Multimedia Tools and Applications (2023) 82:38397–38420
https://doi.org/10.1007/s11042-023-15169-4

# Revisiting predictions of movie economic success: random Forest applied to profits

Thaís Luiza Donega e Souza [1] · Marislei Nishijima [2] · Ricardo Pires [3]

Received: 24 August 2021 / Revised: 22 March 2022 / Accepted: 22 March 2023 /
Published online: 28 March 2023


## Abstract

Previous studies have employed machine learning tools to classify films according to success to guide a reduction in the degree of uncertainty of film production. We revisited the literature to contribute to three relevant issues in classifying films according to economic success. First, we explored the differences between the results of the shortest or longest samples in terms of time to study possible changes in patterns of consumption mainly due to technological changes and between total and wide-released films. Second, we used profits free of price inflation as measures of economic success instead of the usual box office nominal revenues. Third, we employed a smaller set of features, only the ones available at the time of production, to help producers maneuver contingencies since little or nothing can be done by the time a film is in the theaters. We followed the literature to choose the classifiers - Random Forest, Support Vector Machine, and Neural Network - and designed sub-datasets to model and compare the performance of our results. Our dataset includes all films with budgets disclosed at the Box Office Mojo website, resulting in 3167 movies released at theaters worldwide between 1980 and 2019. The Random Forest results outperform previous similar studies with different sampling in time, including results for a less usual larger sample, with the best data sample about 97% both in accuracy and F1-score.

✉ Thaís Luiza Donega e Souza
thais.donega@usp.br

Marislei Nishijima
marislei@usp.br

Ricardo Pires
ricardo_pires@ifsp.edu.br

[1] Information Systems Department, University of São Paulo, 1000, Arlindo Béttio – Ermelino Matarazzo, 03828-000, Room: L1 – 327, São Paulo, SP, Brazil

[2] University of São Paulo, Institute of International Relationships, Av. Prof. Lúcio Martins Rodrigues, Tv. 4 e 5, Cidade Universitária, São Paulo, SP 05508-020, Brazil

[3] Department of Electricity, Federal Institute of São Paulo, R. Pedro Vicente, 625 - Canindé, São Paulo, SP 01109-010, Brazil

**Keywords** Movie success · Machine learning · Classification · Movie market · Profit · Regime change

## 1 Introduction

The film industry has been responsible for about 30% of the total revenue of films since the 2000s [56] and reflects a film's economic success, due to the significant consumption of related goods during and after the release and of the film being consumed with other complementary goods, TV, cable, and others. According to the Motion Picture Association of America, in 2019, ticket revenues alone in the US and Canada were around $11.4 billion, while 76% of their populations could be classified as moviegoers [54]. At the same time, the motion picture and television industries support more than 2.5 million jobs in the United States.[1]

According to economic theory, film is both an information and an experience good. As an information good, it has a high fixed cost (actors, directors, editors, and others) and almost zero reproduction (marginal) cost [70]. As an experience good, its quality is not known until the time of consumption, which explains the uncertainty in its production [64]. These characteristics and recent technological changes make it difficult for an entrepreneur to know in advance whether a new film will be successful as an economic venture [3].

The rapid growth of the Internet and digitization, led by technological innovations in information and communication technologies (ICTs), has reduced production and distribution costs, creating a golden age for creative economic endeavors, such as information goods like music, movies, and books [71]. For example, today, a film can be consumed on any device with Internet access, such as mobile phones and tablets.

In addition, there are multiple substitute ways to consume a movie since it can be watched at home or virtually in any place and at any time just after the theater release, or simultaneously in some situations. Specifically, films' concurrency at movie theaters has increased due to Internet downloads and online streaming platforms [35, 71]. The same ICT development that allowed the reduction in film costs also incentivizes other markets to establish concurrence. Netflix, for example, is using consumer data and artificial intelligence to target consumption tastes to maximize its returns.[2]

Given the effects of these new technologies and the high risk of film production [35, 48, 65], we employ a decision support system to produce guidelines for film producers and their stakeholders such as studios distributors, and their shareholders. A film is a risky endeavor since it is very expensive to produce - including expenses for actors, directors, and marketing among others - and may not find enough viewers to pay for itself. In this sense, an application that allows and indicates how producers can change decisions like budget, distributors, and film duration among others, can reduce the risk of not being profitable. Such a tool, thus, can prevent heavy losses and improve productivity. To try to obtain this tool, we revisit the

[1] According to Motion Picture Association, in 2019 the film and television industry supported 2.5 million jobs, payed out $188 billion in total wages, and comprised over 93,000 businesses in the US alone: https://www.motionpictures.org/what-we-do/driving-economic-growth/ (Last accessed: 09/10/2020)

[2] Netflix do not disclose their rentals or all techniques behind their system recommendation tools, but these articles can give an overview https://insidebigdata.com/2018/01/20/netflix-uses-big-data-drive-success/ (Last accessed: 09/08/2021), https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/ (Last accessed: 10/08/2021) and https://netflixtechblog.com/supporting-content-decision-makers-with-machine-learning-995b7b76006f (Last accessed: 16/08/2021)

literature and focus on at least three main issues to contribute to and improve the performance of the previous studies regarding whether a film will have enough consumers to make it profitable.

First, we follow the economic literature to sample our procedures according to short and long periods to deal with a potential change in the "regime" that models a film's profits. The model that describes film success can change due to ICT evolution over time, which can be seen as an exogenous shock on the model's parameters [10]. In this sense, innovations occurring in a sector could change the model that generates the best classification/prediction, and just increasing the number of observations by using the information of the distant time, as is usual in the case of films, will not necessarily improve accuracy. In this sense, using small samples (near specific date) could produce a more homogeneous sample or one free of outliers. We also explore results using only wide-released film samples since they are more similar (a wide-released movie is very different from a limited-released one in costs, consumers, and so on), thus creating a more homogeneous sample [22].

Second, we measure a film's success based on its profit deflated by the CPI (the US Consumer Price Index). Using profit as a success measure allows us to account for revenues and costs of production since even a colossal box office cannot be profitable if the costs of production are also high. The literature mainly uses total revenues as a measure of economic success and does not control for the effects of inflation (at least, they do not explicitly mention them). Not correcting for the effects of inflation may lead to inaccurate classification of success since the more recent films have higher profits and revenues in current values. Following the still scarce literature, we investigate two measures of success based on profits at theaters as a measure of success in two experiments. The first is a binary measure of film profits, where we consider box office revenues and costs of production (budget) to account for the film's success; the second is a 6-class classification in profit ranges to be closer to reality and more directly comparable with the literature.

Third, we evaluate whether economic success in the theatrical film market can be predicted by a small set of readily observable features available after the film's financial plan and the green lights, that is, before or at the time of film production and release [22]. The literature, on the other hand, tends to not take into account the timing when the features are available, employing features indistinctly observed before and after a film release – such as critic reviews, consumer reviews, the time a film is kept on screens – and there is no room to change features to get better results before a film release. Using variables available at the time of production allows the producers to have a higher degree of freedom in timing to control investment decisions [76].

In connection with our first contribution, we employ a uniquely configured set of data according to the shortest or longest sample in time, and total and wide-released films, to the three most popular machine-learning (ML) algorithms – Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN). The results allow us to properly compare the performance of the methods and datasets with the existing literature. From our knowledge, the three issues addressed jointly, as we propose, offer an additional contribution to the literature.

Employing a dataset scraped from the Box Office Mojo and IMDB sites and features available before a film release, we get about 96% and 97% accuracy and F1-score, respectively, in binary break-even (BE) classification and about 90% of Average Percent Hit Rate (APHR) for profit ranges (PR). Our results indicate an improvement in accuracy compared with the literature. Moreover, the results are more compelling – considering the use of a stricter measure for film economic success (the profits in a constant/deflated dollar value) and a

reduced set of features – and more reliable due to the several tests with different numbers of observations and cuts in time. Finally, the results suggest that our models produce a better performance than those in the literature to date, indicating that our small number of features were appropriately chosen and that RF may be a better tool for predictions of movie profitability.

Given the limited number of movies released per year, the increase in sample size implies an increase in the time window in numbers of years. This, however, means the possibility of ignoring shocks that change the conditions of consumption of films each year.

Therefore, using a larger sample in time should be explored with caution similar to the econometric literature, which can open a new agenda for future studies. The literature of ML applied to film success suggests a trend; small datasets (few years), contrary to expectations, perform as well as or even better than larger datasets (more observations based on longer periods) to classify films by economic success, and our results support this conclusion. We attribute these results to a possible change of the "regime" that could drive the economic performance in theaters. In this sense, technological innovations, changes in individuals' preferences, and other shocks, as COVID-19, could cause these regime changes;

Following this introduction, Section 2 summarizes the literature; Section 3 presents our data and methodological strategy; Section 4 comprises our results and discussion, while the last section summarizes our main findings.

## 2 Literature on movie success

Due to the uncertain returns of films, many scholars have attempted to predict the economic success of a film at theaters aiming to guide producers, studios, distributors, and theater chains. Most of these studies are explanatory, investigating factors and their relations with movie box office performance through regression analysis, and have been published in different fields: Economy [11, 16, 25, 28, 37, 45, 58, 66], Business and Information [41, 52, 55], Marketing [14, 21, 44, 53] and Computer Science [2, 6, 19, 51, 69].

