaviour: Evidence from movie industry. *Enterprise Information Systems*, *11*(10), 1518–1534. https://doi.org/10.1080/17517575.2016.1233458

Watson, F., & Wu, Y. (2022). The Impact of online reviews on the information flows and outcomes of marketing systems. *Journal of Macromarketing*, *42*(1), 146–164. https://doi.org/10.1177/02761467211042552

Wu, X., Liao, H., & Tang, M. (2023). Decision making towards large-scale alternatives from multiple online platforms by a multivariate time-series-based method. *Expert Systems with Applications*, *212*, 1–12. https://doi.org/10.1016/j.eswa.2022.118838

Yang, G., Xu, Y., & Tu, L. (2023). An intelligent box office predictor based on aspect-level sentiment analysis of movie review. *Wireless Networks*, *29*(7), 3039–3049. https://doi.org/10.1007/s11276-023-03378-6

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). *On the (in)fidelity and sensitivity for explanations*. https://doi.org/https://doi.org/10.48550/arXiv.1901.09392

Zhang, C., Tian, Y. X., & Fan, Z. P. (2022). Forecasting the box offices of movies coming soon using social media analysis: A method based on improved bass models. *Expert Systems with Applications*, *191*, 1–16. https://doi.org/10.1016/j.eswa.2021.116241

Zhang, H., Yuan, X., & Song, T. H. (2020). Examining the role of the marketing activity and eWOM in the movie diffusion: The decomposition perspective. *Electronic Commerce Research*, *20*(3), 589–608. https://doi.org/10.1007/s10660-020-09423-2

Zhou, Y., Zhang, L., & Yi, Z. (2019). Predicting movie box-office revenues using deep neural networks. *Neural Computing and Applications*, *31*(6), 1855–1865. https://doi.org/10.1007/s00521-017-3162-x

# 13. Appendix

## 13.1. The Differential Impact of Critic and User Ratings on Box Office Success: Exploring the Roles of Seasonality, Star Power, and Studio Size

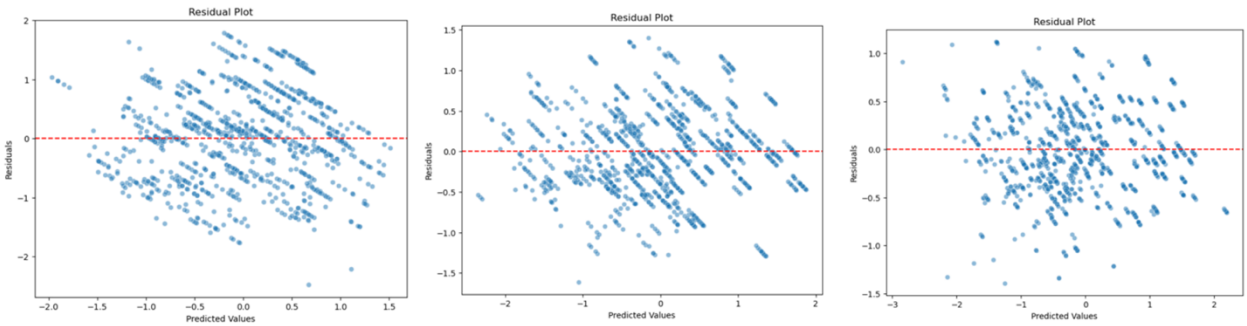



*Figure 13: Fitted-vs-Residual Plot for Model 1,2 and 3*

## 13.2. Understanding Box-Office Dynamics: The Role of Timing and Audience in Movie Revenue Prediction

*Table 13 - Linear Regression for Adult Movies*

**Adult Movies**

| | coefficient | std error | t | P>\|t\| | Significance Level | [0.025 | 0.975] |
|---|---|---|---|---|---|---|---|
| const | 14.3819 | 1.494 | 9.624 | 0 | *** | 11.445 | 17.319 |
| audience_count | 2.853 | 0.627 | 4.55 | 0 | *** | 1.621 | 4.086 |
| fresh_ratio | 3.6463 | 1.535 | 2.376 | 0.02 | ** | 0.631 | 6.662 |
| holiday | -0.1294 | 0.078 | -1.668 | 0.10 | * | -0.282 | 0.023 |
| is_actionandadventure | 0.0972 | 0.082 | 1.184 | 0.24 | | -0.064 | 0.259 |
| is_arthouseandinternational | -0.0938 | 0.171 | -0.549 | 0.583 | | -0.429 | 0.242 |
| is_comedy | 0.1314 | 0.096 | 1.374 | 0.17 | | -0.057 | 0.32 |
| is_drama | -0.1787 | 0.084 | -2.127 | 0.03 | ** | -0.344 | -0.014 |
| is_horror | 0.3268 | 0.109 | 3 | 0.00 | *** | 0.113 | 0.541 |
| is_mysteryandsuspense | 0.0316 | 0.082 | 0.384 | 0.70 | | -0.13 | 0.193 |
| is_romance | 0.0235 | 0.153 | 0.154 | 0.878 | | -0.277 | 0.324 |
| is_sciencefictionandfantasy | 0.1308 | 0.12 | 1.092 | 0.28 | | -0.105 | 0.366 |
| recent | -0.1453 | 0.289 | -0.504 | 0.62 | | -0.712 | 0.422 |
| rotten_ratio | 3.1457 | 1.439 | 2.186 | 0.03 | ** | 0.318 | 5.973 |
| runtime | 1.2488 | 0.262 | 4.771 | 0.00 | *** | 0.734 | 1.763 |
| tomatometer_rating | 0.2954 | 2.35 | 0.126 | 0.9 | | -4.324 | 4.914 |

1% significance level ($p < 0.01$): ***, 5% significance level ($0.01 <= p < 0.05$): **, 10% significance level ($0.05 <= p <= 0.1$): *

*Table 14 - Linear Regression for Non-Adult Movies*

**Non-Adult Movies**

| | coefficient | std error | t | P>\|t\| | Significance Level | [0.025 | 0.975] |
|---|---|---|---|---|---|---|---|
| const | 15.601 | 1.309 | 11.922 | 0 | *** | 13.032 | 18.17 |
| audience_count | 2.663 | 0.311 | 8.554 | 0 | *** | 2.052 | 3.274 |
| fresh_ratio | 6.8528 | 1.297 | 5.284 | 0.00 | *** | 4.307 | 9.399 |
| holiday | 0.1776 | 0.072 | 2.451 | 0.01 | ** | 0.035 | 0.32 |
| is_actionandadventure | 0.391 | 0.074 | 5.282 | 0.00 | *** | 0.246 | 0.536 |
| is_arthouseandinternational | -0.0897 | 0.194 | -0.463 | 0.644 | | -0.47 | 0.291 |
| is_comedy | -0.0569 | 0.083 | -0.681 | 0.50 | | -0.221 | 0.11 |
| is_drama | -0.3872 | 0.075 | -5.16 | 0.00 | *** | -0.534 | -0.24 |
| is_horror | -0.2506 | 0.133 | -1.884 | 0.06 | * | -0.512 | 0.011 |
| is_mysteryandsuspense | -0.1541 | 0.097 | -1.587 | 0.11 | | -0.345 | 0.036 |
| is_romance | -0.1432 | 0.105 | -1.364 | 0.173 | | -0.349 | 0.063 |
| is_sciencefictionandfantasy | 0.3191 | 0.082 | 3.908 | 0.00 | *** | 0.159 | 0.479 |
| recent | -0.7755 | 0.261 | -2.967 | 0.00 | *** | -1.289 | -0.262 |
| rotten_ratio | 2.7748 | 1.618 | 1.715 | 0.09 | * | -0.402 | 5.952 |
| runtime | 1.2853 | 0.291 | 4.423 | 0.00 | *** | 0.715 | 1.856 |
| tomatometer_rating | -3.5842 | 2.126 | -1.686 | 0.092 | * | -7.758 | 0.59 |

1% significance level ($p < 0.01$): ***, 5% significance level ($0.01 \leq p < 0.05$): **, 10% significance level ($0.05 \leq p \leq 0.1$): *

