NOVA IMS Information Management School

MGI

Master's Degree Program in
Information Management

MACHINE LEARNING ON BOX OFFICE PREDICTION AND THE IMPACT OF PROTAGONIST GENDER

José Ricardo Salas Pástor

Master Thesis

presented as partial requirement for obtaining the master's degree in information management.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Machine Learning on Box Office Prediction and the Impact of Protagonist Gender

by

José Ricardo Salas Pástor

Master Thesis presented as partial requirement for obtaining the master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervised by

Carina Albuquerque, PhD., NOVA Information Management School

December 2nd, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, December 2nd, 2024.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my family and friends for their constant belief in me. Your support, especially in moments when time felt fleeting, has been my driving force throughout this journey.

A special thanks to my supervisor, Professor Carina Albuquerque, for her guidance and insightful feedback. It was not an easy journey, and I am very grateful for the continuous support.

I am forever indebted to my parents for their unwavering encouragement and for always pushing me to complete what I start. Your love and support made this work possible.

I would also like to thank my professors at NOVA IMS, as well as the institution itself, for providing me with a rich learning environment and the resources necessary to complete this project.

To my friends and classmates at NOVA IMS— Alisson, Camila, Cristina, María José, and Maytte—thank you for being part of this incredible adventure. Moving to Lisbon together to start our master's was an experience I'll always cherish, and your companionship along the way made all the difference.

Lastly, to all those who contributed to my growth through discussions, challenges, and shared moments of motivation: thank you. Your support has been invaluable and will never be forgotten.

ABSTRACT

The impact of cinema on societies, cultures, and social change is undeniable, and box office prediction plays a crucial role in minimizing risks associated with future film investments. Accurate forecasting models enable better resource allocation, improving the chances of supporting successful films. Historically, it has been assumed that the gender of the lead significantly influences a film's box office success, with female-led films often perceived as less likely to perform well commercially. This perception contributes to the underrepresentation of women in lead roles and persistent gender pay disparities in the film industry. To address these challenges and improve forecasting accuracy, this study develops a machine learning model for accurate box office prediction, analyzing the impact of various features on success, using a dataset of films released between 2012-2019 and 2022-2023. Unlike previous research, this dataset includes post-pandemic years and adjusts for inflation, providing a more up-to-date and comprehensive analysis. After evaluating multiple models, Random Forest emerged as the most effective, achieving 48.82% accuracy for exact matches and 80.2% for matches within one revenue category. Additionally, a comprehensive analysis of gender's impact is conducted by combining machine learning techniques with statistical tests, moving beyond traditional regression models typically used in gender and box office research. The findings demonstrate that while the protagonist's gender is statistically significant, it has a minimal impact on box office performance compared to other key factors such as budget, director ranking, star power, and social media ratings. These results offer actionable insights for film industry stakeholders, suggesting that gender should not be a major consideration when greenlighting films. Instead, greater emphasis should be placed on more substantial predictors of success, thereby contributing to the ongoing conversation about reducing gender inequalities in film representation and pay.

KEYWORDS

Box office prediction; Film gender analysis; Gender and box office; Machine learning; Predictive model; Random Forest

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
APHR	Average Percent Hit Rate
BWT	Bechdel-Wallace Test
CPI	Consumer Price Index
CRISP-DM	Cross Industry Standard Process for Data Mining
EEMD	Ensemble Empirical Model Decomposition
IMDb	Internet Movie Database
KNN	K-nearest Neighbors
IP	Intellectual Property
MANCOVA	Multivariate Analysis of Covariance
MLP-NN	Multi-Layer Perceptron Neural Network
MPAA	Motion Picture Association of America
NN	Neural Networks
OLS	Ordinary Least Squares
p-value	Probability value
RF	Random Forest
RFE	Recursive Feature Elimination
ReLU	Rectified Linear Unit
SD	Standard Deviation
SFE	Sequential Forward Selection
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
WOM	Word-of-Mouth

1. INTRODUCTION

Since its inception in 1895, cinema has been a powerful cultural force, influencing societies and inspiring change. In recent years, however, the film industry has undergone a significant transformation, driven by technological advancements and changing audience preferences. In 2023, the global box office generated a substantial 33.9 billion USD (Mitchell, 2024), a figure that increases to 285.62 billion USD when considering home entertainment and streaming (The Business Research Company, 2024). These figures underscore the financial power of the film industry and highlight the crucial importance of informed investment decisions for studios.

The rise of streaming platforms has disrupted the traditional theatrical model, with studios increasingly opting for direct-to-streaming releases. The COVID-19 pandemic accelerated this shift, as theaters closed, and audiences embraced at-home entertainment (Kim, 2021). While streaming has gained prominence, box office success remains a critical indicator of a film's overall performance, often leading to significant ancillary revenue streams (Gunter, 2018). However, the industry's dynamic nature, compounded by the ongoing effects of the pandemic, requires data-driven approaches to mitigate uncertainties and optimize financial returns.

Gender representation in film has long been a topic of discussion, with women historically underrepresented both in leading roles and behind the camera. Actresses are often paid less than their male counterparts, and films with female leads are fewer in number compared to those starring male protagonists (Heywood et al., 2024). This disparity extends to box office performance, with films featuring female leads often being perceived as less commercially viable, partly due to industry biases and smaller production budgets (Neff et al., 2024). However, in recent years, there has been a gradual shift, with growing recognition of the importance of gender diversity in films (Damico, 2022). Industry trends now highlight that gender representation is not only an issue of equity but also a factor that can influence audience engagement and financial success (Evans et al., 2024). This dynamic change underscores the need to explore the evolving role of gender in determining box office outcomes.

Predictive analytics offers a promising solution to assist studios in forecasting box office performance and informing resource allocation. By developing accurate predictive models, studios can better identify potential hits and optimize their decision-making processes. This study seeks to advance predictive modeling in the film industry, with a particular focus on utilizing machine learning to predict the box office success of feature films.

To achieve this, the study develops a predictive model capable of accurately forecasting box office performance and investigates the influence of key factors, such as budget, genre, and cast, on box office outcomes. It also assesses the performance of the model using real-world film datasets to provide actionable insights that inform industry investment strategies. Additionally, this research examines the role of gender representation in predicting box office performance, using machine learning models and statistical tests to uncover patterns that may not be apparent through traditional linear approaches.

This research makes three significant contributions to the literature on box office prediction. First, it introduces a robust machine learning framework for predicting box office success, with Random Forest emerging as the top-performing model. Achieving an 80.20% accuracy within one classification tier, the Random Forest model outperforms previous methods with similar dimensionality and scope (Lu et al., 2024; Sharda et al., 2006). Second, while many existing studies focus on pre-pandemic data, this research fills a crucial gap by incorporating data from 2022 and 2023, reflecting current industry trends. Additionally, the study accounts for inflation, a factor often overlooked in recent literature (Souza et al., 2023). This adjustment enhances the relevance of the findings in today's financial and market context, providing insights into how inflation and evolving conditions impact box office outcomes. Finally, this study departs from traditional regression-based methods, which have often been used to examine gender's role in box office success while assuming linear relationships (Heywood et al., 2024; Smith et al., 2023; Treme et al., 2019). By employing machine learning techniques, this research uncovers complex, non-linear patterns, revealing that gender is among the least significant predictors when compared to factors such as budget, star power, and director reputation. This approach offers a more nuanced understanding of box office dynamics.

Following this introduction, Section 2 consists of literature review; Section 3 details the methodology, the data, and preprocessing steps; Section 4 discusses results and model comparisons; and Section 5 finalizes with conclusions and future research directions.

2. LITERATURE REVIEW

2.1. BOX OFFICE PREDICTION

Box office forecasting is a critical tool for decision-makers in the film industry, commonly explored in predictive analytics. Two primary approaches guide these predictions: regression models, which estimate specific box office revenue figures, and classification models, which categorize films into performance levels (e.g., high, medium, low). A variety of studies have employed different machine learning and statistical models to predict box office success.

The literature on box office forecasting began with **regression models**. A seminal study on box office forecasting showed that pre-release factors like production budget and director experience are stronger predictors than actor popularity and reviews (Litman, 1983). Later work applied Multiple Linear Regression to analyze the impact of actors and directors on box office revenue in China, finding that a director's reputation had a greater effect than A-list actors (Yongbin & Rongzhao, 2013). More recently, a regression-based model incorporating Ensemble Empirical Mode Decomposition (EEMD) was developed to account for industry volatility by separating data into trend and cyclical components, offering a more comprehensive view by integrating external factors such as public spending and audience interest (Ni & Li, 2023). In the context of sequels, a multi-factor model was used to predict the box office performance of sequel films, employing multiple linear regression for forecasting and logistic regression to address collinearity, based on a dataset of 471 films released in the US between 2000 and 2019 (Ge et al., 2023).

Turning to machine learning models, several studies have explored the use of techniques such as **Naïve Bayes** and **Support Vector Machine (SVM)** to predict box office earnings. For example, an analysis of nearly 5,000 films found Naïve Bayes to be more effective at identifying unprofitable films, with minimal impact from the number of actors involved, suggesting that factors like film content and marketing played a more significant role (Cocuzzo and Wu, 2013). Another study enhanced box office prediction by leveraging microblog data, focusing on both user-count-based and content-based features, and applying machine learning models such as SVM and neural networks, emphasizing the influence of social media word-of-mouth (Du et al., 2014).

Further studies have explored the use of **Neural Networks (NN)** for box office prediction. One study employed neural networks to predict box office success, achieving 75.2% accuracy in predicting a film's success within one category (Sharda & Delen, 2006). The model classified box office performance into nine categories and incorporated features such as star value, MPAA rating, genre, sequel status, competitive months, and number of screens, introducing the concepts of Average Percent Hit Rate (APHR) and 1-Away accuracy, which have since become common metrics in box office prediction. A further study simplified the prediction problem to two categories—Profit or Loss—and incorporated data from social media platforms like IMDb, Rotten Tomatoes, and Metacritic, achieving an accuracy rate of 88.8% (Rhee & Zulkernine, 2016).

Ensemble methods, such as Random Forest (RF), have also been widely applied in box office prediction. One RF study found that word-of-mouth from platforms like IMDb, Metacritic, and YouTube was a key factor influencing revenue, although their dataset was unbalanced, with

71% of films classified as "flops" (Lu et al., 2024). Similarly, another study tested RF, SVM, and NN in predicting box office success, experimenting with subsets of a 2,475-film dataset and ultimately using a 383-film dataset (Souza et al., 2023). RF outperformed the other models, particularly in predicting profitability over break-even success. A different approach developed a stacked model combining RF, extreme gradient boosting, light gradient boosting, and KNN, achieving 86.46% accuracy, with star influence identified as the most powerful predictor in the Chinese film market (Liao et al., 2023).

Despite the progress, existing literature often fails to adjust for inflation when assessing economic success, leading to misleading conclusions, especially when comparing newer films to older ones (Souza et al., 2023). Films released in more recent years tend to show inflated profits due to the increasing nominal value of revenues, which distorts assessments of their success. Only a few studies have included inflation adjustments in their analyses (Ge et al., 2023; Souza et al., 2023). This research aims to address this gap by incorporating inflation-adjusted figures for box office and budget data, ensuring a more accurate and comparable analysis across time periods.

