Measuring "Star Power": Predicting Movie Box Office Revenue Based on Directors' and Leading Actors' Recent Success

Josh Iden
DATA698, Fall 2023

Abstract

This study develops machine learning models to predict the box-office revenue of movies prior to a film's release building on past studies that have used traditional predictors such as genre and MPAA rating, combined with a novel definition of the star power of directors and leading actors based on the cumulative revenues earned and number of films for the prior 5- and 10- years to a movie's release. We use the tree-based methods Bagged Trees, Random Forest, Extreme Gradient Boosting, and Cubist algorithms to make box office revenue predictions based on data collected for 4180 films representing the top 200 movies by domestic (U.S.) box office per year for the years 2002-2022. In our study, Random Forests utilizing the 10-year star power variables produce the lowest Mean Absolute Percentage Error (MAPE) amongst our modeling efforts, with our best model achieving a MAPE of 4.923% and identifying the star power variables as 4 of the top 7 most important predictors to the model. This model provides actionable insights for motion picture stakeholders to promote more reliable box office revenue outcomes.

Introduction

The film industry in the United States is worth $42.5 billion to the US economy as of 2022, contributing 2.5 million jobs (Zane, 2023) – largely due to the profitability of a minority of its movies: between 2000 and 2010, only 36% of movies in the United States had box office revenues higher than their production budgets. (Lash, Zhao, 2016). The decision by film studios and executives regarding which movies to produce (aka *greenlight*), and which to pass, can make or break careers: in the 1940s, the average tenure of executives in charge of production was about 20 years; by the 70s and 80s this figure had decreased to an average of 4. (Ravid, 1999) As William Goldman famously said, "Not one person in the entire motion picture field knows for a certainty what's going to work. Every time out it's a guess—and if you're lucky, an educated one." (Goldman, 1983) The onset of the digital age ushered in an era of unprecedented access to data, and with the development of sophisticated machine learning algorithms for complex modeling, the decisions being made are becoming more and more educated. In 2020, Warner Brothers signed a deal with production company Cinelytic to provide film and talent analytics that can generate film package evaluations or a star's worth to a film in seconds, arming executives and financiers with powerful tools to remove some of the guesswork in producing financially successful films. (Siegel, 2020)

Early efforts to predict a film's success leveraged linear regression to model the number of theaters at which a movie was shown (e.g., "rentals") (e.g., Litman, 1989; Sochay, 1994; Sawhney & Eliashberg, 1996) or box office revenue as a function of certain predictors (Ravid, 1999; Simonoff & Sparrow, 2000; Butler & De'Armand, 2003). Studies could be divided into two prevailing

theories: *economic theory*, involving factors determined before a film's release, such as budget, cast, director, etc., and *communication theory*, which involved word of mouth, critical reviews, awards, etc. (Litman, 1989) These variables could be described as *static* (not changing once the film is released), and *dynamic* (changing upon release). Movies represent a "long succession of creative decisions" (Eliashberg, Elberse, & Leenders, 2006) – the problem with building a predictive model using dynamic factors is that by the time a film is released, there is very little that can be done to affect its key features, so it is our belief that the most useful model to industry executives uses data that is available before the film is released. More recent studies have used neural networks and random forest models, and some decision tree models to classify the degree of success of films and reported better results. (Sharda & Delen, 2006, 2012; Lash & Zhao, 2016; Antipov & Pokryshevskay, 2017; Quader, Gani, Chaki, and Ali, 2017; He & Hu, 2021), however many of these studies focused on the international film industry.

Many studies have sought to quantify the value of directors and stars to a movie's box office success. Early studies used binary variables based on trade publication lists or rankings of top stars (Litman, 1989; Sochay, 1994, Sawhney & Eliashberg, 1996; Ravid, 1999). Other studies used IMDB popularity. (Glotfelty, 2012) Beginning in the mid-2010s, researchers began incorporating the cumulative lifetime earnings of stars and directors into their modeling. (Lash & Zhao, 2016; Antipov & Pokryshevskay, 2017; Quader, Gani, Chaki, and Ali, 2017; He & Hu, 2017) However these studies all positioned their modeling as a classification problem to label outcomes on varying degrees of success, rather than attempting to predict actual revenue. The first true revenue prediction model using the recent box office revenue of actors and directors looked at the previous 3- to 5- years of films in China. (Ni, Dong, Zou & Li, 2022). However, this study included all actors in a cast.

Surveying the literature pertaining to statistical and machine learning models that have been developed in the past 40+ years to predict box office revenue, we introduce a novel approach to quantifying the contributions of each film's director and leading actor(s) using data acquired from BoxOfficeMojo.com, IMDb.com, and OMDb.com, focusing on the recent commercial success of directors and leading actor(s). Combined with features such as Production Budget, Film Rating, Genre, Release Date, etc. we develop regression trees and rule-based regression models to generate accurate predictions of box office revenue that can be made before a film is released, when studios, executives, financiers, and creatives have the runway to make necessary changes to maximum revenue.

Tree-based models are ideal for handling both continuous and categorical independent variables without the need for pre-processing. They consist of nested "if-then" statements for predictors that partition data. They also provide transparent rules for decision-making, enabling useful decision support. Data is split into "terminal nodes" or "leaves" of a tree, with the model formula in each terminal node used to generate a prediction. These models automatically select relevant features and assign variable importance scores, which can help identify which variables are most influential in predicting an outcome. Ensemble models such as bagged trees, random forests, gradient boosting, and cubist models, combine many trees or rule-based models to achieve high predictive accuracy and generalize well to new data.

This study collects data for 4180 films in the United States from 2002-2022, representing the top 200 films per year in domestic box office revenue after removing duplicates and re-released films. We intend to use the data from the years 2002-2012 to engineer features about the economic output of leading actors and directors for preceding 5- to 10- year periods for the years 2013-2022 to develop tree-based models that can accurately predict box office revenue for movies based on data available prior to movie's release.

Review of Literature

Litman (1989) found that research modeling the financial success of movies was scarce through the 80s due largely to the proprietary nature of financial data. He identifies two prevailing theories regarding the success of a film: economic theory, the static factors that are in place before a film is released is determinant of the film's economic success, and communication theory, in which the dynamic factors such as word of mouth, critical reviews, awards, etc., are determinant of a film's success. Litman developed a regression model which used a binary variable to quantify whether there were any box office superstars in the cast of films, determined by a list of top ten box office stars for the two years prior to a film's release per the International Motion Picture Almanac. Rather than using revenue as a dependent variable, he used theater rentals, e.g., the number of theaters that showed the movie. He found that a film's ratings, Oscars, and season of release were not significant predictors, and that production budget was not related to the number of rentals.