*Table 15 - Linear Regression for All Movies*

**Complete (All Movies)**

| | coefficient | std error | t | P>\|t\| | Significance Level | [0.025 | 0.975] |
|---|---|---|---|---|---|---|---|
| const | 17.2587 | 0.376 | 45.957 | 0 | *** | 16.522 | 17.995 |
| audience_count | 2.702 | 0.267 | 10.133 | 0 | *** | 2.179 | 3.225 |
| fresh_ratio | 6.2411 | 0.626 | 9.963 | 0.00 | *** | 5.012 | 7.47 |
| holiday | 0.0625 | 0.054 | 1.153 | 0.25 | | -0.044 | 0.169 |
| is_actionandadventure | 0.3127 | 0.056 | 5.599 | 0.00 | *** | 0.203 | 0.422 |
| is_arthouseandinternational | -0.1422 | 0.131 | -1.085 | 0.278 | | -0.399 | 0.115 |
| is_comedy | 0.0064 | 0.064 | 0.101 | 0.92 | | -0.119 | 0.13 |
| is_drama | -0.3114 | 0.057 | -5.488 | 0.00 | *** | -0.423 | -0.2 |
| is_horror | 0.0728 | 0.085 | 0.86 | 0.39 | | -0.093 | 0.239 |
| is_mysteryandsuspense | -0.0792 | 0.064 | -1.228 | 0.22 | | -0.206 | 0.047 |
| is_romance | -0.134 | 0.086 | -1.55 | 0.121 | | -0.304 | 0.036 |
| is_sciencefictionandfantasy | 0.3247 | 0.067 | 4.859 | 0.00 | *** | 0.194 | 0.456 |
| recent | -0.5206 | 0.196 | -2.657 | 0.01 | ** | -0.905 | -0.136 |
| rotten_ratio | 0.5474 | 0.418 | 1.309 | 0.19 | | -0.273 | 1.368 |
| runtime | 1.2868 | 0.197 | 6.519 | 0.00 | *** | 0.899 | 1.674 |
| tomatometer_rating | -4.6644 | 0.604 | -7.717 | 0 | *** | -5.85 | -3.478 |

1% significance level ($p < 0.01$): ***, 5% significance level ($0.01 \leq p < 0.05$): **, 10% significance level ($0.05 \leq p \leq 0.1$): *

*Table 16 - Advanced Machine Learning Models' Performance*

| | WMAPE | WMAPE | MAPE | MAPE | MAE (M $) | MAE (M $) |
|---|---|---|---|---|---|---|
| **Adult Movies** | 0.594845 | 0.598403 | 0.887715 | 0.889757 | 6.75E+01 | 6.79E+01 |
| **Non-Adult Movies** | 0.54988 | 0.548936 | 1.288298 | 1.282034 | 1.35E+02 | 1.35E+02 |
| **Complete (All Movies)** | 0.549109 | 0.546464 | 1.096107 | 1.106344 | 1.07E+02 | 1.06E+02 |
| *holiday* variable | included | excluded | included | excluded | included | excluded |