Recent ICT developments have reduced the costs of producing films and increased the number of films produced, and this has resulted in more film data available. These data and new computational methods have increased the number of studies predicting movies' success [71]. Most of these ML studies use features available along the whole movie lifecycle to predict a film's success. Yet, as the greater part of data is available only after a film release, most of the studies use these data to predict success. In this case, however, there is no room to change film production decisions.

In this sense, the literature on predicting movie success usually employs post-release features like critic reviews, ratings, nominations, awards, other forms of word-of-mouth (WOM), and awareness information [18]. For example, studies employ social media microblogging to forecast box office revenues using ML in China [63] and the Korean market [34]. There are also similar studies using other methods of classification. For example, one study uses online user reviews applied to Support Vector Machine Regression (SVR) to predict box office revenues according to the genre [33]. Another involves text mining on Twitter to get insights on customer preferences to predict box office revenues with CART and NN regression [47], both in the US movie market. Some authors transform movie box office predictions into a classification problem [23, 38]; in particular, these authors also employ user opinion mining. For example, a study uses critic ratings and visual elements from movie

posters, besides other movie metadata, to classify film success employing deep NN for 6-class box office prediction [78]. Another study uses data extracted from visual elements in trailers and text features from film abstracts, employing a NN to predict box office revenue [73]. Finally, another study explores daily box office patterns through the clustering approach and after-release features [72]. The literature also reports studies using alternative movie success measures, like critic reviews. For instance, some studies implement ML methods and social media to predict movie ratings [1, 5, 17].

Among studies exploring features before the film release, some use the "hype" generated online immediately before the film release through comments, search patterns, and other "buzz" around the movie. Even in this case, however, production and marketing expenditures are already made, leaving no time to reverse decisions. For example, studies utilize social media mentions as proxies for WOM to predict box office returns in the Korean market [39, 40, 43]. Another study mines popularity and purchase intentions from social media in China to predict box office [49]. Yet another uses Gradient Boosting Decision Tree and daily gross revenues to predict daily box office gross [75]. Finally, another study [32] employs ML binary classifiers and Tweet patterns for the US movie gross.

Still considering post-release features, a study that predicts economic success with profit classes instead of gross revenues develops Multilayer Backpropagation NN to predict movie profitability in a binary classification approach [60]. The authors include ratings from users and critics and the volume of reviews by film in their model for 375 movies released in the US and achieve an accuracy of 88.8%. Along the same line, [68] employs SVM and features after release to explore a film's return on investment (ROI) as a 4-class problem – the data were obtained from 138 movies released in 2015 in the US market, and the result is about 56% accuracy.

An ML seminal study reduced the information set to variables observed before a film's release [65]. The authors employ a Multilayer Perceptron NN to solve a 9-class box office problem. Their set of features is composed of competition degree, genre, MPAA rating, star power, number of screens in the first-week release, and a binary feature for a sequel for 834 movies released in the US market. The authors get a performance of 36.9% in APHR accuracy. A comparison study improves [65]'s results with backpropagation, showing 68.1% of APHR in a 6-class 241-sized dataset [76]. In the same way, [24] also improves [65]'s results using a Dynamic NN in a smaller dataset, getting 74.4% Bingo APHR accuracy as a result for the same box office gross 9-class problem. The authors also perform an additional test in an even smaller dataset (354 movies) and add marketing expenditures to the feature set, which resulted in 94.1% Bingo APHR accuracy with the same Dynamic NN.

More recently, other studies have been updating the methods and features for early box office prediction at earlier stages of the film lifecycle. For example, one applies pruned RF and different comparative ML classifiers to predict 8-class first-week box office using Chinese theater-level data and theaters' revenues as the economic success measure [27]. A second study focuses on animated movie gross, with a 3-class NN and basic movie metadata [61]. A third work uses CART to predict 7-class box office revenue in the Chinese market [77]. A fourth study analyzes the differences between movie features while using RF regression, having the early box office prediction as the economic success measure [4]. Lastly, [3] develops an ensemble with several ML classifiers to predict box office revenues in nine classes.

Finally, very few studies explore profit as the success measure and features available before the film's release or during its production simultaneously (Table 1 – bolded). Employing SVM

**Table 1** Literature summary

| Target Feature | Works | Time of features are available | Main Methods | Class | Movie market | Data Size |
|---|---|---|---|---|---|---|
| Machine Learning | | | | | | |
| Box Office | [78] | After | NN | 6 | US | 3807 |
| Box Office | [47] | After | NN-R, RFR, CART | – | US | 22 |
| Box Office | [23] | After | Fuzzy System | 3 | Hindi | 14 |
| Box Office | [72] | After | Cluster | – | China | 68 |
| Box Office | [73] | After | NN | 6 | China | 150 |
| Box Office | [18] | After | SVR, NN-R, LR | – | China | 24 |
| Box Office | [49] | Right Before | LR, SVR | – | China | 57 |
| Box Office | [75] | Right Before | GBDT | – | China | 13,373 |
| Box Office | [76] | Right Before | MBPNN | 6 | China | 241 |
| Box Office | [39] | Right Before | ML Regressions | – | Korea | 212 |
| Box Office | [40] | Right Before | ML Regressions | – | Korea | 175 |
| Box Office | [43] | Right Before | GTB, LD, LR, RF | 6 | Korea | 400 |
| Box Office | [65] | Right Before | MPNN | 9 | US | 834 |
| Box Office | [32] | Right Before | SVM, KNN, BT, AB, NN | 2 | US | 86 |
| Box Office | [24] | Before | Dynamic NN | 9 | US | 354 |
| Box Office | [61] | Before | NN | 3 | US | 120 |
| Box Office | [4] | Before | RFR | – | US | 1672 |
| Box Office | [3] | Before | Ensemble | 9 | US | 5043 |
| Box Office | [27] | Before | Pruned RF, DT, SVM, MLP | 8 | China | * |
| Box Office | [77] | Before | CART | 7 | China | 150 |
| Profit | [60] | After | MLBP NN | 2 | US | 375 |
| Profit | [68] | After | SVM | 4 | US | 138 |
| Profit | [57] | Before | SVM, NN | 5 | US | 755 |
| Profit | [42] | Before | RF, NN | 2 | US | 2506 |
| Profit | Ours | Before | RF, NN, SVM | 2 & 6 | Worldwide | 3167 |

* The authors did not disclose the information

and NN to predict profitability in five range classes, [57] uses budget, the number of screens, release month, MPAA, and star and director power in a 755-observation dataset to get 49.54% of Bingo APHR. The work most similar to ours, however, uses a 2506 sample size to predict who, what, and when a film could be profitable [42]. The authors explore cast relationships, movie abstracts, and release season to classify American movies according to their raw profit and ROI. The authors perform a few experiments, including binary for ROI and profit and 3-class for ROI. Their best result is 90.4% of accuracy for binary profit.

This study distinguishes three main aspects performed simultaneously from the previous closest studies summarized in Table 1.[3] First, we account for the effects of ICT advances or another possible shock in the recent period; then, we design sub-datasets to account for differences in the short and long run as similar as possible to these studied datasets to compare performance. Second, profits were deflated and used as the measure of economic film success. Third, the sets of our features are smaller and more intuitive than the ones used by those studies and available at the time of film production (see the arguments in Section 3.2.)

In addition, considering Table 1, it is notable that we employ a decision support system to classify and forecast film profits using RF, SVM, and NN. Using this set of tools differs from the literature and could also be viewed as a marginal contribution (see Section 4).

[3] Table 1 presents the literature summary of previous ML studies. The previous studies with explanatory regression analysis are summarized in Appendix - Table 9.

## 3 Data and methodological strategy

### 3.1 Data

Around 22% (3167) of the movie releases between 1980 and 2019 (14,510) available at the Box Office Mojo and IMDB sites – the most common data sources used by literature – have budget information.[4] The smaller amount of information on film costs is due to "industry trade secrets" [76]. The collected sample, however, is far larger than the data size average used in the literature, which is 361 observations/movies [39].

All monetary values used in this study were deflated by the 2019 CPI (CPI-2019) to control for inflation over the years; we keep prices of 1980 constant. This procedure is not usual in this literature. Not correcting for inflation, however, can mislead decision makers and compromise results since comparing revenues over time demands control of price inflation to avoid the more recent films being classified wrongly as more profitable or with higher revenues. Figure 1 shows the evolution of revenues and budget by year, between 1980 and 2019, both controlled for inflation.

We collected budgets and worldwide gross revenues to create the profit measures, which means we are considering the box office revenues in all countries where the film premiered. According to Box Office Mojo, all information received from countries is reported. Then, we follow the scarce literature that uses profit measures to create success classes for binary classes [42, 73, 78] and multiclasses [60].

### 3.2 Methods

Figure 2 presents the general workflow of the methodology described in the following sections.