Sochay (1994) studied films in the US and Canada from October 1987 – October 1989 and built a linear regression model that added Length of Run (LOR) to Litman's Theater Rentals as dependent variables. Sochay found that performance based on economic theory could be relevant in terms of prediction and described two phenomenon that impacted the communication theory – the "ripple effect" whereby huge hits expand the box office potential for other films, and the "black hole effect", where blockbusters suck the energy out of competing pictures and create a one- or two- film market that affects other films performance.

Sawhney & Eliashberg (1996) developed forecasts using analysis based on early box office data, and found that with some early sales data, only a few variables were needed for 90% accuracy. However, their stepwise regression efforts did not predict well without any sales data. To quantify "Star Power" in their models, they relied on a list from Variety.

Ravid (1999) identified two competing hypothesis with regards to "Star Power": First, the "rent capture" hypothesis, which states that stars capture most of their expected value added, and quickly adjust their fees to reflect their value, and the "signal effects" hypothesis, which states that the presence of factors such as attached directors, leading actors, writer(s), and the quality of a script may signal superior information. For example, a star might attach to a film based on a director, a director based on a script, etc. Ravid focused his attention on profits, rather than revenues, as a dependent variable in a linear regression model. He also introduced a variable for films that did not contain any stars, as well as assigning a value to the strength of release date.

Simonoff & Sparrow (2000) built linear regression models to predict revenue for movies before they open and immediately after release using a sample from IMDb of 311 films between 1997-1998. They quantified "Star Power" as the number of actors appearing in Entertainment Weekly's best actors list who had appeared in at least 10 movies all time. They stipulated that budget figures may be deceptively low because directors and stars may waive normal salary requirements in exchange for profit participation.

Butler & De'Armond (2003) developed a linear regression model using 500 films from 2001-2003 for information after a film's release and found a relationship between critical approval and an increase in revenue. They also identified a relationship between genre and sequels and revenue.

Eliashberg, Elberse, & Leenders (2006) identify the value chain phases of theatrical motion pictures – production, distribution, exhibition, and consumption, and examine the two research traditions of audience behavior: the "psychological" approach, which considers how individual decisions to attend movies are made, and the "economic" approach, i.e., the variables that influence the financial performance of movies. They note the development of motion pictures is a long succession of creative decisions with far-reaching economic implications, and that budgets are determined based on script, post-production, salaries, financing, etc. They also note that high budgets can mean that movies can employ high-profile stars, which can attract financing, which can lead to wider distribution and higher revenue, and future studies should take these potentially endogenous relationships into account.

Sharda & Delen (2006, 2012) built a 10-fold cross-validation neural network classification model using 5 years of movie data and found that their model performed better than logistic regression, discriminant analysis, and classification and regression trees. Movies were classified from 1 ('flop') to 9 ('blockbuster'). Star Power used three binary variables to represent star power from insignificant to high. This model did not include a variable to represent a director's value. The 2012 follow up study achieved a predictive accuracy of 53%, with a "one-away" accuracy of 87%.

Elberse (2007) developed a cross-sectional regression analysis looking at changes in the Hollywood Stock Exchange (HSX) prices with respect to casting announcements and found evidence that involvement of stars affects a movie's expected theatrical revenue based on trading behavior. However, she also notes that the top 3 movies all-time prior to 2000 contained no stars (at the time) – Star Wars, E.T., Titanic. This modeling is the first to consider a star's economic history, using the average box office revenues of a star's five previous movies at the time of casting announcement. The model also includes the entire cast's economic history for five most recent films, as well as the cumulative number of awards the cast has received.

Simonton (2009) found in a review of empirical research that most existing research focused on defining an outcome variable that falls into one of three categories, he terms as the "success triad": critical evaluations, financial performance, movie awards. Further, he reports that most research focuses on revenue and not profit, as profit a) cannot be determined without expenses and b) tends to be proprietary information. He also identifies the predictors typically found in

most research models as budget, rating, runtime, personnel, and season of release. He finds that the operational definition of the value of personnel is a key facet to reported findings.

Hadida (2009) performed a comprehensive review of empirical studies for motion picture performance between 1977-2006, finding historically, decisions in the motion picture industry have been largely intuitive. Academic research on a film's performance has grown dramatically in this time: between 1977-1987, there were 19 published works research motion picture performance; 12 such works were published in 2006 alone. She identifies certain challenges within developing appropriate models, namely that the processes responsible for producing a film are non-linear involving many different organizations and individuals. She identifies some of these stakeholders and variables as script, writer, producer, development, talent, budget, financing, production, time constraints, directors, distributors, and marketers.

Glotfelty (2012) employed a series of log-linear regression equations for individual countries using international box-office data and quantified the Star Power variables using IMDB STARmeter rankings. He found that the impact of Star Power on box office revenues is a function of whether one controls for BUDGET, further finding that ensemble casts (i.e., multiple stars) performs best, and that directors have little effect.

Lash & Zhao (2016) attempt to model the profit of films as opposed to their revenue defining profit as revenue minus budget, finding only 36% of films had revenues higher than their budget. The modeled the data as a classification problem using logistic regression, naïve Bayes, support vector machines, multilayer perceptron, decision trees, random forest, and the logitBoost. Star Power is measured as the total gross of films over an artist's career. They also introduce a tenure variable that represents the length of an artist's career in years.

Antipov & Pokryshevskay (2017) modeled pre-theatrical release data using Star Power and textual information about its MPAA rating and found a Random Forest model to outperform stepwise regression and multilayer perceptron neural network for 1672 movies between 1999-2012, adjusting dollar values for inflation. They also incorporate a variable that represents the intensity of competition at the time of a movie release by the number of other motion pictures released within two weeks before and after the start of a movie's run. Star Power is modeled as the sum of all actor and director previous revenues.

Quader, Gani, Chaki, and Ali (2017) proposed a decision-support system for movie investment using Support Vector Machines and Neural Network classification models and were able to achieve 84% predictive accuracy using profit as a dependent variable with five categories and calculating Star Power as the sum of income of all movies by actors and directors.

He & Hu (2021) developed a stacked ensemble model using Chinese movie data and calculating Star Power as the cumulative revenue of the combined movies of each star in the cast and incorporating post-release data such as first week box office revenue and used Random Forest, Extreme Random Tree, and Generalized Linear Models to make classification predictions about a film's success.

Ni, Dong, Zou & Li (2022) studied films released in China using a stacking ensemble model including extreme gradient boosting, light gradient boosting machine, categorical boosting, gradient boosting decision tree, random forest, and support vector regression, and introduced market information such as GNP. Their calculation of Star Power is the most comprehensive thus far, calculating the total number of movies an actor/director has been involved with the past 3- and 5- years, the average and total grosses of those films, for all movies, not just those an actor has appeared in as a leading actor. They achieved a Mean Absolute Percentage Error (MAPE) for their final model of 14.49%.