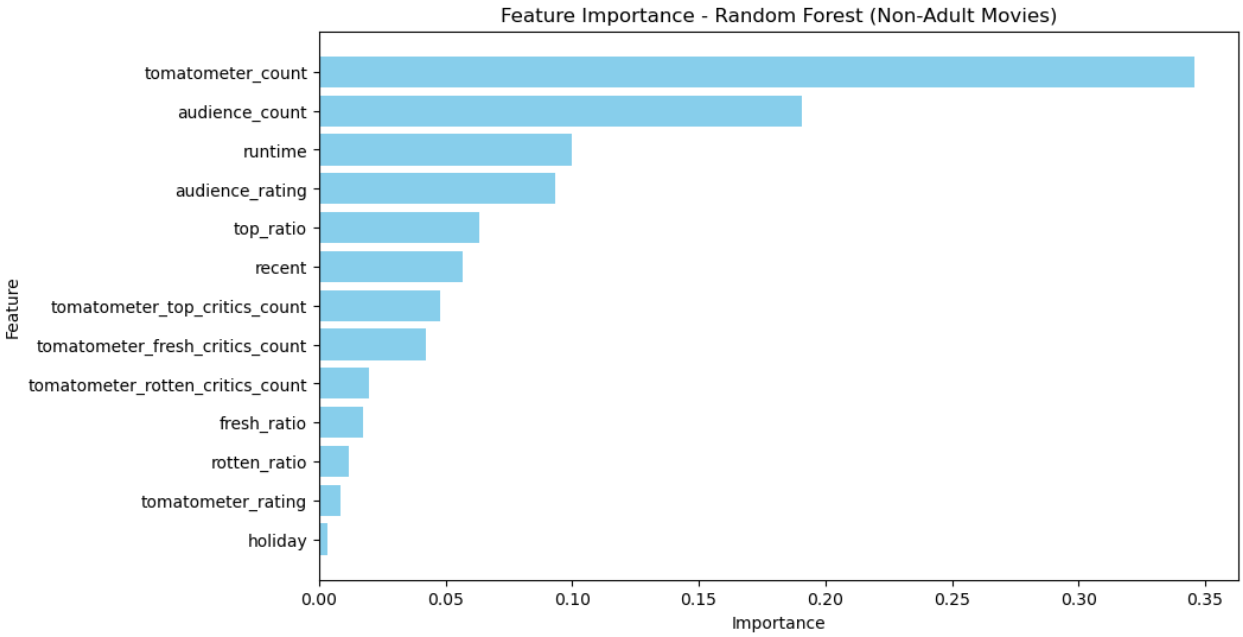



*Figure 14 - Feature Importance for Non-Adult Movies*

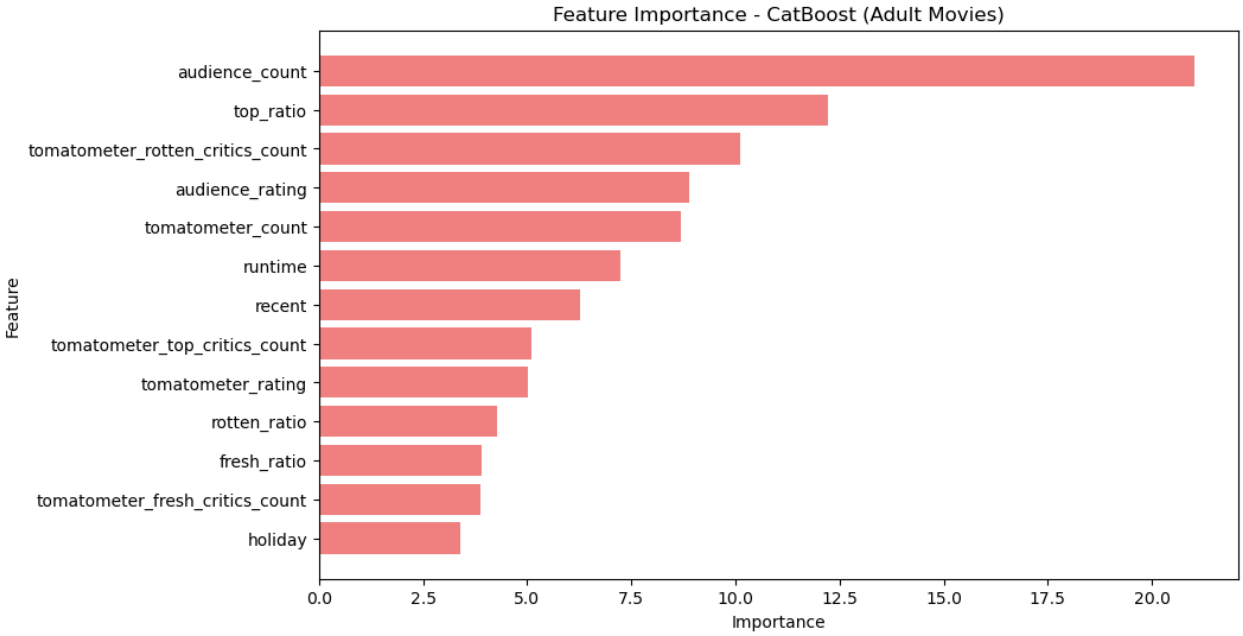



*Figure 15 - Feature Importance for Adult Movies*

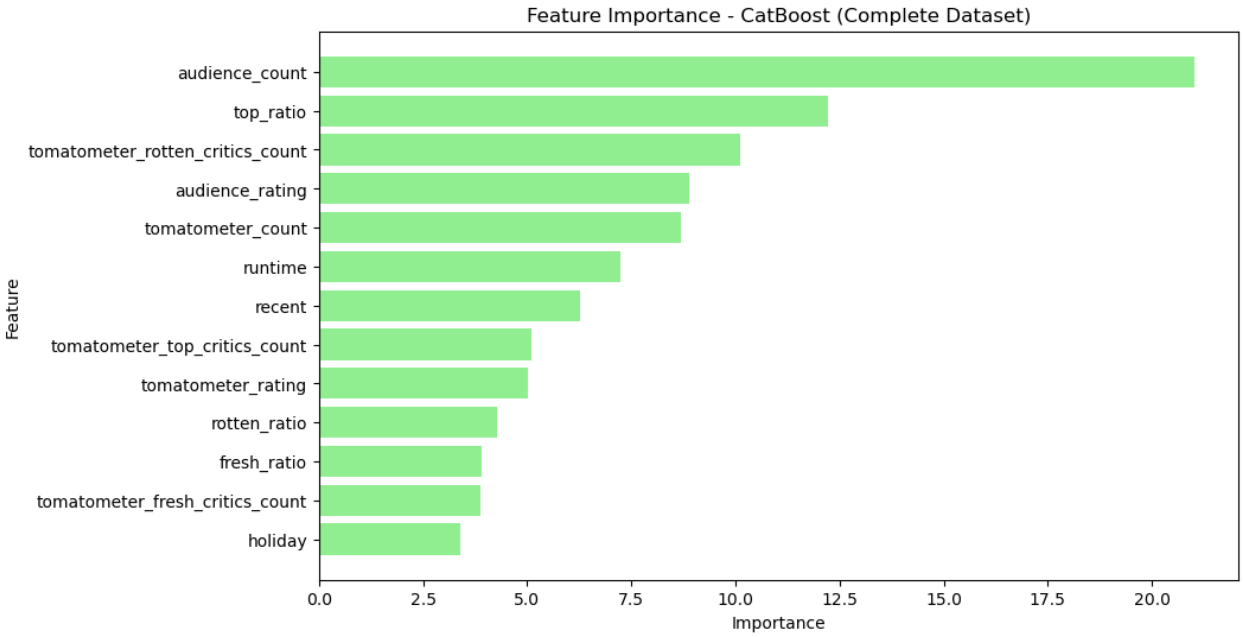



*Figure 16 - Feature Importance for All Movies*

*Table 17 - SHAP Analysis for Adult Movies*

**Adult Movies**

| Feature | SHAP value |
|---|---|
| top_ratio | 1.57E+07 |
| audience_count | 1.47E+07 |
| recent | 5.74E+06 |
| tomatometer_count | 3.73E+06 |
| holiday | 3.54E+06 |
| tomatometer_top_critics_count | 2.91E+06 |
| rotten_ratio | 2.76E+06 |
| tomatometer_rating | 1.46E+06 |
| fresh_ratio | 1.37E+06 |
| tomatometer_rotten_critics_count | 1.33E+06 |
| audience_rating | 1.16E+06 |
| tomatometer_fresh_critics_count | 4.44E+05 |
| runtime | 4.14E+05 |

*Table 18 - SHAP Analysis for Non-Adult Movies*

| Non-Adult Movies | |
|---|---|
| Feature | SHAP value |
| tomatometer_count | 4.07E+07 |
| audience_count | 3.72E+07 |
| recent | 2.95E+07 |
| top_ratio | 1.90E+07 |
| runtime | 1.20E+07 |
| tomatometer_top_critics_count | 6.93E+06 |
| audience_rating | 6.30E+06 |
| fresh_ratio | 4.12E+06 |
| tomatometer_rotten_critics_count | 1.25E+06 |
| rotten_ratio | 1.12E+06 |
| tomatometer_fresh_critics_count | 1.05E+06 |
| tomatometer_rating | 8.77E+05 |
| holiday | 7.45E+05 |

*Table 19 - SHAP Analysis for All Movies*

| Complete (All Movies) | |
|---|---|
| Feature | SHAP Value |
| audience_count | 7.72E+07 |
| recent | 2.64E+07 |
| top_ratio | 2.49E+07 |
| tomatometer_count | 2.22E+07 |
| audience_rating | 1.24E+07 |
| tomatometer_top_critics_count | 1.18E+07 |
| runtime | 1.07E+07 |
| tomatometer_rotten_critics_count | 7.50E+06 |
| fresh_ratio | 7.03E+06 |
| holiday | 6.82E+06 |
| tomatometer_fresh_critics_count | 5.36E+06 |
| rotten_ratio | 4.50E+06 |
| tomatometer_rating | 4.27E+06 |

## 13.3. Quantitative Analysis of Factors Driving Movie Sequel Success: A Predictive Modeling Approach

### 13.3.1. Model Results

#### 13.3.1.1. Model I: Absolute Success

| Rank | Model | WMAPE | MAPE | MAE (in $M) |
|---|---|---|---|---|
| 1 | CatBoost | 0.618 | 43.014 | 89.314 |
| 2 | Random Forest | 0.637 | 55.963 | 91.581 |
| 3 | XGBoost | 0.668 | 29.378 | 96.322 |
| 4 | Gradient Boosting | 0.672 | 75.094 | 97.072 |
| 5 | Linear Regression | 0.830 | 138.417 | 118.228 |

*Table 20: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

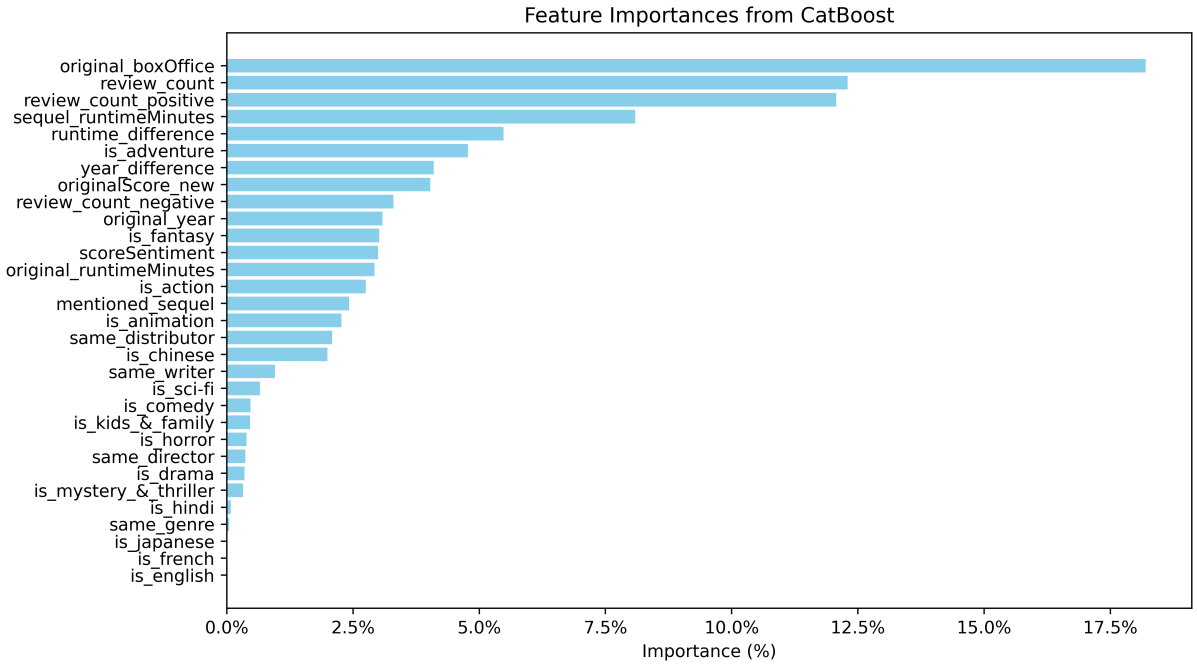
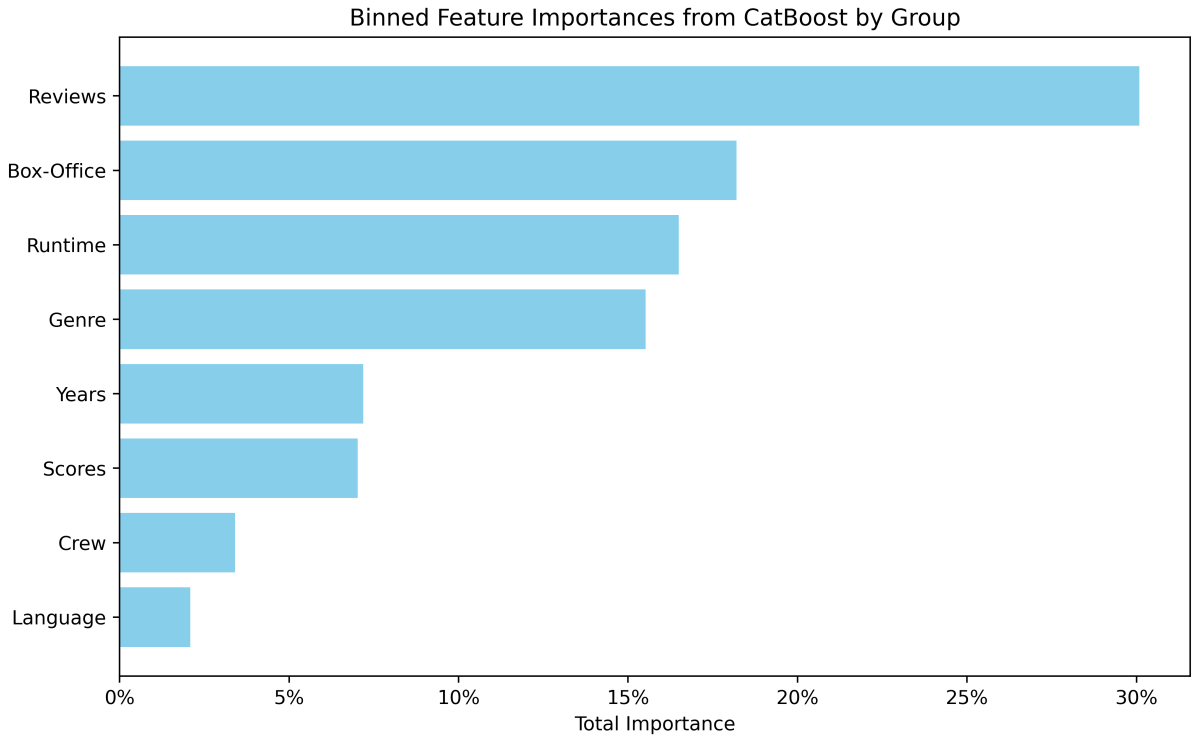



