ItaúCultural

Box office estimator for Brazilian cinematographic productions: a machine learning approach

Julia Taunay Perez,
Everton Rodrigues Reis
e Davi Noboru Nakano

Julia Taunay Perez[1], Everton Rodrigues Reis[2] and Davi Noboru Nakano[3]

Box office estimator for Brazilian cinematographic productions: a machine learning approach

1 Economist, graduated from Universidade Estadual Paulista (Unesp), Perez is currently a PhD student in production engineering at the Polytechnic School of the University of São Paulo (Poli / USP). Her research focuses on designing financial instruments for creative industries. She is also a founding partner of Lumen, consulting firm providing advice on artificial intelligence techniques. She worked as a researcher for the Creativeworks project, undertaken in partnership with Queen Mary University of London, assisting creative ventures in developing their business models.

2 Data scientist with an investment fund and master's degree student at Poli / USP. Reis graduated in science and technology, as well as in management engineering, from the Federal University of ABC (UFABC). He worked as a data scientist for CVC Corp. His research areas are multi-agent systems and machine learning.

3 Associate Professor, Department of Production Engineering, USP Polytechnic School. He is a researcher at the Observatory of Innovation and Competitiveness at the USP Institute of Advanced Studies and a visiting researcher at Queen Mary University of London (Network Centre). Former visiting researcher at the University of Alberta, Canada, and co-editor of the journal Production, of the Brazilian Association of Production Engineering. Nakano is involved in national and international projects with North American and European educational institutions. He published several journal articles and conference papers at both domestic and international levels. His research focuses on creative economy and creative industries, innovation methods for product and service development, and engineering education.

ECONOMIA DA CULTURA

OBSERVATÓRIO ITAÚ CULTURAL

Introduction

The national film market generated a R$ 2.7 billion revenue in 2017, having led more than 181 million viewers to theaters. However, if we look at the share domestic producers receive of that amount things are not so bright. The 160 Brazilian titles released that year attracted about 17 million viewers and generated approximately R$ 240 million, which account for less than 9% of the total box office revenue (ANCINE, 2017).

The difficulty in appealing to audiences in the face of Hollywood competition is not exclusive to the Brazilian scene. In the European Union, titles released by American studios attracted 64% of the audience in 2015 (KANZLER, MILLA, 2016). In this regard, policymakers around the world are discussing the implementation of subsidies and screening quotas as a way to mitigate the effects of US competition on the domestic film industry (MESSERLIN; PARC, 2014).

The value chain of the film industry includes three main activities: production, distribution, and exhibition (LEGOUX et al., 2016, MICHEL, AVELLAR, 2014). The economic relationship between these links is mediated by contracts, whose terms must mirror the way parties perceive and mitigate risks.

Because production and distribution activities are vertically integrated in the North American market, the largest studios (Disney, Sony, Fox, Paramount, Universal, Warner and DreamWorks) produce and distribute their films (LEGOUX et al., 2016) and only a few studies have sought to understand the conflict of interests between such agents. In Brazil, production and distribution activities do not seem to be vertically organized. According to data from the National Film Agency (Agência Nacional do Cinema - Ancine), only two of the production companies that attracted together 80% of the domestic feature film audience (Rede Record, DiamondBack, Paris Produções, Lereby, Glaz Entretenimento, Camisa Listrada, Fraiha Produções, and FilmlandInternational) are active in distribution as well[4].

4 DiamondBack has co-distribution agreements with American studios, and Paris Produções distributes its films in partnership with Downtown Filmes..

Upstream, relationships between distributors and exhibitors are fraught with conflicts. When drafting contracts, distributors aim at maximizing the revenue earned by a film, whereas exhibitors intend to maximize the revenue from a portfolio standpoint. Thus, while distributors increase their revenue through longer runs in theaters, exhibitors, who collect box office participation, wish to release a larger number of films to attract audiences (CHISHOLM et al., 2015).

Several studies point out that the performance of the film in its pre-release and/or first week of release determines its success (CHEN; XU; ZHANG, 2016; KARNIOUCHINA, 2011; LEGOUX et al., 2016; SAWHNEY; ELIASHBERG, 1996). This is a problem in Brazil as well; on average, domestic films bring in 48% of their revenue in the first screening week. In addition, concentration of distribution on majors decreases domestic players' competitiveness, as Michel and Avellar pointed out (2014).

Therefore, a model that would estimate the future box office performance of a film would lead to a better understanding of determinants and dynamics of the Brazilian demand for domestic films, and would allow designing strategies in the sphere of both chain players (production companies, distributors and exhibitors) and public policy, thus enhancing the competitiveness of domestic productions.

Nevertheless, as the high uncertainty of the demand in this sector leads to a box office behavior with unlimited variance (WALLS, 2005), linear prediction models, such as the least square method, do not quite fit the characteristics of cinematographic demand. This scenario led to the development of studies aimed at finding more robust methods of analysis, based on computer learning (CHEN; XU; ZHANG, 2016; DELEN; SHARDA, 2010; GHIASSI; LIO; MOON, 2015; GUO; ZHANG; HOU, 2015; HUR; KANG; CHO, 2016; LEE et al., 2018; LIU; ZHAO, 2016; SHARDA; DELEN, 2006; ZHANG; LUO; YANG, 2009).

Thus, this study applies computer-learning techniques to a database including 403 films released and screened between 2009 and 2016 and containing information from different sources (Ancine, IMDb, AdoroCinema, Ministry of Justice and Grupo Globo) to develop a box office prediction model. The selected explanatory variables are projection time, genre, age rating, production company type, production time, seasonality, sequel, release theaters, and cast star power.

Understanding such parameters should be helpful for production companies in the initial development stages of domestic productions (PACKARD et al., 2016). This instrument allows distributors to achieve greater efficiency in choosing films to be distributed and helps them include better terms in their contract with exhibitors.

These latter will also benefit by achieving greater safety and reducing uncertainty as to their portfolio of movies to be screened.

1. Determinants of demand

In empirical studies, cinema demand is usually represented by two variables, audience and box office. Audience refers to the audience watching a movie in a theater, whereas box office refers to the amount collected from ticket sales.

Several factors help us understand demand behavior in terms of both audience and revenue: movie characteristics, economic and demographic conditions, competitive scenario, financial conditions, distribution and release conditions, cast and crew network and reputation, criticism and word of mouth, as shown in **Chart 1**.

Chart 1: List of explanatory variables related to cinema demand

	AUDIENCE	BOX OFFICE
MOVIE CHARACTERISTICS		
Age Rating	CLEMENT; WU; FISCHER (2014)	SAWHNEY; ELIASHBERG (1996) COLLINS; HAND; SNELL (2002) BASUROY; CHATTERJEE; RAVID (2003) TERRY; BUTLER; ARMOND (2003) CHANG; KI (2005) DELMESTRI; MONTANARI; USAI (2005) WALLS (2005) SHARDA; DELEN (2006) BREWER; KELLEY; JOZEFOWICZ (2009) NELSON; GLOTFELTY (2012) KIM; PARK; PARK (2013) CLEMENT; WU; FISCHER (2014) DERRICK; WILLIAMS; SCOTT (2014) CRAIG; GREENE; VERSACI (2015) PACKARD et al. (2016)