Hypothesis

We attempt to show that by including our calculation of star power with other known variables prior to a film's release (the point at which decisions that can materially affect a film's economic success are made, e.g., runtime, budget, rating, etc.), we can improve on the predictive accuracy of past studies modeling a film's box office revenue as an outcome using tree-based models.

Our hypothesis is that star power is an important predictor of a film's box office revenue.

The justification for this hypothesis is the basic heuristic that past success, represented by the star power of actors and directors, is the best predictor of future success.

The null hypothesis is that star power is not an important predictor of a film's box office revenue.

To measure the importance of star power variables, we build multiple tree-based regression models to develop the model that achieves the highest predictive accuracy. We then determine if the model selects one or more of the star power variables within the top 5 most important variables to its model and describe how such importance is measured.

We define a highly predictive model as one that achieves a Mean Absolute Percentage Error (MAPE) below 5%. More on how this calculation is obtained in the "Methodology" section below.

Data and Variables

Data was acquired using python from three sources: IMDb.com, the Internet Movie Database, owned by Amazon, BoxOfficeMojo.com, a subsidiary of IMDb which tracks box office revenue, and the Open Movie Database (OMDb), a RESTful web service containing information from IMDb.com. First, we deployed a web scraper to obtain unique IMDb IDs for the top 200 movies by domestic box office revenue per year from Box Office Mojo. Next, using these IDs, and removing duplicates (as some movies that were released late one year may appear in the following year's top earners as well), we scraped each film's IMDb page to extract production budget and unique IDs for their respective director(s) and leading actor(s). Finally, we queried the OMDb API by each unique ID, and parsed from JSON the following pieces of data for each film: Title, Year, Rating, Release Date,

Runtime, Genre, Director, Writer, Actors, Plot, Language, Country, Awards, Poster, MetaRatings, and Domestic Box Office Revenue.

Combining data from the three sources and removing columns that were not pertinent to this study, resulted in a master dataset containing the following pieces of information:

Title – The film's title.
Year – The year of release.
Country – The country of origin.
Language – The language(s) in which the film is available.
Rated – The film's MPAA rating.
Released – The film's release date.
Runtime – The film's length in minutes.
Genre – The film's genre.
Director(s) – The film's director(s).
Director ID – The director(s) IMDB ID.
Writer – The film's writer.
Actors – The film's leading actor(s) as reported by IMDB.
Actor ID – The actor(s) IMDB ID.
Box Office Revenue – The film's total box office revenue.
Production Budget – The film's production budget.

From this data, we wrote python scripts that used the Director IDs and Actor IDs to engineer the following features:

Director Films, 5 years – The director's total number of films, 5 years prior to release.
Director Films, 10 years – The director's total number of films, 10 years prior to release.
Director Revenue, 5 years – The director's total box office revenue, 5 years prior to release.
Director Revenue, 10 years – The director's total box office revenue, 10 years prior to release.
Actor Films, 5 years – The combined total of films in which the leading actors appear as leading actors, 5 years prior to release.
Actor Films, 10 years – The combined total of films in which the leading actors appear as leading actors, 10 years prior to release.
Actor Revenue, 5 years – The combined revenue of films in which the leading actors appear as leading actors, 5 years prior to release.
Actor Revenue, 10 years – The combined revenue of films in which the leading actors appear as leading actors, 10 years prior to release.
Number of Actor(s) – The number of leading actors appearing in the film.
Number of Director(s) – The number of directors of the film.

Some additional cleaning steps were necessary to prepare our dataset for modeling. First, we removed all films released prior to 2002, as the dataset contained some re-released classic films that were skewing our data. We derived the *Month of Release* from the *Release Date* variable. The

Country, *Language*, and *Genre* variables in many instances contained multiple values – for example, some genres contained fields such as "*Drama, Romance*" or "*Action, Adventure*". In these cases, we utilize the first given genre. The *Country* variable, likewise, contained every country in which a film had some element of production, with a few instances of no data provided at all. We converted this column to a three-class categorical variable, with the categories "United States", if the film contained the United States as one of its countries of origin, "Foreign", if it did not, and "None", if no information was provided. Similarly, for the *Language* variable, if the observation contained English as a language, we labeled it "English". If it did not, we labeled it "Foreign". Observations that did not contain a language were labeled "None". We observe that nearly all films in our dataset are in English (93.2%) or from the United States (87.1%). However, we retain these variables as we observe that the non-English and non-United States observations may provide useful information to our models, as we can see in Fig. 1 below, there is a clear distinction in the distribution of our response variable, box office revenue, between the distinct categories.

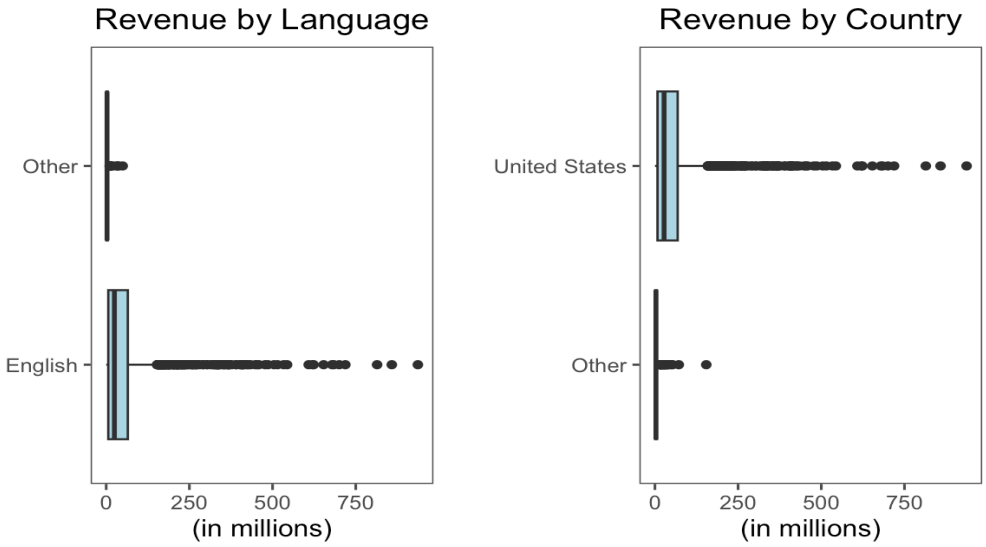

Figure 1. *Revenue Distribution by Language and Country.*

For our *Rated* variable, we observed that 91.6% of all films in our dataset were rated "R", "PG", or "PG-13", with 80% rated "R" or "PG-13" alone. We observe that a distinction between the ratings and the distribution of revenues as seen in Fig. 2. R-rated movies represent the largest percentage (39.7%) of films in our dataset, as well as the narrowest distribution of revenue, with a median revenue ($15,841,514) nearly half that of PG-13 films (37.4% of all films, median revenue: $32,015,231) and nearly a *third* of those rated PG (14.6% of films, $46,700,633). We observe that films not rated R, PG, or PG-13 may potentially contain valuable information for our model, so we leave those in. Action movies comprise the most frequent genre in our dataset (27.9%), followed by Comedy (18.9%) and Drama (15.9%). No other genre represents more than 10% of our dataset.

