Nordic Journal of Media Management

Research article

Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success

Christophe Bruneel [1] 🆔, **Jean-Louis Guy** [2] 🆔, **Dominique Haughton**[3] 🆔, **Nicolas Lemercier** [4] 🆔, **Mark-David McLaughlin** [5] 🆔, **Kevin Mentzer** [6] 🆔, **Quentin Vialle** [7] 🆔, **Changan Zhang**[8] 🆔, **Paul Clemens Murschetz**[9],* 🆔 and **Barira Bakhtawar**[10] 🆔

[1] Toulouse School of Economics, Toulouse, France. Email: Christophe.bruneel@gmail.com

[2] Universite´ Toulouse 1, Toulouse, France. Email: Jean-Louis.Guy@ut-capitole.fr

[3] Bentley University, Waltham, MA, USA. Email: dhaughton@bentley.edu

[4] Universite´ Toulouse 1, Toulouse, France. Email: Nicolas.lemercier@hotmail.fr

[5] Bentley University, Waltham, MA, USA. Email mmclaughlin@bentley.edu

[6] Bryant University, Smithfield, RI, USA. Email: kmentzer@bryant.edu

[7] Universite´ Toulouse, Toulouse, France. Email: Vialle.quentinalexandre@gmail.com

[8] CTrip, Shanghai, China. Email: hellozca@gmail.com

[9] University of Digital Science, Germany. Email: murschetz@berlin-university.digital   (Corresponding author *)

[10] Lahore College for Women University, Pakistan. Email: barira.bakhtawar@lcwu.edu.pk

Abstract:

Purpose: Data analytics techniques can help to predict movie success, as measured by box office sales or Oscar awards. Revenue prediction of a movie before its theatrical release is also an important indicator for attracting investors. While measures for predicting the success of a movie in box office sales and awards are widely missing, this study uses data analytics techniques to present a new measure for prediction of movies' financial success.

Methodology: Data were collected by web-scraping and text mining. Classification and Regression Tree (CART), Random Forests, Conditional Forests, and Gradient Boosting were used and a model for prediction of movies' financial success proposed. Content strategy and generating high profile reviews with complex themes can add to controversy and increase the chance of nomination for major movie awards, including Oscars.

Findings/Contribution: Findings show that data analytics is key to predicting the success of movies. Although predicting sales based on data available *before* the release remains a difficult endeavor, even with state-of-the-art analytics technologies, it potentially reduces the risk of investors, studios and other stakeholders to select successful film candidates and have them chosen before the production process starts. The contribution of this study is to develop a model for predicting box office sales and the chance of nomination for winning Oscars.

🕊 Aalborg University Journals

Practical Implications: Cinema managers and investors can use the proposed model as a guide for predicting movies' financial success.

Keywords: Media Economics; Cinema Economics; Film Financing; Hollywood Economics; Box-Office Revenues; Data Mining; Text Mining; Movie Analytics; Oscars; Prediction Markets; Measurement.

1. Introduction

The movie industry is one of the most profitable sectors of media, generating multi-billion dollars of revenue each year. The global market reached 101 billion USD in 2019, with an 8% increase from the previous year. The U.S market alone reached 36.6 billion USD with a 4% percent increase from the previous year and 25% from 2015 (MPA, 2019); while it was expected to have raised to 50 billion USD in 2020 (Film Industry Statistics, 2018). The figures are a combination of cinema and home/mobile entertainment. Such a huge market size has been created by various types of investors who finance the production of movies and other activities in the value chain, including distribution and exhibition. Research shows that the average investment to produce a movie is 65 million USD (Mueller, 2011). However, the success of a movie is highly uncertain, and a large number of failed movies reflect the risk of investment for investors. Between 2000 and 2010, only one-third of movies in the US were profitable (Lash & Zhao, 2016). So, the prediction of movies' performance is a critical factor to encourage investment in the movie industry and to support growth in this market.

In addition to investors, owners of cinema theatres also face the risk of releasing unsuccessful movies. These owners of movie theatres largely depend on box-office revenue to turn a profit, and selecting the right movies, preferably US blockbusters, is critical for their business survival. However, figures on revenues of movies from box office sales and home/mobile consumption are strongly affected by market changes. The global income from box office revenues is reported as 42.2 billion USD (down from 101 billion USD mentioned above), only increasing by 1% from the previous year. U.S. and Canada box office revenue reached 11.4 billion USD in 2019, a decrease of 4% from 2018 and was nearly equal to 2016 (MPA, 2019). This decline resulted in downsizing the number of cinemas. Digital access and new ways of home/mobile consumption made downstream influenced cinema theaters in the U.S. and the number of cinemas dropped from 7,480 in 1997 to 5,750 in 2017 (Film Industry Statistics, 2018). This is not surprising as home viewing had been considered a major trend in movie consumption long ago (Gomery, 2004: 202).

Other statistics support the fact that cinema owners face a decline in the number of visitors. 54% of Americans prefer to consume the movies in their home and not to go to cinemas anymore, while 19% watch movies once a month, 8% several times a month, and only 4% of Americans buy tickets to watch a movie in the cinema every week. The U.S. film market ranks third for box office revenues in the world, lagging China and India (Film Industry Statistics, 2018). Such statistics reflect the economic challenge of cinema owners in the age of digital distribution (Ulin, 2010).

The economics of the movie shows that the money generated from a film comes from a sequence of profit or release windows. As stressed by Vogel (2014), market control is enacted by these "windowing" practices. These come as revenue-maximization strategies of feature film exploitation which begin their marketing life in domestic theatres and then go on to maximize revenue streams in the ancillary markets, such as global distribution in theatrical and subsidiary markets to pay cable, pay-per-view, commercial TV and home video (Vogel, 2014).

Gomery confirmed that windowing for Hollywood studios is a process that "begins in theaters and then goes 'downstream' to the 'ancillary' markets of pay-TV, home video, and DVD." (2004: 201). The seminal importance of cinema theatres in film economics is further reflected by Gomery (2004) who argued that cinemas are 'voting booths' for the return of investment to be structurally

determined. A good sale of a movie that turns into a blockbuster is likely to bring in millions of USD sales in the downstream windows, as shown in Figure 1. Consequently, advancing cinema ticket sales stimulates success in the downstream windows.