Genre	CLEMENT; WU; FISCHER (2014)	SAWHNEY; ELIASHBERG (1996) COLLINS; HAND; SNELL (2002) TERRY; BUTLER; ARMOND (2003) CHANG; KI (2005) DELMESTRI; MONTANARI; USAI (2005) WALLS (2005) SHARDA; DELEN (2006) BREWER; KELLEY; JOZEFOWICZ (2009) TREME (2010) NELSON; GLOTFELTY (2012) KIM; PARK; PARK (2013) CLEMENT; WU; FISCHER (2014) DERRICK; WILLIAMS; SCOTT (2014) CRAIG; GREENE; VERSACI (2015) PACKARD et al. (2016)
Production origin	CLEMENT; WU; FISCHER (2014)	MCKENZIE; WALLS (2013) CLEMENT; WU; FISCHER (2014)
Sequel	CLEMENT; WU; FISCHER (2014)	SAWHNEY; ELIASHBERG (1996) COLLINS; HAND; SNELL (2002) BASUROY; CHATTERJEE; RAVID (2003) TERRY; BUTLER; ARMOND (2003) CHANG; KI (2005) WALLS (2005) SHARDA; DELEN (2006) BREWER; KELLEY; JOZEFOWICZ (2009) KIM; PARK; PARK (2013) MCKENZIE; WALLS (2013) CLEMENT; WU; FISCHER (2014) CRAIG; GREENE; VERSACI (2015) PACKARD et al. (2016)
Special effects		SAWHNEY; ELIASHBERG (1996) SHARDA; DELEN (2006)
ECONOMIC AND DEMOGRAPHIE CONDITIONS		
Population		NELSON; GLOTFELTY (2012)
Consumer price index		BREWER; KELLEY; JOZEFOWICZ (2009)
Per capita income		BREWER; KELLEY; JOZEFOWICZ (2009) NELSON; GLOTFELTY (2012)

Seasonality	CLEMENT; WU; FISCHER (2014)	TERRY; BUTLER; ARMOND (2003) KARNIOUCHINA (2011) CLEMENT; WU; FISCHER (2014) KARNIOUCHINA (2011) DERRICK; WILLIAMS; SCOTT (2014)
COMPETITIVE SCENARIO		
Competition for screen and/or audience		SHARDA; DELEN (2006) CLEMENT; WU; FISCHER (2014) KARNIOUCHINA (2011) KIM; HONG; KANG (2017)
FINANCIAL CONSTRAINTS		
Advertising spending	CLEMENT; WU; FISCHER (2014)	MCKENZIE; WALLS (2013) CLEMENT; WU; FISCHER (2014)
Budget		BASUROY; CHATTERJEE; RAVID (2003) CHANG; KI (2005) BREWER; KELLEY; JOZEFOWICZ (2009) TREME (2010) NELSON; GLOTFELTY (2012) KIM; PARK; PARK (2013) MCKENZIE; WALLS (2013) CRAIG; GREENE; VERSACI (2015)
Negative cost		WALLS (2005)
DISTRIBUTION AND RELEASE CONDITIONS		
Release date		BASUROY; CHATTERJEE; RAVID (2003) CHANG; KI (2005) DELMESTRI; MONTANARI; USAI (2005) WALLS (2005) BREWER; KELLEY; JOZEFOWICZ (2009) TREME (2010) KIM; PARK; PARK (2013)
Distributor		KIM; PARK; PARK (2013) DERRICK; WILLIAMS; SCOTT (2014) PACKARD et al. (2016)

CAST AND CREW NETWORK AND REPUTATION		
Director	CLEMENT; WU; FISCHER (2014)	CHANG; KI (2005) DELMESTRI; MONTANARI; USAI (2005) KIM; PARK; PARK (2013) NELSON; GLOTFELTY (2012) CLEMENT; WU; FISCHER (2014) PACKARD et al. (2016)
Cast	CLEMENT; WU; FISCHER (2014)	SAWHNEY; ELIASHBERG (1996) BREWER; KELLEY; JOZEFOWICZ, (2009) CHANG; KI (2005) COLLINS; HAND; SNELL (2002) BASUROY; CHATTERJEE; RAVID (2003) DELMESTRI; MONTANARI; USAI (2005) TREME (2010) KARNIOUCHINA (2011) KIM; PARK; PARK (2013) MCKENZIE; WALLS (2013) CLEMENT; WU; FISCHER (2014) DERRICK; WILLIAMS; SCOTT (2014) NELSON; GLOTFELTY (2012) PACKARD et al. (2016) SHARDA; DELEN (2006) WALLS (2005) CRAIG; GREENE; VERSACI (2015)
CRITICISM AND WORD OF MOUTH		
Social media	CLEMENT; WU; FISCHER (2014)	KARNIOUCHINA (2011) KIM; PARK; PARK (2013) CLEMENT; WU; FISCHER (2014) CRAIG; GREENE; VERSACI (2015) CHEN; XU; ZHANG (2016) KIM; HONG; KANG (2015, 2017)
Audience review		CHANG; KI (2005) NELSON; GLOTFELTY (2012) PACKARD et al. (2016)

Critic review	CLEMENT; WU; FISCHER (2014)	SAWHNEY; ELIASHBERG (1996) BREWER; KELLEY; JOZEFOWICZ (2009) BASUROY; CHATTERJEE; RAVID (2003) CHANG; KI (2005) COLLINS; HAND; SNELL (2002) TREME (2010) KIM; PARK; PARK (2013) MCKENZIE; WALLS (2013) CLEMENT; WU; FISCHER (2014) PACKARD et al. (2016) TERRY; BUTLER; ARMOND (2003)
Nominations and awards		TERRY; BUTLER; ARMOND (2003) PACKARD et al. (2016)

1.1 Demand prediction model

The basis for this study is a set of 403 films released and screened between 2009 and 2016. Nine attributes are used as predictor variables. These latter were selected for two reasons: they are widely used in empirical studies, as shown in **Chart 1**, and there is information available about them. These variables are genre, age rating, type of production company, seasonality, sequel, release theaters, and star power rating. Projection time and production time, although not used in other studies, were added in this case and have proven significant, as we will discuss below.

The variable to be predicted was based on films' gross revenue, published by Ancine, indexed by IPCA (Brazilian inflation index)-Monthly Cinema and then converted into eight classes divided into 12.5% quartiles. **Table 1** shows the relevant intervals. For the purposes of this model, a movie is considered a blockbuster when its revenue exceeds R$5.8 million.

Table 1: Film distribution among box office performance classes (in Reais, R$)

INTERVAL	CLASS	NUMBER OF FILMS
[5.800.000,00)	A	51
[920.000,5.800.000)	B	50
[170.000,920.000)	C	49
[62.000,170.000)	D	52
[29.000,62.000)	E	50
[13.000,29.000)	F	50
[5.500,13.000)	G	51
[0,5.500)	H	50

1.2 Predictor variables

Genre

The impact of specific genre categories, i.e. whether the film is a romance or a drama, for instance, varies from country to country due to cultural factors that end up affecting consumers' taste and preferences (NELSON, GLOTFELTY, 2012). As empirical studies do not use standard genre classification, it is impossible to compare results for this variable obtained with different models found in the literature.

In this study, we aimed to overcome such shortcoming by resorting to cross-classification from different sources. Accordingly, the genre variable was constructed through a frequency analysis between IMDb, Ministry of Justice, and AdoroCinema ratings.

There were only a few, or none, examples in the sample for a number of genre classifications, as shown in **Table 2**, which is a problem for machine learning techniques. First, because there are not enough examples to train and test the algorithm, then because each classification becomes a different model dimension. Thus, we chose to group all genres for which we had less than 20 films in the "other" class. **Table 2** shows the classification used and the relevant frequencies.