Another significant gap in the existing literature is the exclusion of post-pandemic years in box office prediction models. One of the few studies to include post-pandemic years analyzed movies from 2012 to 2022 but did not adjust for inflation and did not address the issue of underperforming years caused by the pandemic (Zheng et al., 2023). Most studies published after 2020, acknowledging COVID-19's effect on ticket sales, limit their datasets to pre-2020 data, opting to exclude more recent years due to their outlier nature caused by the pandemic (Ge et al., 2023; Lu et al., 2024; Ni & Li, 2023; Yang et al., 2023). Recent studies have further explored the impact of COVID-19 on box office performance, with one examining 187 films released between January 2020 and November 2021, finding significant impacts from COVID-19 conditions at the time of release, alongside other factors like movie features and audience sentiment (Ni et al., 2022). Further research emphasized that the global COVID-19 outbreak marked a turning point for the movie industry, leading to major changes in trends and industry drivers (Ni & Li, 2023). These findings suggest that forecasting models should incorporate these changes when analyzing post-pandemic data. To bridge this gap, the present study includes data from 2022 and 2023, adjusting for inflation and excluding 2020 and 2021 as outliers, providing a more up-to-date and relevant understanding of the industry's trends.

2.2. GENDER REPRESENTATION AND ITS IMPACT ON BOX OFFICE

Gender representation in film has long been a significant area of research, with women historically underrepresented in leading roles and behind the camera. Actresses earn, on average, 56% less than male actors, or 25% less when accounting for factors such as prior success and genre, with an unexplained gap exceeding $2 million per film (Heywood et al., 2024; Izquierdo Sanchez & Navarro Paniagua, 2017). In 2022, the percentage of female leads and co-leads in top films peaked at 44%, but dropped sharply to 30% in 2023, reflecting a return to levels seen a decade earlier (Neff et al., 2024). This imbalance is not merely a representation issue but also affects the financial viability of films, as female-led films often struggle to secure production funds and distribution, partly due to industry biases and

stereotypes suggesting audiences are less likely to support films with women at the helm (Treme et al., 2019). However, recent successes suggest shifting dynamics in the financial potential of female-led films. Films such as *Wonder Woman* (2017) and *Captain Marvel* (2019) demonstrated the marketability of female protagonists, leading to increased interest in gender-diverse narratives (Weiß, 2024).

Several studies have employed quantitative methods to examine the impact of gender on box office performance. A multivariate regression analysis of 974 films from 2000 to 2009 found that films with significant female presence often underperformed at the box office, with the primary factor being smaller production budgets rather than audience interest, highlighting the role of industry investment practices (Lindler et al., 2015). Another study revealed that male celebrity exposure positively correlated with higher box office revenues, while female celebrity exposure had little to no effect, and older actresses struggled to secure lead roles, unlike their male counterparts (Treme & Craig, 2013). Further regression analysis showed that male stars contributed a 12% revenue premium to a film's box office, while female stars had no significant impact, reinforcing the financial dominance of male-driven narratives (Treme et al., 2019).

More recently, a study found that, after controlling for production costs and marketing spending, protagonist gender and race were not significant predictors of box office success, with financial backing, particularly for films with white male leads, being the key driver (Smith et al., 2023). In terms of the pay gap, another study revealed an unexplained earnings disparity of over $2 million per film for actresses, particularly in action movies, with smaller gaps in other genres and among older stars, indicating potential industry bias or consumer preferences (Heywood et al., 2024). Additionally, research on films passing the Bechdel-Wallace Test (which requires at least one conversation between two female characters that is not about men) found that such films generated higher international and total revenues, with post-#MeToo films performing better, particularly in international markets (Evans et al., 2024).

All these cited studies have used multivariate regression analysis to examine the influence of gender on box office outcomes, often focusing on factors like star power and production budgets (Heywood et al., 2024; Lindler et al., 2015; Smith et al., 2023; Treme et al., 2019). However, these models have limitations in capturing the complex and non-linear interactions among variables. By applying machine learning techniques, this study aims to better account for these intricate relationships and to provide a more robust understanding of how gender, alongside other factors, impacts a film’s financial success.

Recent box office successes challenge long-held assumptions about the commercial viability of female-led films. *Barbie*, the highest-grossing film of 2023, and *Inside Out 2*, the highest-grossing film of 2024 so far, demonstrate that female-led films can be both commercially successful and critically acclaimed (Box Office Mojo, 2023; The Numbers, 2024). These examples, combined with shifting audience preferences and evolving industry practices, suggest that gender representation is increasingly a crucial factor in a film's ability to attract diverse audiences and achieve financial success (Damico, 2022). Despite these advancements, systemic challenges persist. Women remain underrepresented in key behind-the-scenes roles, such as directors and producers, limiting the range of stories told on screen (Lauzen,

2023). Furthermore, the production budgets for female-led films are still often lower than their male counterparts, which can limit their commercial potential, even when the films are well-received by audiences (Smith et al., 2023). As the industry increasingly embraces inclusivity, addressing these barriers is essential not only for achieving equity but also for unlocking the full creative and financial potential of diverse storytelling.

3. METHODOLOGY

This study follows the Crisp-DM methodology (Fig. 3.1). As supported by previous studies, primary data sources include BoxOfficeMojo for gross revenue (Hunter et al., 2016; Nishijima et al., 2024), IMDb for film attributes (Cocuzzo et al., 2013; Ge et al., 2023; Hunter et al., 2016; Lu et al., 2024; Yang et al., 2023), and Metacritic for critic reviews (Rhee & Zulkernine, 2016). These sources provide extensive data on box office performance and various film attributes. The research employs a variety of machine learning models, aiming to discover the one that can predict box office success with the highest accuracy. These models include KNN algorithm, Artificial Neural Networks, Support Vector Machine, and Random Forest. A comprehensive impact analysis investigates the relevance of the protagonist's gender in predicting box office revenues, highlighting its significance within the context of revenue forecasting.

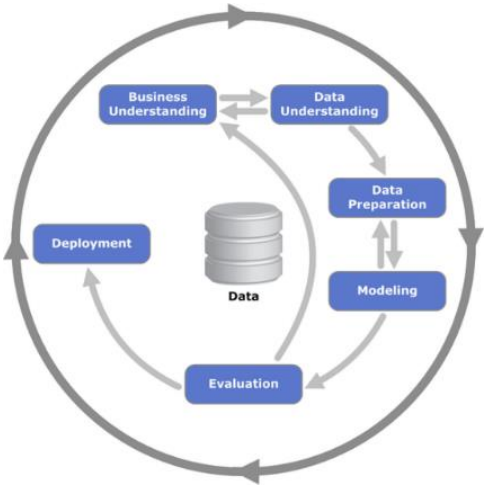

Figure 3.1. Phases of the CRISP-DM model (Chapman et. al., 2000)

3.1. DATA COLLECTION AND INTEGRATION

For this research project, data was collected regarding worldwide box office revenue for the top 200 highest grossing films of the following 10 years: 2012 to 2023, excluding pandemic years 2020 and 2021.

The exclusion of the pandemic years (2020 and 2021) is a necessary methodological choice due to the significant disruptions caused by COVID-19. The pandemic marked a turning point for the industry, leading to fewer fluctuations and a change in the key factors influencing box office performance (Ni & Li, 2023). As such, we focus on data from 2012 to 2019 and 2022 to 2023, consistent with other research (Ge et al., 2023; Lu et al., 2024), which also excluded pandemic years due to their outlier effects on box office trends.

This leads to a database of 2000 films. This data comes from the webpage Box Office Mojo, which is a website that tracks box office performance and other film industry data. As mentioned earlier, Box Office Mojo has been used in several academic studies for collecting

box office data on a large scale. A web scraping algorithm (written on Python with *Beautifulsoup*) has been created to automate the process of collecting data from multiple pages from Box Office Mojo and then consolidating it into a single data frame.

Afterwards, IMDb and Metacritic were used as additional data sources to get more information about each film. IMDb provided extensive information on film characteristics such as Genre, Distributor, Director, Runtime, and Cast. Additionally, IMDb and Metacritic also provide an aggregated rating from user and critics' data, respectively, that was integrated into our data to enrich it with social media and word-of-mouth (WOM) parameters. As previously noted, IMDb has been used extensively as a data source on many box-office prediction studies. The Python *Cinemagoer* library (previously known as *IMDbPY*) is used as a built-in method to complement the primary database and aggregate this additional information. *Cinemagoer* is a library that provides a programmatic interface to access IMDb's data (Alberani & Uyar, 2018). It is a reliable, well-maintained library that simplifies the process of retrieving data from IMDb.com in a structured format, allowing for efficient integration with Python code. Whenever *Metascore* (Metacritic score) was not available at IMdB, a similar web scrapping algorithm to the one used for Box Office Mojo was applied to obtain missing scores from the Metacritic website.

3.2. FEATURE DEFINITIONS

Target variable (Box office revenue): The target variable for this study is worldwide box office revenue. The first transformation that was implemented for this variable was to adjust these values for inflation to the current year, 2024, using the 2024 CPI. This procedure is not usually done in previous box office prediction studies, which can lead to misclassifying films as more profitable or with bigger revenues. For this research, forecasting box office is solved using classification algorithms by discretizing the target variable into 7 different classes as seen in Table 3.1.

Table 3.1. Discretization of target variable (box office revenue)

Class No.	1	2	3	4	5	6	7
Range	<100 M	>100 M <200 M	>200 M <300 M	>300 M <400 M	>400 M <600 M	>600 M <1 B	<1 B

M = Million USD dollars*
B= Billion USD dollars*
*After adjusting for inflation (2024 CPI)

Several studies prefer using discrete variables rather than continuous ones in developing prediction models because this simplifies and reduces data, making features easier to comprehend, use, and explain (Liu et al., 2002). Additionally, discretization enhances the speed and accuracy of many models (Han et al., 2023).

The independent features utilized in this model are outlined below and summarized in Table 3.2. The studies supporting the selection of these features are referenced in Table 3.3 in the following chapter.

1. Theaters: The number of theaters during the domestic run of the film. This data was scraped separately from Box Office Mojo since we are referring to US Domestic Box Office. Most of our films come from the US.

2. Distributor: The distributor variable was chosen from 3 categories: Disney, major film studio (Paramount, Universal Pictures, 20th Century Studios, Warner Bros., and Sony Pictures), and other distributors. This categorization was done considering the correlation between each studio and the box-office receipts and the number of films distributed by each studio. A recent study categorizes this feature in 2 classes: big distributor and others (Souza et al., 2023), while an even more recent approach categorizes it in 3: high, medium influence distributor, and others (Lu et al., 2024).

3. Runtime: Is the duration of the film in minutes.

4. Budget: The production budget is also a variable that appears regularly as an important feature to predict box office. This monetary value was also adjusted for inflation to the year 2024 to maintain consistency, using the 2024 CPI.

5. MPAA Rating: (G, PG, PG-13; R). Because in the database there is a very small number of G-rated films, PG and G films were conflated into a single category, due to their similar correlation with the target variable. NR rated films were not present in our database.

6. IMDb Rating: Feature that appears regularly in studies that analyze the effect of word-of-mouth on box office revenue. It is a user-generated rating that reflects the overall viewer satisfaction with a movie. Values range from 0-10.

7. Metascore: Another WOM variable, Metascore is the aggregated numerical rating of a film derived from critic reviews on Metacritic. Values range from 0-100.

8. Existing IP: This variable denotes whether the film comes from an existing IP (videogame adaptation, sequel, prequel, remake). There are several studies that have used this variable as an important feature for box office prediction. A database from Kaggle used by Ge et al. (2023) with information about sequels and prequels was used to determine this variable.

9. Release Month: Seasonality also plays a significant role in box-office success. Previous literature has used this variable as a predictive feature. May, June, July, November, and December are considered *high months* and the other months as *low months* (Sharda & Delen, 2006).