#### 3.2.1 Variable selection

Unlike most previous studies, we use a reduced set of features easily observable during film production to classify film success. Furthermore, we limit the features to include only those available before the film release, particularly at the film production stage. Thus, differently from previous studies, we can offer a policy guide for producers and stakeholders that allows changes while a movie is still in production. Additionally, the variables were chosen carefully and based on the literature to bring the most meaningful features for an optimal classification given the dimensionality course. Table 2 summarizes the features and their preprocessing step based on the literature.

Compared to [42], we employ more straightforward, less costly, and directly observable features or at least the ones that industry agents have to bet on. For instance, during the production process of a film, it is possible to know the planned runtime, the season to be released, the distributor, and the genres. Thus, if the proposed tool's prediction is faulty, there is time to change characteristics to increase the chances of success. Among the studies, [42] is

[4] Over the last two decades, movies with budget information had better revenue performances in the movie market than movies without disclosed budgets. This may indicate a bias in the total population towards wide released movies – a characteristic that is considered in the dataset slices described in Table 3.

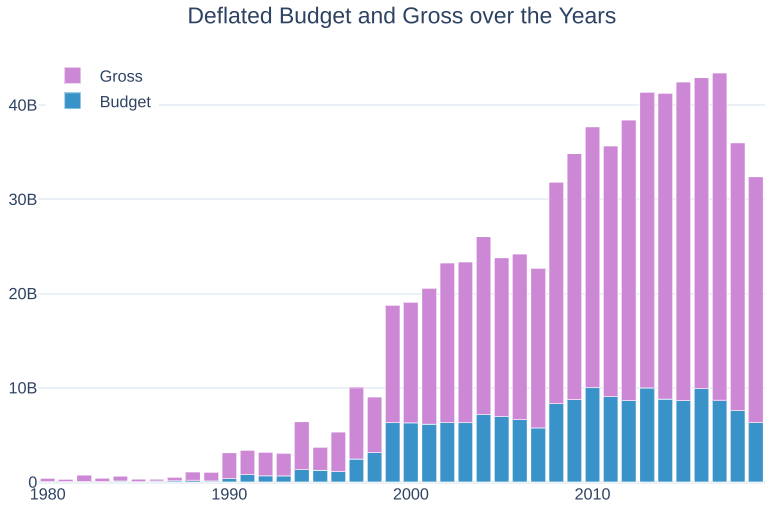


**Fig. 1** Distribution of deflated gross and budget over 40 years of data

the most similar to ours regarding using binary profit, but not deflated, as a success measure and variables before a film's release (see Table 1).

### 3.2.2 Predicting methods

We choose the three most popular ML classifier algorithms in the film literature –SVM, Multilayer Perceptron Neural Network (MLP-NN), and RF – to conduct our experiments.

SVM is a supervised classifier based on the statistical framework proposed by Vapnik and Chervonenkis (VC Theory). It aims to find the best hyperplane to maximize the separation between data points; it can perform linear and nonlinear classification by applying kernel tricks. For further information and the math behind it, please refer to [7, 12].

The MLP-NN is also a supervised classifier that approximates functions that lead the entered data to the output class by adjusting weights between layers (forwards and backwards). For further information about MLP, see [30].

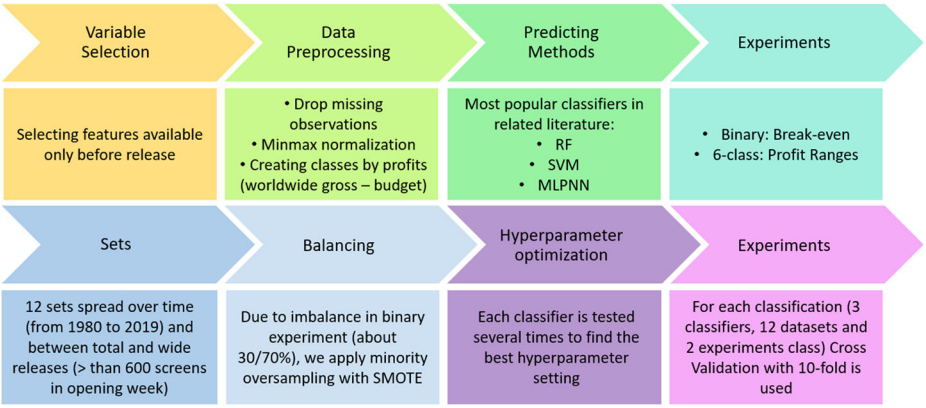


**Fig. 2** Methodology workflow

**Table 2** Variable names and description

| Name | Description |
|---|---|
| Target | |
| Profit | To create classes, we utilize the variable Profit, which is the worldwide gross revenues minus budget. Thus, we manage different class's arrangements. |
| Features | |
| Release Season * *REL_SEASON* | The movie market is naturally seasonal [20]. Such characteristics can be observed from data, revealing bigger earnings in some periods like holidays and summer. Therefore, a movie's release period can affect the box office returns, and probably a film would win higher incomes if released in hot seasons. This variable also plays a competition proxy since more movies and bigger ones are released and hot seasons. Also, the competition is unobservable information during production; it relies on other companies' schedules. We then use this information as a binary variable, being 1 when the movie is released in May, June, July, November, or December, and 0 otherwise. |
| Runtime * *RUNTIME* | The length of a movie, in minutes, could also affect the consumer's decision to watch a movie. On average, longer and shorter movies have lower box office returns. |
| Big Distributor * *BIG_DISTRIBUTOR* | The movie industry has big companies responsible for movie production and distribution. Although they are more than thousands, the joining of few ones holds the largest market share. Moreover, being a large studio can benefit movie revenues by having its name associated with it. Another factor is that the bigger the company, the more money is available, the better is the production in terms of publicity, staff, and cast, and special effects [31]. To control for studios, we separate the twenty biggest companies according to their gross earnings and number of productions; then, we apply one-hot encoding. |
| First Week Theaters * *FIRSTWEEK_THEATERS* | In general, movies make more money in their release week/weekend than other weeks alone, so the number of theaters in the first week is a number to overthink and planned. This measure gives studios and movie theaters an idea of how much the producers expect to collect overall since the very movie theater seats limit the film sale. Also, the number of theaters that a movie is released is highly correlated with its earnings and can also be used as a proxy for star power and advertising [58, 65]. |
| Budget * *BUDGET* | Budget is the most correlated information with gross, but having big budgets does not guarantee higher incomes and better profits. Literature has discussed it and claims that bigger budgets can only serve as an insurance policy not to have greater losses [15, 28]. |
| Genres * | The genre of a movie is also an important characteristic. Genres decide the movie content and, consequently, audience basis and movie influence, once each moviegoer has its own cultural background and preferences in genre consumption. For example, in our data, it is possible to observe that some genres tend to have higher earnings, and the disparity between them is also high. The dataset has the 10 more frequent genres as dummies: Action, Adventure, Biography, Comedy, Crime, Drama, Fantasy, Horror, Music, and Sci-Fi. As a multilabel problem, one movie can belong in one or more genres. |
| MPAA * | The Motion Picture Association of America (MPAA) classification is used mainly as an age restriction for consumers. This classification can restrict audiences and, therefore, limiting the movie's possible earnings [9]. We built 5 binary variables to represent each MPAA category: G, PG, PG-13, R, and NC-17. There are also some Not Classified observations in our sample, not included in the model to prevent linear dependency. |
| Domestic + *DOMESTIC* | Another metric is where the movie is first released. If it was designed to launch only in the domestic market or overseas; it can be a measure of how big is a movie financial plan, and its use is a novelty in literature. In this case, the binary variable was set to 1 for domestic only and 0 for any other market(s). |
| Number of Markets + *NMARKETS* | To precise the Domestic feature, we also created a novel discrete variable counting the number of markets the movie first released, since the movie may be released worldwide simultaneously with the domestic opening. |

Source: * Box office; + IMDB

The RF classifier [8] is an ensemble of decision trees and can perform very well in different tasks [26, 36, 59], in particular, regarding the heterogeneity of data, including continuous and discrete variables, as the binary/dummy features employed. Besides being versatile in binary and multiclass classification, RF is also simple to build, train, tune, and the method is robust and less sensitive to noise [29]. Additionally, RF can outperform other non-ensemble methods [62]. Finally, since most previous movie prediction studies focused on NN, this makes the RF method still little explored (Table 1); we can thus consider its use as a marginal contribution to the domain. In addition, RF is very well suited to preview movie financial success thanks to its capacity to handle mixed data (dummies/binaries and continuous/discrete) Fig. 4.

In our samples, RF is less sensitive to noise (giant blockbusters or flops) and is explainable, allowing us to assess the feature's importance in the models and evaluate whether samples of different ranges of time matter to predict success. Thus, it works as an indirect measure of shocks effects. We brought this idea from the economic literature in time series [10, 50], which states that a process generating a model, in this case, film profits, can change its regime throughout time due to shocks. To implement and test this, we created different data sets using different timing and a complete full dataset including year dummies to test the Gini importance effect of the years on RF (Fig. 5). RF's lower sensitivity to noise is also suitable to compare total and wide-released film sets to preview success.

### 3.2.3 Experiments

Following the literature, the prediction problem was transformed into a classification problem, aiming to classify the movie into its profit success or failure based on its worldwide gross revenues and budget. Two different class arrangements were designed: Break-even (BE) and Profit Ranges (PR).