Gaustad (2019) stretched these windows into Cinema; DVD/BluRay; Electronic Sale Through (EST), Video on Demand (VoD), Pay-TV; and Free-TV. He also classified Video-on-Demand into three types of Subscription-based Video on Demand (SVoD), Advertising-based Video on Demand (AVoD) and Transactional Video on Demand (TVoD). Based on this, the full set of profit windows for the refunding film is outlined below (Figure 1).
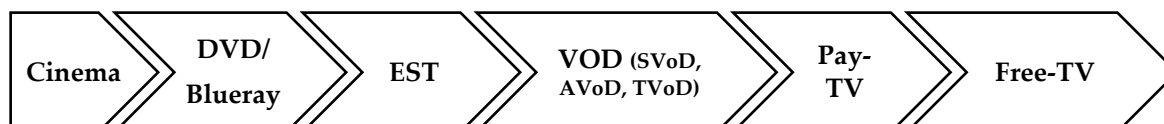
Cinema > DVD/ Blueray > EST > VOD (SVoD, AVoD, TVoD) > Pay-TV > Free-TV

Figure 1: Profit windows for the movie industry (adapted from Gaustad, 2019)

Film exploitation across these windows is becoming progressively more important as a source of (re)financing increasingly expensive film productions, which today can hardly be financed from the receipts generated from the cinema alone (Ravid, 2018). However, the relationship between the various segments still stresses the importance of box office success which largely determines the attractiveness of films. Indeed, as asserted by Debande (2018), under this system, the information generated in the domestic theatrical exhibition market—in terms of box office revenues and word-of-mouth transmission of film quality assessment—has a great influence on consumer demand in the ancillary and foreign exhibition markets.

Given the impact of box office revenues on the whole success of a movie project, prediction measures are a critical tool for investors to choose the right project in advance. Seen this way, the successful prediction of movie performance has become indispensable within the movie industry, since it "immunes" investment for the film (Chakraboty et al, 2006) and, hence, developing effective prediction measures can encourage investment in movies and boost the movie market in principle (Ashenafi et al, 2016; Lash and Zhao, 2016). This study addresses two separated but connected windows: box office sales and Oscars awards. The first window exclusively deals with cinema owners, while sales and awards are likely to affect investors. For this purpose, this research aims at developing a scale to measure the success of movies as part of efforts to assist these two major groups to take the right choice.

2. Literature Review

Before the arrival of new technologies of data mining and customer behavior tracking, the prediction of movie success was not possible. In 1978, Jack Valenti, then CEO of Motion Picture Association of America (MPAA), noticed that before a movie had been shown on theatre screens audience reaction was not measurable, no one could predict a movies' performance in the marketplace. But today, the game has changed. With the emergence of analytics and mining technologies, investors can predict the possible chance of success for a movie. There are several scholarly works on the prediction of movies' success.

Lash and Zhao (2016) developed MIAS, a decision support system to help investors and other stakeholders with models of early prediction of movie success. They used social network analysis and text mining for analysis of historical data of eleven years extracted from various sources. They extracted features such as the cast, the subject of a movie, the time of release and the combination of

such attributes of a movie and suggested a framework for a decision support system, MIAS, to assist investors to decide whether a proposed movie is worth investing. In their definition, the success of movies was best reflected by profitability instead of revenue. They suggested several directions for future research, considering variables such as script, formal collaboration networks among actors, informal friendship among the cast, internet advertising. These could impact the success of movies beyond the features of who, what, and where.

Delan and Sharda (2012) used machine learning methods for forecasting the financial success of Hollywood movies. They focused on box office sales as a predictor of the success of a movie and the determinant of the success of a movie. Based on IMDB and ShowBiz data, they selected 386 movies released in 2009 and 2010. Box office gross revenue was considered as a dependent variable. Other revenues such as home video, commercialization and international markets were excluded. They also used a nine-category system to rank the movies from flop to blockbuster. They developed performance metrics, APHR to predict the success of each movie in comparison with the other movies in the same category. Considering the limitation of their metrics, they suggested that other methods can improve the accuracy of the system.

Henning-Thurau et al (2007) studied determinant factors for the profitability of movies at the box office. They suggested variables such as production cost, advertising expenses, review, consumer perceived quality, the number of awards won, short-term theatrical success and long-term theatrical success. Summer release, cultural familiarity, level of star power, and the level of director power. By using path analysis, they developed a box office model and a profitability model. Their profitability model showed that except for star power and director power all other factors have a significant impact on movie profit. In the case of box office revenue, their model showed that production cost, cultural familiarity, summer release, and the customers' perception of quality impact on short term box office revenue, while production cost, cultural familiarity, and awards impact on long term box office revenue.

Ashenafi et al (2016), studied the top ten trending movies from 2013 to 2015, a total of 30 movies, and analyzed the critics' reviews, budget, and domestic box office performance. They suggested a multiple linear regression model and showed that critics review and budget can explain part of the performance of the box office. They also suggested using a larger sample and more extraneous variables for future research. They suggested that actual returns on investment can be a direct measure for the profitability of a film.

Smith and Pangarker (2013) studied the determinants of box office performance in the film industry. They identified certain film genres, MPAA ratings, the size of the budget, major studio involvement, academy award nominations or awards, time of release, and critics review as factors that impact on box office sales. Their findings showed that production cost is the most significant factor that impacts box office revenue. The major studio that produced the film and award nominations are the factors that influence the success of the movie at the box office. Other factors also contributed to box-office performance such as film sequel. The findings show that there is no significant relationship between holiday release and critics' reviews.

Bhave et al (2015) categorized the success factors of a movie into two classes of classical factors and social factors. The classical factors include producer, production house, director, cast, the runtime of the movie, the genre, the script, the time of release and marketing. Social factors include the IMDb ratings, the viewer and critics reviews, and ongoing social, cultural, political and economic trends. They argued that factors from both categories are required for a movie to hit the box office. Chakraboty et al (2006) analyzed the factors that are involved in the prediction of movie success, including budget, actors, director, producer, IMDb rating, IMDb Metascore, IMDb vote count, actors and director social fan following, Wikipedia views, and trailer views.