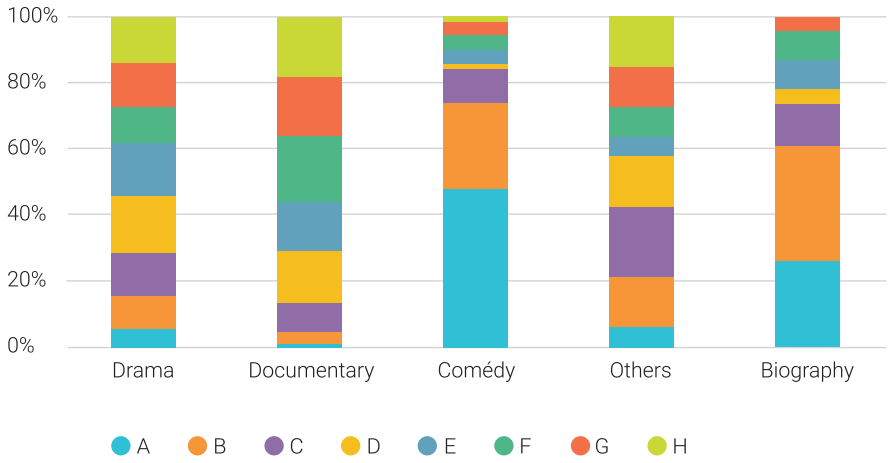
Table 2: Genre classification

INITIAL CLASSIFICATION	
GENRE	NUMBER
Drama	151
Documentary	127
Comedy	69
Biography	23
Thriller	8
Adventure	5
Action	4
AnimAction	4
Musical	3
Horror	2
Crime	1
Historical	1
Family	1
Romance	1
Western	1
Fantasy	1
Detective	1

FINAL CLASSIFICATION	
GENRE	NUMBER
Drama	151
Documentary	127
Comedy	69
Others	33
Biography	23

Figure 1 shows the percentage of films by genre and class: 48% of the comedies and 26% of the biographies in the sample are blockbusters. In other words, these, put together, are the genres that achieve better box office performance.

Figure 1: Percentage of films by genre and class



Age rating

Age rating aims to indicate film content, guiding parents and guardians as to possible sex and violence content. In market terms, age rating indicates the size of the potential audience (CHANG; KI, 2005). Thus, more stringent ratings point to smaller potential audience and, therefore, to slimmer chances of achieving a good box office performance. While a few studies confirmed this hypothesis empirically (CHANG, KI, 2005; NELSON; GLOTFELTY, 2012; WALLS, 2005), others could not compare results for this variable in different models found in the literature (SHARDA; DELEN, 2006; BREWER; KELLEY; JOZEFOWICZ, 2009; KIM; PARK; PARK, 2013).

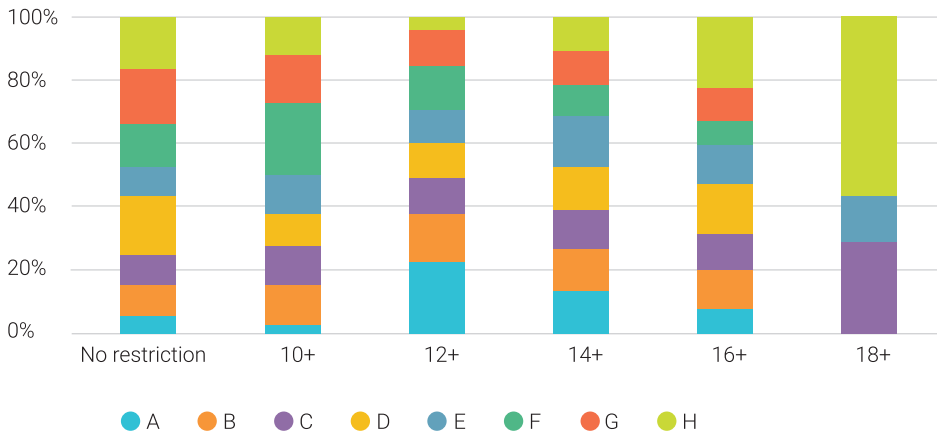
In Brazil, the Coordination for Age Rating (Coordenação de Classificação Indicativa - Cocind), of the Justice Policy Division (Departamento de Promoção de Políticas de Justiça – DPJUS), under the National Department of Justice (Secretaria Nacional de Justiça – SNJ) of the Ministry of Justice, is in charge of assigning age ratings. **Table 3** shows the number of films by age rating (AR) in the sample.

Figure 2 presents the performance of films in each category. While the share of blockbusters is quite well distributed, movies rated as NC-17 stand significantly stronger chances of being a true box office flop.

Table 3: Distribution of age ratings in the sample

AGE RATING	NUMBER
No restrictions	73
10+	40
12+	120
14+	112
16+	51
18+	7
Total	**403**

Figure 2: Percentage of films by class and by age rating



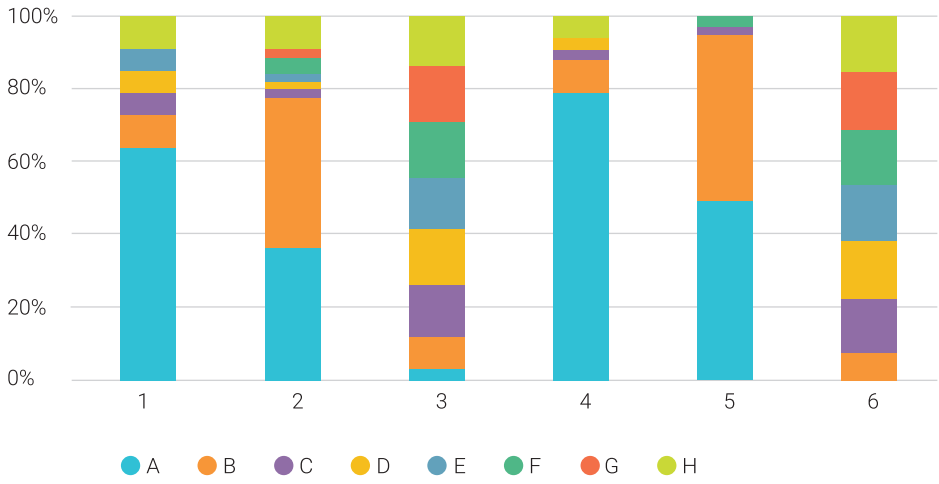
Type of production company

A majority of the studies seek to understand how strong the impact of studios is on movie success. As they are both producers and distributors, studios often evaluate "distributor's power." In Brazil, film production and distribution are not vertically integrated activities. Thus, we tried to evaluate how production companies tend to influence box office success. Therefore, we analyzed production agents, as distributors' influencing ability ought to be reflected in the number of release theaters, another variable in the model.

The operationalization of agents in the chain's influence is usually carried out by determining classes of economic power; majors are those with greater economic power. In this study, we looked at three classes of influence: major, medium, and minor. Major includes production companies that accumulate approximately 85% of total box office in the study period; medium companies accumulate 10% of total box office; and minor production companies are those that take the "fringe", that is, the remaining 5% of box office.

We considered two periods to determine producers' influence and retained the one with better results. For the first, we tried to categorize production companies based on previous year performance. For the second, based on the average performance per movie of last three years. **Figure 3** shows the results.