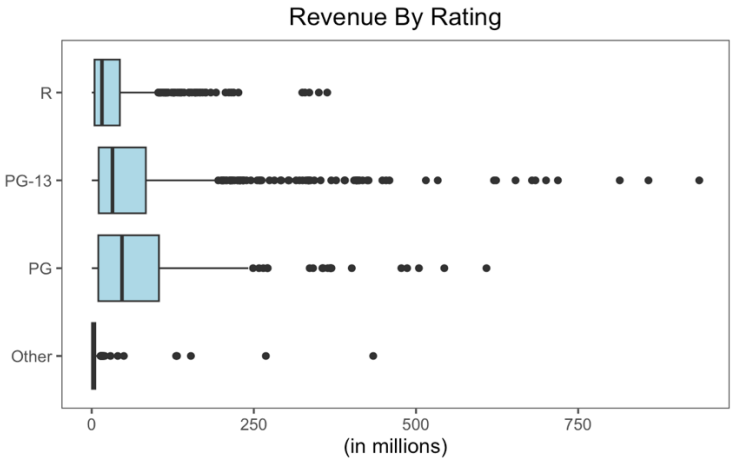

Figure 2. *Revenue Distribution by Rating.*

R-rated movies represent the largest percentage (39.7%) of films in our dataset, as well as the narrowest distribution of revenue, with a median revenue ($15,841,514) nearly half that of PG-13 films (37.4% of all films, median revenue: $32,015,231) and nearly a *third* of those rated PG (14.6% of films, $46,700,633). We observe that films not rated R, PG, or PG-13 may potentially contain valuable information for our model, so we leave those in. Action movies comprise the most frequent genre in our dataset (27.9%), followed by Comedy (18.9%) and Drama (15.9%). No other genre represents more than 10% of our dataset.

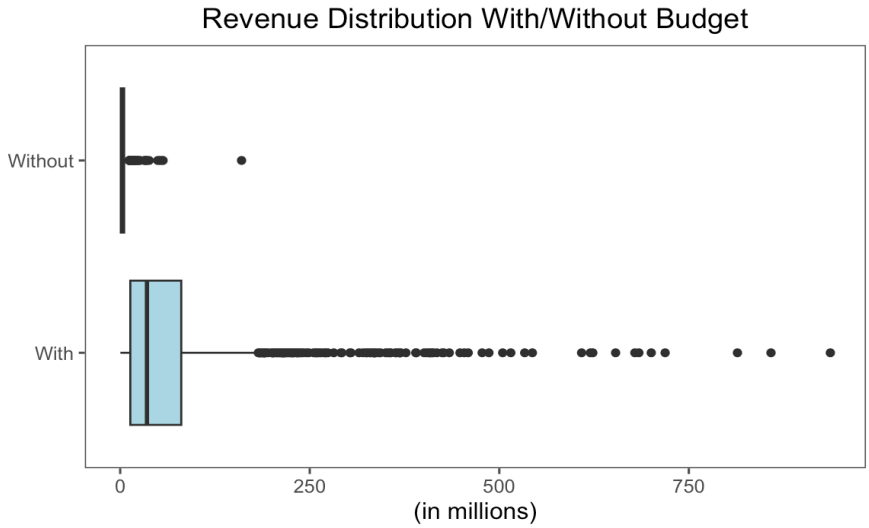

Figure 3. *Distribution of Revenue for missing and non-missing Budget Data.*

Missing values are observed for the budget variable in our data, with 24% of our observations absent of values for this variable. This is to be expected, as budget figures are not reported for every film. However, a boxplot indicates that this missingness might be informative for our model (Fig. 3), so we consider this data missing not at random (MNAR) and do not discard these observations. In fact, tree-based models are good at dealing with missing data.

	Min	Q1	Median	Mean	Q3	Std.Dev	Max
actor_films10	0	2	8	10.09	16	9.28	44
actor_films5	0	1	4	5.45	9	5.16	27
actor_rev10	0	48091713	450255348	759167254.42	1168551880	897987469.29	7316416397
actor_rev5	0	10726630	213778031	432617771.15	644587426	571830292.26	4469634688
boxoffice	159014	4936819	20777061	54189922.78	60311495	92859139.66	936662225
budget	100000	12000000	30000000	52157266.12	68000000	58391659.06	356000000
director_films10	0	0	1	1.53	2	1.99	14
director_films5	0	0	0	0.76	1	1.07	9
director_rev10	0	0	12727256	124370528.27	150394119	230007052.55	2693332806
director_rev5	0	0	0	61377871.21	65565279	135630420.74	2173799662
runtime	39	97	108	111.27	122	19.84	195

Figure 4. *Summary Statistics.*

We observe amongst our numeric predictors (Fig. 4) that only the *runtime* variable demonstrates anything approximating a normal distribution. The remaining variables are highly right-skewed, with most of observations concentrated at the lower end of the numeric distribution. The means for each of the numeric variables are greater than the medians. This makes sense, as there are fewer higher-earning movies, while those that do earn highly will have an outsized effect on the means for all revenue columns. We also see some interesting data in terms of the number of recent films by actors and directors that may have an impact on our modeling: directors are involved in fewer films than actors – we see that the median number of films for directors in the past 5 years is less than 1. In fact, 53.3% of films in our dataset were made by directors with no films in the previous 5 years that had placed in the top 200 domestic box office earners. This may be an indication that the *revenue* generated by a director is a greater predictor than the number of films to their credit.

Plotting the numeric variables against domestic box office revenue demonstrated no visible patterns. In fact, the only variable that appear to have a linear relationship, although faint, with box office revenue is the *budget* variable. Again, this makes sense. We assume that films with higher budgets will generally have higher returns. Lastly, before moving on to modeling, we reduce our dataset to only films released after 2012, leaving us with a dataset of 1859 observations.