10. Director Ranking, 11. Star Ranking, and **12. Co-star Ranking:** Directors and actors were ranked in 3 categories depending on their total box office gross over the last 12 years. A high ranking was assigned if their aggregated box office totals over 1B USD (adjusted for inflation). Medium ranking was assigned for total gross between 500M and 1B USD, and low ranking for

less than 500M USD. For films with more than one director, the highest-ranking director was chosen to represent the variable.

13. Genre: The genre of a film is also a feature that has been used regularly for box office prediction. A film can have more than one genre, and this variable was treated using one-hot-encoding. War, sport, western, and documentary genres were dropped because of their low frequency, and Short for its common misclassification. Music and musical were conflated into one category.

14. Protagonist Gender: The gender of the film's leading actor was determined to see if it has relevancy in predicting box office success.

Table 3.2. Independent variables

Feature	Type	Values
Theaters	Continuous	Positive integer
Budget	Continuous	Float (USD adjusted for inflation 2024)
IMDb Rating	Continuous	Float (0-10)
Metascore	Continuous	Positive integer (0-100)
Runtime	Continuous	Positive integer
MPAA Rating	Ordinal	3 (0: G, PG; 0.5: PG-13; 1: R)
Distributor	Ordinal	3 (0: Other distributors; 0.5: Major film studio; 1: Disney)
Star Ranking	Ordinal	3 (0: Low; 0.5: Medium; 1: High)
Co-star Ranking	Ordinal	3 (0: Low; 0.5: Medium; 1: High)
Director Ranking	Ordinal	3 (0: Low; 0.5: Medium; 1: High)
Existing IP	Binary	2 (0: No; 1: Yes)
Release Month	Binary	2 (0: Jan, Feb, Mar, Apr, Aug, Oct; 1: May, Jun, Jul, Nov, Dec)
Protagonist Gender	Binary	2 (0: Male; 1: Female)
Genre	Categorical	17 (Action, Adventure, Sci-Fi, Fantasy, Comedy, Music, Biography, History, Animation, Family, Superhero, Crime, Drama, Mystery, Thriller, Horror, Romance)

All continuous variables were normalized using Min-Max scaling, transforming the data into a range between 0 and 1. Min-Max scaling ensures that all features are on the same scale, preventing features with larger numerical ranges from dominating those with smaller values (Zheng & Casari, 2018). Compared to other scaling techniques like Standardization, which centers data around a mean of 0, Min-Max scaling is preferred here because it retains the original distribution of the data and scales it to a fixed range, making it particularly well-suited for algorithms like NN, KNN, and SVM where feature magnitudes can disproportionately influence the model's performance (Han & Kamber, 2023). By scaling the variables in this way, each feature contributes equally to the model's learning process. The ordinal features were encoded as 0 (Low), 0.5 (Medium), and 1 (High), while binary features were encoded to 0 and 1. Lastly, genre was encoded using one-hot encoding. After these transformations, all features were scaled between 0 and 1 (Fig. 3.2), ensuring consistent input representations across all variable types and improving the performance of the classification algorithms employed.

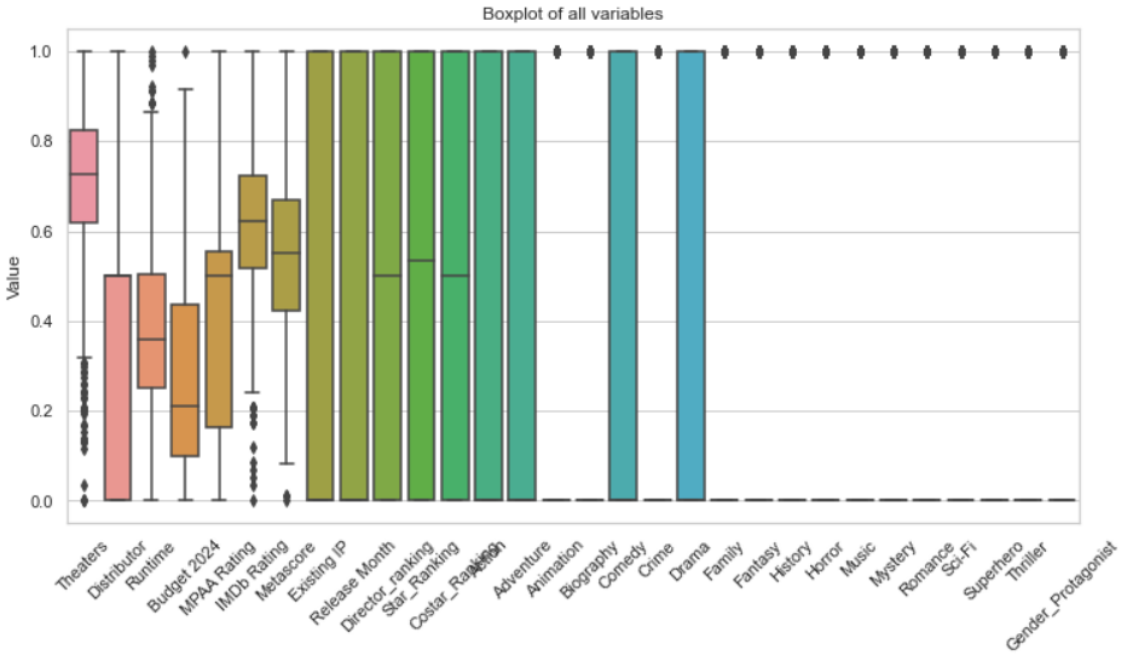

Figure 3.2. Box-plot distribution of normalized features

3.3. FEATURE SELECTION

The initial selection of features to be used by this research study was extensively supported by previous literature (Table 3.3). Discarded variables include production companies, year of release, and country because of the lack of literature supporting these features, extremely high dimensionality (production companies), dominance of majority class (country: US), and very low correlation with the target variable (all of them).

Table 3.3. Feature selection supported by literature review

Reference	MPAA Rating	Release Month	Distributor	Star	Director	Genre	Sequel	No. Theaters	Runtime	Budget	IMDb Rating	Metascore
Sharda et al. (2006)	■	■		■		■	■	■				
Rhee et al. (2016)	■	■		■	■		■			■	■	■
Lu et al. (2023)	■		■			■	■				■	
Ni et al. (2022)												
Yang et al. (2023)				■	■	■			■			
Cocuzzo, Wu (2013)	■	■		■	■	■			■	■		
Ge et al (2023)	■					■	■		■	■	■	
Ni, Li (2023)												
e Souza et al. (2023)								■				

For continuous features, dispersion charts can demonstrate their relationship with box office revenue before its discretization (Fig 3.3). Furthermore, Pearson, Spearman Rank coefficient and their corresponding p-values can provide us with more insights into these relationships (Table 3.4).

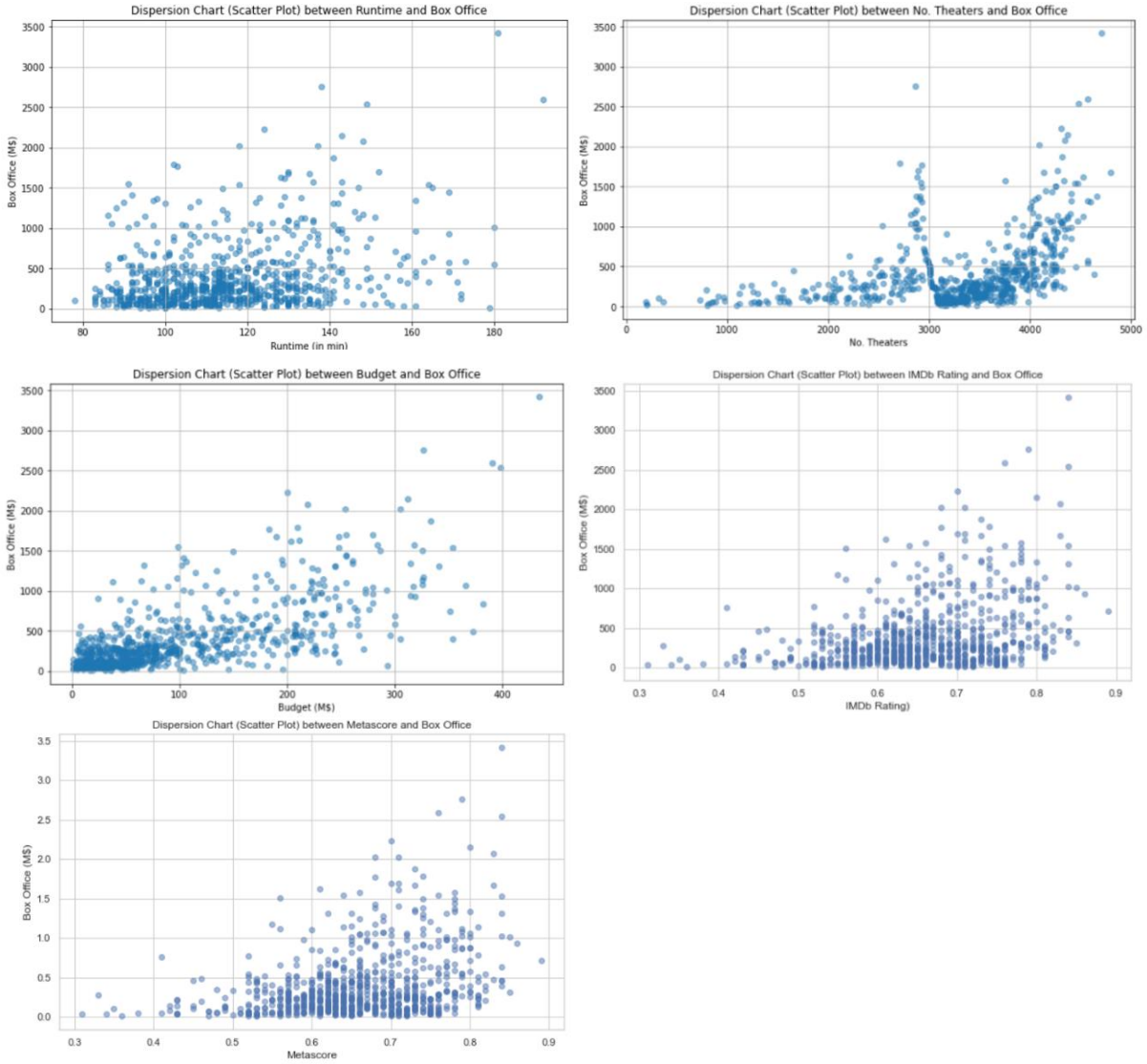

Figure 3.3. Dispersion charts between continuous features and box office revenue

As observed in Table 3.4, all p-values are extremely low, indicating that the correlations are statistically significant and not likely to have occurred by chance. The feature *Theaters* has statistically significant Pearson (58%) and Spearman (77%) coefficients. Since Spearman rank is higher, this signifies a monotonic relationship with the target, not a linear one. Conversely, *Budget* has similar Pearson and Spearman coefficients which implies a linear relationship to box office. Both relationships can be observed in their respective dispersion chart (Fig. 3). Therefore, all these variables were considered statistically related to the target, and thus selected for building the machine learning models.