*Break-Even (BE):* Similar to [42, 60], the output is binary, 1 when the film's profit is zero or positive – the worldwide gross is equal or greater than its budget – and 0 in the contrary case. In this sense, a movie only has to collect (in terms of box office gross) the exact amount spent in production (announced budget).

*Profit Ranges (PR):* To get results closer to actual profit values and comparable with previous literature, we created a 6-class problem considering the total amount of profits of a given movie following [73, 78].

### 3.2.4 Sets

Although the results for the full dataset (1980–2019) were good (see Section 4), we noticed that the literature uses much smaller datasets. Therefore, we also analyze different slices of the dataset to explore possible heterogeneous results among the smallest and greatest samples in time, which could capture changes in consumer behavior over time due to technical changes, for example, and between and within datasets. We also explore wide-released film subsets since they are more homogeneous in box office revenues. Thus, we created 12 subsets of data, considering the years of film releases and wide and total releases.[5] Tables 3 and 4 present the thresholds for classification and the rules to separate data in these subsets.

[5] According to [83] and the Box Office Mojo website, wide-released movies are those that have their opening in 600 or more screens.

**Table 3** Classifications thresholds for each class arrangement (break-even and profit ranges)

| Class | 0 | 1 | | | | |
|---|---|---|---|---|---|---|
| BE | Profit <0 | Profit >=0 | | | | |
| Class | 1 | 2 | 3 | 4 | 5 | 6 |
| PR | < 0 | 0 – 10 M | 10 M – 30 M | 30 M – 80 M | 80 M – 200 M | >=200 M |

M = Millions of dollars

Figure 3 shows the class distributions of the full sample (A) over the years for BE (panel a) and PR (panel b), while Fig. 3 presents similar class distributions for the wide-released movies (B) over the years.

It is necessary to sort out the imbalanced class problem as observed in Fig. 3a (862 unsuccessful vs. 2305 successful films for dataset A) to classify a film according to the profitability's BE classes to avoid biased results toward the success/positive class. This imbalance in our sample is mainly due to budget information, a feature generally disclosed only from big studios. We use SMOTE to oversample the minority (negative) class and address the imbalance. SMOTE is an algorithm that creates, by mimicking, synthetic new observations. The new observations are not duplicated; they are similar to the examples by selecting records and altering one column in that record by a random amount within the difference to the neighboring records [13]; note that the synthetic instances are used only in training folds. Thus, we balanced all BE experiment datasets and the SMOTE proved, through tests, to be better than class weight and near-miss methods.

To obtain the best hyperparameter set, we use the Grid Search tool to optimize all experiments**,** models and sets. We start with big ranges and different configurations of hyperparameters and refine them to get the best scenario. The best sets of hyperparameters are in the footnotes following the results.

For both experimental setups (BE and PR) and all datasets (A to L), we use 10-fold cross-validation. This validation method allows a decrease in the train dependency and creates a more fairly comparable method [67]. Therefore, the results are presented based on the average of these 10 executions.

Finally, to evaluate, present, and discuss the results properly, we use accuracy (Eq. 1) and F1-score (Eq. 2) metrics for both binary (BE) and 6-class (PR) experiments. In

**Table 4** Datasets, slice rules and number of observations for binary class

| Dataset | Rule | Size | 0 | 1 |
|---|---|---|---|---|
| A | Full Sample (1980–2019) | 3167 | 862 | 2305 |
| B | Wide Release Movies | 2475 | 509 | 1966 |
| C | Movies released in or after 1990 | 3091 | 834 | 2257 |
| D | Movies released in or after 1999 | 2732 | 687 | 2045 |
| E | Wide release and released in or after 1990 | 2430 | 499 | 1931 |
| F | Wide release and released in or after 1999 | 2164 | 411 | 1753 |
| G | Set A without outliers detected by Isolation Forest (0.8) | 634 | 493 | 141 |
| H | Wide release and released in or after 2010 | 990 | 104 | 886 |
| I | Wide release and released between 2015 and 2018 | 383 | 45 | 338 |
| J | Wide release and released between 1990 and 1994 | 120 | 39 | 81 |
| K | Wide release and released between 2000 and 2004 | 567 | 166 | 401 |
| L | Wide release and released between 2010 and 2014 | 536 | 54 | 482 |

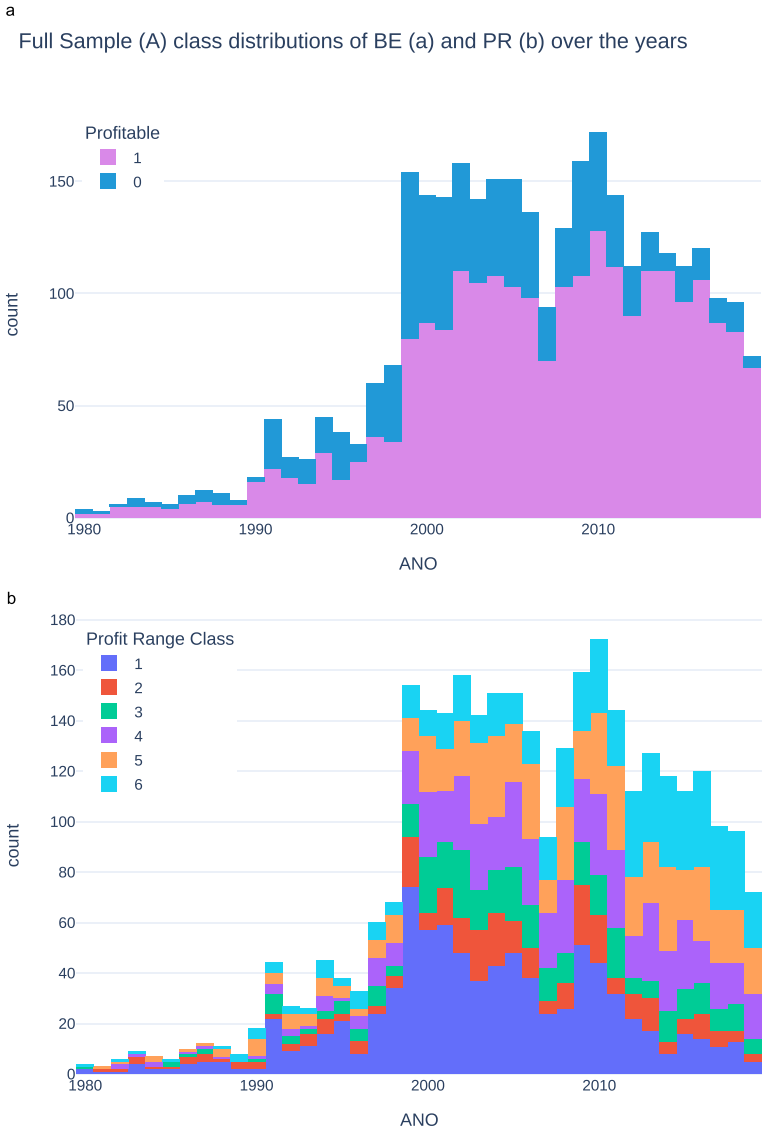


**Fig. 3** **a** Break-Even Classes' distributions of dataset A (full sample) over years. Profitable movies are majority, especially in last two decades. **b** Profit Ranges Classes' distributions of dataset A (full sample) over years. Class 1 diminishes over time, while class 6 grows

addition, APHR is used for multiclass sets (PR), following the most common literature approaches. APHR (Eq. 3) is the total correct classifications to the total number of samples, averaged for all classes in the classification problem – or precision in multiclass problems.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (1)$$