*Figure 17: Feature Importance Scores CatBoost*




Language: is_hindi, is_english, is_japanese, is_chinese, is_french
Genre: is_action, same_genre, is_drama, is_comedy, is_mystery_&_thriller, is_adventure, is_horror, is_kids_&_family is_fantasy, is_sci-fi, is_animation
Crew: same_writer, same_director, same_distributor
Runtime: sequel_runtimeMinutes,'runtime_difference, original_runtimeMinutes
Reviews: review_count, review_count_negative, review_count_positive, mentioned_sequel
Scores: originalScore_new, scoreSentiment
Box-Office: original_boxOffic
Years: year_difference, original_year

*Figure 18: Binned Feature Importance Scores CatBoost*

### 13.3.1.2. Model II: Relative Success

| Rank | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 1 | CatBoost | 0.777 | 0.768 | 0.777 | 0.771 | 0.863 |
| 2 | Gradient Boosting | 0.775 | 0.769 | 0.773 | 0.769 | 0.846 |
| 3 | Logistic Regression | 0.771 | 0.760 | 0.773 | 0.764 | 0.844 |
| 4 | XGBoost | 0.763 | 0.765 | 0.742 | 0.752 | 0.838 |
| 5 | Random Forest | 0.741 | 0.741 | 0.718 | 0.728 | 0.846 |

*Table 21: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

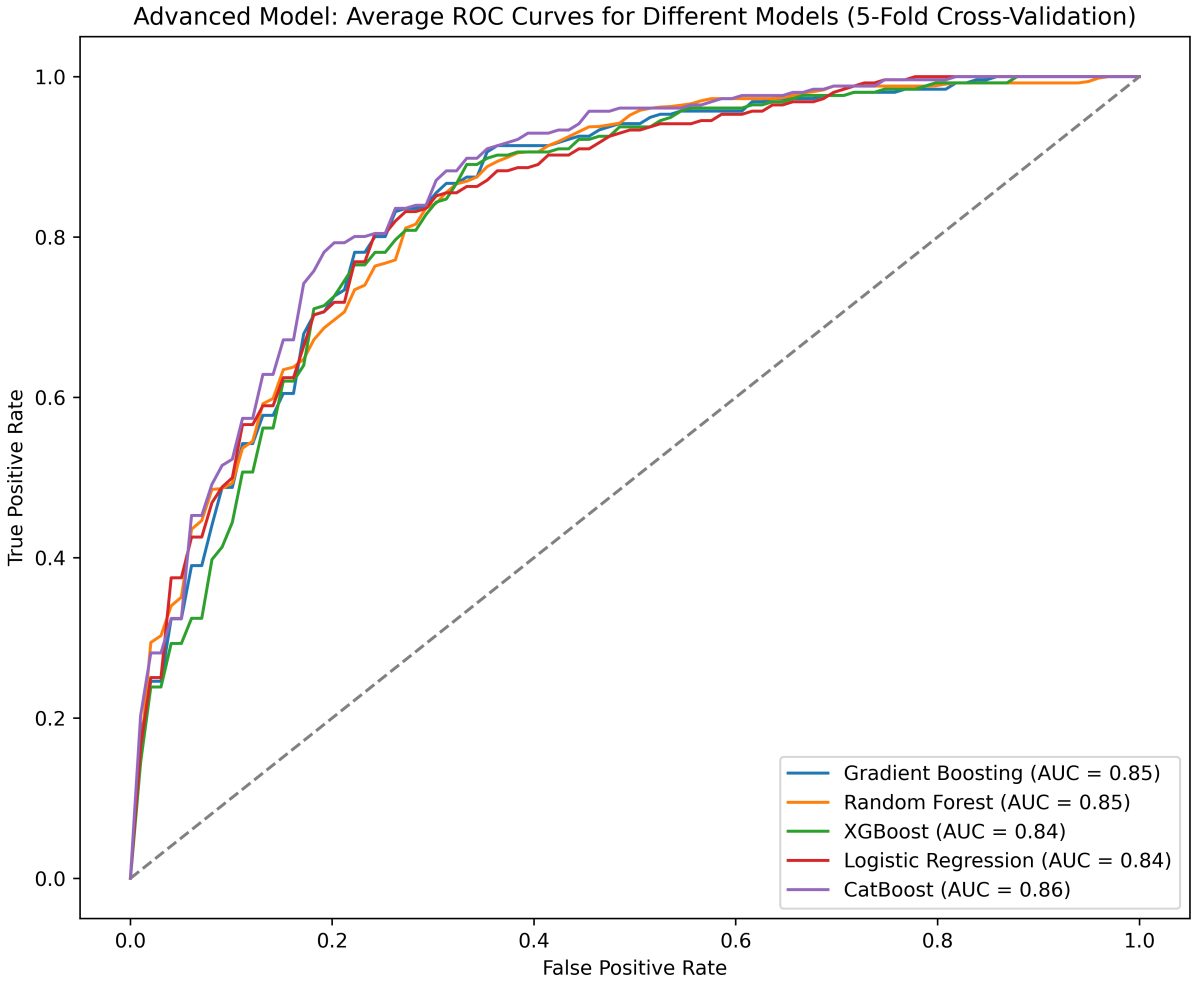



*Figure 19: Comparison of ROC Curves (Average Metrics across 5-Fold Cross-Validation)*

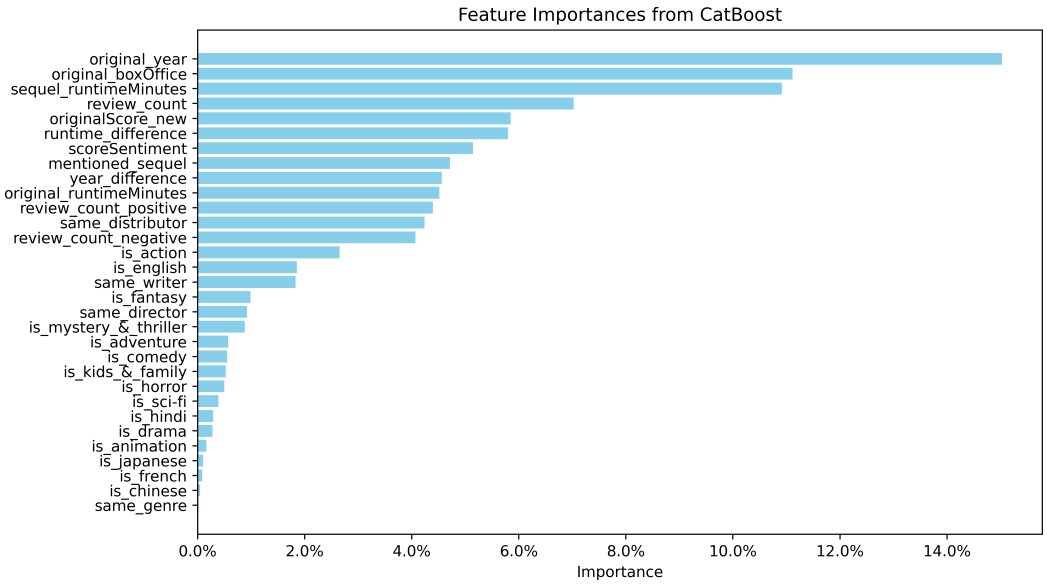
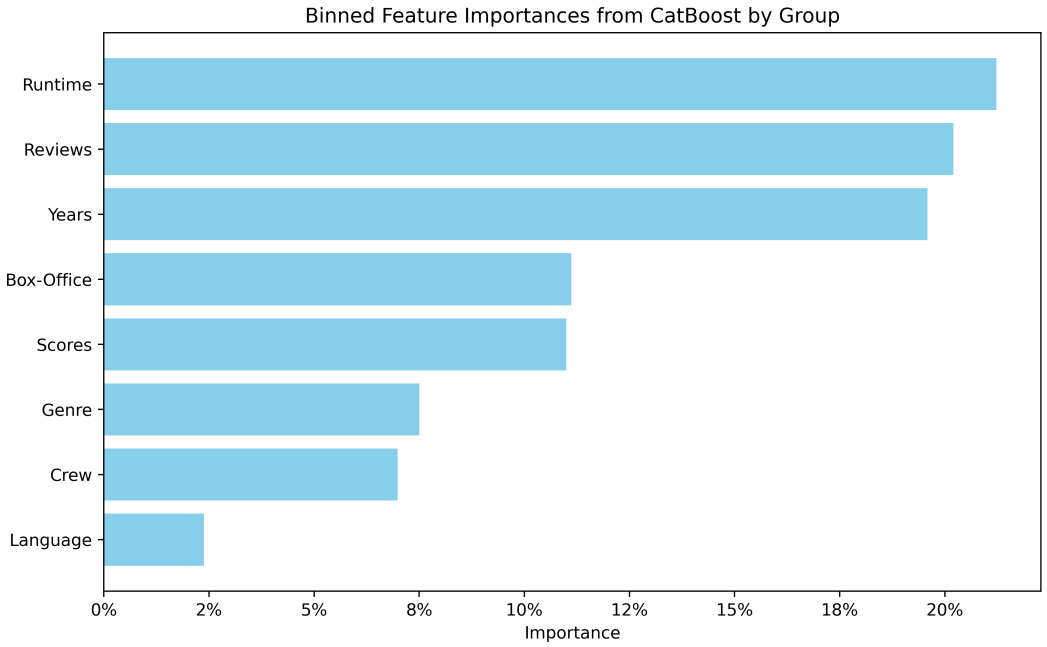