Table 3.4. Correlation between continuous features and box office revenue

Correlation with Box Office	Pearson Coefficient	P-value (Pearson)	Spearman Coefficient	P-value (Spearman)
Runtime	32.5%	6.49E-21	27.8%	1.89E-15
Theaters	58.2%	1.57E-65	76.7%	1.14E-138
Budget	69.5%	9.04E-115	69.0%	1.03E-112
IMDb Rating	33.8%	1.57E-22	31.1%	3.58E-19
Metascore	22.9%	7.14E-11	20.1%	1.20E-08

Next, a chi-square test (Appendix A) was performed between categorical features and box office (discretized in 7 classes). Features like distributor, director ranking, star ranking, and costar ranking show very high Chi-Square statistics and extremely low p-values, indicating strong associations with box office. Features like family, fantasy, history, and music show high p-values and low Chi-Square statistics, suggesting they do not have a significant association with the target, but were kept anyways, since they belong to the original genre feature.

Furthermore, Pearson and Spearman correlation matrices were created to show the relationship between variables (Figure 3.4). Budget and director ranking have a strong association with the target, which can imply that they are good predictors. Metascore and IMDb Rating show a strong correlation between each other. However, since their Pearson coefficient is below 80% both variables were kept enriching the data with WOM information.

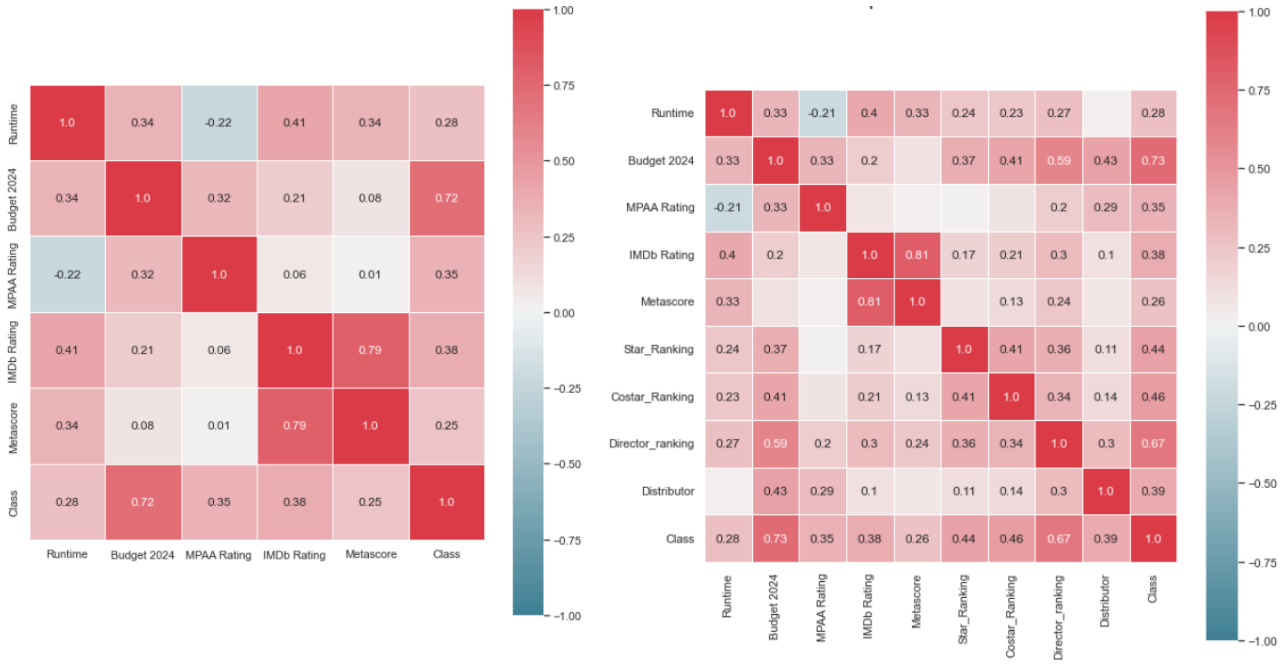

Figure 3.4. Pearson and Spearman correlation matrices

3.4. DATA REDUCTION

The original database consisted of 2000 films released from 2012 to 2019, and 2022 to 2023. Like previous research, an initial data cleansing was performed to decrease training time and increase the accuracy of predictions (Cocuzzo & Wu, 2023). Movie titles that were omitted include:

- Re-releases of films, including 3D re-releases
- Concert films
- Short films
- Stand-up specials
- Films missing budget data
- Films with missing values in several of its features

After pruning the film database, we were left with 1562 films. However, this left an unbalanced dataset favoring Class 1. To balance the dataset, a mixture of under sampling and oversampling was applied. First, the majority classes 1, 2, 3 were randomly under-sampled with a 150-film-limit. This resulted in a database of 789 films (Table 3.5). After the train-test split, an oversampling algorithm was applied to the training data, which is discussed in Chapter 3.6.

Table 3.5. Number of films per class before and after under sampling

Class	Interval	No. Films	
		Before	After
1	<100M	747	150
2	100-200M	324	150
3	200-300M	152	150
4	300-400M	84	84
5	400-600M	107	107
6	600M-1B	74	74
7	>1B	74	74
	Total	1562	789

3.5. EVALUATION METHODOLOGY AND PERFORMANCE METRICS

For model evaluation, the dataset was split 85%-15% between training and test data. A mixture of the holdout method and 10-fold cross validation was used to evaluate model performance. 10-fold cross validation is the most widely used methodology in box office prediction (Cocuzzo & Wu, 2013; Lu et al., 2024; Ni et al. (2022); Rhee & Zulkernine, 2016; Sharda et al. (2016); Yang et al., 2023). A previous study stands out by applying a mixture of both model evaluation techniques (Ni et. al., 2022). By combining the strengths of 10-fold cross validation and the hold-out method, a more comprehensive model evaluation can be

achieved. Applying only 10-fold cross validation could result to overfitted results on training data, without testing unseen data.

The evaluation metrics for this thesis are Accuracy for Bingo hit rates and Average Percent Hit Rate (APHR) Accuracy for 1-Away hit rates. The APHR indicates the rate at which testing data samples are accurately classified into their correct classes (Sharda & Delen, 2006). There are two distinct hit rates: the exact (bingo) hit rate, which only considers correct classifications into the exact same class, and the within 1 class (1-Away) hit rate. The hit rate measures the average accuracy of the predictions compared to the desired output. APHR can be formulated as shown in Equations 1, 2, and 3 (Sharda & Delen, 2006) where g is the total number of classes, n is the total number of samples, and p_i represents the total number of samples classified as class i.

$$\text{APHR} = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples}} \tag{1}$$

$$\text{APHR}_{\text{Bingo}} = \frac{1}{n}\sum_{i=1}^{g} p_i \tag{2}$$

$$\text{APHR}_{1-\text{Away}} = \frac{1}{n}\left((p_1 + p_2) + \sum_{i=2}^{g-1} p_{i-1} + p_i + p_{i+1} + (p_{g-1} + p_g)\right) \tag{3}$$

3.6. MISSING VALUES, OUTLIERS, AND OVERSAMPLING

The only feature that presented missing values was *Theaters* with 10% of absent data. KNN imputer was used to fill missing values on the training data, and the same configuration was used for filling missing values on the test data. The KNN imputer was selected for its ability to accurately predict missing values by considering similarities between data points, providing a robust solution compared to mean or median imputation. This method preserves the data distribution and relationships between variables, ensuring better prediction accuracy and avoiding biases (Troyanskaya et al., 2001). Using the same configuration for both training and test data prevents data leakage and overfitting.

Outliers were identified in continuous variables using histograms, boxplots, and the interquartile range method (Fig. 3.5), with all outlier percentages remaining below 4%. These outliers were retained because they represent valuable information—particularly in the context of high-budget films or large-scale releases, which often correlate with significant box office returns. Removing them would risk losing critical data, especially since they account for a small portion of the dataset and do not excessively skew the data. A/B testing was conducted with different methods for treating outliers, including clipping extreme values and adjusting skewness using log and Box-Cox transformations, but the models consistently performed better without treating outliers. Extreme values, such as high budgets or large theater counts, often correlate with substantial box office revenue, and removing them could

limit the model's ability to generalize effectively. Studies suggest that models like Random Forest handle outliers effectively, and excluding them could reduce variance, hindering the model's predictive power (Breiman, 2001). Additionally, while the budget variable displayed expected left skewness, its linear relationship with box office revenue justified not applying transformations, as this preserved the integrity of the data for more accurate prediction.

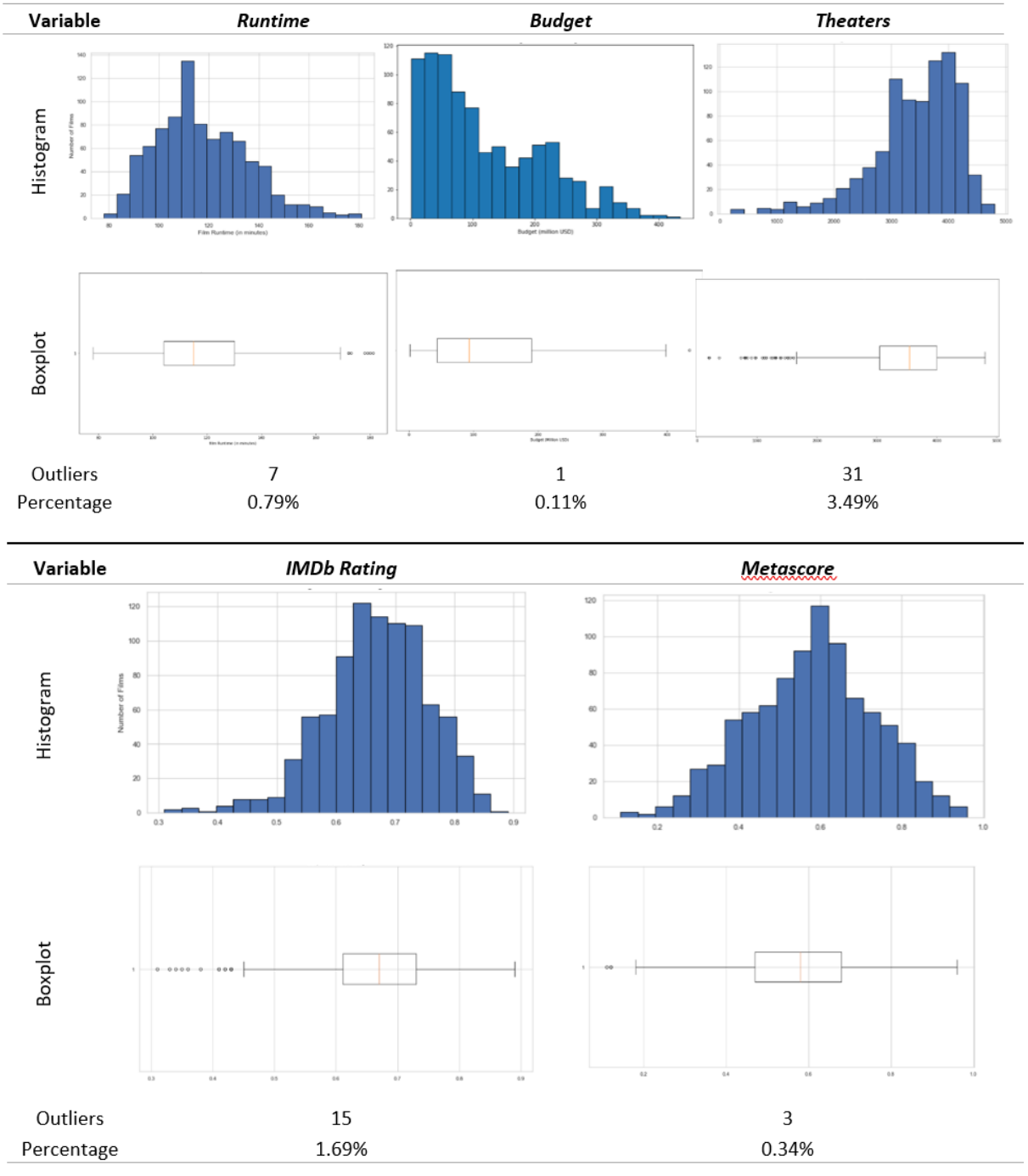

Figure 3.5. Histograms and boxplots of continuous variables for outlier detection

Since the dataset exhibited class imbalance, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was applied to the underrepresented classes, generating synthetic samples to enhance class representation and reduce model bias toward the majority class. This approach improves model performance by ensuring balanced training data, which is crucial for algorithms sensitive to class distribution (Batista et al., 2004; Chawla et al., 2002). Importantly, SMOTE was applied exclusively to the training dataset, preventing data leakage and ensuring that the test set remained representative of real-world distributions, maintaining the integrity of model evaluation. The final training dataset consists of 889 films, with 670 representing real data points and the remaining films generated through the application of SMOTE, creating synthetic data points (Table 3.6). The main data preprocessing pipeline is summarized in Table 3.7.