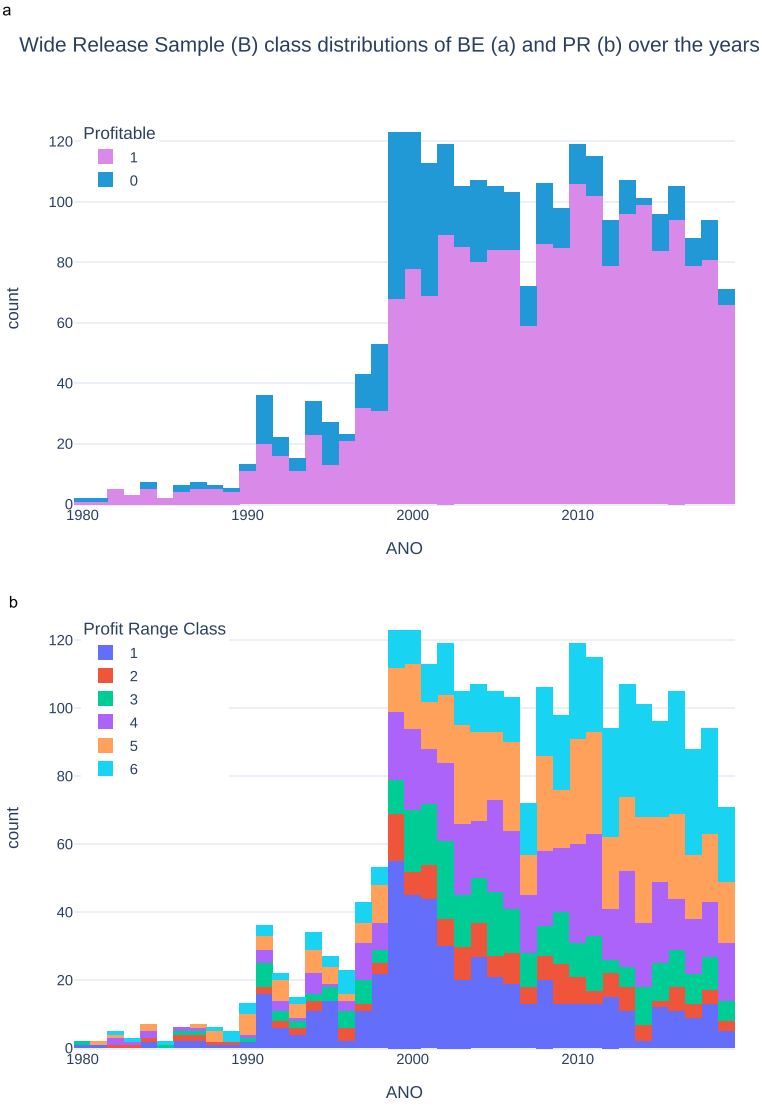


**Fig. 4** **a** Break-Even Classes' distributions of dataset B (wide releases) over years. Profitable movies are majority. **b** Profit Ranges Classes' distributions of dataset B (wide releases) over years. Distributions are similar to those shown in the Fig. 3b

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \tag{2}$$

$$APHR_{Bingo} = \frac{number\ of\ a\ class\ samples\ correctly\ classified}{total\ number\ of\ a\ class\ samples} \tag{3}$$


**Fig. 5** Average accuracy performance (10-cv) of the three classifiers (RF, MLP and SVM) in binary experiment (BE) for sets B, G, H, I and L. Confidence interval shows that RF has the largest both lower and upper bound

## 4 Results and discussion

Table 5 presents the BE results under all datasets (A to L) for the three ML methods: RF, SVM, and NN.

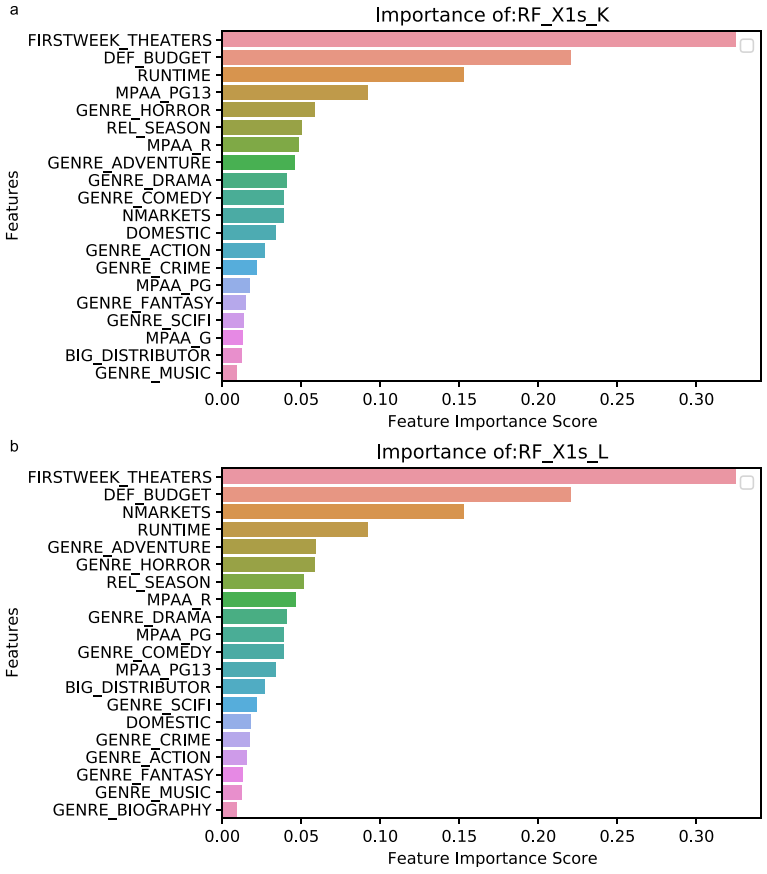


**Fig. 6** Random Forest feature importance for break-even experiment in dataset K (left figure) and in dataset L (right figure). In both First Week Theaters feature is the most important, while other features vary (genres, budget and runtime among others)

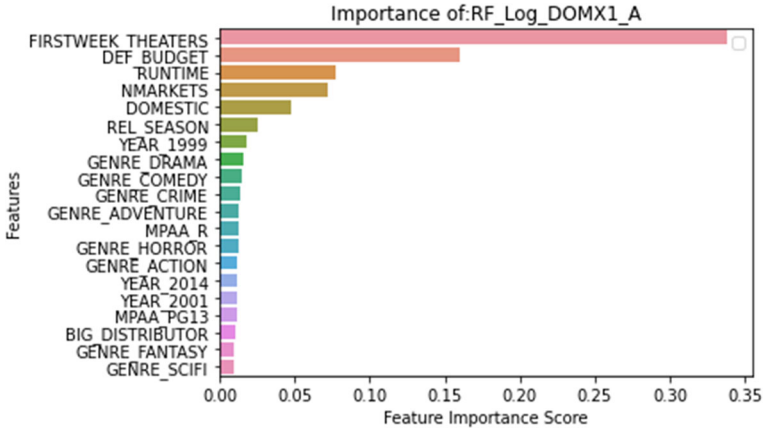


**Fig. 7** Random Forest feature importance for break-even experiment and Full sample (A) added of years' dummies to catch their importance. Years 1999, 2014 and 2001 figured in the top 20 features

The results show a good performance of the model in predicting whether a movie will pay its production costs compared with the literature. The best accuracy result for BE is 96.7% in the B dataset (wide release only) and datasets G, H, I, and L with 95%, 93.3%, 92.1%, and 94.2%, respectively – all with RF. For more details of parameters, see Table 10 in the Appendix. The referenced datasets also have an F1-score above 95%. Except for set J, RF performed better than MLP and SVM. Figure 5 presents the performance of the three classifiers along with their confidence intervals for the best result sets; the confidence intervals reinforce the superiority of RF for the cases presented.

Most studies classify film success employing their revenues as the main measure of success; thus, regarding studies that use revenue net of costs, our best binary experiment result, 96.77%, outperforms the literature with significant margins, 88% in [60], and 90.4% in [42].

**Table 5** Break Even (BE) experiment 10-fold cross validation median Accuracy (Acc) and F1 score average results for RF, MLP and SVM and for each dataset

| Break-Even | | Accuracy | | F1-Score | | | |
|---|---|---|---|---|---|---|---|
| Set | Size | RF | MLP | SVM | RF | MLP | SVM |
| A | 3167 | 86.8 | 80.39 | 79.54 | 90.37 | 85.55 | 84.41 |
| B | 2475 | 96.77 | 94.1 | 94.79 | 97.97 | 96.27 | 96.69 |
| C | 3091 | 87.12 | 82.43 | 79.55 | 90.61 | 87.14 | 84.54 |
| D | 2732 | 88.03 | 81.44 | 79.83 | 91.55 | 86.54 | 85.01 |
| E | 2430 | 89.05 | 84.28 | 80.74 | 92.77 | 89.5 | 86.56 |
| F | 2164 | 89.42 | 85.95 | 82.12 | 93.17 | 90.73 | 87.9 |
| G | 1900 | 95 | 92.37 | 89.11 | 96.58 | 94.81 | 92.6 |
| H | 990 | 93.33 | 90.51 | 88.99 | 96.19 | 94.4 | 93.51 |
| I | 383 | 92.15 | 87.44 | 89.55 | 95.34 | 92.36 | 93.8 |
| J | 120 | 80 | 79.17 | 82.50 | 82.25 | 80.48 | 84.27 |
| K | 567 | 86.23 | 83.76 | 79.88 | 89.37 | 87.16 | 83.8 |
| L | 536 | 94.23 | 91.97 | 90.49 | 96.64 | 95.29 | 94.43 |

**Table 6** Profit Range (PR) experiment 10-fold cross validation median Accuracy (Acc) and F1 score average results for RF, MLP and SVM and for each dataset

| Profit Range | | Accuracy | | | F1-Score | | |
|---|---|---|---|---|---|---|---|
| Set | Size | RF | MLP | SVM | RF | MLP | SVM |
| A | 3167 | 46.19 | 44.33 | 38.71 | 36.96 | 35.66 | 33.83 |
| B | 2475 | 45.94 | 35.07 | 37.58 | 38.78 | 31.93 | 34.04 |
| C | 3091 | 46.36 | 41.83 | 36.04 | 37.6 | 33.09 | 31.81 |
| D | 2732 | 46.16 | 40.7 | 36.82 | 36.84 | 33.34 | 33.13 |
| E | 2430 | 45.72 | 38.4 | 36.26 | 37.65 | 34.15 | 33.11 |
| F | 2164 | 45.29 | 41.55 | 44.78 | 38.46 | 36.02 | 34.18 |
| G | 1900 | 43.94 | 36.42 | 34 | 36.25 | 32.74 | 29.48 |
| H | 990 | 46.87 | 42.63 | 45.25 | 35.46 | 33.56 | 33.15 |
| I | 383 | 49.1 | 37.86 | 39.95 | 35.73 | 28.99 | 29.67 |
| J | 120 | 45.85 | 28.94 | 30.77 | 20.53 | 20.82 | 17.9 |
| K | 567 | 42.68 | 34.39 | 40.04 | 33.13 | 29.88 | 32.24 |
| L | 536 | 46.45 | 34.90 | 35.08 | 33.43 | 27.3 | 28.53 |

For the multiclass experiment, PR, the best average accuracy is from dataset I, with roughly 50% accuracy with RF, followed by sets A, C, D, H and L – all with about 46% accuracy. As shown in Table 6, RF has the best performance for all datasets. The APHR results for PR-I are presented in Table 7.