*Figure 20: Feature Importance Scores CatBoost*




Language: is_hindi, is_english, is_japanese, is_chinese, is_french
Genre: is_action, same_genre, is_drama, is_comedy, is_mystery_&_thriller, is_adventure, is_horror, is_kids_&_family is_fantasy, is_sci-fi, is_animation
Crew: same_writer, same_director, same_distributor
Runtime: sequel_runtimeMinutes,'runtime_difference, original_runtimeMinutes
Reviews: review_count, review_count_negative, review_count_positive, mentioned_sequel
Scores: originalScore_new, scoreSentiment
Box-Office: original_boxOffic
Years: year_difference, original_year

*Figure 21: Binned Feature Importance Scores CatBoost*

### 13.3.2. Ablation Study

#### 13.3.2.1. Model I: Absolute Success

| Rank | Model | WMAPE | MAPE | MAE (in $M) |
|---|---|---|---|---|
| 1 | Random Forest | 0.652 | 52.624 | 93.785 |
| 2 | CatBoost | 0.654 | 43.952 | 94.650 |
| 3 | Gradient Boosting | 0.676 | 60.024 | 97.631 |
| 4 | XGBoost | 0.680 | 34.529 | 98.585 |
| 5 | Linear Regression | 0.847 | 146.142 | 120.876 |

*Table 22: Without Sequel-Runtime: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

| Rank | Model | WMAPE | MAPE | MAE (in $M) |
|---|---|---|---|---|
| 1 | CatBoost | 0.608 | 46.560 | 87.994 |
| 2 | Random Forest | 0.636 | 53.576 | 91.635 |
| 3 | XGBoost | 0.666 | 28.826 | 95.486 |
| 4 | Gradient Boosting | 0.684 | 70.310 | 98.716 |
| 5 | Linear Regression | 0.826 | 132.688 | 117.482 |

*Table 23: Without Year-Difference: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

| Rank | Model | WMAPE | MAPE | MAE (in $M) |
|---|---|---|---|---|
| 1 | CatBoost | 0.616 | 44.309 | 88.984 |
| 2 | Random Forest | 0.621 | 60.296 | 89.737 |
| 3 | Gradient Boosting | 0.663 | 66.631 | 96.019 |
| 4 | XGBoost | 0.678 | 35.355 | 98.178 |
| 5 | Linear Regression | 0.841 | 142.191 | 119.274 |

*Table 24: Without Review-Sentiment: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

| Rank | Model | WMAPE | MAPE | MAE (in $M) |
|---|---|---|---|---|
| 1 | CatBoost | 0.654 | 48.452 | 94.796 |
| 2 | Random Forest | 0.703 | 109.129 | 100.009 |
| 3 | Gradient Boosting | 0.714 | 94.823 | 103.379 |
| 4 | XGBoost | 0.754 | 69.329 | 107.159 |
| 5 | Linear Regression | 0.870 | 184.585 | 123.417 |

*Table 25: Without Review-Volume: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

#### 13.3.2.2. Model II: Relative Success

| Rank | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 1 | Random Forest | 0.764 | 0.764 | 0.746 | 0.753 | 0.838 |
| 2 | Gradient Boost | 0.758 | 0.767 | 0.722 | 0.743 | 0.826 |
| 3 | Logistic Regression | 0.756 | 0.744 | 0.757 | 0.749 | 0.825 |
| 4 | CatBoost | 0.754 | 0.750 | 0.742 | 0.745 | 0.846 |
| 5 | XGBoost | 0.733 | 0.737 | 0.699 | 0.717 | 0.796 |

*Table 26: Without Sequel-Runtime: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

| Rank | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 1 | CatBoost | 0.777 | 0.774 | 0.769 | 0.770 | 0.853 |
| 2 | Logistic Regression | 0.769 | 0.755 | 0.777 | 0.763 | 0.844 |
| 3 | Gradient Boost | 0.766 | 0.761 | 0.762 | 0.759 | 0.839 |
| 4 | XGBoost | 0.762 | 0.755 | 0.753 | 0.753 | 0.823 |
| 5 | Random Forest | 0.752 | 0.741 | 0.757 | 0.747 | 0.846 |

*Table 27: Without Year-Difference: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

| Rank | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 1 | CatBoost | 0.779 | 0.768 | 0.781 | 0.773 | 0.865 |
| 2 | Logistic Regression | 0.771 | 0.767 | 0.757 | 0.760 | 0.842 |
| 3 | XGBoost | 0.771 | 0.759 | 0.773 | 0.765 | 0.848 |
| 4 | Random Forest | 0.769 | 0.757 | 0.773 | 0.764 | 0.858 |
| 5 | Gradient Boosting | 0.767 | 0.763 | 0.761 | 0.761 | 0.850 |

*Table 28: Without Review-Sentiment: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

| Rank | Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|---|
| 1 | CatBoost | 0.771 | 0.758 | 0.781 | 0.767 | 0.862 |
| 2 | Logistic Regression | 0.769 | 0.765 | 0.758 | 0.759 | 0.838 |
| 3 | Random Forest | 0.765 | 0.753 | 0.777 | 0.763 | 0.852 |
| 4 | XGBoost | 0.760 | 0.747 | 0.765 | 0.755 | 0.831 |
| 5 | Gradient Boosting | 0.758 | 0.750 | 0.754 | 0.751 | 0.845 |

*Table 29: Without Review-Volume: Model Comparison (Average Metrics across 5-Fold Cross-Validation)*

### 13.3.3. Hypothesis Testing

#### 13.3.3.1. Model I: Absolute Success

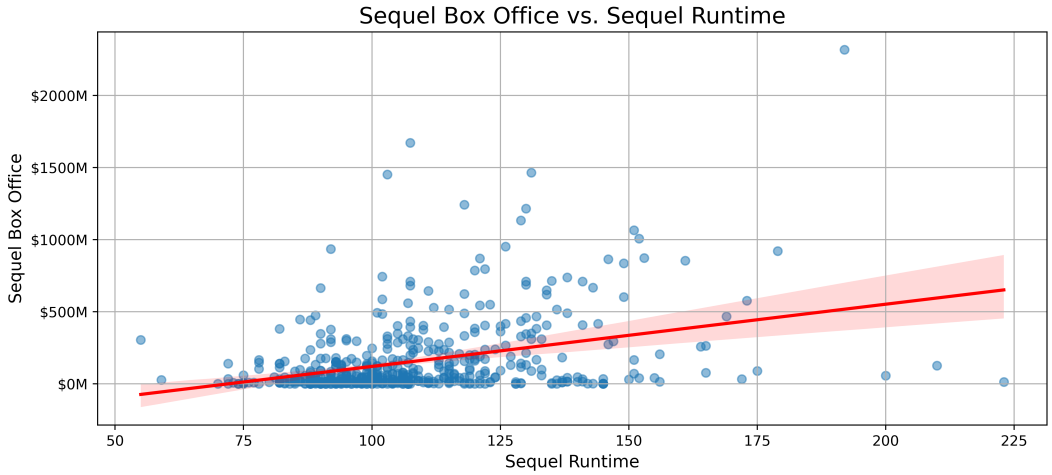



*Figure 22: Linear Regression Plot with 95% Confidence Interval for Sequel Runtime and Absolute Sequel Box Office Success*

| | Coefficient | Std error | t | P>\|t\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | -310.7 M | 53.0 M | -5.858 | 0.000 | -415.0 M | -206.0 M |
| Sequel Runtime | 4.3 M | 484,000 | 8.913 | 0.000 | 3.3 M | 5.3 M |

Spearman Correlation: 0.340
P-Value (Spearman): 0.000
$R^2$: 0.131

*Table 30: Linear Regression and Spearman Correlation Results for Sequel Runtime and Absolute Sequel Box Office Success*