Table 3.6. Number of films per class in training data set before and after oversampling

Class	Interval	No. Films	
		Before	After
1	<100M	127	127
2	100-200M	127	127
3	200-300M	127	127
4	300-400M	91	127
5	400-600M	71	127
6	600M-1B	64	127
7	>1B	63	127
	Total	670	889

Table 3.7. Data preparation pipeline

Stage	**Method**	**Objective**	**Data**
Data Collection	Extract data from Box Office Mojo and enrich it with IMDb	Obtain film database	2000 films
Data Pruning	Exclude re-releases, shorts, concert films, and films with missing data	Decrease training time and increase prediction accuracy	1572 films
Undersampling	Undersample majority classes	Address unbalanced dataset and improve model performance	789 films
Test-Train Split	85% training, 15% test	Split data for model evaluation	Train: 670 Test: 119
Oversampling	SMOTE algorithm on training data	Address unbalanced dataset and improve model performance	Train: 889* Test: 119

*Includes synthetic data points

3.7. PREDICTIVE MODELS

Four machine learning classifiers, widely supported by box office prediction literature—KNN Classifier, Support Vector Machine, Neural Networks, and Random Forest—were evaluated in this thesis. Each algorithm was fine-tuned and optimized to identify the best-performing configuration for predicting box office revenue.

The **K-Nearest Neighbors** algorithm is a straightforward, effective supervised learning method for classification and regression. KNN makes predictions based on the majority class (for classification) or average value (for regression) of the 'k' closest neighbors to a query point in the feature space. Its simplicity and ease of interpretation make KNN particularly effective in tasks where the decision boundary is highly irregular (Cover & Hart, 1967; Hastie et al., 2009).

Multilayer Perceptron Neural Network is a type of feedforward artificial neural network that maps input data to output classes by adjusting weights between layers through forward and backward propagation. This iterative weight adjustment enhances its ability to learn and generalize from training data, allowing it to capture complex, non-linear relationships (Haykin, 2009; Rumelhart et. al., 1986).

The **Support Vector Machine** classifier finds an optimal hyperplane to maximize class separation. By using kernel tricks, SVM can perform both linear and non-linear classification effectively, which is useful for handling complex datasets (Cortes & Vapnik, 1995; Scholkopf & Smola, 2002).

Random Forest is an ensemble learning method that builds multiple decision trees, combining their predictions for better accuracy and reduced overfitting. It's particularly well-suited for high-dimensional data and non-linear relationships due to its robust aggregation of multiple models (Breiman, 2001; Liaw & Wiener, 2002).

To ensure the highest predictive performance, hyperparameter optimization was performed for each model using GridSearch, a systematic method for exploring parameter combinations (Bergstra & Bengio, 2012). The final optimized hyperparameters for each model are provided in Appendix B.

Feature selection was conducted to refine the input features for each algorithm, enhancing model performance and interpretability. The following wrapper methods were utilized:

- *Sequential Feature Selection (SFS)* was used for the KNN Classifier.
- *Recursive Feature Elimination (RFE)* was applied for SVM, while RFE with Logistic Regression was applied for MLP-NN.
- *Random Forest*, as an ensemble model, inherently handles feature importance and did not require additional wrapper methods for optimization (Breiman, 2001).

Each model underwent multiple iterations of parameter fine-tuning, feature selection techniques, and wrapper methods to achieve optimal configurations. This iterative process resulted in four final optimized models—one for each algorithm—that formed the basis for predictive analyses.

3.8. RELATIONSHIP BETWEEN PROTAGONIST'S GENDER AND BOX OFFICE

This study aims to examine the relationship between the gender of the protagonist and box office performance. Several statistical methods were employed to assess the strength and significance of this relationship, including the Chi-Square test and Random Forest Algorithm. The following methodology outlines the procedures used to analyze the impact of protagonist gender on box office revenue.

To examine potential factors associated with box office revenue categories, a series of **Chi-Square Tests of Independence** were conducted for all features, including distributor, existing IP, director ranking, star ranking, co-star ranking, and gender protagonist. Each test was performed separately to assess whether the observed distribution of each feature across revenue categories significantly deviates from the expected distribution under the assumption of independence. This approach helps identify features that may have a statistically significant association with box office performance.

Focusing specifically on the gender protagonist feature, an additional Chi-Square Test, referred to as the **Detailed Gender-Box Office Association Test**, was conducted. This test evaluates the observed and expected frequencies of male and female protagonists across seven box office revenue categories. By examining these frequencies, the test allows for a closer analysis of potential patterns in the distribution of protagonist gender across different revenue levels, without assuming any prior association between gender and box office outcomes.

Subsequently, a **Feature Importance Analysis** using Random Forest was applied to rank the importance of all features, including protagonist gender, in predicting box office performance. Protagonist gender's relative importance was compared to other features such as director ranking, star ranking, and genre classifications. Additionally, the feature's inclusion in the four optimal predictive models (KNN, SVM, NN, RF) was evaluated to determine whether gender consistently contributed to model accuracy.

Lastly, A comprehensive **Impact Study** was performed on the four optimal machine learning algorithms by adding the protagonist's gender and analyzing the variation in model performance. By including gender in the predictive models, the analysis determined whether its presence influenced model performance. A significant improvement in accuracy would indicate that gender is a strong predictor, while its exclusion would suggest it is not crucial for predicting box office revenue.

These methodologies provide a comprehensive analysis of the role of protagonist gender in box office performance. By combining statistical tests and machine learning, the study evaluates whether gender influences revenue and its relative importance compared to other factors, offering insights into the impact of gender on box office success and industry representation.

4. RESULTS AND DISCUSSION

4.1. BOX OFFICE PREDICTION

This section presents the performance analysis of the optimal models for each machine learning algorithm evaluated in this study: K-Nearest Neighbors, Neural Networks, Support Vector Machine, and Random Forest. Performance metrics were assessed using 10-fold cross-validation and test data, focusing on two primary accuracy measures: Bingo (exact match) and 1-Away APHR (Table 4.1; Fig. 4.1).

Table 4.1. Results comparison between optimal models

Model	**KNN**		**NN**		**SVM**		**RF**	
Experiment	**Bingo**	**1-Away**	**Bingo**	**1-Away**	**Bingo**	**1-Away**	**Bingo**	**1-Away**
10-folds	47.86%	80.10%	44.31%	79.30%	43.31%	79.42%	**48.82%**	**80.20%**
SD* (10 folds)	4.90%	5.17%	3.32%	5.26%	**3.17%**	4.38%	6.12%	**3.26%**
Test Data	43.70%	78.99%	44.54%	78.99%	42.02%	80.67%	**44.54%**	**80.67%**
SD* (Train-test)	4.16%	1.11%	**0.23%**	**0.31%**	1.29%	1.25%	4.28%	0.47%
Robustness Indicator	0.893	0.874	1.319	0.859	**1.381**	1.031	0.715	**1.386**
Generalization Error	4.16%	1.11%	**0.23%**	**0.31%**	1.29%	1.25%	4.28%	0.47%
Stability Score (10-fold)	95.10%	94.83%	96.68%	94.74%	**96.83%**	95.62%	93.88%	**96.74%**
Stability Score (Train-Test)	95.84%	98.89%	**99.77%**	**99.69%**	98.71%	98.75%	95.72%	99.53%

*SD = Standard Deviation

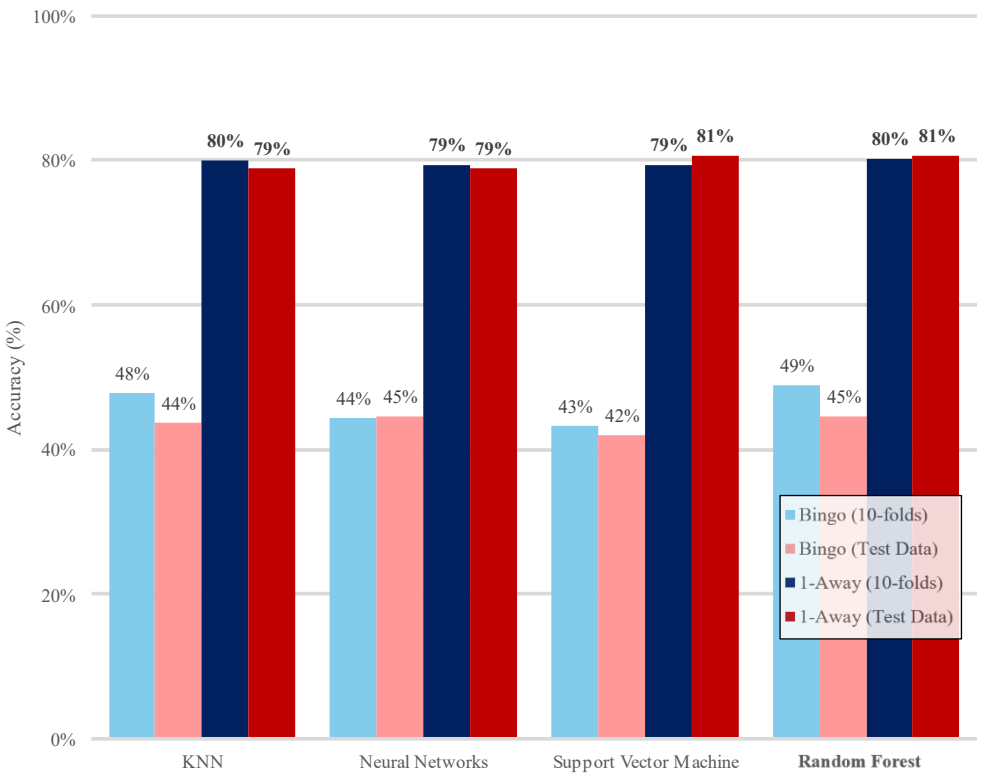

Figure 4.1. Results comparison across machine learning models

Among the evaluated algorithms, **Random Forest** emerged as the top performer for predicting box office success, excelling in both Bingo and 1-Away accuracy metrics. During 10-fold cross-validation, RF achieved the highest Bingo accuracy of 48.82% and the best 1-Away accuracy of 80.20%. On test data, RF maintained its lead, achieving the top Bingo accuracy of 44.54% and the strongest 1-Away accuracy of 80.67%. RF consistently outperformed other models in both metrics, showcasing its strong predictive capabilities. Focusing on the 1-Away experiment (Fig. 4.2), RF exhibited excellent stability with the top score on cross-validation (96.74%) and the second highest on test data (99.53%). RF also achieved the strongest Robustness Indicator (1.386) and the second-best Generalization Error (0.74%), further reinforcing its effectiveness under varying conditions. These results highlight RF as a reliable, robust, consistent, and highly effective model for box office prediction.