As Table 7 shows, we obtained 89.8% of the APHR-Bingo average, therefore being better than APHR 56% from [68] and APHR 49.5% from [57]. Broadly comparing these results with literature that uses information before film release to classify, since their measure of success is raw revenues and we use deflated profits, our models also have better performance in prediction than 54.4% from [78], 36.9% from [65], and 68.1% from [76]. Considering that these authors utilize some NN architecture as a predictor method in a multiclass problem, we conclude that RF performs better to support movie stakeholders' decisions. Table 8 shows a better view of the comparison between our results and the literature.

Overall, the four BE results (B, H, I, and L – Table 5) have excellent scores in predicting the profitability of movie theaters since their metrics are better. In addition, results suggest that profits can be more adequate measures of a film's success because they account for the tradeoff between revenues and costs. Yet, as the exclusive use of features available before the film release or during its production process significantly reduces the number of features available

**Table 7** APHR of experiment PR in set I with RF

| Actual Categories | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| Predicted | | 1 | 2 | 3 | 4 | 5 | 6 | |
| Categories | 1 | 40 | 0 | 0 | 2 | 1 | 2 | |
| | 2 | 1 | 15 | 0 | 1 | 0 | 0 | |
| | 3 | 2 | 1 | 31 | 4 | 2 | 1 | |
| | 4 | 0 | 0 | 0 | 59 | 4 | 8 | |
| | 5 | 1 | 0 | 0 | 1 | 70 | 11 | |
| | 6 | 0 | 0 | 0 | 4 | 4 | 118 | |
| | Bingo | 90.9 | 93.8 | 100 | 83.1 | 86.4 | 84.3 | 89.8 |

**Table 8** Summary table of results and comparison

| Ref | Dependent Feat | Independent Features | Class | Best Result | Data Size | Data Time |
|---|---|---|---|---|---|---|
| [78] | BO | Score, Ratings, Comments, Star Value, Budget, Duration, Genres, Poster | 6 | 53.2% APHR Acc | 3807 | – |
| [32] | BO | Wide release, genre, MPAA and a feature collection of mined Tweets | 2 | ~70% F-score | 86 | 2013–2014 |
| [65] | BO | MPAA, Competition, Star Value, Genre, Special Effects, Sequel, Number of Screens | 9 | 36.9% APHR Acc | 834 | 1998–2002 |
| [76] | BO | Nation, Star Value, Propaganda, Content Category, Showing Time, Competition, Cinema Information | 6 | 68.1% APHR Acc | 241 | 2005–2006 |
| [24] | BO | MPAA, Competition, Star Value, Genre, Special effects, Sequel, Number of screens, Budget, Marketing Expenditures, Runtime, Seasonality | 9 | 94% APHR Acc | 354 | 1999–2010 |
| [61] | BO | MPAA, Star Value, Genre, Studio, Sequel | 3 | 58.1% APHR Acc | 120 | 1995–2013 |
| [3] | BO | Director, Actor1, Actor2 and Actor3 scores, experiences, ratings and Facebook likes; Cast total Facebook likes; genre score. | 9 | 82% APHR Acc | 5043 | 1915–2015 |
| [60] | Profit | Actor and Director Star Power; Competition; Seasonality; MPAA; Sequel; Budget; IMDB Rating, Votes and Metascore; Rotten Tomatoes User Rating, Tomatometer, User Reviews and User Votes. | 2 | 88% Acc | 375 | 2010–2015 |
| [68] | Profit | Budget; Trailer and Wikipedia metadata; Rotten Tomatoes Score; Studio; Cast and Crew; Genre. | 3 | 56.5% Acc | 138 | 2016 |
| [57] | Profit | IMDB and Rotten Tomatoes user ratings and scores and sentiment analysis scores; MPAA; Star and Director Power; Seasonality; Budget; Number of Screens | 5 | 56.1% APHR Acc | 755 | 2012–2015 |
| [42] | Profit | Star Power, Seasonality, Genre, MPAA, Movie metadata content, budget | 2 | 86.3% Acc | 2506 | 2000–2010 |
| Ours | Profit | MPAA, Seasonality, Big Distributor, Runtime, Budget, Genre, Number of Screens, Domestic and Markets | 2 | 96.7% Acc | 2475 | 1980–2019 |
| Ours | Profit | MPAA, Seasonality, Big Distributor, Runtime, Budget, Genre, Number of Screens, Domestic and Markets | 6 | 89.8% APHR Acc | 383 | 1980–2019 |

in classifiers, the results are much more significant since we are not using information like critics, user reviews, and WOM data.

These best datasets – H, I, and L – include only wide-released films and brief periods after 2000, explaining their similarities (See Table 3). These datasets perform better than dataset F, which contains all wide-released movies after 1999. The difference may shed light on the timing in which a window slice is designed, consequently on the sample size, where smaller samples and more recent datasets had better performances. Another way to discuss these findings is the homogeneity underlying the data slices, since set B covers all wide-released movies, with no time slice, and the model got the best performance. The same occurs for set G, which has no outliers (Isolation Forest), reinforcing the importance of homogeneity in the predictions.

These results suggest the model generating data might have changed due to structural breaks [10]. Shocks – like technological innovations, changes in consumer preferences, political and economic interventions, and natural shocks like COVID-19 – cause a structural break. To evaluate this possibility, we explore the feature importance generated by RF, via Gini Index, for BE experiments with different sample sized datasets to check whether there are changes in their relative feature importance since such changes indicate a different model. Using datasets distant in time – K (wide releases between 2000 and 2004, 567 observations) and L (wide releases between 2010 and 2014, 536 observations) – we extract the Gini Feature Importance for each case, as Fig. 6 shows.

Comparing the features in K and L datasets, it is possible to notice a clear change in the relative importance of budget, runtime, crime and adventure genres, number of markets, and other features. This change in theatrical consumption can result from technological innovations, as an alternative way to consume a movie brought by the streaming videos or the availability of other new goods, like games, leading to a change in consumption behavior. To check the robustness of these changes and better comprehend a possible shift in the "regime" that governs the data generation, we also included another BE experiment with year dummy variables (from 1981 to 2019) as features to dataset A and performed RF classification. Note that we included all years because we are using worldwide revenues, having many countries, and it is difficult to define a specific year shock. If the year dummies are relevant determinants to film success, however, it means evidence of the regime's change since it is supposed that the time would not affect the classification. The relative importance score of the first 20 features is shown in Fig. 7.

Additionally, by exploring different data samplings and the importance of features in each dataset, we find that the number of theaters, budget, runtime, and the number of market releases are the main features to explain a movie's economic success. Note that the number of theaters, however, may bias the results to be suitable only for wide-opening films since these types of movies disclose budget information more commonly. Alternatively, the two least significant are the MPAA ratings of NC-17 and G; this may be because of their low representativeness in data. Apart from these last two, our models were able to classify very well by using a few variables easily observable or available in a movie's planned production/pre-production period.

## 5 Concluding remarks

Uncertainty in new film production is high, with failure rates ranging between 25 and 45% [46]. Therefore, a large portion of movies are unprofitable, and productions with large budgets and impressive star power are not guarantees of profit [15]. We, thus, evaluate three classifiers to determine the economic success, measured by profits free of price inflation, of film release at theaters using few and simple observable types of information at the time of production stage. We consider economic success as the movie revenue over its costs (or profits) in two different approaches; binary classification (BE) and 6-class classification (PR). For binary classification, we use SMOTE to solve problems of class imbalance.

Forecasting film profitability based only on the early stages of film production is a complex task, mainly due to eliminating several other relevant determinants of film quality and economic performance available only at or after the release. Nevertheless, our results show

better performance than the previous studies, mainly using RF and small datasets (accuracy of 96% and F1-score of 97% for binary and about 50% APHR for 6-class).

In addition, the analysis of feature importance suggests that the movie market model changes over time. The theoretical literature in ML and statistics [74] indicates that more data (more instances/information) improves performance. Our findings, similar to the literature on applied film success, show that limiting data to brief periods of timing supports patterns of similarity over time, thus resulting in better learning. We, therefore, argue that shocks like technological innovations, which change supply and demand behaviors, and other shocks can alter a model regime to classify film success.