Figure 3: Comparison between methods of production company classification



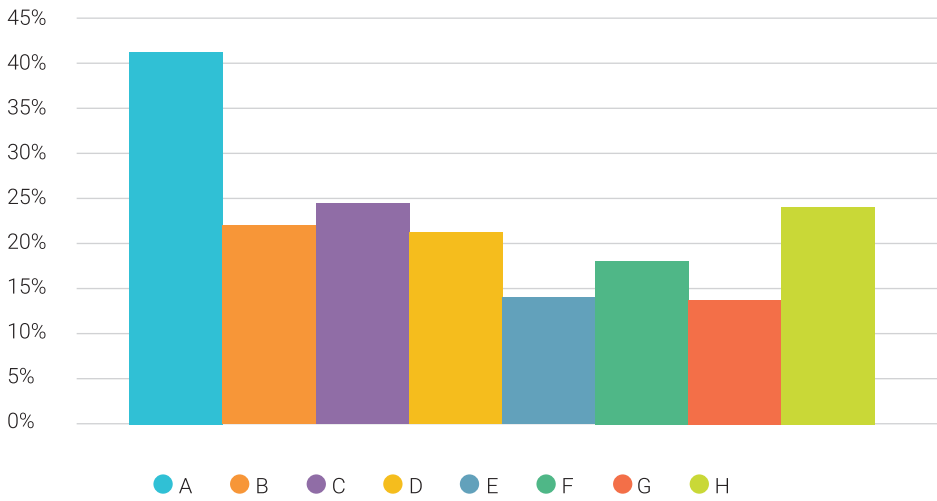
It is clear that a better separation between classes is achieved by using the average of last 3 years. When comparing the average of last 3 years for Medium, most films are in classes A and B (96% of movies), whereas with the previous year method, Medium has only 78% of its movies in classes A and B. For Minor there are no class A movies when using the last 3 years method, whereas with the previous year method 4% of the films are in this category, which generates more noise for the machine learning algorithm, increasing classification error.

Seasonality

The film industry is characterized by high seasonality, with peak demand during holidays and school vacations (DELMESTRI; MONTANARI; USAI, 2005; EINAV, 2007). Therefore, films released in these periods tend to achieve better box office performance.

January, July, and December are considered as school vacations. **Figure 4** shows the percentage of movies released on vacations or holidays.

Figure 4: Percentage of films released on vacation or holidays by class



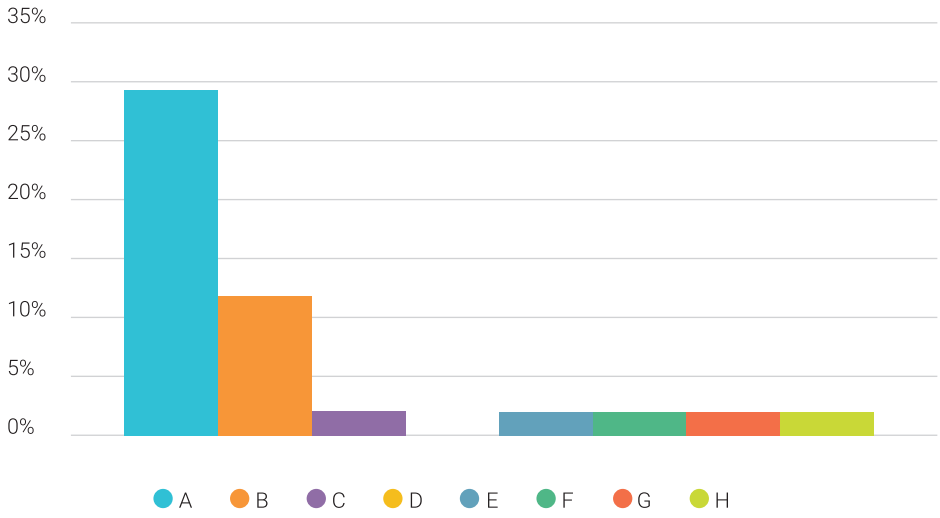
A certain correlation between seasonality and class is apparent. For instance, 41% of blockbusters (class A) were released on these special dates. In the other classes, this percentage falls almost linearly, were it not for class H, which reverses the trend and brings noise to the variable. We hypothesize that, as major studios, such as Disney and Fox, tend to release their productions on these dates, there might be a tendency to relegate domestic films to secondary importance.

Sequel

Sequels are films whose materials and content are based on other films, television shows, videogames or books (BREWER; KELLEY; JOZEFOWICZ, 2009). Due to their pre-existing fan base, it is assumed that sequels face lower risks and stand a lower chance of failure (CRAIG; GREENE; VERSACI, 2015).

Films in this sample were sorted manually and Figure 5 shows results by performance class. At first glance, this proves the hypothesis that sequels face lower risks. However, the small number of films classified as sequels -26, or 6.5% of the base-, makes the number of learning examples insufficient, even in classes with larger number of sequels (A and B).

Figure 5: Sequel percentage by class

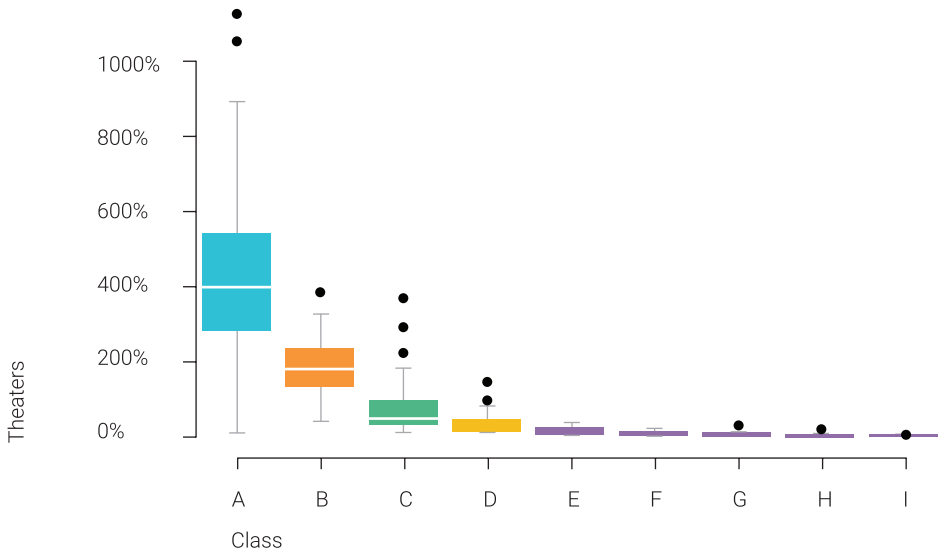


Release theaters

The number of release theaters reflects distributor's expectations about films' performance (CHANG; KI, 2005; CLEMENT; WU; FISCHER, 2014; MCKENZIE; WALLS, 2013). As the box office total is largely determined by films' performance in the first week, the larger the number of release theaters, the wider the audience reach, and hence the greater the chance of success.

As shown in **Figure 6**, the number of release theaters is positively related with box office sales.