Figure 4.2. Generalization, stability, and robustness evaluation for the 1-Away experiment

The RF model utilized twelve features in its optimal configuration: theaters, budget, director ranking, IMDb rating, co-star ranking, Metascore, existing IP, runtime, distributor, star ranking, adventure, and drama. The relative importance of these features is depicted in Figure 4.3.

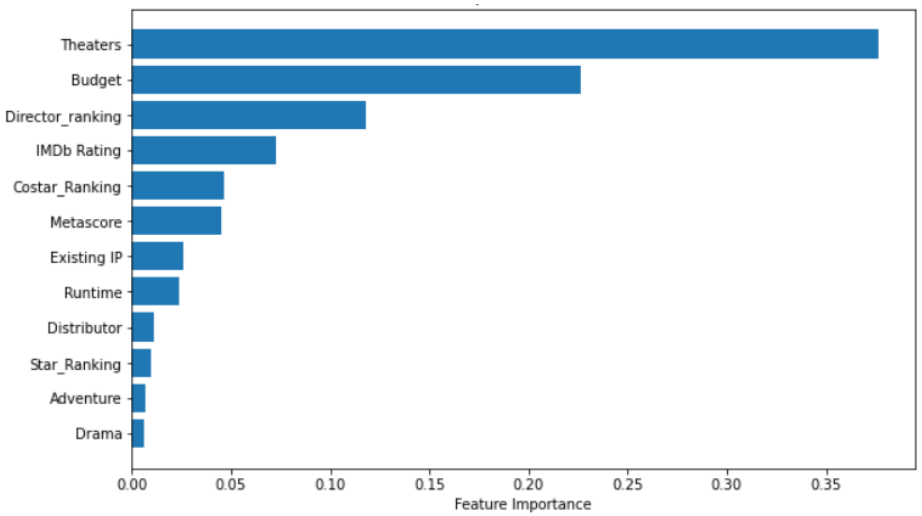

Figure 4.3. Feature importance of optimal model (RF)

To further contextualize its performance, the RF model was compared with results from previous studies (Table 4.2). This study's approach stands out for its inclusion of post-pandemic years, 2022 and 2023. Additionally, the classification task addresses a 7-class problem, reflecting the complexity of real-world box office success, compared to binary classification tasks (e.g., profit or loss/break-even) in some prior studies.

Table 4.2. Results comparison with previous studies

Reference	Target Variable	Features	Classes	Accuracy (Bingo)	APHR (1-Away)	Data Size	Time Range	Market
Lu et al. (2023)	Box Office	13	7	50%	80%	498	2017-2018	Taiwan
Yang et al (2023)	Profit	9	5	-	86.9%-89.7%	197	2018	China
Liao et al. (2023)	Box Office	37	7	69.10%	86.46%	1182	2010-2019	China
Sharda & Delen (2006)	Box Office	7	9	36.90%	75.20%	834	1998-2002	Worldwide
Rhee & Zulkernine (2016)	Profit	14	2	88.80%	-	375	2011-2015	Worldwide
Ahmed et al. (2020)	Box Office	17	9	82.00%	95.00%	5043	1915-2015	Worldwide
Galvão & Henriques (2018)	Profit	12	9	40.00%	-	1920	2000-2016	Worldwide
Souza et al. (2023)	Profit	9	6	49.10%	89.80%	383	2015-2018	Worldwide
This model	Box Office	14	7	48.82%	80.20%	792	2012-2019; 2022-2023	Worldwide

With a Bingo accuracy of 48.82%, the RF model outperforms Sharda & Delen (2006) at 36.9% and Galvão & Henriques (2018) at 40%. It is slightly below Lu et al. (2024) at 50% (that focuses on the Taiwanese market) and Souza et al. at 49.1%. While Rhee & Zulkernine (2016) reported a higher accuracy of 88.8%, their simpler binary classification task inherently reduces complexity compared to this study's 7-class structure. Liao et al. (2023) also outperforms this model at 69.1%, however its focus is on the Chinese market.

In terms of 1-Away APHR, the RF model achieved 80.2%, outperforming Sharda et al. (2006) at 75.2% and Lu et al. (2024) at 80%. While Yang et al. (2023) and Souza et al. (2023) reported higher APHR scores of 86.9% and 89.8%, respectively, these studies tackled less complex classification tasks, with fewer box office revenue classes (5 in Yang et al., 6 in Souza et al.). Furthermore, this study utilized a significantly larger and more diverse dataset, spanning 10 years—double the time range of Souza et al. (2023) and far broader than Yang et al. (2023), which was limited to a single year. This extended dataset introduces added complexity but provides a richer context for generalizable predictions.

The confusion matrix for the RF model (Table 4.3) reveals variable performance across the seven box office revenue classes, with notable strengths in predicting extreme categories like class 1 (flops) and class 7 (blockbusters). For class 1, the model achieved a Bingo accuracy of 66.14% and a 1-Away accuracy of 91.34%, demonstrating its reliability in identifying underperforming films. Similarly, for class 7, which includes films grossing over $1 billion (after adjusting for inflation), the model excelled, with a Bingo accuracy of 78.74% and a 1-Away accuracy of 94.49%, showcasing its capability to predict major box office hits with high precision.

Table 4.3. Confusion matrix for the Random Forest model

		True Classes							**Avg.**
		1	**2**	**3**	**4**	**5**	**6**	**7**	
Predicted Classes	**1**	**84**	37	25	6	5	0	0	
	2	32	**52**	36	16	8	3	2	
	3	4	17	**27**	11	11	2	1	
	4	5	14	12	**43**	22	8	0	
	5	2	7	18	29	**50**	9	4	
	6	0	0	8	20	27	**78**	20	
	7	0	0	1	2	4	27	**100**	
Bingo		66.14%	40.94%	21.26%	33.86%	39.37%	61.42%	78.74%	**48.82%**
1-Away		91.34%	83.46%	59.06%	64.57%	77.95%	89.76%	94.49%	**80.20%**

Overall, the Random Forest model outperformed other algorithms in predicting box office success, with superior performance in both Bingo and 1-Away accuracy metrics. Its robust and stable performance, particularly in predicting both underperforming films and blockbusters, underscores its effectiveness. Compared to previous studies, this model benefits from a larger, more diverse dataset and a more complex 7-class classification, enhancing its predictive accuracy and real-world applicability.

4.2. IMPACT OF GENDER ON BOX OFFICE

The relationship between the gender of the protagonist and box office performance was evaluated using multiple statistical techniques, including the Chi-Square test and machine learning models. The results reveal important insights into the role of the protagonist's gender in predicting box office outcomes.

The **Chi-Square Test of Independence** (Appendix A) was conducted to assess the association between the discrete features (including the protagonist's gender) and box office revenue categories. The analysis reveals that while gender has a statistically significant impact on box office performance ($p=0.027$), it ranks very low (19^{th} out of 25 features) compared to other features. For instance, its p-value is several orders of magnitude higher than those of director ranking ($p=2.51\times10^{-61}$), co-star ranking ($p=6.95\times10^{-34}$), and existing IP ($p=4.90\times10^{-33}$). Similarly, in terms of Chi-Square values, the protagonist's gender ($\chi^2=14.21$) ranks 19^{th}, demonstrating a far weaker association with box office performance compared to director ranking ($\chi^2=320.32$), co-star ranking ($\chi^2=188.70$), and genre categories like Adventure ($\chi^2=128.34$) and Superhero ($\chi^2=127.56$). This highlights that production and marketing factors like high-profile collaborations and franchises play a larger role in a film's success, with gender's influence, though significant, being comparatively minor due to industry biases favoring star power, genre, and marketable IP.

The **Detailed Gender-Box Office Association Test** (Table 4.4) confirmed a significant relationship between protagonist gender and box office revenue. The test shows that male protagonists are expected to perform better in higher revenue categories (4-7), while female protagonists are expected to be more common in lower revenue categories (1-3). This suggests potential gender biases in how films are marketed, produced, or positioned. Although statistically significant, the relationship suggests gender alone doesn't determine box office success, with other factors likely playing a bigger role.

Table 4.4. Detailed gender-box office association test

Statistic				**Value**			
Chi-Square Test Statistic				14.21915915			
p-value				0.027281579			
Degrees of Freedom				6			
Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Male (0)	94.39701	94.39701	94.39701	52.77313	67.63881	46.82687	47.57015
Female (1)	32.60299	32.60299	32.60299	18.22687	23.36119	16.17313	16.42985

The **Feature Importance Analysis** conducted using the Random Forest model (including all features) ranked the protagonist's gender relatively low in importance (1.09%) compared to other factors (Fig. 4.4, Appendix C). While gender does contribute to a film's commercial success, its impact is much smaller than more influential features like theatres (10.08%), budget (8.57%), and IMDb rating (6.13%). These higher-ranked features suggest that the scale of distribution, financial backing, and audience reception outweighs gender in influencing box

office outcomes, with star power, genre, and marketing playing a larger role in commercial success.

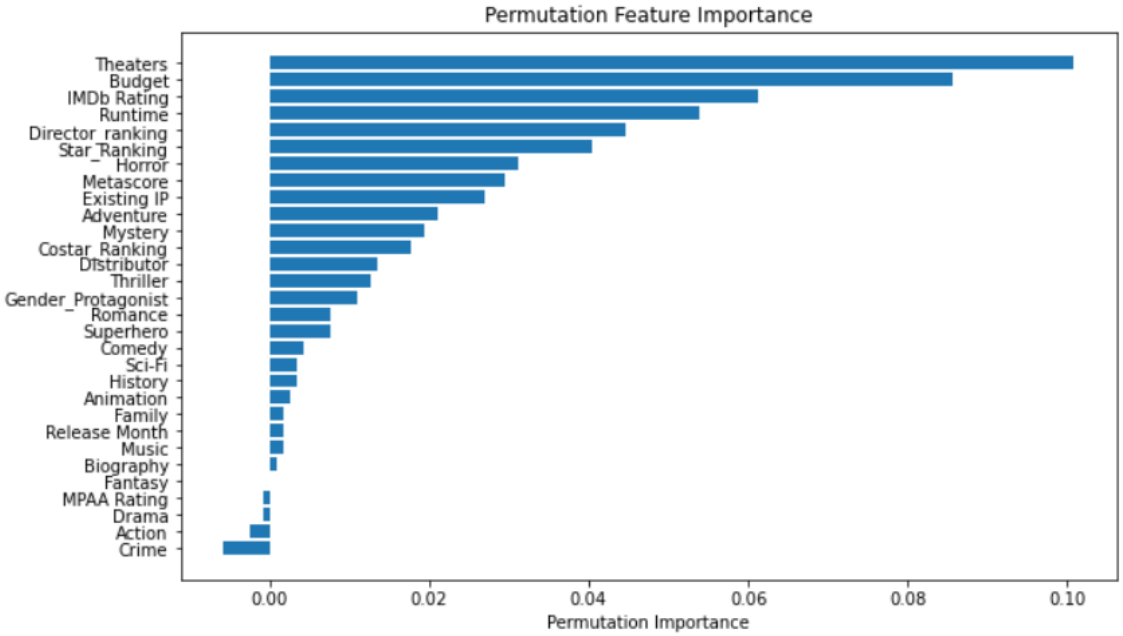

Figure 4.4. Feature importance of all variables using Random Forest

Further analysis of optimal models (Appendix D) reveals that gender was not selected as a feature in any of the best-performing models. This indicates that gender may not be a crucial factor in predicting box office revenue. Its exclusion from optimal models could reflect either a lesser impact of gender in predicting the target variable or the influence of cultural and contextual factors in the data.