3. Materials and Methods

This study aims at forecasting box office revenues obtained during the first week of the movie release. Such prediction is useful as cinema owners and their managers can now better decide what movies are to be shown in their theaters. As illustrated in figure 1, the largest portion of a movie's revenue (40% on average) is obtained from the box office sale during the first week of a release, when other cinema managers are yet to decide whether to show the movie or not.

Due to the importance of the cinema managers' decision about what movie to show, we had put ourselves in the shoes of managers to consider a real-life situation, based on all information available, and see how we can make decisions.
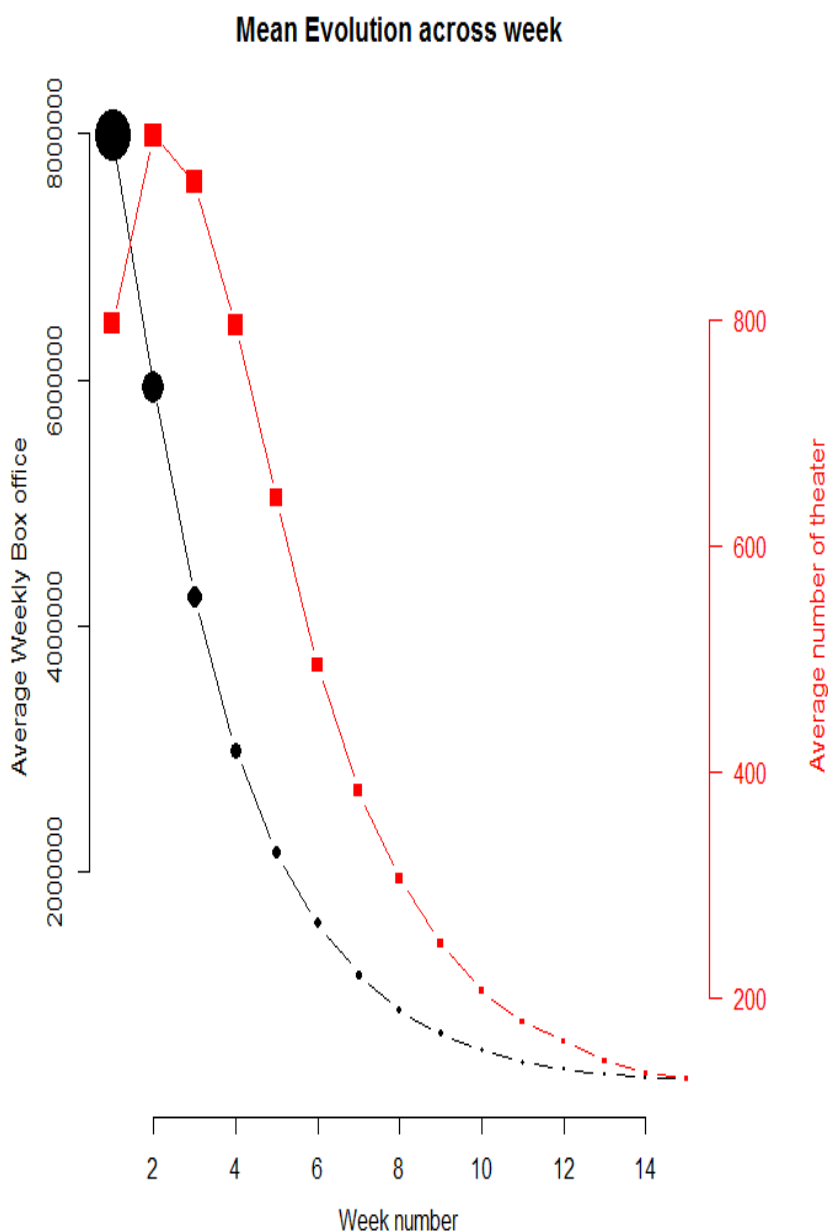

Figure 1: Average weekly box office performance and average number of movie theaters showing a movie (Source: computed by the authors, reprinted with permission)

3.1. Sample

The sample of this study consists of movies that were released from 2000 to the end of 2011. More specifically, all films whose first week of screening was over by January 1st, 2012 are considered (2166 films) from models built on data available ex-ante, i.e. based on films released before January 2012. We use different time learning samples for our model: 1) films released from 2000 to 2012: 5874 movies; 2) films released from 2006 to 2012: 3392; 3) films released from 2008 to 2012: 2193; 4) films released from 2010 to 2012: 1090; and 5) films released in 2012 (ex-post sample): 566 movies.

Dividing the sample in the above time series allowed us to test the respective box office revenues and to select the most suitable sample. This also helped to control the 'optimal temporal horizon' which refers to an unobserved economic conjuncture.

3.2. Model

The following predictive models are used: 1) Linear regression, 2) Decision tree, 3) Random forest, 4) Conditional forest, 5) Gradient boosting. Also, to optimize the result a so-called 'stacking method' was used. These models were applied to ex-ante data (as of above) to predict the financial success of movies that were released after January 2012 (566 movies). The advantage of this sampling is that no movies below the average revenue are excluded and all movies in the selected period are included. Finally, the predictions of the model were compared with the actual revenues of the movies that were released in the years 2013 to 2015. In all, the objective of this research is to identify the optimal period a manager must observe as a test period to construct his/her estimation model and the modeling technique with the highest predictive power.

3.3. Data

The data of this research were collected via a web-scraping technique on www.boxofficemojo.com. The database built from the data collection included detailed information about the daily, weekly, and total box office revenue of 15,459 movies in the U.S. Also, the database contains information about the genres, the production budget, distributor company, cast, and other relevant details.

R software was used for extracting the data from the above-mentioned website. The process consists of 1) extraction of all the links to the movies' pages, 2) creation of a 'scraping function' to execute on the extracted links. As a result, a rich database was created. However, several data transformations were necessary to be able to make use of it.

First, box office revenues were deflated by the monthly CPI of the weeks the movies were released to make the intertemporal comparison possible. If information was unavailable, we deflated global box office revenues by CPI of the release year. The deflated figures of the top 10 of the movies with the highest box office revenues in normal, as well as real terms, are presented in Table 1.