Figure 6: Number of release theaters by performance class



Star power

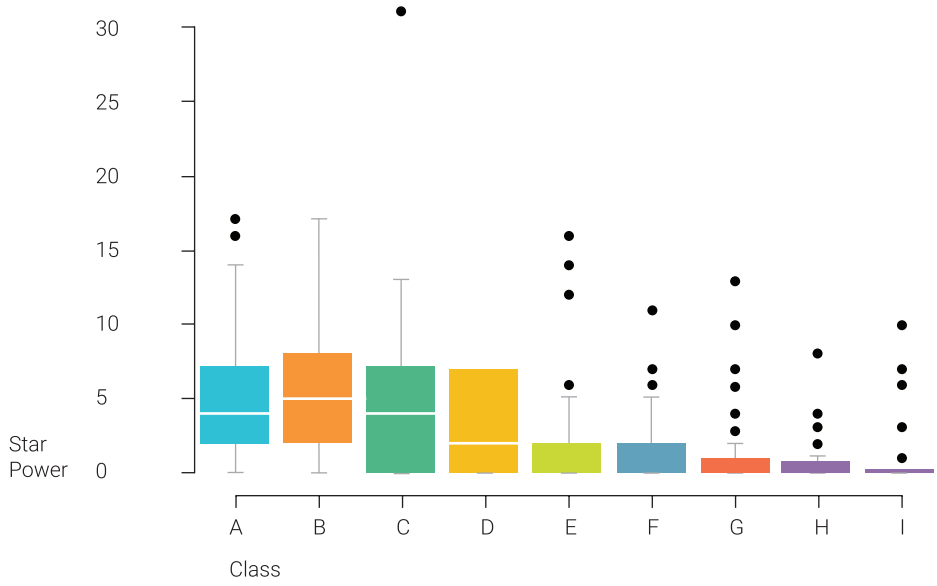
As shown in **Table 1**, empirical studies widely dealt with famous actor's power to attract audiences. Once again, results are hard to compare due to lack of standardization in the way variables are constructed in different models. While a number of studies define as famous actors those having previously achieved expressive performance in terms of audience (HUR, KANG, CHO, 2016) or box office (PACKARD et al., 2016), others do so based on inclusion in popular ranks and magazines (BREWER; KELLEY; JOZEFOWICZ, 2009; TREME, 2010) or nominations and/or awards (BASUROY; CHATTERJEE; RAVID, 2003).

In Brazil, due to *Rede Globo* soap opera's great popularity, the operationalization of this variable (I) took into account the appearance of cast members in this kind of production in the five years preceding the release.

(I) *Star Power*= $\sum_{1}^{n} \alpha_n$, where α represents the number of soap operas in which actor *n* appeared in the last five years.

Therefore, the final value of the variable is the sum of the result for each actor. **Figure 7** shows the relationship between actors' star power and box office performance. Although star power averages for classes with the best economic performance are the highest, many outliers can be spotted in classes with worse economic performance. Thus, featuring a famous actor does not decrease the risk of being a box office flop.

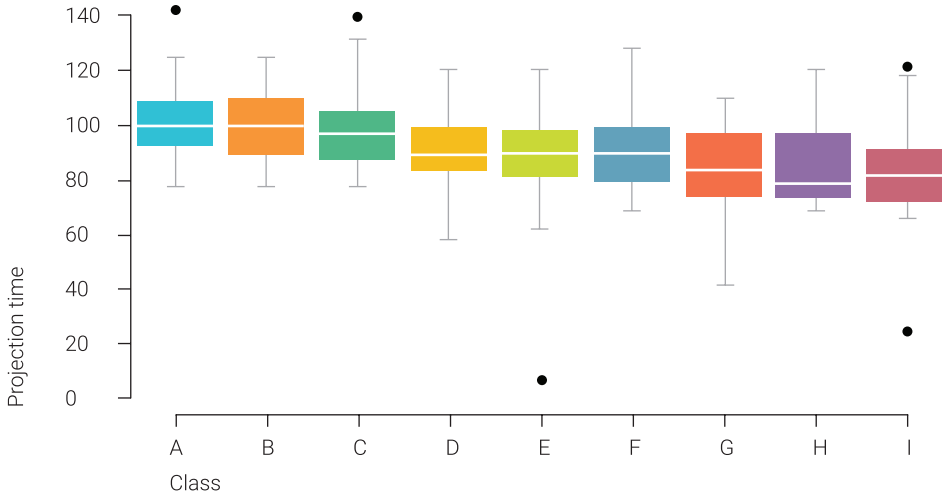
Figure 7: Cast's star power by performance class



Projection time

The influence of projection time on box office performance is poorly established (GHIIASI, LIO, MOON, 2015). In this model, the variable is an integer that indicates film duration in minutes. The best-grossing movies are, on average, one hundred minutes long. With shorter projection times, box office performance also tends to decrease, as shown in **Figure 8**.

Figure 8: Projection time per performance class



Production time

"Production time" has not yet been explored in the literature. It means the time elapsed between the year of production and the film release[5]. The rationale for including this variable is that it is a proxy for entrepreneurial ability.

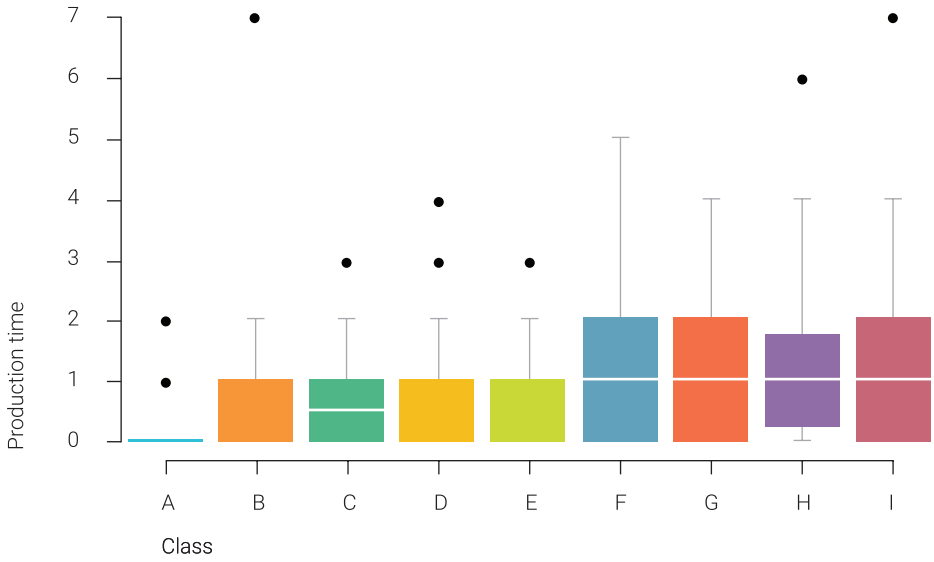
Thus, the shorter the time elapsed between beginning of production and release, the higher the cultural entrepreneur's ability to mobilize resources. This entrepreneurial ability, in turn, increases the chances that audiences will perceive the quality of the film.

Such a hypothesis only makes sense for the domestic context, largely based on the production of dramas, documentaries, and comedies with very little special effects. Therefore, the post-production stage is naturally short.

Figure 9 shows the relationship between production time and box office performance classes. It is apparent that the longer the production time, the worse the performance on average.

5 Production years were obtained from the Filmow website using Scraper, whereas years of release came from the Ancine database.

Figure 9: Relationship between production time and box office performance



2. Methodology

Out of the 403 films in the sample, 80% were used for training (322 films) and 20% for testing (81 films). To divide this base, we used the random_state = 1 parameter, which is important for reproducibility. Thus, every time the experiment is repeated, the division will be the same. The training and testing bases were constructed in a stratified manner, i.e., class percentages were the same in both training and testing base, as shown in **Table 4**.