Statistical Methods

This study uses tree-based models to predict box-office revenues. Tree-based models have high predictive accuracy, can handle categorical and continuous variables, do not require linear relationships predictors and the dependent variable, and do not make assumptions about the underlying data distribution of the predictors. However, tree models are susceptible to overfitting and bias. To address these shortcomings we utilize ensemble methods, which combine and weight the predictions of multiple models to improve generalizability. Tree-based ensemble methods take a variety of approaches to building individual decision trees and are often good

choices for predicting a continuous outcome, such as box office revenue, in situations where the predictors are a mix of categorical and continuous data, the relationships between the predictors and the response are non-linear, and the variables are not normally distributed. These models are less sensitive to outliers as well, recursively splitting data relative to their distribution, rather than by mean or variance. As such, no assumptions are made about data distribution.

In this study, we deploy bagged trees, random forests, extreme gradient boosting, and cubist models to generate predictions of our continuous response variable and measure their performance. Each model has its own benefits and drawbacks, which we survey below.

Bagged tree models create multiple decision trees by bootstrapping (sampling with replacement) samples from the training dataset, with each sample the same size as the original training set, resulting in different training sets for each sample. Multiple decision trees are trained on each sample using all features, resulting in a collection of diverse trees, the predictions of which are then aggregated and averaged, reducing variance and overfitting. K-fold cross-validation is used to assess the model's performance and make hyperparameter tuning decisions. A benefit of this approach is that the processing can be parallelized, improving computational efficiency. While bagged tree models can reduce overfitting, since each tree uses all features from the sample, appropriate pruning or limiting the depth of the individual trees is still necessary to prevent overfitting. And while the (B)ootstrap (Agg)regation (hence, BAGGing) approach diversifies the data on which individual trees are being built, information loss is a consequence of this diversification, meaning each tree will not necessarily contain the maximum amount of information available in the dataset. This can be more pronounced with smaller datasets, such as the data in this study. Variable importance scores are calculated using an out-of-the-bag (OOB) estimation technique, in which the model measures the decrease in prediction accuracy when each variable is randomly permuted across samples of unused data from each bootstrap sample, known as out-of-the-bag data. Each predictor is assigned a score based on the magnitude of the increase in OOB error. (James, et al. 2013)

Random forests extend the bagged tree approach, bootstrapping the original dataset and averaging the predictions, but rather than using every feature of each sample, random features are selected (hence, random forest) at each split of each decision tree, and recursive splits are made to minimize the mean squared error of each feature and split. This creates trees with potentially different sets of important features, reducing the dependence on pruning to prevent overfitting. K-fold cross-validation is used along with grid search to evaluate the model and select the best hyperparameters, and while random forests can be parallelized like bagged trees, cross-validation and hyperparameter tuning can quickly escalate the computational expense associated with these models. Variable importance scores are calculated by measuring the decrease in prediction accuracy for each predictor, normalized to 1, with higher scores indicating greater importance. (James, et al. 2013)

Extreme Gradient Boosting (XGBoost) takes a different approach to decision tree building than bagged trees or random forests, using information from previously grown trees to correct errors made by the ensemble's current predictions. (James, Witten, Hastie & Tibshirani, 2013) These

models build a series of “weak learner” decision trees, or stumps – trees that consist of single features split in two, sequentially, in which each stump selects the feature that minimizes the mean squared error between predicted and actual values by computing the gradient of this function. The initial predictions are generally the mean value of the response variable, while each subsequent stump uses the errors from the previous iteration as the target variable. In so doing, the algorithm sequentially selects features that minimize the error. XGBoost uses regularization techniques to penalize large coefficients to prevent overfitting and improve generalizability. A learning rate is used to control the contribution of each tree to the final prediction. A lower learning improves the accuracy of the model’s predictions but increase computational expense. Cross-validation and grid search can be used to find the model’s optimal hyperparameters and ensure generalizability. These models can be highly predictive. Variable importance is determined by the number of times a variable is used to split the data across all trees in the ensemble combined with a “gain” metric indicating the improvement in model accuracy when each variable is used for splitting. (Kuhn & Johnson, 2019)

Cubist models use decision trees to generate a set of rules to describe the relationship between features and response, with each rule corresponding to a specific tree. These rules are used to segment the data, ensuring that each rule applies to a specific subset of the dataset. Linear models are generated at each split of each tree, smoothed by the predictions of the linear model in the previous split, resulting in a linear model at the terminal node of each tree that represents a smoothed model of each split in the tree, effectively representing each rule as a smoothed linear model that is eventually combined across segments to create a unified model. Multiple individual models known as committees are trained on different random bootstraps of the data that consider the result of each previous model fit, generating a new set of rules, and averaging the predictions of each rule. This approach provides high generalization but at the expense of computational efficiency. Variable importance is represented as a linear combination of the usage in the final rule conditions and the model. (Kuhn & Johnson, 2019)

To assess the performance of each model, the data is shuffled and split into training and testing sets, using an 80/20 split (80% for training, 20% for testing). 10-fold cross-validation is used on the training set to estimate each model’s performance over multiple subsets of the training data. This approach improves model generalizability on unseen data by reducing variance, preventing overfitting, and allowing for optimal hyperparameter tuning. Splitting our data into training and testing sets resulted in training and testing sets of 1378 and 344 observations, respectively.

Each model was trained and tested in R using the caret library (Kuhn, 2008) alongside the necessary complementary packages. Grid search was performed to obtain the optimal hyperparameters for the Random Forest, XGBoost, and Cubist models. The Bagged CART algorithm used (“treebag”) does not allow for hyperparameter tuning.

Model accuracy is measured by calculating the Mean Absolute Percentage Error (MAPE) of the predictions made by each model against the validation data. MAPE measures the absolute percentage difference between predicted and actual values, summing all errors and taking the

mean across all data points. MAPE is expressed as a percentage, with 0% indicating a perfect prediction.

Findings & Discussion

We began our modeling by testing and comparing approaches to deal with the missing budget data – 333 observations contained missing data for this variable. First, we attempted no imputation to determine if each algorithm could inherently account for the missing data: none could. As the data appears to be Missing Not at Random (MNAR), meaning there is some reason the data was not provided, and the missingness has some informative value based on the boxplot of budget vs. revenue, we tried the following approaches to deal with the missing data. First, we dropped the missing data. Next, we encoded a categorical variable, budget provided, to store whether a budget was provided ("Yes") or not ("No). This is an important step because without it, we lose a degree of information in the model. Next, we tried two imputation methods: first, we used K-nearest neighbors to impute the missing values, setting K to 5. Second, as the minimum budget in the dataset is 100,000, we set all missing budget values to zero. Lastly, we discretized the budget variable by its interquartile range, with missing values as a fifth category. Of these, replacing missing values with zero along with the categorical "budget provided" variable yielded the most accurate results in terms of MAPE. However, our overall models were highly inaccurate.

The reason for this is that while tree-based models are not explicitly based on linear relationships, the individual decision trees in ensemble methods often make decisions based on linear splits. In this way, the assumption of normality in the response variable distribution can affect the performance and accuracy of tree-based models. As we observed, the distribution of the box office revenue data to be data to be highly right-skewed.