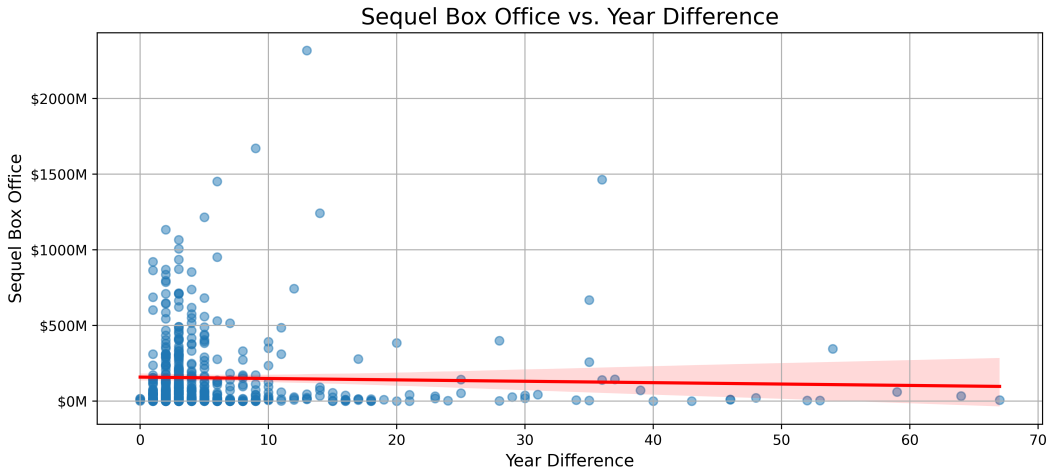



*Figure 23: Linear Regression Plot with 95% Confidence Interval for Year Difference and Absolute Sequel Box Office Success*

| | Coefficient | Std error | t | P>\|t\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | 158.9 M | 13.6 M | 11.645 | 0.000 | 132.0 M | 186.0 M |
| Year Difference | -918,400 | 1.2 M | -0.756 | 0.450 | -3.3 M | 1.5 M |

Spearman Correlation: -0.127
P-Value (Spearman): 0.004
$R^2$: 0.001

*Table 31: Linear Regression and Spearman Correlation Results for Year Difference and Absolute Sequel Box Office Success*

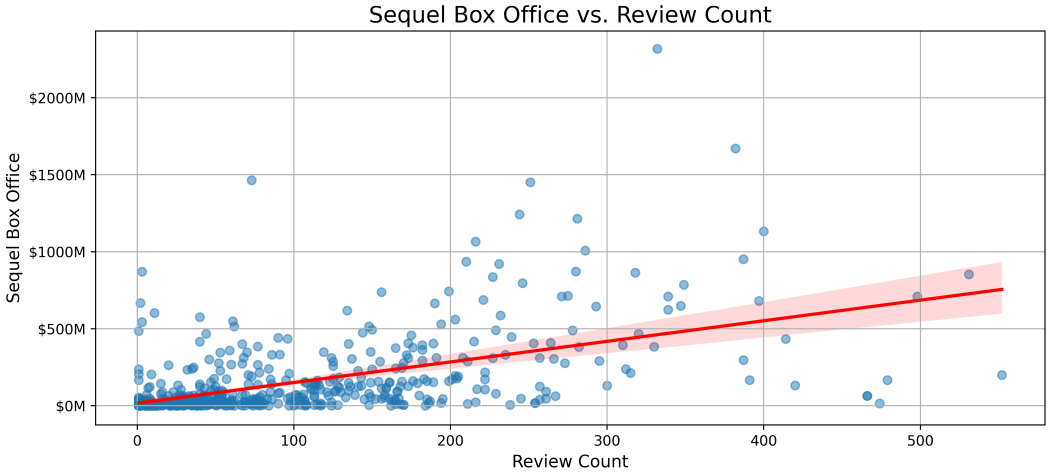



*Figure 24: Linear Regression Plot with 95% Confidence Interval for Review Count and Absolute Sequel Box Office Success*

| | Coefficient | Std error | t | P>\|t\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | 17.4 M | 13.4 M | 1.292 | 0.197 | -9.0 M | 43.7 M |
| Review Count | 1.3 M | 94,000 | 14.229 | 0.000 | 1.2 M | 1.5 M |

Spearman Correlation: 0.580
P-Value (Spearman): 0.000
$R^2$: 0.278

*Table 32: Linear Regression and Spearman Correlation Results for Review Count and Absolute Sequel Box Office Success*

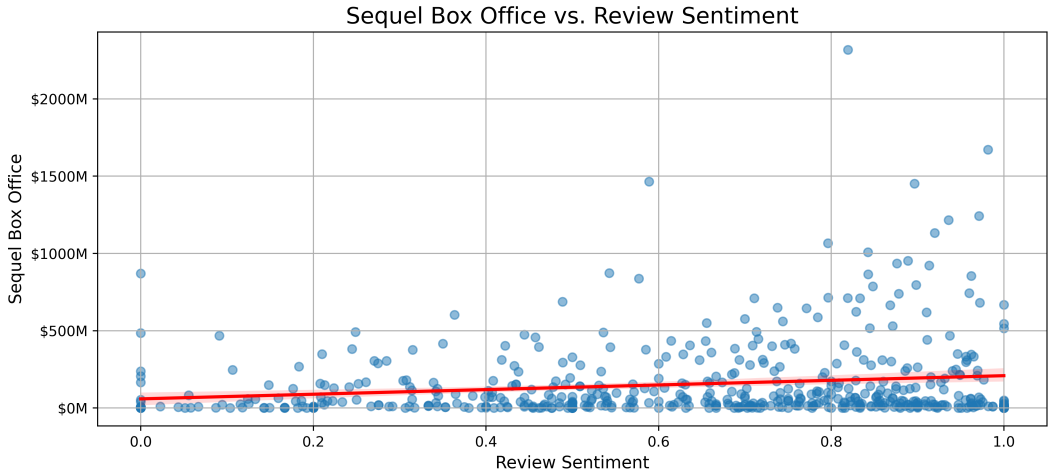



*Figure 25: Linear Regression Plot with 95% Confidence Interval for Review Sentiment and Sequel Box Office*

| | Coefficient | Std error | t | P>\|t\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | 58.7 M | 27.4 M | 2.142 | 0.033 | 4.9 M | 113.0 M |
| Review Sentiment | 150.1 M | 40.0 M | 3.750 | 0.000 | 71.4 M | 229.0 M |

Spearman Correlation: 0.127
P-Value (Spearman): 0.004
$R^2$: 0.026

*Table 33: Linear Regression and Spearman Correlation Results for Review Sentiment and Absolute Sequel Box Office Success*

### 13.3.3.2. Model II: Relative Success

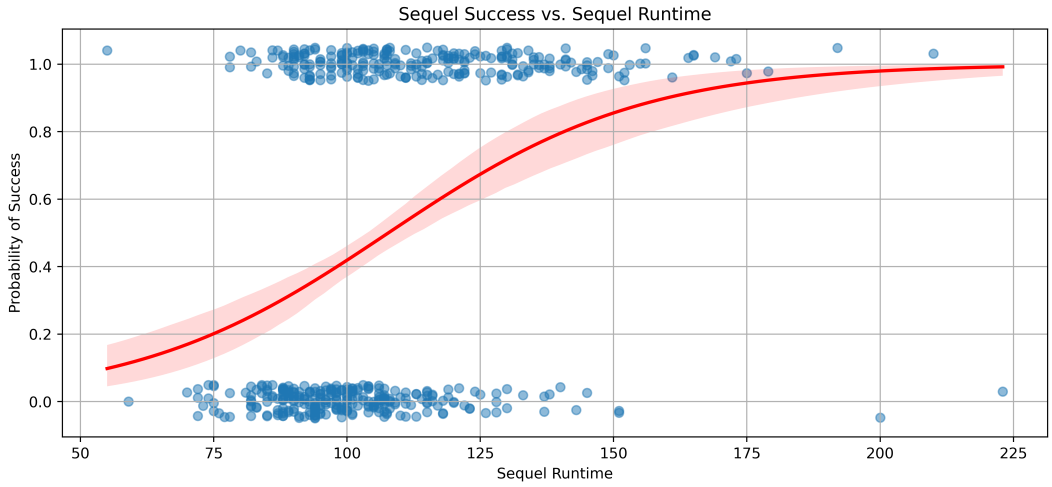



*Figure 26: Logistic Regression Plot with 95% Confidence Interval for Sequel Runtime and Relative Sequel Box Office Success*

| | Coefficient | Std error | z | P>\|z\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | -4.5914 | 0.595 | -7.718 | 0.000 | -5.757 | -3.425 |
| Sequel Runtime | 0.0425 | 0.006 | 7.618 | 0.000 | 0.032 | 0.053 |

Pseudo $R^2$ (McFadden): 0.105

*Table 34: Logistic Regression Results for Sequel Runtime and Relative Box Office Success*