Second, only the production budget of less than 25 percent of movies was available. We split this information into two sub-variables: 1) a first binary variable to indicate whether the information is available or not, 2) the second variable of interaction between the binary availability variable and the production budget. This variable is, therefore, zero when the information is not available and equals the production budget otherwise. This 'dichotomy' is the best way to maintain this variable, which is positively correlated with the box office, thereby avoiding a non-negligible bias.

Table 1: Top 10 movies in case of the highest box office revenue, based on January 2010 (Source: computed by the authors, reprinted by permission)

Rank	Top Films (nominal)	Year	BO nominal (in $ millions)	Top Films (real)	Year	BO real (in $ millions)
1	Avatar	2009	750	Star Wars 4	1977	1, 099
2	Avengers	2012	623	Titanic	1997	808
3	Titanic	1997	601	E.T.	1982	805
4	Batman: Dark Knight	2008	533	Avatar	2009	750
5	Batman : DKR	2012	448	Avengers	2012	593
6	Avengers 2	2015	445	Star Wars 1	1999	565
7	Shrek 2	2004	441	Star Wars 6	1983	554
8	Star Wars 1	1999	431	Star Wars 5	1980	552
9	Hunger Games: Catching Fire	2013	425	Jurassic Park	1993	538
10	Pirates of the Caribbean 2	2006	423	Grease	1978	535

4. Results

4.1. Descriptive Statistics

4.1.1. Box Office Revenue

As illustrated in Figure 2, the distribution of the logarithm of box office revenue in the first week follows a bimodal distribution, with the main mode representing movies with a relatively low (< USD 100,000) box office revenue in the first week. The second mode represents movies with an average to high box office revenue. As shown, the box office revenue is not normally distributed, and the problem lies in this bimodality. The usual literature which implicitly excludes low-performing films, that is most of the films, concentrates only on the 'second mode' of the distribution.

Such a problem can be seen in most of the previous research that excluded financially unsuccessful movies and merely included movies with average to high revenue. Of course, if we do not have an ex-ante method of knowing which mode the box office revenue will belong to, working only on higher revenue films, does not inform cinema managers if the model for higher revenue movies does not apply to lower revenue movies.
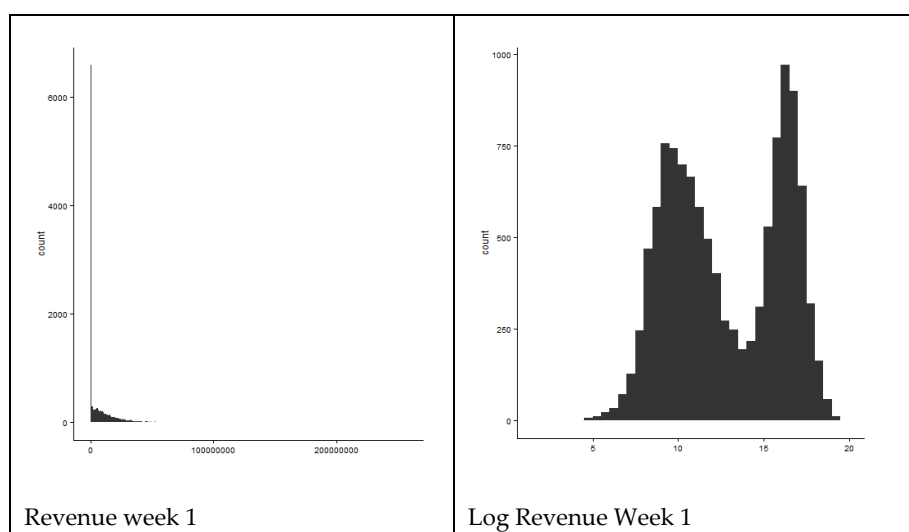

Revenue week 1   Log Revenue Week 1

Figure 2: Distribution of operating revenue in the first week (Source: Computed by the authors, reprinted with permission)

4.1.2. Production Budget

Tables 2 and 3 present summary statistics with deflated values of production budgets.

Table 2. Description: Production budget (Source: Computed by the authors)

Statistic	Min	Median	Max	Mean	Standard Deviation
Is the production budget available? (0=no, 1=yes)		0		0.24	
Deflated Production budget	0	0	315 573 505	11 678 181	30 955 336

4.1.3. Star power: actors, directors, producers, and distributors

Table 3: Description: Star Power (Source: Computed by the authors, reprinted with permission)

Statistic	Median	Mean	Standard Deviation
Actors			
Star power of actors in the film	4 274 693	972 238 048	1 888 323 059
Sum of number of previous films actors have acted in	1	18.97	33.70
Sum of Oscar nominations of actors in the film	0	0.17	0.52
Director			
Star power of the director	0	79 763 117	303 180 322
Number of previous films directed by director	0	1.37	3.50
Number of nominations to Oscars for the director	0	0.07	0.42
Distributor			
Star power of the distributor of the film	351 598 009	5 786 861 069	9 724 837 497
Number of previous films distributed by the distributor	77	182.86	218.13
Number of films distributed by the distributor nominated for a Best Movie Oscar	0	4.09	6.30
Producers			
Star power of producers of the film	0	463 742 505	1 428 304 802
Sum of the number of previous films produced by the producers	0	7.37	19.86
Sum of the number of previous films produced by the producers which were nominated for a Best Movie Oscar	0	0.26	1.06

4.1.4. Other explanatory variables and controls

Tables 4 to 8 present the other variables that cinema managers access for prediction, including genre and MPAA rating. There are also other variables such as limited release before wide launch and a variable for the days of such limited release. Another variable considers the day of the week that the movie was launched as well as the month of release and seasonal effect on it.