Each model was trained and tested using the original response variable distribution, with no model achieving a best MAPE of less than 200, and at worst, well over 400. We applied the Box-Cox transformation in R to the response variable (box office revenue) to satisfy the assumption of normality. The Box-Cox transformation is a family of power transformations which finds a value, lambda, that maximizes the log-likelihood of the transformed data to achieve normality. For the box office revenue variable, the Box-Cox function selected a lambda of 0.1, which raises to the original values to the power of 0.1. In order to restore final predictions to their original value scales, an inverse Box-Cox transformation using the selected lambda must be performed on the predictions.

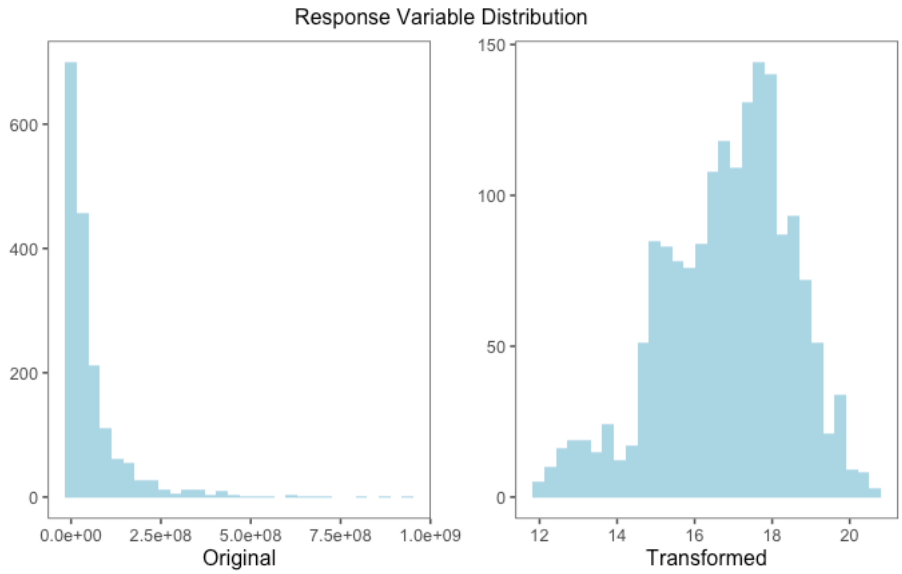

Figure 5. Response variable distribution, before and after Box-Cox transformation.

We ran each model using the Box-Cox transformed response variable, and the resulting MAPEs improved to between 5.4 and 9, depending on which approach was used for handling the missing budget data. Zero imputation and a categorical variable for budget availability yielded the highest accuracy amongst imputation methods. Lastly, for each model, we observed 4 near zero variance variables: the number of actors, the number of directors, language = None, and country = None. We tested each model with and without these variables and observed an improvement in accuracy when we removed the number of actors and the number of directors. However, the models were less accurate when we removed the language = None and country = None observations, so we move forward with these features included.

With our final feature engineering in place, we trained and tested five "focus" versions of each model comprising specific predictor sets which we outline below. In addition to these "focus" versions, models were trained and tested withholding specific star power variables based on variable importance to determine if model accuracy would improve – except for the five "focus" models, withholding any predictors from the model produced inferior model accuracy that does not provide meaningful contribution to this study, so we "focus" our discussion on the five models below.

All models were trained and tested with and without feature scaling (reducing the range of each feature by dividing each value by its standard deviation resulting in a standard deviation of 1) and observed that feature scaling improved speed but with no change in accuracy. Centering was not employed for these models as all values are greater than zero. The five versions of each model trained and tested were as follows:

1. A baseline version of each model ("BASE"), consisting of all prior known variables except the star power variables to gauge the accuracy of each algorithm using the available data before introducing Star Power.

2. The full dataset ("FULL") including all star power variables. The accuracy of each model improved after introducing the star power variables.
3. All prior known variables except the 10-year Star Power variables ("SP-5") to compare model accuracy between the 5- and 10- year Star Power variables.
4. All prior known variables except the 5-year Star Power variables ("SP-10").
5. A recalculated 10-year star power variable, containing the difference between the 10-year and 5-year variables, to represent the values between years 5-10.

Grid search on the "FULL" data yielded the optimal parameters used for subsequent modeling presented in Figure 6.

PARAMETER	DETAIL	DEFAULT	RF	XGBOOST	CUBIST
mtry	Number of predictors at each split yielding the smallest RMSE	# predictors / 3	21		
eta	Shrinkage/Learning Rate	0.3		0.05	
max_depth	Maximum Tree Depth / Controls model complexity	6		3	
colsample_bytree	Fraction of features to be randomly selected per tree	1		0.6	
nrounds	Number of trees to build	100		150	
gamma	Minimum Loss Reduction / Regularization penalty	0		0	
min_child_weight	Minimum Sum of Instance Weight	1		10	
subsample	Fraction of training data to be sampled at each iteration	1		1	
committees	Number of rulesets	0			60
neighbors	Nearest neighbors to consider	0			0

Figure 6. Tuning Hyperparameters

The *mtry* parameter selected for the Random Forest model is rather large; the default parameter is the square root of the number of predictors. Allowing for each categorical variable, our model contains a total of 40 predictors (sqrt = 6.3). The value of *mtry* affects the trade-off between the individual tree correlations and the overall accuracy of the model. Smaller values may lead to less correlated trees but may be too shallow and lack accuracy, whereas larger values may have correlation between trees but higher accuracy. (Kuhn, 2008)

For the XGBoost model, an *eta* (learning rate) of 0.05 represents a lower learning rate that reduces the impact of each individual tree that requires more iterations to reach optimal performance and helps prevent overfitting and improves generalization. A *max_depth* of 3 limits each tree in

the model to a depth of 3, resulting in simpler trees that improve generalization. The *colsample_bytree* introduces randomness by using a random subset of 60% of the features for each tree in the ensemble, reducing the likelihood of trees relying too heavily on a specific subset of features and helps overfitting. Grid search produced an optimal *nrounds* parameter of 150 trees, slightly larger than the default setting, which may allow the model to capture more complex patterns but risks overfitting the data. No *gamma,* or minimum loss reduction, was used, so splits are made whenever they result in a positive gain. The *min_child_weight* parameter represents the minimum sum of instance weight for a child node to split. Larger values result in smaller, simpler trees. The optimal *subsample* (1) is the entire dataset.

The optimal Cubist model hyperparameters were 60 *committees* (see: Statistical Methods) and 0 *neighbors*.

A table containing the results of each model is presented in figure 7.