Therefore, our study contributes to the productive sector and related academic studies. It can guide studios, producers, and other stakeholders to make better investments and decisions when there is room to change plans. In this case, they can count on the low cost of obtaining inputs to make predictions (directly observable features), excellent accuracy in prediction, and time enough to make changes in movie plans in case a poor contingent prediction occurs.

Regarding the literature contribution, we envisage five novelties that can be summarized in three main issues. First, we use deflated profits as the measure of film success instead of non-deflated film revenues as in most literature, which allows us to balance the trade-off between film revenues and costs. In addition, the few studies that employ profits as a film success measure do not deflate them, which can mislead the classification towards considering the most recent films as the most profitable. Second, to preview success, the proposed tool uses a small number of simple features that are not pre-processed and are directly observable. In addition, the features are available mainly at film production time; thus, when some bad results are predicted, there is room to change the production course to increase the chances of the film's success. Third, it calls attention to the regime's potential changes that describe a model over time due to shocks like technological innovations. In this sense, considering all the previous items and the cuts of sample to compare with the literature, the use of RF, and the higher scores obtained, we believe we have contributed to the literature.

Regarding the potential "regime" changes, more investigation is needed. In this sense, structural breaks should be analyzed through specific statistical tests – a future work to be explored – to develop exogenous tests to guarantee the future predictability of the film market and other social time-related domains. Another line of investigation is to exploit the differences between more homogeneous and heterogeneous samples employed to predict film success. For instance, eliminating outliers is a way to make a sample more homogeneous and improve binary predictions. In this sense, reducing a film sample in a shorter time makes films more homogeneous and improves results on success prediction, as we and other authors have found. Also, using samples with only wide-released films, a more homogeneous dataset, resulted in better prediction in our results. In addition, we might improve feature selection for future works, removing those that are minimally informative and adding others like a sequel and/or star power – for example, in agreement with the literature – and experiment with different computational models to estimate missing budget data to enhance data size.

# APPENDIX

**Table 9** Summary of previous studies (regression analysis)

| Dependent variable (economic success measure) | Works | Time of determinants are available | Main Method | Movie market | Data Size |
|---|---|---|---|---|---|
| Regressions | | | | | |
| Box Office | [58] | After | Regression Analysis | US | 609 |
| Box Office | [28] | After | Regression Analysis | US | 2080 |
| Box Office | [37] | After | Regression Analysis | US | 169 |
| Box Office | [16] | After | Regression Analysis | US | 135 |
| Box Office | [45] | After | Regression Analysis | US | * |
| Box Office | [52] | After | Regression Analysis | US | 27 |
| Box Office | [55] | After | Regression Analysis | US | 106 |
| Box Office | [21] | After | Regression Analysis | US | 56 |
| Box Office | [53] | After | Regression Analysis | US | 246 |
| Box Office | [14] | After | Regression Analysis | US | 148 |
| Box Office | [51] | After | Regression Analysis | US | 312 |
| Box Office | [19] | After | Regression Analysis | US | 188 |
| Box Office | [69] | After | Regression Analysis | US | 474 |
| Box Office | [6] | After | Regression Analysis | Hindi | 7 |
| Box Office | [2] | After | Regression Analysis | Hindi | * |
| Ratings & Awards | [25] | After | Regression Analysis | * | 368 |
| Awards | [11] | After | Regression Analysis | US | 463 |
| Awards | [41] | After | Regression Analysis | * | 25 |
| Survival | [66] | After | Regression Analysis | US | 4700 |
| Survival | [44] | After | Regression Analysis | CA | 788 |

**Table 10** Hyperparameters used in the best results experiments

| Test | Set | Hyperparameters (Random Forest) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Criterion | Max_ features | Min_ samples_leaf | Min_ samples_split | N_ estimators | Max_ depth |
| BE | B | Entropy | 6 | 2 | 5 | 500 | None |
| BE | G | Entropy | 8 | 2 | 5 | 1000 | None |
| BE | H | Entropy | Auto | 2 | 2 | 500 | None |
| BE | I | Entropy | 8 | 2 | 2 | 100 | None |
| BE | L | Gini | 4 | 2 | 2 | 500 | None |
| PR | I | Gini | Auto | 2 | 5 | 1000 | None |

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11042-023-15169-4.

**Funding** This work was supported by CAPES (Higher Education Improvement Coordination); and FAPESP (São Paulo Research Foundation).

**Data availability** https://data.mendeley.com/datasets/s36kp8rc4h/draft?a=b191f4c8-d0ba-4798-9c44-58f88d0231d7

## Declarations

**Conflict of interest** none.

## References


1. Abidi SMR, Xu Y, Ni J, Wang X, Zhang W (2020) Popularity prediction of movies: from statistical modeling to machine learning techniques. Multimed Tools Appl 79:35583–35617. https://doi.org/10.1007/s11042-019-08546-5
2. Ahmad J, Duraisamy P, Yousef A, Buckles B (2017) Movie success prediction using data mining. In: 8th International Conference on computing, communications and networking technologies, ICCCNT 2017
3. Ahmed U, Waqas H, Afzal MT (2020) Pre-production box-office success quotient forecasting. Soft Comput 24:6635–6653. https://doi.org/10.1007/s00500-019-04303-w
4. Antipov EA, Pokryshevskaya EB (2017) Are box office revenues equally unpredictable for all movies? Evidence from a Random forest-based model. J Revenue Pricing Manag 16:295–307. https://doi.org/10.1057/s41272-016-0072-y
5. Basu S (2019) Movie rating prediction system based on opinion mining and artificial neural networks. In: Advances in Intelligent Systems and Computing
6. Bhattacharjee B, Sridhar A, Dutta A (2017) Identifying the causal relationship between social media content of a Bollywood movie and its box-office success - a text mining approach. Int J Bus Inf Syst 24:344. https://doi.org/10.1504/IJBIS.2017.082039
7. Boser BE, Guyon IM, Vapnik VN (1992) Training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory
8. Breiman L (2001) Random forests. Mach Learn https://doi.org/10.1023/A:1010933404324
9. Brewer SM, Kelley JM, Jozefowicz JJ (2009) A blueprint for success in the US film industry. Appl Econ 41:589–606. https://doi.org/10.1080/00036840601007351
10. Casini A, Perron P (2019) Structural breaks in time series. In: Oxford Research Encyclopedia of Economics and Finance
11. Chang BH, Ki EJ (2005) Devising a practical model for predicting theatrical movie success: focusing on the experience good property. J Media Econ 18:247–269. https://doi.org/10.1207/s15327736me1804_2
12. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:1–27. https://doi.org/10.1145/1961189.1961199
13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
14. Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: accounting for sequential. Mark Sci
15. De Vany A, Walls WD (1999) Uncertainty in the movie industry : Does star power reduce the terror of the box office ? J Cult Econ 23:285–318. https://doi.org/10.1023/a:1007608125988
16. Derrick FW, Williams NA, Scott CE (2014) A two-stage proxy variable approach to estimating movie box office receipts. J Cult Econ 38:173–189. https://doi.org/10.1007/s10824-012-9198-y
17. Dhir R, Raj A (2018) Movie success prediction using machine learning algorithms and their comparison. ICSCCC 2018 - 1st Int Conf Secur cyber Comput Commun 385–390. https://doi.org/10.1109/ICSCCC.2018.8703320
18. Du J, Xu H, Huang X (2014) Box office prediction based on microblog. Expert Syst Appl 41:1680–1689. https://doi.org/10.1016/j.eswa.2013.08.065
19. Duan J, Ding X, Liu T (2015) A Gaussian copula regression model for movie box-office revenue prediction with social media. In: Communications in Computer and Information Science
20. Einav L (2007) Seasonality in the U.S. motion picture industry. RAND J Econ. https://doi.org/10.1111/j.1756-2171.2007.tb00048.x
21. Eliashberg J, Shugan SM (1997) Film critics: influencers or predictors? J Mark 61:68–78. https://doi.org/10.2307/1251831
22. Eliashberg J, Elberse A, Leenders MA (2006) The motion picture industry: Critical issues in practice, current research, and new research directions. Mark Sci 25:638–661. https://doi.org/10.1287/mksc.1050.0177