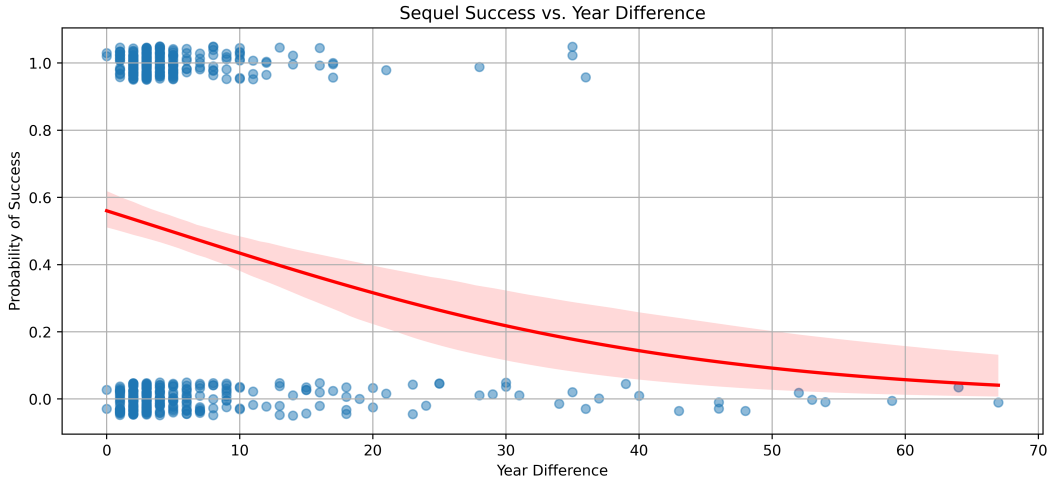



*Figure 27: Logistic Regression Plot with 95% Confidence Interval for Year Difference and Relative Sequel Box Office Success*

| | Coefficient | Std error | z | P>\|z\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | 0.2489 | 0.115 | 2.162 | 0.031 | 0.023 | 0.475 |
| Year Difference | -0.0518 | 0.014 | -3.829 | 0.000 | 0.078 | -0.025 |

Pseudo $R^2$ (McFadden): 0.028

*Table 35: Logistic Regression Results for Year Difference and Relative Box Office Success*

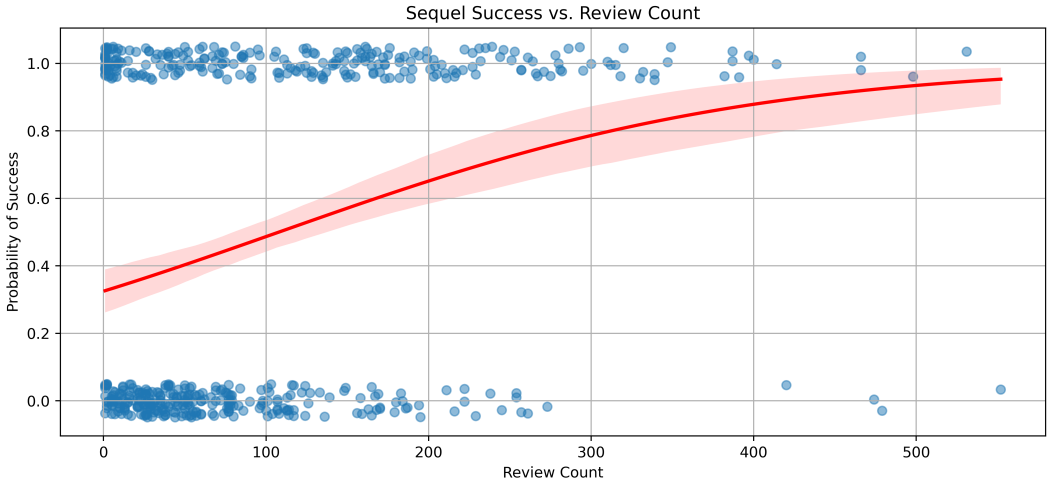



*Figure 28: Logistic Regression Plot with 95% Confidence Interval for Review Count and Relative Sequel Box Office Success*

| | Coefficient | Std error | z | P>\|z\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | -0.7342 | 0.134 | -5.461 | 0.000 | -0.998 | -0.471 |
| Review Count | 0.0068 | 0.001 | 6.361 | 0.000 | 0.005 | 0.009 |

Pseudo $R^2$ (McFadden): 0.068

*Table 36: Logistic Regression Results for Review Count and Relative Box Office Success*

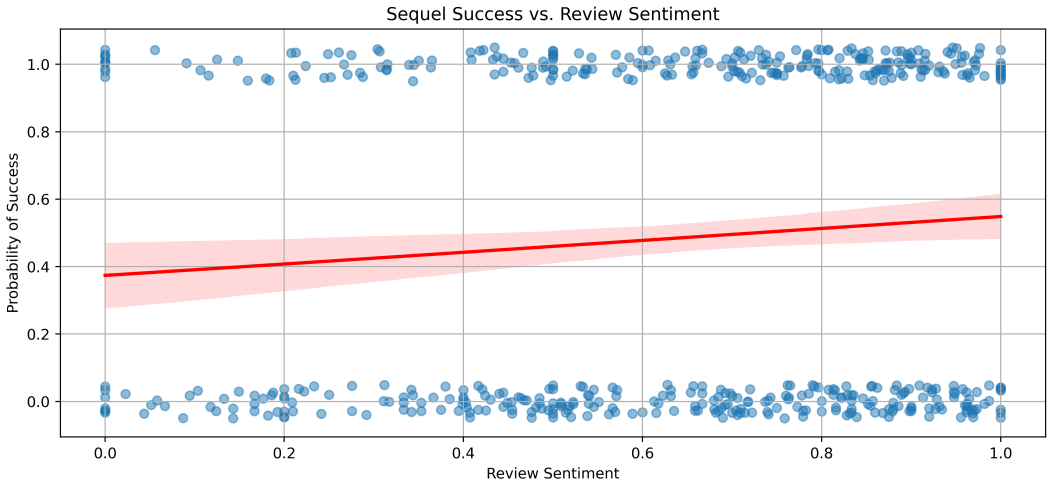



*Figure 29:Logistic Regression Plot with 95% Confidence Interval for Review Sentiment and Relative Sequel Box Office Success*

| | Coefficient | Std error | z | P>\|z\| | [0.025 | 0,975] |
|---|---|---|---|---|---|---|
| Constant | -0.5053 | 0.221 | -2.289 | 0.022 | -0.938 | -0.073 |
| Review Sentiment | 0.7008 | 0.321 | 2.182 | 0.029 | 0.071 | 1.330 |

Pseudo $R^2$ (McFadden): 0.007

*Table 37: Logistic Regression Results for Review Sentiment and Relative Box Office Success*

## 13.4. Temporal Modeling for Movie Success Prediction: A Comparative Study on the Applicability of different Deep Learning Models in Entertainment Analytics

*Equation 1: Movies Return-on-Invest*

$$RoI_i = \frac{Revenue_i - Budget_i}{Budget_i}$$

Where $i \in M$: $M\ is\ the\ set\ of\ movies\{i, \dots, n\}$.

$Revenue_i$: Total revenue generated by the movie $i$ over its lifetime.

$Budget_i$: The production cost or budget for movie $i$.

*Table 38: Summary Statistic for Time Series Daatsets before and after Preprocessing.*

| **Dataset**<br>*Source* | **Raw TS Dimensionality**<br>*(num rows/num cols)* | **Aggregated + Filtered TS Dimensionality**<br>*(num rows/num cols)* | **Number Movies**<br>*#* | **Seq. Length**<br>*Median* | **Target Distribution**<br>*(hit; avg; flop)* |
|---|---|---|---|---|---|
| **Rotten** | (1009126, 7) | (16687, 15) | 450 | 34 | (0.47; 0.29; 0.24) |
| **IMDB** | (26024289, 6) | (96971, 12) | 1365 | 57 | (0.44; 0.35; 0.21) |
| **Twitter** | (921398, 6) | (26848, 11) | 393 | 59 | (0.67; 0.28; 0.05) |
| **BoxOffice** | (309561, 4) | (309561, 4) | 5111 | 60 | - |

*Table 39: Features included in each of the three datasets.*

| **Rotten Tomatoes** | **IMDB** | **Twitter** |
|---|---|---|
| *Revenue* | *Revenue* | *Revenue* |
| *Budget* | *Budget* | *Budget* |
| *MovieID* | *MovieID* | *MovieID* |
| *Date* | *Date* | *Date* |
| *rating_score_encoded_mean* | *rating_score_encoded_mean* | *rating_score_encoded_mean* |
| *rating_score_encoded_var* | *rating_score_encoded_var* | *rating_score_encoded_min* |
| *rating_score_encoded_min* | *rating_score_encoded_min* | *rating_score_encoded_max* |
| *rating_score_encoded_max* | *rating_score_encoded_max* | *review_word_count_mean* |
| *review_word_count_mean* | *daily_number_of_ratings* | *daily_number_users* |
| *review_sentiment_polarity_mean* | *daily_box_office* | *daily_box_office* |
| *review_sentiment_polarity_min* | *daily_theaters_avg_gross* | *daily_theaters_avg_gross* |
| *review_sentiment_polarity_max* | *daily_num_theaters* | *daily_num_theaters* |
| *review_sentiment_polarity_var* | | |
| *daily_box_office* | | |
| *daily_theaters_avg_gross* | | |
| *daily_num_theaters* | | |