Modeling Results

MODEL	BASE	FULL	SP-5	SP-10	SP-ADJ
Bagged Trees	5.429	5.353	5.363	5.323	5.367
Random Forest	5.200	4.949	4.999	4.923	4.947
XGBoost	5.249	5.085	5.100	5.121	5.184
Cubist	5.254	5.137	5.138	5.142	5.162

Figure 7. MAPE per model

We constructed the full models for each tuned algorithm and examined the most important variables to identify possible feature selection for model optimization. Each algorithm identifies its most important variables in different ways: the Bagged Trees and Random Forest algorithms used measure variable importance as the average decrease in Mean Squared Error (MSE) when a predictor is excluded, suggesting the predictor's relevance in explaining the variability in the response variable. The XGBoost algorithm calculates variable importance as the number of times a variable is used to split the data across all trees in the ensemble. Variable importance in the Cubist algorithm is a linear combination of the percentage of times where each variable was used in a rule or linear model. (Kuhn, 2008)

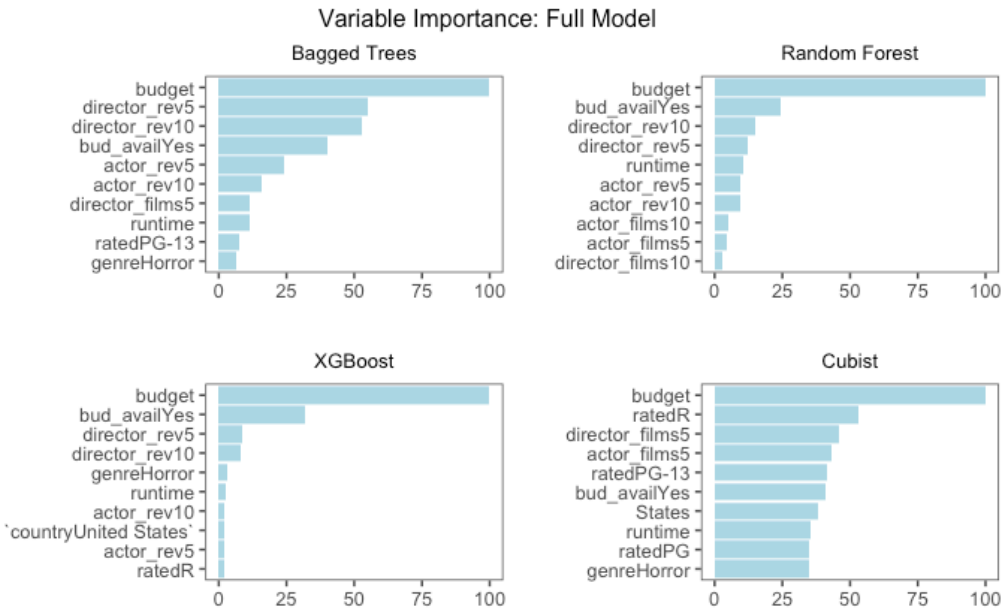

Figure 8. Variable importance measures, full model by algorithm

Each algorithm identified budget as the most important variable to its model (Figure 8), with the subsequent variables identified, and the extent to which importance decays varying between algorithms. In the full model, the Bagged Trees, Random Forest, and XGBoost algorithm each selected the 5- and 10- year director revenue variables amongst the top 5 most important variables, while neither of those variables appear in the Cubist algorithm top 10. Surprisingly, no actor or director revenue variables appear in the Cubist model's top 10. The Random Forest and XGBoost models exhibit the largest decrease between variables, from 100 to 24 and 100 to 31, respectively. This is a possible indication of redundancy or collinearity amongst the lower importance predictors, consistent with the observed collinearity. It can also be an indication that the algorithms are relying on specific interactions involving the top variables. We observed consistent improvement in the accuracy of each algorithm when the star power variables are added to the base model.

We compared the variable importance results from the full model and observed how each model performed when only the 5- or 10- year star power was included. We considered that the 10-year variables included information already contained in the 5-year variables, so we recalculated the 10-year variables to represent only the difference between years 5-10. We observed that all model accuracy decreased when presented with only the 5 year data ("SP-5"), while results for the 10 year data models ("SP-10") was split: the Bagged Trees and Random Forest SP-10 models were the most accurate of all models, with the Random Forest version producing a MAPE of 4.923, representing the most accurate model produced, while the XGBoost and Cubist models were less accurate when any star power variables were removed from the full data. Surprisingly, the recalculated 10-year star power ("SP-ADJ") performed no better (Random Forest) or worse than each algorithm, which can be attributable to some information loss occurring in the transformation, consistent with what we observed removing predictors in the non-focus versions of each model.

Random Forest models outperformed each algorithm in every version, consistent with the findings of Lash & Zhao (2016) and Antipov & Pokryshevskay (2017), suggesting that the randomization of feature selection subsets at each split produces the most effective variance reduction for this dataset amongst the algorithms used. This approach tends to be robust to outliers and noise, which are present in this data and that the XGBoost algorithm may not be handling as well. Additionally, the Cubist incorporation of linear modeling may be failing to capture the complex non-linear relationships in the data. Using MAPE as the accuracy metric, our best model, 4.923%, outperforms Ni, Dong, Zou & Li (2022), 14.49%.

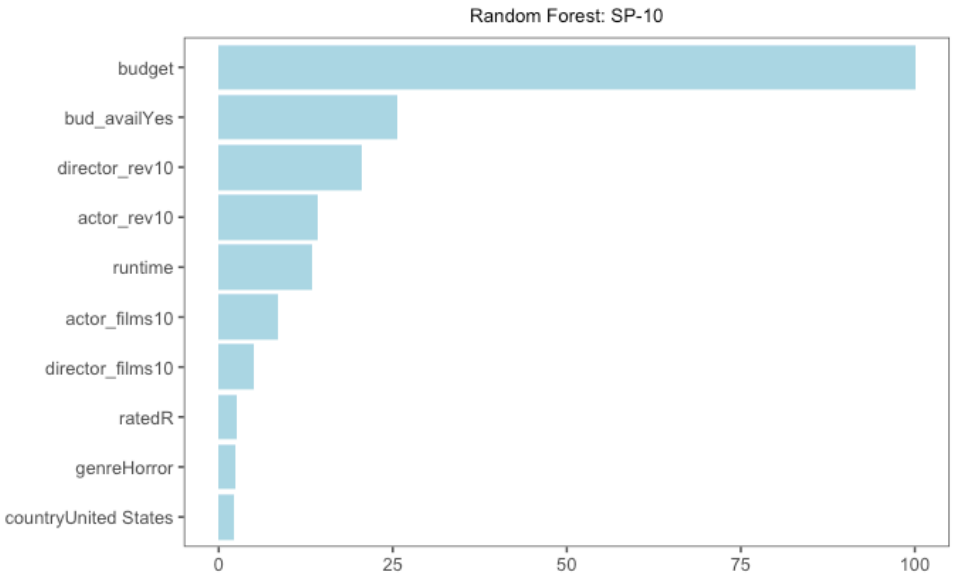

Figure 9. Variable importance, "best" Random Forest model (SP-10)

The most accurate Random Forest model, SP-10, selects the four "Star Power" variables (10- year actor and director revenues, 10- year actor and director total films) amongst the top 7 most important variables to its model (Figure 9), so we reject the null hypothesis that "Star Power" is not an important predictor of a film's revenue.