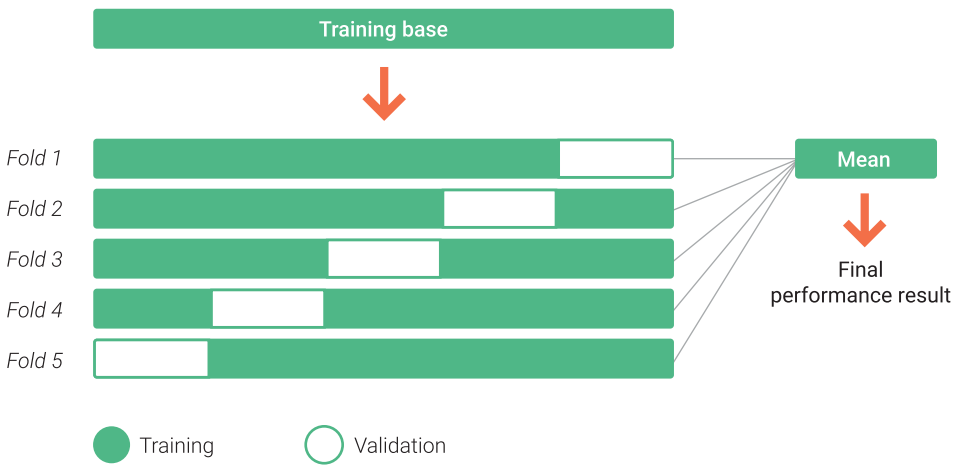
In order to train and validate the model, as well as optimize its parameters, we used "k-fold" cross validation, in which the training base was divided into 5 folds for training and validation, and the final result is the mean for each fold. **Figure 10** illustrates the concept.

.

Table 4 – Class percentages in training and testing bases

CLASS	TRAINING	TESTING
A	41	10
B	40	10
C	39	10
D	41	11
E	40	10
F	40	10
G	41	10
H	40	10
Total	**322**	**81**

Figure 10: "K-fold"cross - validation



When a model is trained, it can suffer for either being oversimplified (high bias), thus underfitting data, or very complex (high variance), thus overfitting data. In both cases, when new data is introduced into the model, this latter will not have enough capabilities to generalize its recommendations. This problem is known as bias-variance trade-off (RASCHKA, 2015). Cross-validation helps estimate the model's generalization error in dealing with data that it had never seen, used to calibrate the model parameters.

In order to calibrate the model, we used the grid search technique, recommended when the parameter space is not very large, that is, when an exhaustive search was performed between parameter combinations. **Chart 2** shows parameters, values to be optimized, and metrics used in the logistic regression model:

Chart 2 – Parameters, values to be optimized and metrics in the logistic regression model

CLASSIFIER	HYPERPARAMETER	VALUES	METRIC
Logistic regression (LR)	C	10^x, $x \in \{-7,-6.82,...,2\}$	Accuracy
	solver	lbfgs	
	multi_class	multinomial	

3. Results

In order to compare results, a DummyClassifier was trained to be used as a baseline. The DummyClassifier is a classifier that makes predictions based on simple rules. We use the majority class strategy. The algorithm takes from the training base the class with more cases and predicts this class for all cases in the testing base. For example, since classes A, D, and G have 41 cases in the training base, the algorithm chooses the first one (A) and predicts that on the testing base all observations are "A", which generates a hit score of 10/81 (12.35%). This number of hits is the model's baseline accuracy.

The model was initially trained without optimizing its hyperparameters to be compared with the baseline. The mean accuracy was 39.42 (± 3.35%) in cross--validation with 5 folds and 48.15% in the testing base. Compared to the baseline of 12.35%, it was almost four times higher.

After optimizing the parameters, the mean accuracy was 42.47 (± 4.65%) in the training base and 46.91% in the testing base. While there was improvement in the training base, there was no improvement in the testing base. Thus, without optimizing the parameters, the model achieved a better result than otherwise.

Table 5 shows each model variable's coefficient for each class. Based on it, **Table 6** was constructed, which includes indications of the importance each variable has for each class. To do so, we used each coefficient's absolute value and analyzed whether their influence on the definition of each class is negative or positive. Then, in order to understand the influence of each variable on class definitions, we developed **Table 7**. This table shows, for each class, the position of the variable; then, the median and the mean of the position of each variable is calculated. The variables were first sorted by median and then, as a tiebreaker criterion, by mean.

From **Table 7** it is possible to gain better understanding of how each variable generally influenced class classification. It is apparent that the most important variable is the number of release theaters, and, surprisingly, the second one is production time, which was not used in previous studies - this is an important contribution.

Table 5 – Logistic regression coefficients

VARIABLE	CLASS							
	A	B	C	D	E	F	G	H
Projection time	-0.05461	-0.03466	-0.01590	0.00301	0.01725	0.01854	0.02251	0.04387
Production time	-0.00274	-0.04151	-0.07159	-0.12149	-0.15833	0.17523	0.17653	0.04389
Seasonality	0.03471	0.00802	0.08062	-0.06886	-0.06327	-0.00862	-0.05029	0.06768
Sequel	-0.00784	0.00653	0.01145	-0.01759	0.00629	0.00975	-0.00575	-0.00285
Release theaters	0.26543	0.25686	0.22728	0.16278	0.04478	-0.03485	-0.11161	-0.81069
Star power	-0.02274	0.03620	0.07242	0.02940	-0.00166	-0.09813	0.03948	-0.05497
Biography genre	0.00820	0.07231	-0.03177	-0.01798	-0.00200	0.00183	-0.01744	-0.01314
Comedy genre	-0.00304	0.01365	0.03166	-0.02266	-0.01451	0.00908	0.00468	-0.01887
Documentary genre	-0.01156	-0.02605	0.02153	-0.03572	-0.04546	0.03220	0.00977	0.05528
Drama genre	0.02954	-0.08244	-0.01898	0.06655	0.06455	-0.02585	-0.00259	-0.03079
Other genres	-0.04474	0.01103	0.01494	-0.00874	-0.01943	0.00751	0.00633	0.03310
No restriction	-0.02680	0.02479	0.02932	0.06272	-0.07037	-0.03487	-0.00139	0.01661
14+	0.02526	-0.04289	-0.02233	-0.04641	0.10019	-0.01112	0.03856	-0.04128
10+	-0.02021	0.00193	0.04046	-0.04551	-0.01020	0.03432	0.01694	-0.01774
16+	-0.00142	0.01795	-0.01791	0.02902	-0.03442	-0.03276	-0.05408	0.09361
18+	-0.00020	-0.00088	0.03815	-0.01352	-0.00080	-0.02333	-0.03072	0.03130
12+	0.00176	-0.01239	-0.05031	-0.00485	-0.00127	0.09253	0.03143	-0.05691
Type of production company: major	0.05833	-0.05146	-0.00816	0.00561	-0.01029	-0.00878	-0.01214	0.02689
Type of production company: medium	0.08457	0.03785	-0.08434	-0.02940	-0.00478	-0.00202	-0.00102	-0.00086
Type of production company: minor	-0.16451	0.00212	0.10988	0.00525	-0.00178	0.03557	0.01391	-0.00044

4. Discussion

The model yielded significant results, with about 48% accuracy, and parameters had a different impact on each performance class.

As expected, the number of release theaters is the variable with the strongest positive impact on box office success. Insofar as it reflects distributor's expectations and stems from negotiations between exhibitors and distributors, we suggest a more in-depth analysis of this relationship should be conducted in future studies.

As to types of film production companies, minors are negatively related to blockbusters. Thus, it is possible to infer that a small company will hardly be able to produce a big box office success.

Despite the positive relation between projection time and box office performance, as seen in **Figure 6**, it should be stressed that this relation is inversed in the best performing classes, for which the longer the projection time, the poorer the box office performance.