Table 4: Description: Genres of films where a movie may have several genres (Source: Computed by the authors, reprinted with permission)

Genre	Proportion
Romance	0.05
Adventure	0.03
Family	0.03
Comedy	0.23
Documentary	0.10
Action	0.08
Drama	0.20
Fantasy	0.02
Foreign	0.13
Horror	0.06
Thriller	0.08
Musical	0.02
Crime	0.03
Western	0.005
Science fiction	0.03
War	0.01
Animation	0.03
Sport	0.01
Histoire	0.004
Epic	0.001
Period	0.02

Table 5: MPAA rating: proportions (Source: Computed by the authors, reprinted with permission)

GP	0.0001
NC-17	0.002
PG	0.128
PG-13	0.212
R	0.383
Unrated	0.251

Table 6: Description of other controls (Source: Computed by the authors, reprinted with permission)

Other controls	Mean
Dummy limited release	0.04
Length in weeks of limited release	1.02
Dummy Remake	0.02
Dummy book adaptation	0.02
Dummy prequel	0.003
Dummy series	0.05
Dummy foreign language	0.13
Dummy palm at Cannes	0.002

Table 7: Control variables: seasonality and day of release (Source: Computed by the authors, reprinted with permission)

Month of release	Proportion
01	0.067
02	0.076
03	0.094
04	0.094
05	0.082
06	0.074
07	0.075
08	0.092
09	0.094
10	0.103
11	0.080
12	0.070

Table 8: Day of release (Source: Computed by the authors, reprinted with permission)

Day of release	Proportion
Sunday	0.001
Thursday.	0.009
Monday.	0.001
Tuesday	0.004
Wednesday	0.081
Saturday	0.003
Friday	0.900

4.2. Model

As explained, four different models are used to predict the box office revenue in terms of ex-ante information. These four models are explained as follows:

4.2.1. Classification And Regression Tree (CART)

Classification And Regression Tree (CART) has been used as the first model. This technique is an improved nonlinear and entirely non-parametric statistical learning technique first introduced by Breiman, Friedman, Stone, & Olshen (1984). This technique enables researchers to predict a dependent variable from independent variables. To do this, a decision tree should be built, and graphical illustrations need to be used to interpret it. CART proceeds as follows. at each node, the algorithm splits the dataset into two subsets, using any possible predictor and any cut-off point for continuous predictors, in such a way that the two subsets be as homogeneous as possible concerning the dependent variable. This technique has the advantage of being non-parametric, thus not postulating any a priori assumption on the distribution of the data, being robust to outliers, and supporting all types of variables. Also, the CART algorithm effectively handles missing values. When the learning sample is as large as in the present case for most reference periods, the CART algorithm has properties that are similar to the nearest neighbor algorithm. On the other hand, limitations include the inability to detect combinations of variables as effective predictors, and the need for a large sample (which may be problematic for the period 2010-2012 and 2011-2012).

4.2.2. Random Forests

Random Forest has been known as the most powerful predictor among the statistical learning techniques. Breiman (2001) developed this technique by adapting the decision tree for bootstrap aggregating (bagging); which is used to reduce the variance of an estimated prediction function with law bias. Random Forest is a powerful statistical learning technique (often considered as the most powerful predictor available) developed by Breiman in 2001 (Breiman, 2001) that adapts decision trees for Bootstrap Aggregating (bagging). Bagging is a technique used to reduce the variance of an estimated prediction function while maintaining a relatively low bias. Here, this technique is particularly well suited since the variance of the box-office revenue variable is very large. It is therefore expected that this method will be more efficient than decision trees. On the other hand, as for all models built by aggregation, there is no direct interpretation. The *Random Forest* algorithm proceeds with a double random selection of both predictors and data (via a bootstrap of the learning sample), and the majority vote on the resulting CART trees (hence the name of *Random Forest*).

4.2.3. Conditional Forests

The *Conditional Forest* algorithm developed by Hothorn et al (2006) makes it possible to remedy the problems faced by Random Forest such as selection bias or overfitting. It is therefore expected that this technique will perform at least as well as Random Forest. One of the main disadvantages of this method is that the underlying algorithm takes much longer to run than Random Forest since it performs tests to select the variables.

4.2.4. Gradient Boosting

This algorithm first introduced by Freund and Schapire (1996) is "a method of prediction that minimizes several types of the loss function to a prediction function" (Bruneel et al, 2018: 563). This method can adapt to any type of data. Even in a situation where the number of variables exceeds the number of observations, good results are generated.

4.3. Results

4.3.1. Results of four models

The results of each of the four above-mentioned models are presented in Table 9. It shows that regardless of the estimation method, an increase in the number of observations that serve as reference leads to an increase in the R-square; therefore, the use of more movies to construct the model does better explains the variance. The results are similar when using movies released between 2000 and 2012 or those released between 2006 and 2012 (the best-explained variance of 79% is for this latter period). Therefore, arguably, our optimal time range of films to be considered for estimation is 6 years (with a preference for the 2006-2012 range), if we wanted to maximize the explained variance. Random forests, conditional forests and gradient boosting seem to be the three methods giving marginally better results. This makes sense given their complexity. However, the difference in performance between the best models and classical linear regression remains marginal.

For obtaining higher coefficients of determination, other variables that indicate quality expectations of movies are needed. Neither use nor cinema managers have access to such information. Root Mean Square Error (RMSE) is large, (10 million US dollars in January 2010), and the average error rate is very high too, between 596.99% to 22.30%. Such figures are due to the presence of extreme values and explain the reason why the use of Root Median Square Error and the median error rate for this matter. Since the RMSE fluctuated between 110,000 and 2,000,000 USD, the indicators of predictive power are preferable. These indicators support the idea that Random Forests and Gradient Boosting are the best models.

Gradient Boosting is dominated when it has the longest range of data available (2000-2012 or 2006-2012) with a median error rate ranging between 100 and 110%. Such very large rates show that even the most advanced predictive techniques, which only apply ex-ante information, cannot correctly recognize which movies will fail at the box office. The stronger predictive models require more exact data, for example, data from social networks. However, with the available data, it is possible to identify the more important variables in predicting movies' income from the box office as presented in Table 9.