Finally, a comprehensive **Impact Study** was conducted to evaluate how the inclusion of protagonist gender affects the performance of the four optimal machine learning algorithms. The results (Table 4.5) show that the inclusion of gender led to a decline in model accuracy across all algorithms. This decline was observed in both the "Bingo" accuracy and 1-away accuracy metrics for training and test data sets. The addition of the gender feature appears to introduce complexity or noise, rather than improving the model's ability to make correct predictions.

Table 4.5. Impact of protagonist's gender on model performance

Model	**KNN**		**NN**		**SVM**		**RF**	
Experiment	Without Gender	With Gender	Without Gender	With Gender	Without Gender	With Gender	Without Gender	With Gender
Bingo (Train)	47.9%	45.6%	44.3%	41.1%	41.5%	43.6%	48.8%	46.0%
1-away (Train)	80.1%	78.5%	79.3%	78.6%	79.6%	79.6%	80.2%	77.6%
Bingo (Test)	43.7%	**37.8%**	44.5%	40.3%	42.0%	37.8%	44.5%	**37.8%**
1-away (Test)	79.0%	79.8%	79.0%	78.2%	80.7%	79.8%	80.7%	75.6%

Variation in performance

-7% +2%

Figure 4.5 further illustrates this conclusion, with most models showing reduced performance when gender was included. Notably, the K-Nearest Neighbors model was the only one where the inclusion of gender improved the 1-away accuracy on the test set, albeit marginally. In contrast, the other models saw a significant decline in performance ranging from -1% to -7%. Overall, the results suggest that while gender might provide some value in certain models, it does not significantly enhance predictive accuracy across the board.

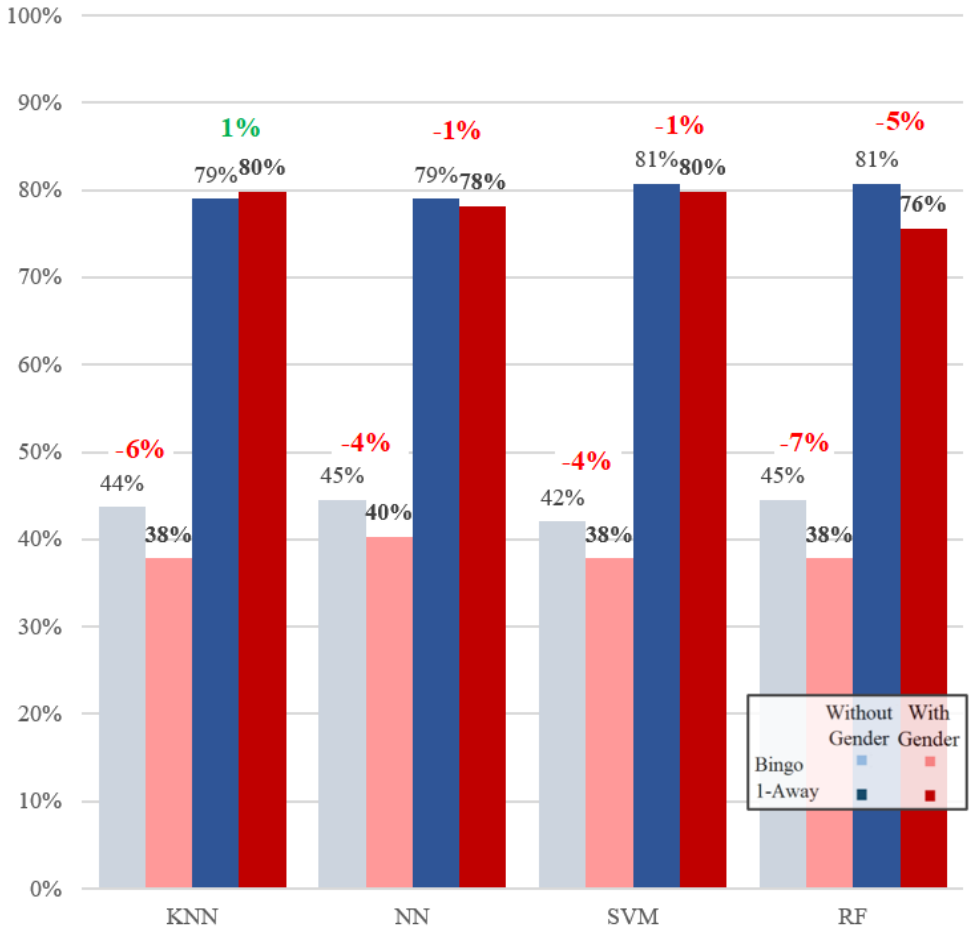

Figure 4.5. Performance comparison with vs. without gender (test data)

Ultimately, these statistical and machine learning analyses reveal that while the protagonist's gender is statistically significant, its predictive power is limited compared to factors like director ranking, star power, budget, and genre. Gender's impact is secondary, influenced by other variables and industry practices. The decline in model accuracy when gender was included underscores its limited role, highlighting the need to assess gender alongside other factors to understand box office outcomes.

CONCLUSIONS AND FUTURE RESEARCH

Machine learning models, particularly Random Forest, prove to be highly effective tools for predicting box office success. Among the evaluated algorithms, RF consistently emerged as the most robust and reliable, leveraging a comprehensive feature set and a decade-long dataset to outperform models like K-Nearest Neighbors, Neural Networks, and Support Vector Machines. Specifically, RF achieved a Bingo accuracy of 48.82% and a 1-Away APHR accuracy of 80.20% during 10-fold cross-validation, with similar results on test data. These findings highlight RF's ability to handle the complexity of a 7-class classification problem, reflecting the nuanced realities of box office outcomes.

The RF model's superiority stems from its robustness, generalizability, and stability. It exhibited the highest Robustness Indicator, the strongest Stability Score in cross-validation, and a low Generalization Error, confirming its reliability under diverse scenarios. Notably, RF excelled at predicting extreme categories, such as flops and blockbusters, where feature distributions are distinct. However, it struggled more with middle-range classes, underscoring the need for more nuanced features to reduce ambiguity in these predictions. This highlights the model's ability to capture trends at the extremes while revealing areas for further refinement.

Compared to previous studies, the RF model demonstrated competitive accuracy, even while tackling a more complex classification problem involving seven revenue classes. Unlike binary classification tasks in earlier research, this approach incorporates greater complexity, providing a nuanced perspective on box office performance. Moreover, this study stands out by including post-pandemic years (2022–2023), and inflation adjustments, enriching its relevance and adaptability to evolving industry trends. While some studies reported higher accuracy rates, they often relied on smaller time frames, fewer revenue classes, and local markets, without considering post-pandemic data, making direct comparisons less meaningful.

Key predictors of box office success—such as theater count, budget, director ranking, and star power—proved far more significant than demographic factors like protagonist gender. Gender ranked low in importance, with machine learning analyses consistently excluding it from optimal models due to its minimal contribution to prediction accuracy. Both the Chi-Square and Association Tests confirmed a statistically significant but limited relationship between protagonist gender and box office revenue, contributing only 1.09% to model predictions in the RF test. Moreover, including gender in the optimal models consistently reduced their performance, suggesting it introduces noise rather than improving predictive power.

Industry stakeholders should prioritize factors such as ratings, genre, budget, and star power when predicting box office success, as these are far more influential than demographic variables like the protagonist’s gender. While gender remains culturally significant, it is not a key determinant of commercial outcomes. This challenges the underrepresentation of female-led films, as gender-based box office predictions offer no justification for such disparities. By demonstrating that the protagonist’s gender should not heavily influence production decisions, this research contributes to the broader conversation on gender

equality in film and supports the United Nations' goals for promoting balance and reducing disparities. As such, it calls for a shift in industry practices, where women should be given more opportunities to lead films and paid equitably, reflecting their contribution to a film's success—just as their male counterparts are.

4.3. LIMITATIONS AND FUTURE WORK

While this study offers valuable insights into box office prediction, it also highlights limitations and areas for further exploration. Incorporating additional features such as social media sentiment, audience reviews, and marketing data could enhance the models' predictive accuracy and capture emerging trends in consumer behavior. Additionally, exploring advanced machine learning techniques, like ensemble methods or deep learning, could improve prediction capacity, especially for middle-range revenue categories.

Future studies could also benefit from testing different datasets with varying time spans and sizes, as well as exploring the dynamics of international markets where box office drivers differ by region and cultural context. Such research could provide valuable insights into how different global markets influence box office success and which predictors are most relevant in diverse environments. Additionally, examining the long-term impact of gender representation on audience reception and profitability could offer deeper insights into demographic factors over time. With the rise of digital platforms and streaming, integrating real-time data, such as pre-release sentiment or trailer engagement, could improve predictive capabilities.

Refining the model could also involve adopting a pre-release feature selection approach to capture early market indicators (Souza et al., 2023), such as first-week theater count and early ratings. The model's challenge in predicting middle-range revenues highlights the need for more nuanced features, including social media trends and real-time feedback. Future work should also explore gender's role in prediction by examining its intersection with other factors like genre, director, and star power.

Another limitation is the dataset's focus on Hollywood films, which may not fully capture global dynamics. The importance of features such as director ranking, star power, and genre may differ in non-Western markets where audience preferences and industry practices vary. Additionally, the role of gender as a predictor may differ in regions where diversity and gender representation are more emphasized in media. Regional economic factors, such as box office economics and marketing strategies, could also influence the relevance of certain features across different markets. To improve the generalizability of the model, future research should incorporate a more diverse dataset that includes films from a broader range of cultural and economic contexts.

Finally, the inclusion of post-pandemic films presents new opportunities for research. Given the profound shifts in audience behavior, marketing strategies, and distribution models since the pandemic, further examination of these changing dynamics could help refine box office prediction models. By adapting the current framework to account for these changes, future research could deepen our understanding of box office success in a rapidly evolving film industry.