Production time, i.e., the entrepreneurial ability, bears a negative relation to classes: the longer the production time, the worse the class. Therefore, this variable's coefficients were negative for the best classes, penalizing films with longer production times. For the worst classes, coefficients are positive, further stressing that, the longer the time, the worse the class.

This coefficient's value is high for some classes, such as D, E, F and G, and it can be the largest one (for E, F, G), which is more important than the number of release theaters. Time is in years, which led to little variation among blockbusters produced and released the same year. In this regard, if values were in months, the results could be even more significant.

As to seasonality, it is apparent that coefficients' values decrease with each class, as shown in this variable descriptive analysis.

There were a few constraints on this study, such as the size of the database and the fact that some characteristics were not defined based on expert evaluation. However, the introduction of the "production time" variable is an important contribution to the literature, as it was not used in previous works and proved an important variable for class separation. Other variables require more in-depth study, such as genre, age rating, and sequel, which in other studies proved more important than in this one, opening opportunities for future research.

6 – Logistic Regression Variables sorted by absolute value and class

VARIABLE	A	INFLU-ENCE
Release theaters	0.26543	+
Type of production company: minor	0.16451	-
Type of production company: medium	0.08457	+
Type of production company major	0.05833	+
Projection time	0.05461	-
Other genres	0.04474	-
Seasonality	0.03471	+
Drama genre	0.02954	+
No restrictions	0.0268	-
14+	0.02526	+
Star power	0.02274	-
10+	0.02021	-
Documentary genre	0.01156	-
Biography genre	0.0082	+
Sequel	0.00784	-
Comedy genre	0.00304	-
Production time	0.00274	-
12+	0.00176	+
16+	0.00142	-
18+	0.0002	-

VARIABLE	B	INFLU-ENCE
Release theaters	0.25686	+
Drama genre	0.08244	-
Biography genre	0.07231	+
Type of production company: major	0.05146	-
14+	0.04289	-
Production time	0.04151	-
Type of production company: medium	0.03785	+
Star power	0.0362	+
Projection time	0.03466	-
Documentary genre	0.02605	-
No restrictions	0.02479	+
16+	0.01795	+
Comedy genre	0.01365	+
12+	0.01239	-
Other genres	0.01103	+
Seasonality	0.00802	+
Sequel	0.00653	+
Type of production company: minor	0.00212	+
10+	0.00193	+
18+	0.00088	-

VARIABLE	C	INFLU-ENCE
Release theaters	0.22728	+
Type of production company: minor	0.10988	+
Tipo produtora: medium	0.08434	-
Seasonality	0.08062	+
Star power	0.07242	+
Production time	0.07159	-
12+	0.05031	-
10+	0.04046	+
18+	0.03815	+
Biography genre	0.03177	-
Comedy genre	0.03166	+
No restrictions	0.02932	+
14+	0.02233	-
Documentary genre	0.02153	+
Drama genre	0.01898	-
16+	0.01791	-
Projection time	0.0159	-
Other genres	0.01494	+
Sequel	0.01145	+
Type of production company: major	0.00816	-

VARIABLE	D	INFLU-ENCE
Release theaters	0.16278	+
Production time	0.12149	-
Seasonality	0.06886	-
Drama genre	0.06655	+
No restrictions	0.06272	+
14+	0.04641	-
10+	0.04551	-
Documentary genre	0.03572	-
Star power	0.0294	+
Type of production company: medium	0.0294	-
16+	0.02902	+
Comedy genre	0.02266	-
Biography genre	0.01798	-
Sequel	0.01759	-
18+	0.01352	-
Other genres	0.00874	-
Type of production company: major	0.00561	+
Type of production company: minor	0.00525	+
12+	0.00485	-
Projection time	0.00301	+

Table 6 – Logistic Regression Variables sorted by absolute value and class (continuation)

VARIABLE	E	INFLU-ENCE	VARIABLE	F	INFLU-ENCE	VARIABLE	G	INFLU-ENCE	VARIABLE	H	INFLU-ENCE
Production time	0.15833	-	Production time	0.17523	+	Production time	0.17653	+	Release theaters	0.81069	-
14+	0.10019	+	Star power	0.09813	-	Release theaters	0.11161	-	16+	0.09361	+
No restrictions	0.07037	-	12+	0.09253	+	16+	0.05408	-	Seasonality	0.06768	+
Drama genre	0.06455	+	Type of production company: minor	0.03557	+	Seasonality	0.05029	-	12+	0.05691	-
Seasonality	0.06327	-	No restrictions	0.03487	-	Star power	0.03948	+	Documentary genre	0.05528	+
Documentary genre	0.04546	-	Release theaters	0.03485	-	14+	0.03856	+	Star power	0.05497	-
Salas de lançamento	0.04478	+	10+	0.03432	+	12+	0.03143	+	Production time	0.04389	+
16+	0.03442	-	16+	0.03276	-	18+	0.03072	-	Projection time	0.04387	+
Other genres	0.01943	-	Documentary genre	0.0322	+	Projection time	0.02251	+	14+	0.04128	-
Projection time	0.01725	+	Drama genre	0.02585	-	Biography genre	0.01744	-	Other genres	0.0331	+
Comedy genre	0.01451	-	18+	0.02333	-	10+	0.01694	+	18+	0.0313	+
Type of production company: major	0.01029	-	Projection time	0.01854	+	Type of production company: minor	0.01391	+	Drama genre	0.03079	-
10+	0.0102	-	14+	0.01112	-	Type of production company: major	0.01214	-	Type of production company: major	0.02689	+
Sequel	0.00629	+	Sequel	0.00975	+	Documentary genre	0.00977	+	Comedy genre	0.01887	-
Type of production company: medium	0.00478	-	Comedy genre	0.00908	+	Other genres	0.00633	+	10+	0.01774	-
Biography genre	0.002	-	Type of production company: major	0.00878	-	Sequel	0.00575	-	No restrictions	0.01661	+
Type of production company: minor	0.00178	-	Seasonality	0.00862	-	Comedy genre	0.00468	+	Biography genre	0.01314	-
Star power	0.00166	-	Other genres	0.00751	+	Drama genre	0.00259	-	Sequel	0.00285	-
12+	0.00127	-	Type of production company: medium	0.00202	-	No restrictions	0.00139	-	Type of production company: medium	0.00086	-
18+	0.0008	-	Biography genre	0.00183	+	Type of production company: medium	0.00102	-	Type of production company: minor	0.00044	-

Table VII – Final classification of variables

VARIABLE	A	B	C	D	E	F	G	H	MEDIAN	MEAN	FINAL CLASSIFICATION
Release theaters	1	1	1	1	7	6	2	1	1.0	2.5	1
Production time	17	6	6	2	1	1	1	7	4.0	5.1	2
Seasonality	7	16	4	3	5	17	4	3	4.5	7.4	3
Star power	11	8	5	9	18	2	5	6	7.0	8.0	4
14+	10	5	13	6	2	13	6	9	7.5	8.0	5
Drama genre	8	2	15	4	4	10	18	12	9.0	9.1	6
Documentary genre	13	10	14	8	6	9	14	5	9.5	9.9	7
16+	19	12	16	11	8	8	3	2	9.5	9.9	8
Projection time	5	9	17	20	10	12	9	8	9.5	11.3	9
No restrictions	9	11	12	5	3	5	19	16	10.0	10.0	10
12+	18	14	7	19	19	3	7	4	10.5	11.4	11
10+	12	19	8	7	13	7	11	15	11.5	11.5	12
Type of production company: medium	3	7	3	10	15	19	20	19	12.5	12.0	13
Type of production company: major	4	4	20	17	12	16	13	13	13.0	12.4	14
18+	20	20	9	15	20	11	8	11	13.0	14.3	15
Biography genre	14	3	10	13	16	20	10	17	13.5	12.9	16
Comedy genre	16	13	11	12	11	15	17	14	13.5	13.6	17
Type of production company: minor	2	18	2	18	17	4	12	20	14.5	11.6	18
Other genres	6	15	18	16	9	18	15	10	15.0	13.4	19
Sequel	15	17	19	14	14	14	16	18	15.5	15.9	20

References

ANCINE. *Anuário estatístico do cinema nacional brasileiro*. 2017. Disponível em: <https://oca.ancine.gov.br/sites/default/files/repositorio/pdf/anuario_2017.pdf>. Acesso em: 4 jan. 2018.