We observe from our modeling that a film's budget is by far the most important predictor of box office revenue. After all, any measurable Star Power will to some degree contribute to budget by way of an actor or director's salary; the more successful a leading actor's past films have been, the higher the salary. The same goes for directors. However, we can see by a simplified model containing the total films and revenues of all leading actors and the director of the film for the prior 10- year period to the film's release can improve the predictive accuracy of a model over the inclusion of shorter-term 5-year data to achieve a best Mean Absolute Percentage Error of less than 4.923%.

Conclusion

Studies designed to predict the box office revenue of films over the past 40 years have deployed a variety of statistical approaches as we surveyed in our literature review. In this study, we developed a novel approach to measuring the Star Power of leading actors and the director of a film as predictive features in tree-based machine learning models using the cumulative prior 5- and 10-year revenue totals of the films of leading actors and director and found that a Random Forest model utilizing 10-year Star Power metrics produces the most accurate predictions of box office revenue amongst the algorithms used, achieving a Mean Absolute Percentage Error (MAPE) of 4.923%. Future studies may focus on developing or engineering new supplemental star power measures to quantify the predictive value of leading actors and directors, or perhaps combining our measures not included in our study, such as the power of writers or producers based on prior revenue of their films. By modeling our data using prior-known variables, this approach provides actionable insight into film production that can be employed industry-wide to promote more reliable box office revenue outcomes.

References

1. Goldman, & Goldman, W. (1984). Adventures in the screen trade: a personal view of Hollywood and screenwriting (First trade edition.). Grand Central Publishing
2. Litman, & Kohl, L. S. (1989). Predicting financial success of motion pictures: The ’80s experience. *Journal of Media Economics*, *2*(2), 35–50.
3. Sochay. (1994). Predicting the Performance of Motion Pictures. *Journal of Media Economics*, *7*(4), 1–20. https://doi.org/10.1207/s15327736me0704_1
4. Sawhney, & Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science (Providence, R.I.)*, *15*(2), 113–131. https://doi.org/10.1287/mksc.15.2.113
5. Ravid. (1999). Information, Blockbusters, and Stars: A Study of the Film Industry. *The Journal of Business (Chicago, Ill.)*, *72*(4), 463–492. https://doi.org/10.1086/20962
6. Simonoff & Sparrow (2000) Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers, CHANCE, 13:3, 15-24, DOI: 10.1080/09332480.2000.10542216
7. Terry, Butler, M., & De’Armond, D. (2003). Determinants of the Box Office Performance of Motion Pictures. Allied Academies International Conference. Academy of Marketing Studies. Proceedings, 8(2), 23–.
8. Eliashberg, Elberse, A., & Leenders, M. A. (2006). The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions. *Marketing Science*, *25*(6), 638–661. https://doi.org/10.1287/mksc.1050.0177
9. Sharda, & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, 30(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018
10. Elberse. (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

11. Kuhn, Max (2008). “Building Predictive Models in R Using the caret Package.” *Journal of Statistical Software*, 28(5), 1–26. doi:10.18637/jss.v028.i05, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.
12. Simonton. (2009). Cinematic success criteria and their predictors: The art and business of the film industry. Psychology & Marketing, 26(5), 400–420. https://doi.org/10.1002/mar.20280
13. Hadida. (2009). Motion picture performance: A review and research agenda. International Journal of Management Reviews: IJMR, 11(3), 297–335. https://doi.org/10.1111/j.1468-2370.2008.00240.x
14. Delen, D., & Sharda, R. (2012). Forecasting Financial Success of Hollywood Movies - A Comparative Analysis of Machine Learning Methods. International Conference on Informatics in Control, Automation and Robotics.
15. Nelson, R. & Glotfelty, R. (2012). Movie stars and box office revenues: an empirical analysis. Journal of Cultural Economics, 36(2), 141–166. https://doi.org/10.1007/s10824-012-9159-5
16. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (2nd ed.). New York, NY: Springer.
17. Lash, & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. https://doi.org/10.1080/07421222.2016.1243969
18. Antipov, & Pokryshevskaya, E. B. (2017). Are box office revenues equally unpredictable for all movies? Evidence from a Random forest-based model. Journal of Revenue and Pricing Management, 16(3), 295–307. https://doi.org/10.1057/s41272-016-0072-y
19. Quader, Gani, M. O., Chaki, D., & Ali, M. H. (2017). A machine learning approach to predict movie box-office success. 2017 20th International Conference of Computer and Information Technology (ICCIT), 1–7. https://doi.org/10.1109/ICCITECHN.2017.8281839
20. Kuhn, M., & Johnson, K. (2019). *Applied predictive modeling*. Springer.
21. Nwanganga, F., & Chapple, M. (2020). *Practical machine learning in R*. Nashville, TN: John Wiley & Sons.
22. Wang, Zhang, J., Ji, S., Meng, C., Li, T., & Zheng, Y. (2020). Predicting and ranking box office revenue of movies based on big data. Information Fusion, 60, 25–40.
23. Siegel, T. (2020). Warner Bros. signs deal for AI-driven film management system, *Hollywood Reporter*, 2020. https://www.hollywoodreporter.com/business/business-news/warner-bros-signs-deal-ai-driven-film-management-system-1268036/
24. He, & Hu, B. (2021). Research on the Influencing Factors of Film Consumption and Box Office Forecast in the Digital Era: Based on the Perspective of Machine Learning and Model Integration. Wireless Communications and Mobile Computing, 2021, 1–10. https://doi.org/10.1155/2021/6094924
25. Ni, Dong, F., Zou, M., & Li, W. (2022). Movie Box Office Prediction Based on Multi-Model Ensembles. Information (Basel), 13(6), 299–. https://doi.org/10.3390/info13060299
26. Zane, M. (2023). Zippia. "25+ Striking U.S. Film Industry Statistics [2023]: Facts About the Video Production Industry in The U.S." *Zippia.com*. Jun. 14, 2023, <https://www.zippia.com/advice/us-film-industry-statistics/>
27. *Press Room – IMDb Statistics* (2023). Imdb.com, as of June 2023. https://www.imdb.com/pressroom/stats/