BIBLIOGRAPHY

Ahmed, U., Waqas, H., & Afzal, M.T. (2020) *Pre-production box-office success quotient forecasting. Soft Comput 24,* 6635–6653. https://doi.org/10.1007/s00500-019-04303-w

Alberani, D., & Uyar, H. T. (2018). *Cinemagoer*. https://imdbpy.readthedocs.io/en/latest/

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*. *SIGKDD Explorations*, 6(1), 20–29

Bergstra, J., & Bengio, Y. (2012). *Random search for hyper-parameter optimization*. *Journal of Machine Learning Research*, 13, 281-305

Box Office Mojo (2023). *2023 Box Office Results*. https://www.boxofficemojo.com/year/world/2023/

Breiman, L. (2001). *Random forests*. *Machine Learning*, 45(1), 5–32

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 9*(13), 1-73.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. *Journal of Artificial Intelligence Research*, 16, 321–357

Cocuzzo, D., & Wu, S. (2013). *Hit or flop: Box office prediction for feature films.* Stanford University

Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. *Machine Learning*, 20(3), 273-297

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Damico, A. M. (2022). Problems Controversies and Solutions. In *Women in media: A reference handbook*. ABC-CLIO. ISBN 978-1-4408-7605-9

Du, J., Xu, H., & Huang, X. (2014). Box office prediction based on microblog. *Expert Systems with Applications, 41*(4, Part 2), 1680-1689. https://doi.org/10.1016/j.eswa.2013.08.065

Evans, R. D., Karabas, I., Andonova, Y., & Nochebuena-Evans, L. (2024). Let's not talk about men: When meaningful female-to-female interaction and dialogue drive higher box office sales. *Journal of Global Scholars of Marketing Science: Bridging Asia and the World, 34*(1), 57–70. https://doi.org/10.1080/21639159.2023.2248758

Galvão, M. & Henriques, R. (2018). Forecasting Movie Box Office Profitability. *Journal of Information Systems Engineering & Management, 3*(3), 22. https://doi.org/10.20897/jisem/2658

Ge, J., Chen, H., Qiu, K., & Sun, J. (2023). Beyond the Box Office Performance: A Multi-Factor-Based Prediction Model for Sequel Movie. *ACM International Conference Proceeding Series*, 42–46. https://doi.org/10.1145/3640872.3640879

Gunter, B. (2018). Is Box Office Still Relevant. In *Predicting movie success at the box office*. Springer International Publishing. (pp. 1-20). https://doi.org/10.1007/978-3-319-71803-3

Han, J., Pei, J., & Tong, H. (2023). Data, measurements and data preprocessing. In *Data mining: concepts and techniques*. Morgan Kaufmann. (4th ed., pp. 23-84)

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

Haykin, S. (2009). *Neural Networks and Learning Machines* (3rd ed.). Prentice Hall

Heywood, J. S., Izquierdo Sanchez, S., & Navarro Paniagua, M. (2024). The Hollywood gender gap: The role of action films. *Journal of Cultural Economics.* https://doi.org/10.1007/s10824-024-09514-0

Izquierdo Sanchez, S., & Navarro Paniagua, M. (2017). *Hollywood's wage structure and discrimination* (Economics Working Paper Series, 2017/005). *Lancaster University Management School*. https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/lums/economics/working-papers/LancasterWP2017_005.pdf

Kim, I.K. (2021) The impact of social distancing on box-office revenue: Evidence from the COVID-19 pandemic. *Quantitative Marketing and Economics 19,* 93–125. https://doi.org/10.1007/s11129-020-09230-x

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

Lauzen, M. M. (2023). *25th anniversary: The celluloid ceiling: Employment of behind-the-scenes women on top grossing U.S. films in 2022*. SDSU Center for the Study of Women in Television and Film. https://womenintvfilm.sdsu.edu/wp-content/uploads/2023/01/2022-celluloid-ceiling-report.pdf

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,* 321–338

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

Lindner, A. M., Lindquist, M., & Arnold, J. (2015). Million Dollar Maybe? The Effect of Female Presence in Movies on Box Office Returns. *Sociological Inquiry, 85*(3), 407–428. https://doi.org/10.1111/soin.12081

Litman, B. R. (1983). Predicting Success of Theatrical Movies: An Empirical Study. *The Journal of Popular Culture, 16*(4). https://doi.org/10.1111/j.0022-3840.1983.1604_159.x

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery, 6*(4), 393–423. https://doi.org/10.1023/A:1016304305535

Lu, S. H., Wang, H. J., & Nguyen, A. T. (2024). Machine Learning Applications on Box-Office Revenue Forecasting: The Taiwanese Film Market Case Study. *Studies in Systems, Decision and Control, 483,* 384–402. https://doi.org/10.1007/978-3-031-35763-3_49

Mitchell, R. (2024, January 4). *2023 Worldwide Box Office*. Gower Street Analytics. https://gower.st/articles/gower-street-analytics-estimates-2023-global-box-office-hit-33-9-billion/#:~:text=

Neff, K. L., Smith, S. L., & Pieper, K. (2024). *Inequality across 1,700 popular films: Examining gender, race/ethnicity & age of leads/co-leads from 2007 to 2023* (Report). USC Annenberg Inclusion Initiative. https://assets.uscannenberg.org/docs/aii-inequality-1700-films-2024-02-21.pdf

Ni, Y., Dong, F., Zou, M., & Li, W. (2022). Movie Box Office Prediction Based on Multi-Model Ensembles. *Information 2022, 13*(6) 299. https://doi.org/10.3390/info13060299

Ni, Y., & Li, S. (2023). Decomposition-integration-based prediction study on the development trend of film industry. *Heliyon, 9*(11), e21211. https://doi.org/10.1016/j.heliyon.2023.e21211

Nishijima, M., & Souza, T. L. D. e (2024). Do American Critic Reviews Affect Film Consumption Abroad? The Brazilian Case. *Empirical Studies of the Arts, 42*(1). https://doi.org/10.1177/02762374231196836

Rhee, T. G., & Zulkernine, F. (2016). Predicting movie box office profitability: A neural network approach. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 547-553). IEEE. https://doi.org/10.1109/ICMLA.2016.0117

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors*. *Nature*, 323(6088), 533-536

Sawhney, M. S., & Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science, 15*(2), 113–131. https://doi.org/10.1287/mksc.15.2.113

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press

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

Smith, S. L., Pieper, K. M., Granados, A., & Choueiti, M. (2010). Assessing gender-related portrayals in top-grossing g-rated films. *Sex Roles, 62*(11), 774–786. https://doi.org/10.1007/s11199-009-9736-z

Souza, T. L. D. e, Nishijima, M., & Pires, R. (2023). Revisiting predictions of movie economic success: random Forest applied to profits. *Multimedia Tools and Applications, 82*(25). https://doi.org/10.1007/s11042-023-15169-4

Starovoitov, V. V., & Golub, Y. I. (2021). Data normalization in machine learning. *Informatics 18* (3), 83-96. https://doi.org/10.37661/1816-0301-2021-18-3-83-96

The Business Research Company. (2024, November). *Film and video market 2024*. https://www.thebusinessresearchcompany.com/report/film-and-video-market

The Numbers. (2024) *Top 2024 Movies at the Worldwide Box Office*. Retrieved December 2, 2024, from https://www.the-numbers.com/box-office-records/worldwide/all-movies/cumulative/released-in-2024

Treme, J., & Craig, L. A. (2013). Celebrity star power: Do age and gender effects influence box office performance? *Applied Economics Letters, 20*(5), 440–445. https://doi.org/10.1080/13504851.2012.709594

Treme, J., Craig, L. A., & Copland, A. (2019). Gender and box office performance. *Applied Economics Letters, 26*(9), 781–785. https://doi.org/10.1080/13504851.2018.1495818

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). *Missing value estimation methods for DNA microarrays.* Bioinformatics, 17(6), 520–525

Weiβ, A.-N. (2024). Portrayals of the shero: A critical discourse analysis on the representation of Wonder Woman and Captain Marvel. *Journal of Gender Studies*, *29*(1), 1-18. https://doi.org/10.1080/09589236.2024.2360499

Xun, Y., Yin, Q., Zhang, J., Yang, H., & Cui, X. (2021). A novel discretization algorithm based on multi-scale and information entropy. *Applied Intelligence, 51*(2), 991–1009. https://doi.org/10.1007/S10489-020-01850-W

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

Yongbin, Y., & Rongzhao, O. (2013). A study on the relationship among the leading actors, directors, and the box office income of a film - Based on multiple linear regression model. In *Proceedings of the 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2013*, (p. 1). https://doi.org/10.1109/ICIII.2013.6702975

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

Zheng, D., Sun, Y., & Yu, F. (2024). Research on Movie Box Office Prediction Method Based on Blending Model Fusion. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing* (pp. 140-146)

APPENDIX A.- CHI-SQUARE TEST

Chi-Square Test between categorical variables and box office (discretized)

Rank	Feature	Chi-Square Test Statistic	p-value	Degrees of Freedom
1	Director_ranking	320.3235608	2.51143E-61	12
2	Costar_Ranking	188.7000081	6.95383E-34	12
3	Existing IP	165.1064701	4.90393E-33	6
4	Star_Ranking	140.0399715	5.87913E-24	12
5	Distributor	137.3277258	2.07224E-23	12
6	Adventure	128.3475641	2.86351E-25	6
7	Superhero	127.5620744	4.19003E-25	6
8	MPAA Rating	66.87591639	1.22449E-09	12
9	Release Month	65.127971	4.06203E-12	6
10	Drama	47.12535308	1.76645E-08	6
11	Sci-Fi	38.64684061	8.39381E-07	6
12	Thriller	35.24139929	3.86937E-06	6
13	Horror	33.83664786	7.23404E-06	6
14	Mystery	29.76693304	4.35283E-05	6
15	Action	22.37393762	0.001035707	6
16	Animation	20.45233776	0.002299836	6
17	Crime	20.07778942	0.002682465	6
18	Biography	15.8348736	0.014668251	6
19	**Protagonist Gender**	**14.21915915**	**0.027281579**	**6**
20	Romance	11.96673634	0.062715003	6
21	Family	11.77340563	0.06721901	6
22	Fantasy	11.28531449	0.079948682	6
23	History	4.567466755	0.600356992	6
24	Comedy	4.547382691	0.603027154	6
25	Music	3.435033299	0.752592211	6

APPENDIX B.- HYPERPARAMETERS OF OPTIMAL MODELS

Optimal Models	**KNN**	**NN**	**SVM**	**RF**
No. Features	5	5	28	12
Hyperparameters	No. Neighbors	Layer sizes	Kernel	No. Estimators
	20	*(25,25)*	*Linear*	*120*
	Weights	Max. Iterations	C	Max. Features
	Distance	*300*	*1.2*	*12*
	Metric	Activation	Gamma	Min. Samples Split
	Minkowski	*ReLU*	*Scale*	*12*
		Solver	Random State	Max Depth
		Adam	*42*	*6*

APPENDIX C.- FEATURE IMPORTANCE OF ALL VARIABLES USING RANDOM FOREST

Rank	Feature	Importance
1	Theatres	10.08%
2	Budget	8.57%
3	IMDb Rating	6.13%
4	Runtime	5.38%
5	Director Ranking	4.45%
6	Star Ranking	4.03%
7	Horror	3.11%
8	Metascore	2.94%
9	Existing IP	2.69%
10	Adventure	2.10%
11	Mystery	1.93%
12	Co-star Ranking	1.76%
13	Distributor	1.34%
14	Thriller	1.26%
15	Gender Protagonist	1.09%
16	Romance	0.76%
17	Superhero	0.76%
18	Comedy	0.42%
19	Sci-Fi	0.34%
20	History	0.34%
21	Animation	0.25%
22	Family	0.17%
23	Release Month	0.17%
24	Music	0.17%
25	Biography	0.08%
26	Fantasy	0.00%
27	MPAA Rating	-0.08%
28	Drama	-0.08%
29	Action	-0.25%
30	Crime	-0.59%

APPENDIX D.- FEATURE SELECTION ANALYSIS

Feature Selection Analysis: Selected features in the optimal models

Rank	Optimal Model	KNN	NN	RF	SVM
	No. Features	5	5	12	28
1	Theaters				
2	Budget				
3	Director Ranking				
4	IMDb Rating				
5	Co-star Ranking				
6	Metascore				
7	Family				
8	Existing IP				
9	Distributor				
10	MPAA Rating				
11	Adventure				
12	Star Ranking				
13	Runtime				
14	Sci-Fi				
15	Animation				
16	Drama				
17	Action				
18	Music				
19	Mystery				
20	Romance				
21	Comedy				
22	Horror				
23	Thriller				
24	History				
25	Crime				
26	Biography				
27	Fantasy				
28	Superhero				
29	Release Month				
30	**Gender Protagonist**				