BASUROY, S.; CHATTERJEE, S.; RAVID, S. A. How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, v. 67, n. 4, 2003, p. 103-117.

BREWER, S. M.; KELLEY, J. M.; JOZEFOWICZ, J. J. A blueprint for success in the US film industry. *Applied Economics*, v. 41, n. 5, 2009, p. 589-606.

CHANG, B. H.; KI, E. J. Devising a practical model for predicting theatrical movie success: focusing on the experience good property. *Journal of Media Economics*, v. 18, n. 4, 2005, p. 247-269.

CHEN, R.; XU, W.; ZHANG, X. Dynamic box office forecasting based on microblog data. *Filomat*, v. 30, n. 15, 2016, p. 4.111-4.124.

CHISHOLM, D. C.; FERNÁNDEZ-BLANCO, V.; RAVID, S. A.; WALLS, W. D. Economics of motion pictures: the state of the art. *Journal of Cultural Economics*, v. 39, n. 1, 2015, p. 1-13.

CLEMENT, M.; WU, S.; FISCHER, M. Empirical generalizations of demand and supply dynamics for movies. *International Journal of Research in Marketing*, v. 31, n. 2, 2014, p. 207-223.

COLLINS, A.; HAND, C.; SNELL, M. C. What makes a blockbuster? Economic analysis of film success in the United Kingdom. *Managerial and Decision Economics*, v. 23, n. 6, 2002, p. 343-354.

CRAIG, C. S.; GREENE, W. H.; VERSACI, A. E-word of mouth: early predictor of audience engagement: How pre-release "E-WOM" drives box-office outcomes of movies. *Journal of Advertising Research*, v. 55, n. 1, 2015, p. 62-72.

DELEN, D.; SHARDA, R. Predicting the financial success of Hollywood movies using an information fusion approach. *EndüstriMühendisliğiDergisi*, v. 21, n. 1, 2010, p. 30-37.

DELMESTRI, G.; MONTANARI, F.; USAI, A. Reputation and strength of ties in predicting commercial success and artistic merit of independents in the italian feature film industry. *Journal of Management Studies*, v. 42, n. 5, 2005, p. 975-1.002.

EINAV, L. Seasonality in the U. S. motion picture industry. *The Rand Journal of Economics*, v. 38, n. 1, 2007, p. 127-145.

GHIASSI, M.; LIO, D.; MOON, B. Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, v. 42, n. 6, 2015, p. 3.176-3.193.

GUO, Z.; ZHANG, X.; HOU, Y. Predicting box office receipts of movies with pruned random forest. International Conferenceon Neural Information Processing. 22. 2015, Istambul. Springer. 758p. Disponível em: <https://www.springer.com/gp/book/9783319265544. Acesso em: 5 dez. 2018.

HUR, M.; KANG, P.; CHO, S. Box-office forecasting based on sentiments of movie reviews and Independent subspace method. *Information Sciences*, v. 372, 2016, p. 608-624.

KANZLER, M.; MILLA, J. Focus 2016 – *world film market trends*. [s.l: s.n.].

KARNIOUCHINA, E. V. Impact of star and movie buzz on motion picture distribution and box office revenue. *International Journal of Research in Marketing*, v. 28, n. 1, 2011, p. 62-74.

KIM, S. H.; PARK, N.; PARK, S. H. Exploring the effects of online word of mouth and expert reviews on theatrical movies' box office success. *Journal of Media Economics*, v. 26, n. 2, 2013, p. 98-114.

KIM, T.; HONG, J.; KANG, P. Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting*, v. 31, n. 2, 2015, p. 364-390.

KIM, T.; HONG, J.; KANG, P. Box office forecasting considering competitive environment and word-of-mouth in social networks: a case study of korean film market. *Computational Intelligence and Neuroscience*, 2017.

LEE, K.; PARK, J.; KIM, I.; CHOI, Y. Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*, v. 20, n. 3, 2018, p. 577-588.

LEGOUX, R.; LAROCQUE, D.; LAPORTE, S.; BELMATI, S.; BOQUET, T. The effect of critical reviews on exhibitors' decisions: do reviews affect the survival of a movie on screen? *International Journal of Research in Marketing*, v. 33, n. 2, 2016, p. 357-374.

LIU, L.; ZHAO, Y. Research of box-office prediction based on rough set and support vector machine. *International Journal of Hybrid Information Technology*, v. 9, n. 2, 2016, p. 417-426.

MCKENZIE, J.; WALLS, W. D. Australian films at the australian box office: performance, distribution, and subsidies. *Journal of Cultural Economics*, v. 37, 2013, p. 247-269.

MESSERLIN, P.; PARC, J. The effect of screen quotas and subsidy regime on cultural industry: a case study of french and korean film industries the effect of screen quotas and subsidy regime on cultural industry. *Journal of International Business and Economy*, v. 1.517, n. 2, 2014, p. 57-73.

MICHEL, R. C.; AVELLAR, A. P. Indústria cinematográfica brasileira de 1995 a 2012: Estrutura de mercado e políticas públicas. *Nova Economia*, v. 24, n. 3, 2014, p. 491-516, 2014.

NELSON, R. A.; GLOTFELTY, R. Movie stars and box office revenues: an empirical analysis. *Journal of Cultural Economics*, v. 36, 2012, p. 141-166.

PACKARD, G.; ARIBARG, A.; ELIASHBERG, J.; FOUTZ, N. Z. The role of network embeddedness in film success. *International Journal of Research in Marketing*, v. 33, n. 2, 2016, p. 328-342.

RASCHKA, S., MIRJALILI, V. Python Machine Learning. 2. ed., 2017.

SAWHNEY, M. S.; ELIASHBERG, J. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, v. 15, n. 2, 1996, p. 113.

SHARDA, R.; DELEN, D. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, v. 30, n. 2, 2006, p. 243-254.

TERRY, N.; BUTLER, M.; ARMOND, D. A. DE. Determinants of the box office performance of motion pictures. *Proceedings of the Academy of Marketing Studies*,v. 8, n. 2, 2003, p. 23-28.

TREME, J. Effects of celebrity media exposure on box-office performance. *Journal of Media Economics*, v. 23, n. 1, 2010, p. 5-16.

WALLS, W. D. Modeling movie success when "nobody knows anything": conditional stable-distribution analysis of film returns. *Journal of Cultural Economics*, v. 29, n. 3, 2005, p. 177-190.

ZHANG, L.; LUO, J.; YANG, S. Forecasting box office revenue of movies with BP neural network. *Expert Systems with Applications*, v. 36, n. 3 (II), 2009, p. 6.580-6.587.