Table 9: Estimation results for box office revenue in the first week (Source: Computed by the authors, reprinted with permission)

Reference Period	Method	R2 (adjusted)	Root Mean Square Error	Average relative error	Root Median Square Error	Median relative error
2000-2012	OLS	0.74	10901374.00	253.91	1494037.00	26.99
2000-2012	decision tree	0.72	11265019.00	143.44	743540.80	21.77
2000-2012	random forest	0.78	9986626.00	36.91	188348.30	1.56
2000-2012	conditional forest	0.77	10218146.00	43.90	218966.50	6.86
2000-2012	gradient boosting	0.76	10414030.00	35.92	153258.20	1.00
2006-2012	OLS	0.74	10909403.00	264.01	1276029.00	26.29
2006-2012	decision tree	0.71	11517285.00	189.99	943572.10	28.13
2006-2012	random forest	0.79	9867366.00	27.79	180777.40	1.41
2006-2012	conditional forest	0.76	10507541.00	47.14	216554.30	7.97
2006-2012	gradient boosting	0.76	10531481.00	38.15	175234.30	1.10
2008-2012	OLS	0.73	10958782.00	309.64	1349485.00	29.10
2008-2012	decision tree	0.66	12396390.00	244.44	1119748.00	33.23
2008-2012	random forest	0.76	10378329.00	37.97	168820.90	1.43
2008-2012	conditional forest	0.74	10813495.00	41.55	190900.50	6.88
2008-2012	gradient boosting	0.76	10486929.00	42.28	156116.70	1.23
2010-2012	OLS	0.72	11344506.00	382.79	1749335.00	28.29
2010-2012	decision tree	0.31	17701598.00	80.72	375038.40	11.47
2010-2012	random forest	0.74	10769061.00	22.30	113640.10	1.31
2010-2012	conditional forest	0.71	11478915.00	27.28	126784.80	4.33
2010-2012	gradient boosting	0.74	10835496.00	34.14	161450.40	1.73
2011-2012	OLS	0.63	13014399.00	596.99	2007622.00	40.80
2011-2012	decision tree	0.58	13845035.00	66.04	336170.90	9.77
2011-2012	random forest	0.70	11709300.00	66.24	156481.00	1.48
2011-2012	conditional forest	0.63	12921583.00	32.81	156024.20	5.24
2011-2012	gradient boosting	0.69	11891466.00	31.08	153910.30	2.01

Figure 3 depicts which variables are more important than others in predicting box office revenue, based on the period of 2006-2012. It reveals that the movie production budget is the most important variable. This figure also shows that seasonality is an important variable in the prediction model.
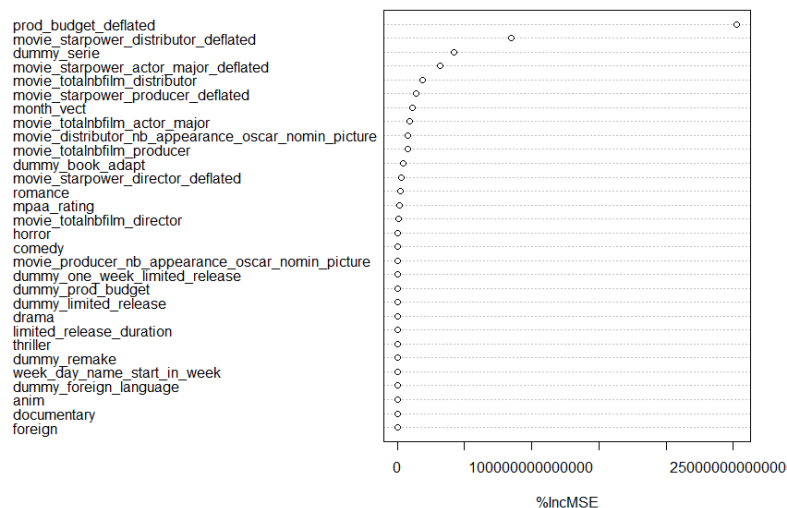

Figure 3: Importance of variables (Source: Computed by the authors, reprinted with permission)

4.3.2. Prediction Markets for Predicting Oscars

An estimation of the probability of a win in the prediction market such as, for example, an Oscar needs to be accurate (Haughton et al, 2015). Prediction markets have successfully estimated the wins in different entertainment events such as Grammy Awards and Oscars (Gold, McClarren, & Gaughan, 2013), elections (Saxon, 2010; Rothschild & Wolfers, 2008; Erikson & Wlezien, 2008), and in research on the probability of a U.S. recession (Leamer, 2008). Prediction markets did also well in the 2015 Academy Awards (Leonhardt, 2015).

The price per contract for each of the nominees winning the Oscar for the 2013 Best Picture award is presented in Figure 4. As shown, until December it was not clear who is the front runner. In early December, *Lincoln* took advantage, but in January the film *Argo* surpassed *Lincoln*, kept position and finally won the award. Analysis of the average contract price for each of the movies reveals that the top contenders for the best picture award, including *Argo*, *Les Misérables*, *Lincoln*, *Silver Lining Playbook*, and *Zero Dark Thirty*. Seeking to understand the reason why *Argo* surpassed *Lincoln* in late January, we found that on January 26, 2013, *Los Angeles Time* published an article with the headline "The Gold Standard; now for real insight into Oscars – by the guilds", arguing that the Producer Guilds of America Awards (PGA) is a reliable predictor of success. That evening *Argo* won the Zanuck Award for Outstanding producer of theatrical motion pictures.
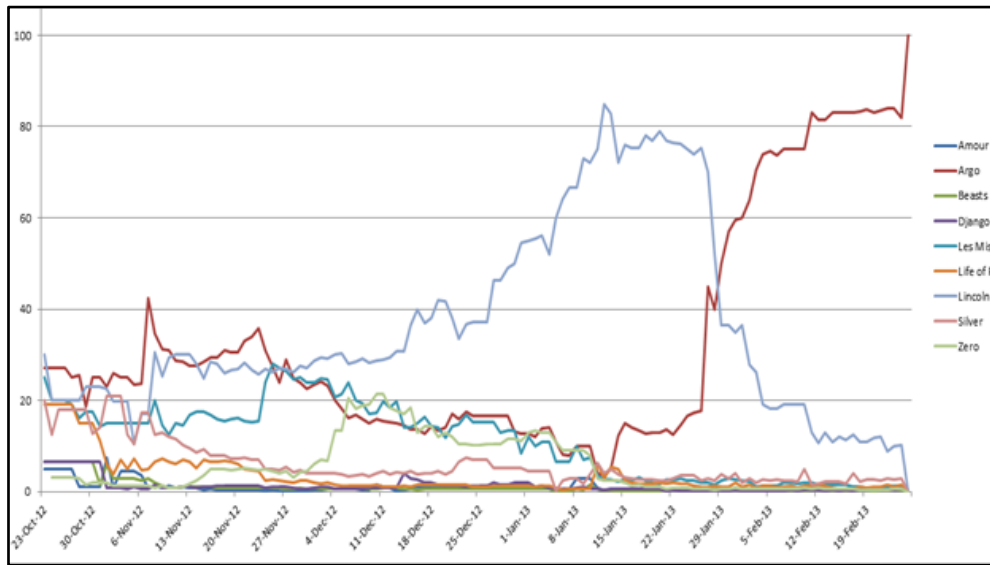
Figure 4: *Close Intrade* contract prices for each nominated movie to win the 2013 *Best Picture Award*. (Source: Extracted by the authors, reprinted with permission)