*Table 40: Mean AUC scores across models and Datasets*

| *Datasets/Models* | *TimeMIL* | *TodyNet* | *LSTM* | *Rank* TimeMIL |
|---|---|---|---|---|
| ***Rotten Tomatoes*** | 0.658596 | **0.749291** | 0.698412 | 3 |
| ***IMDB*** | 0.651499 | **0.763592** | 0.742273 | 3 |
| ***Twitter*** | **0.750273** | 0.750155 | 0.729751 | 1 |

*Table 41: Mean Precision scores across models and Datasets.*

| *Datasets/Models* | *TimeMIL* | *TodyNet* | *LSTM* | *Rank* (MIL/Tody/LSTM) |
|---|---|---|---|---|
| ***Rotten Tomatoes*** | 0.456646 | 0.547320 | 0.508332 | 3 |
| ***IMDB*** | 0.476054 | 0.585437 | 0.551034 | 3 |
| ***Twitter*** | 0.511275 | 0.500420 | 0.457373 | 1 |

*Table 42: Mean Recall scores across models and Datasets.*

| *Datasets/Models* | *TimeMIL* | *TodyNet* | *LSTM* | *Rank* (MIL/Tody/LSTM) |
|---|---|---|---|---|
| ***Rotten Tomatoes*** | 0.469753 | 0.554691 | 0.510556 | 3 |
| ***IMDB*** | 0.462770 | 0.569559 | 0.557993 | 3 |
| ***Twitter*** | 0.506270 | 0.534921 | 0.484365 | 3 |

*Table 43: Mean Accuracy scores across models and Datasets.*

| *Datasets/Models* | *TimeMIL* | *TodyNet* | *LSTM* | *Rank* (MIL/Tody/LSTM) |
|---|---|---|---|---|
| ***Rotten Tomatoes*** | 0.479412 | 0.524510 | 0.504902 | 3 |
| ***IMDB*** | 0.467317 | 0.559675 | 0.551545 | 3 |
| ***Twitter*** | 0.650847 | 0.566667 | 0.509040 | 1 |

*Table 44: Mean Loss scores across models and Datasets.*

| *Datasets/Models* | *TimeMIL* | *TodyNet* | *LSTM* | *Rank* (MIL/Tody/LSTM) |
|---|---|---|---|---|
| ***Rotten Tomatoes*** | 1.039801 | 0.843371 | 1.011542 | 3 |
| ***IMDB*** | 2.854342 | 0.896940 | 0.915404 | 3 |
| ***Twitter*** | 0.612022 | 0.461628 | 0.567411 | 3 |

*Table 45: Shapiro Wilk Test for pairwise differences distributions of F1 Score and AUROC*

| *MIL-Model* | *Metric* | *Dataset* | *Test statistic* | *p-value* | *Conclusion* |
|---|---|---|---|---|---|
| *TimeMIL vs. LSTM* | *F1* | *Rotten* | *0.9648* | *0.4081* | *Normal Distribution* |
| | | *IMDB* | *0.9608* | *0.3245* | *Normal Distribution* |
| | | *Twitter* | *0.981* | *0.8519* | *Normal Distribution* |
| | *AUROC* | *Rotten* | *0.9843* | *0.9244* | *Normal Distribution* |
| | | *IMDB* | *0.9654* | *0.4228* | *Normal Distribution* |
| | | *Twitter* | *0.9678* | *0.4799* | *Normal Distribution* |
| *TimeMIL vs. TodyNet* | *F1* | *Rotten* | *0.9607* | *0.3222* | *Normal Distribution* |
| | | *IMDB* | *0.9821* | *0.8772* | *Normal Distribution* |
| | | *Twitter* | *0.9658* | *0.4318* | *Normal Distribution* |
| | *AUROC* | *Rotten* | *0.9746* | *0.6696* | *Normal Distribution* |
| | | *IMDB* | *0.9775* | *0.7546* | *Normal Distribution* |
| | | *Twitter* | *0.983* | *0.8979* | *Normal Distribution* |

*Table 46: Mean attribution entropy for each model-dataset and attribution method.*

| ***Models*** | ***TimeMIL*** | | ***TodyNet*** | | ***LSTM*** | |
|---|---|---|---|---|---|---|
| ***Datasets\Entropy*** | ***IntGrad*** | ***DeepLift*** | ***IntGrad*** | ***DeepLift*** | ***IntGrad*** | ***DeepLift*** |
| ***Rotten Tomatoes*** | 2.344884 | 2.334574 | 2.207041 | 2.213954 | 2.118640 | 2.132280 |
| ***IMDB*** | 2.020013 | 1.920931 | 1.885764 | 1.872083 | 1.737910 | 1.725997 |
| ***Twitter*** | 1.848057 | 1.841669 | 1.693297 | 1.662531 | 1.677392 | 1.686569 |
| ***Cross-data-mean*** | 2.07098 | 2.032391 | 1.92870 | 1.916189 | 1.84464 | 1.84828 |

*Table 47: Results of Shapiro-Wilk test for the infidelity metric on IntGrad and Deeplift*

| **Dataset** | **Infidelity** | **Compared Models** | **Shapiro_stat** | **Shapiro_p-value** | **Shapiro_interpretation** |
|---|---|---|---|---|---|
| rotten | IntGrad | TimeMIL vs. LSTM | 0.820378672 | 0.006795486 | Non-Normal distribution |
| rotten | IntGrad | TimeMIL vs. TodyNet | 0.786436363 | 0.002475413 | Non-Normal distribution |
| rotten | DeepLift | TimeMIL vs. LSTM | 0.760036879 | 0.001180044 | Non-Normal distribution |
| rotten | DeepLift | TimeMIL vs. TodyNet | 0.7241101 | 0.000455507 | Non-Normal distribution |
| imdb | IntGrad | TimeMIL vs. LSTM | 0.825790202 | 0.008033388 | Non-Normal distribution |
| imdb | IntGrad | TimeMIL vs. TodyNet | 0.825815397 | 0.008039683 | Non-Normal distribution |
| imdb | DeepLift | TimeMIL vs. LSTM | 0.77623358 | 0.001850902 | Non-Normal distribution |
| imdb | DeepLift | TimeMIL vs. TodyNet | 0.776327965 | 0.001855838 | Non-Normal distribution |
| twitter | IntGrad | TimeMIL vs. LSTM | 0.659769637 | 9.52703E-05 | Non-Normal distribution |
| twitter | IntGrad | TimeMIL vs. TodyNet | 0.647835889 | 7.25173E-05 | Non-Normal distribution |
| twitter | DeepLift | TimeMIL vs. LSTM | 0.684341223 | 0.000169846 | Non-Normal distribution |
| twitter | DeepLift | TimeMIL vs. TodyNet | 0.672167277 | 0.000127182 | Non-Normal distribution |

*Table 48: Results of Wilcoxon signed-rank test for the infidelity metric on IntGrad and Deeplift*

| Dataset | Infidelity | Compared Models | Test_used | Test_stat | Test_p-value |
|---|---|---|---|---|---|
| rotten | IntGrad | TimeMIL vs. LSTM | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| rotten | IntGrad | TimeMIL vs. TodyNet | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| rotten | DeepLift | TimeMIL vs. LSTM | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| rotten | DeepLift | TimeMIL vs. TodyNet | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| imdb | IntGrad | TimeMIL vs. LSTM | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| imdb | IntGrad | TimeMIL vs. TodyNet | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| imdb | DeepLift | TimeMIL vs. LSTM | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| imdb | DeepLift | TimeMIL vs. TodyNet | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| twitter | IntGrad | TimeMIL vs. LSTM | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| twitter | IntGrad | TimeMIL vs. TodyNet | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| twitter | DeepLift | TimeMIL vs. LSTM | Wilcoxon signed-rank test | 0 | 6.10352E-05 |
| twitter | DeepLift | TimeMIL vs. TodyNet | Wilcoxon signed-rank test | 0 | 6.10352E-05 |

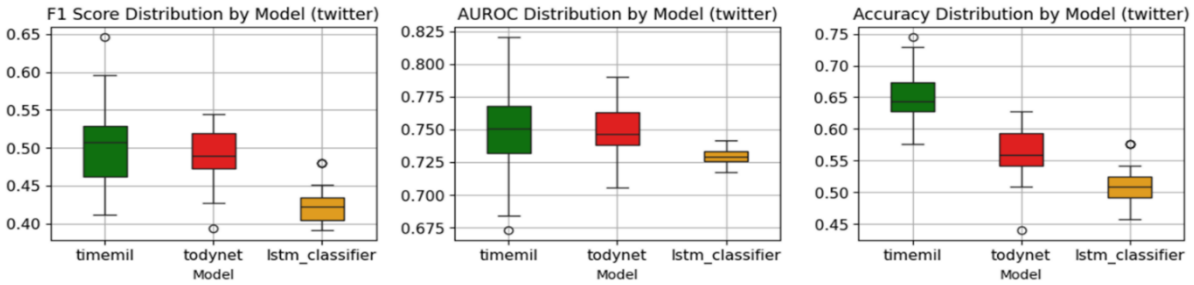



*Figure 30: Distribution of Performance Metrics for each model across experiments.*