23. Gaikar DD, Marakarkandy B, Dasgupta C (2015) Using twitter data to predict the performance of bollywood movies. Ind Manag Data Syst 115:1604–1621. https://doi.org/10.1108/IMDS-04-2015-0145
24. Ghiassi M, Lio D, Moon B (2015) Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Syst Appl 42:3176–3193. https://doi.org/10.1016/j.eswa.2014.11.022
25. Ginsburgh V (2003) Awards, success and aesthetic quality in the arts. In: Journal of Economic Perspectives
26. Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. In: Pattern Recognition Letters
27. Guo Z, Zhang X, Hou Y (2015) Predicting box office receipts of movies with pruned random forest. In: lecture notes in Computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)
28. Hadida AL (2010) Commercial success and artistic recognition of motion picture projects. J Cult Econ 34: 45–80. https://doi.org/10.1007/s10824-009-9109-z
29. Hastie T, Tibshirani R, Friedman J (2009) Elements of statistical learning 2nd ed.
30. Hecht-Nielsen R (1992) Theory of the Backpropagation Neural Network**Based on "nonindent" by Robert Hecht-Nielsen, which appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593–611, June 1989. © 1989 IEEE. In: Neural Networks for Perception
31. Honthaner EL (2013) The complete film production handbook
32. Hossein N, Miller DW (2018) Predicting motion picture box office performance using temporal tweet patterns. Int J Intell Comput Cybern 11:64–80. https://doi.org/10.1108/IJICC-04-2017-0033
33. Hu YH, Shiau WM, Shih SP, Chen CJ (2018) Considering online consumer reviews to predict movie box-office performance between the years 2009 and 2014 in the US. Electron Libr 36:1010–1026. https://doi.org/10.1108/EL-02-2018-0040
34. Hur M, Kang P, Cho S (2016) Box-office forecasting based on sentiments of movie reviews and independent subspace method. Inf Sci (Ny) 372:608–624. https://doi.org/10.1016/j.ins.2016.08.027
35. Husak W (2004) Economic and other considerations for digital cinema. Signal Process Image Commun 19: 921–936. https://doi.org/10.1016/j.image.2004.06.006
36. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11. https://doi.org/10.1186/1472-6947-11-51
37. Kim SH, Park N, Park SH (2013) Exploring the effects of online word of mouth and expert reviews on theatrical movies' box office success. J Media Econ 26:98–114. https://doi.org/10.1080/08997764.2013.785551
38. Kim D, Kim D, Hwang E, Choi HG (2013) A user opinion and metadata mining scheme for predicting box office performance of movies in the social network environment. New Rev Hypermedia Multimed 19:259–272. https://doi.org/10.1080/13614568.2013.835450
39. Kim T, Hong J, Kang P (2015) Box office forecasting using machine learning algorithms based on SNS data. Int J Forecast 31:364–390. https://doi.org/10.1016/j.ijforecast.2014.05.006
40. Kim T, Hong J, Kang P (2017) Box office forecasting considering competitive environment and word-of-mouth in social networks: a case study of Korean film market. Comput Intell Neurosci 2017:1–16. https://doi.org/10.1155/2017/4315419
41. Krauss J, Nann S, Simon D, et al (2008) Predicting movie success and academy awards through sentiment and social network analysis. In: 16th European Conference on information systems, ECIS 2008
42. Lash MT, Zhao K (2016) Early predictions of movie success: the who, what, and when of profitability. J Manag Inf Syst 33:874–903. https://doi.org/10.1080/07421222.2016.1243969
43. Lee K, Park J, Kim I, Choi Y (2018) Predicting movie success with machine learning techniques: Ways to improve accuracy. Inf Syst Front 20:577–588. https://doi.org/10.1007/s10796-016-9689-z
44. Legoux R, Larocque D, Laporte S, Belmati S, Boquet T (2016) The effect of critical reviews on exhibitors' decisions: do reviews affect the survival of a movie on screen? Int J Res Mark 33:357–374. https://doi.org/10.1016/j.ijresmar.2015.07.003
45. Lehrer S, Xie T (2017) Box office buzz: Does social media data steal the show from model uncertainty when forecasting for Hollywood? Rev Econ Stat 99:749–755. https://doi.org/10.1162/REST_a_00671
46. Leung TC, Qi S, Yuan J (2020) Movie industry demand and theater availability. Rev Ind Organ.https://doi.org/10.1007/s11151-019-09706-5
47. Lipizzi C, Iandoli L, Marquez JER (2016) Combining structure, content and meaning in online social networks: the analysis of public's early reaction in social media to newly launched movies. Technol Forecast Soc Change 109:35–49. https://doi.org/10.1016/j.techfore.2016.05.013
48. Litman BR (1983) Predicting success of theatrical movies: An empirical study. J Pop Cult 16:159–175. https://doi.org/10.1111/j.0022-3840.1983.1604_159.x
49. Liu T, Ding X, Chen Y, Chen H, Guo M (2016) Predicting movie box-office revenues by exploiting large-scale social media content. Multimed Tools Appl 75:1509–1528. https://doi.org/10.1007/s11042-014-2270-1

50. Lucas RE (1976) Econometric policy evaluation: A critique. Carnegie-Rochester Confer Ser Public Policy 1:19–46. https://doi.org/10.1016/S0167-2231(76)80003-6
51. Mestyán M, Yasseri T, Kertész J (2013) Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data PLoS One 8: https://doi.org/10.1371/journal.pone.0071226
52. Mohanty S, Clements N, Gupta V (2018) Investigating the effect of eWOM in movie box office success through an aspect-based approach. Int J Bus Anal 5:1–15. https://doi.org/10.4018/IJBAN.2018010101
53. Moon S, Bergey PK, Lacobucci D (2010) Dynamic effects among movie ratings, movie revenues and viewer satisfaction. J Mark 74:108–121. https://doi.org/10.1509/jmkg.74.1.108
54. MPA - Motion Picture Association (2019) 2019 THEME Report - Motion Picture Association
55. Oh C, Roumani Y, Nwankpa JK, Hu HF (2017) Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. Inf Manag 54:25–37. https://doi.org/10.1016/j.im.2016.03.004
56. Pokorny M, Sedgwick J (2010) Profitability trends in Hollywood, 1929 to 1999: somebody must know something. Econ Hist Rev 63:56–84. https://doi.org/10.1111/j.1468-0289.2009.00488.x
57. Quader N, Gani MO, Chaki D, Ali MH (2017) A machine learning approach to predict movie box-office success. In: 2017 20TH INTERNATIONAL CONFERENCE OF COMPUTER AND INFORMATION TECHNOLOGY (ICCIT)
58. Reinstein DA, Snyder CM (2005) The influence of expert reviews on consumer demand for experience goods: a case study of movie critics. J Ind Econ 53:27–51. https://doi.org/10.1111/j.0022-1821.2005.00244.x
59. Ren Y, Zhang L, Suganthan PN (2016) Ensemble classification and regression-recent developments, applications and future directions [review article]. IEEE Comput Intell Mag 11:41–53
60. Rhee TG, Zulkernine F (2016) Predicting movie box office profitability: a neural network approach. 2016 15TH IEEE Int Conf Mach learn Appl (ICMLA 2016) 665–670. https://doi.org/10.1109/icmla.2016.0117
61. Riwinoto MT, Zega SA, Irlanda G (2015) Predicting animated film of box-office success with neural networks. J Teknol. https://doi.org/10.11113/jt.v77.6693
62. Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33:1–39. https://doi.org/10.1007/s10462-009-9124-7
63. Ru Y, Li B, Liu J, Chai J (2018) An effective daily box office prediction model based on deep neural networks. Cogn Syst Res 52:182–191. https://doi.org/10.1016/j.cogsys.2018.06.018
64. Shapiro C, Varian HR (1999) Information rules
65. Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks. Expert Syst Appl 30:243–254. https://doi.org/10.1016/j.eswa.2005.07.018
66. Souza TLD, Nishijima M, Fava ACP (2019) Do consumer and expert reviews affect the length of time a film is kept on screens in the USA? J Cult Econ 43:145–171. https://doi.org/10.1007/s10824-018-9332-6
67. Stone M (1974) Cross-Validatory choice and assessment of statistical predictions. J R Stat Soc Ser B https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
68. Subramaniyaswamy V, Vaibhav MV, Prasad RV, Logesh R (2018) Predicting movie box office success using multiple regression and SVM. Proc Int Conf Intell Sustain Syst ICISS 2017:182–186. https://doi.org/10.1109/ISS1.2017.8389394
69. Tadimari A, Kumar N, Guha T, Narayanan SS (2016) Opening big in box office? Trailer content can help. ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc 2016-May:2777–2781. https://doi.org/10.1109/ICASSP.2016.7472183
70. Varian HR (2000) Buying, sharing and renting information goods. J Ind Econ 48:473–488. https://doi.org/10.1111/1467-6451.00133
71. Waldfogel J (2017) How digitization has created a golden age of music, movies, books, and television. J Econ Perspect 31:195–214
72. Wang Y, Ru Y, Chai J (2019) Time series clustering based on sparse subspace clustering algorithm and its application to daily box-office data analysis. Neural Comput & Applic 31:4809–4818. https://doi.org/10.1007/s00521-018-3731-7
73. Wang Z, Zhang J, Ji S, Meng C, Li T, Zheng Y (2020) Predicting and ranking box office revenue of movies based on big data. Inf Fusion 60:25–40. https://doi.org/10.1016/j.inffus.2020.02.002
74. Wooldridge JM (2002) Econometric analysis of cross section and panel data. Booksgooglecom 58:752. https://doi.org/10.1515/humr.2003.021
75. Wu S, Zheng Y, Lai Z, et al (2019) Movie box office prediction based on ensemble learning. ISPCE-CN 2019 - IEEE Int Symp prod compliance Eng 2019 1–4. https://doi.org/10.1109/ISPCE-CN48734.2019.8958631
76. Zhang L, Luo J, Yang S (2009) Forecasting box office revenue of movies with BP neural network. Expert Syst Appl 36:6580–6587. https://doi.org/10.1016/j.eswa.2008.07.064
77. Zhang Z, Chai J, Li B, et al (2016) Movie box office Inteval forecasting based on CART. In: proceedings - 2015 8th International symposium on computational intelligence and design, ISCID 2015

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