Following PGA, in the 19th Annual Screen Actors Guild (SAG) *Argo* was awarded the Outstanding Performance. Therefore, the results of these two awards can be used to predict the Oscar for the best picture award. We also considered the award winners of the past decade. Over the past decade, five times, the PGA and SAG were awarded to the same movie; and 4 out of that 5 times the Oscars also went to that movie. In the remaining five years that the PGA and SAG awarded different movies, the Oscars selected one of those awarded movies and only one time, in 2004, Oscars selected a movie different from both PGA and SAG.

Table 10: PGA awards, SAG awards and *Oscars* (Reprinted with permission)

	PGA	**SAG**	**Oscars**
2012	Argo	Argo	Argo
2011	The Artist	The Help	The Artist
2010	The King's Speech	The King's Speech	The King's Speech
2009	The Hurt Locker	Inglourious Basterds	The Hurt Locker
2008	Slumdog Millionaire	Slumdog Millionaire	Slumdog Millionaire
2007	No Country for Old Men	No Country for Old Men	No Country for Old Men
2006	Little Miss Sunshine	Little Miss Sunshine	The Departed
2005	Brokeback Mountain	Crash	Crash
2004	The Aviator	Sideways	Million Dollar Baby

4.3.2. Movie Review Data for Predicting Oscars

In this section we study the reviews for the movies which were nominated for the Best Picture Award, to see if there is any indicator for prediction of the award-winning movie. We assume that 'controversy' is a measure for prediction of the chance of a movie to win the award. This measure can be extracted by text analysis of the reviews of movies. We used the IMDB reviews as the data for the extraction of the value of controversy.

4.3.3. IMDB Review

The use of the IMDB database for text mining of ideas and opinions of movie watchers has several advantages over Twitter, which is usually considered as a source of data for this method (Kolli & Khajeheian, 2020). The first advantage is the length of IMDB reviews compared with 280

characters of tweets; which generates a richer and more complex source of opinions. The second advantage is the quality of review writers in IMDB, which can be obtained by filtering of phrase 'Prolific Author' on the review page (look at figure 5). The third advantage is the feature of voting in favor of or against a review. This can be used as a sign of the quality of the respective review. The disadvantage, on the other hand, is that IMDB does not provide an API or other possibilities for download reviews. Therefore, we had to crawl the raw HTML pages to extract the review data.

Figure 5: "Argo" IMDb reviews including prolific authors only (Source: extracted by the authors, reprinted with permission)

4.3.4. Review themes and Predicting the Chance for Oscar

The text mining of the IMDB reviews which are collected before the Oscars reflects the different themes from watchers of each movie and provide a preliminary indicator of controversy. So, the question is what is the optimal level of controversy for a movie to win the Oscar? Chang & Li (2010) discussed the use of controversy in marketing. They suggest that the standard deviation of the numerical ratings is one possible quantitative measure of controversy. This perspective is properly applicable to our study for measuring the chance of movies to win an Oscar. To extract themes from movie reviews, we use the text mining algorithms proposed by SAS *Text Miner* within the *Enterprise Miner* platform. Details of the algorithm are published elsewhere, but the algorithms work essentially as follows: Each review is defined to be a document, and a very large but sparse matrix is constructed with documents as rows and all possible terms (words in documents and their grammatical relatives, such as begin, began, beginning, etc.) as columns. *Singular Matrix Decomposition* (SVD) techniques are used to reduce the matrix without losing too much information and cluster analysis is applied to the reduced matrix, yielding for each set of reviews, a set of clusters of documents. The list of most common terms in these documents is then obtained and gives an idea of the main themes in that cluster. Here, cluster 3 is related to perceived Oscar chances for the movie, director and Ben Affleck, the leading actor.

Table 11: Clusters and main terms for *Argo* reviews. (Source: Computed by the authors, reprinted with permission)

Cluster	Main terms	No. of documents
1	tony +ambassador +plan +embassy Mendez Canadian six +hostage +crisis chambers CIA fake john goodman Arkin	142
2	+movie people watching +good movies great +world first +end characters +fact +country history historical +time	95
3	best +picture acting +great +oscar well +good affleck +actor +director argo alan ben +film +movie	149
4	+feel +seat +edge characters +little especially few films +know +thriller suspense +end +fact +film fake	22
5	Canadians shah airport history Iranians Americans +country Canadian people Iranian events historical +fact CIA American	72
6	chambers Bryan +ambassador Cranston +plan +crisis john Mendez Iranian +actor tony fake +thriller alan especially	44

To understand how this controversial method works, a comparison of *Argo* and *Amour* can be used as an example. The text analysis of Argo generated 6 clusters, while for Amour, a movie by Heneke, a controversial director, with a complex theme of death and euthanasia, 23 clusters generated. As Zhang & Li (2010) discussed, the number of themes is important not their positive or negative position. Thus, reasonably, the number of issues such a complex movie rises may be simply too large for a group to rally on. A scatter plot of the standard deviation of ratings against the number of clusters that is extracted by the text analysis of the nine movies is presented in Figure 6.
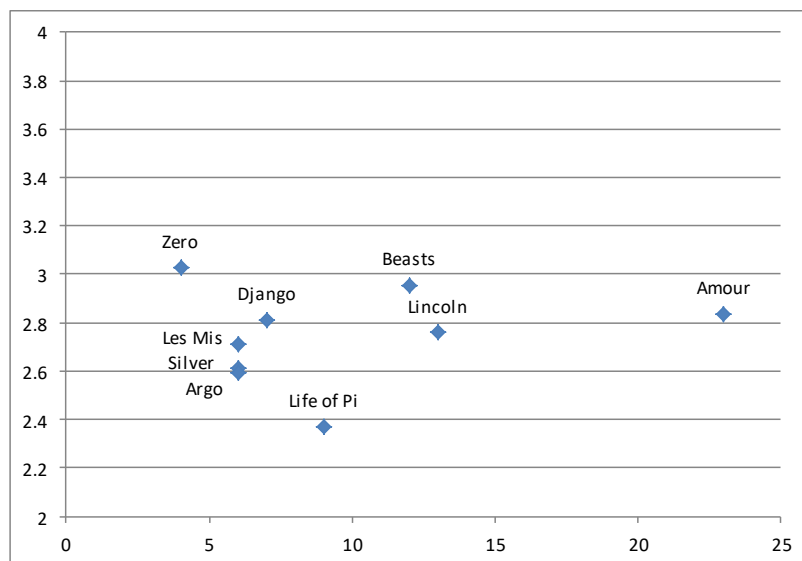

Figure 6: Standard deviation of ratings and the number of clusters for each nominated movie to win the 2013 *Best Picture Award*. (Source: Computed by the authors, reprinted with permission)

As the standard deviation of ratings is small for all movies, and *Zero* and *Amour* act as outliers, the standard deviation of ratings tends to increase with the number of clusters. It is noteworthy that the five serious contenders for the best picture award, including *Argo, Les Misérables, Lincoln, Silver Lining Playbook*, and *Zero Dark Thirty* tend to yield a moderate number of clusters.

Table 12: Number of extracted themes and statistics for the nine movies nominated for a *Best Picture Award*.

(Source: Computed by the authors, reprinted with permission)

	Mean rating	Number of themes (of clusters)	Mean *Intrade* close	Last *Intrade* close	St dev of ratings	Profit
Amour	7.1	23	0.99	0.4	2.84	$ (2.16)
Beasts	6.8	12	1.07	0.4	2.96	$ 10.98
Django	7.4	7	1.22	0.5	2.81	$ 62.80
Zero	6.3	4	5.53	0.7	3.03	$ 55.72
Les Mis	7.3	6	11.27	1.2	2.71	$ 87.78
Life of Pi	7.8	9	3.43	1.5	2.37	$ 4.98
Silver	7.5	6	6.05	3	2.61	$ 111.09
Lincoln	7.2	13	36.24	10.3	2.76	$ 117.20
Argo	7.4	6	32.27	82	2.60	$ 91.52

5. Conclusions

Early prediction of movie performance is critical for investors, film producers, cinema owners and other stakeholders in film exploitation. Considering a large amount of required investment to produce a movie, the risk of investing in a failed product is significant. For this reason, research has been conducted to develop models for prediction of movies' success to assist the stakeholders in their decision-making.

This article explained models to predict box office revenue for movies by the set of variables identified, and then discussed how these could correlate with Oscars awards. It could also be shown that data analysis is key to predicting movie success and hence may reduce the risk of investors as well as cinema owners to select the right movies in advance.

The findings suggest that more research into controversy indicators in movie reviews can provide an estimation of the chance of movies to win Oscars. The controversy is strongly associated with WOM theory (O'Leary & Sheehan, 2008). WOM could be positive or negative. The question is, do customers' negative opinions always fall on the bad side of the coin, or is there any advocacy for a brand coming from negative WOM or mixed WOM (so-called "controversy")? Some research indicates the possibility that controversy arising from consumers' opinions might have a positive impact. Liu (2006) suggested that box office revenue is correlated with the volume of WOM activity, but *not* correlated with the percentage of negative critical reviews. Zhang & Li (2010) argued that controversy attracts market attention and promotes box office sales. However, some studies show that controversy can undermine the chance of a movie to win an award. For example, Hayao Miyazaki's animated historical drama film "The Wind Rises" (2013) lost the 2014 Best Animated Feature Award, largely because of the level of controversy (Bruneel et al, 2018).

Findings of this research also support Hennig-Thurau et al's (2007) proposition that movie reviews impact income, although opposing the findings of Ashenafi et al (2013) which shows that there is no relationship with critics' reviews. Due to such findings, data-enhanced movie development can improve the chance of increased box office sales (Bhattacharjee et al, 2017) and winning awards (Yfantido et al, 2017). The movie studios must invest in creating controversial content and reflect the complexity of a theme in a way that increases the attention and debate of the movie. This also encourages WOM, engages prospective consumers, and influences ticket purchase behavior. As the cost of WOM is low, this may also positively affect profitability. However, it should not be neglected that WOM may also negatively affect box office sales. Still, the movie primer determines the success of forthcoming weeks, thus investing in films with a predicted success in advance and during the first weeks of release is advisable. Begging on those strategies before release

can increase the chance of good box office sales, which potentially multiplies revenues in the subsequent release windows.

To conclude, this study analyzed several approaches to predicting box office revenue using data analytics methods using variables available before the release of a movie and further presented several correlates of *Oscar* awards. Data analysis, coupled with strong human judgment, is likely to be the key contributor to reducing investor risk and enhancing revenue planning. That the movie industry is complex and that it operates under high risk and uncertainty are standard inferences for anyone who has been even a casual observer of, or participant in, the process of financing, making, and marketing films. As Vogel (2014) wryly noticed, "seemingly sure-bet, big-budget films with "bankable" stars flop, low-budget titles with no stars sometimes inexplicably catapult to fame, and some releases perform at the box office inversely to what the most experienced professional critics prognosticate" (pp. 144–5). Yet, amid those paradoxes, prediction technologies may help refund the start-up costs, get the film production processes going, create and safeguard jobs, and enhance value in any stage of the industry chain. In that respect, investment in the entertainment industry, for all the passion it may entail, shares many common features with those in other areas of business activity.

References

Ashenafi, Y., Chea, S., Chen, K., Hanson, J., & Jones, B. (2016). Analysis of Factors Affecting the Success of a Movie: Critic Reviews, Budget and Domestic Box Office Performance. http://rstudio-pubs-static.s3.amazonaws.com/233939_bbeb292c0c20440f97d31b616662c06f.html

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. *International Journal of Business Information Systems*, *24*(3), 344-368.

Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015, January). Role of different factors in predicting movie success. In *2015 International Conference on Pervasive Computing (ICPC)* Pune, pp. 1-4, doi: 10.1109/PERVASIVE.2015.7087152

Bone, P. F. (1995). Word-of-mouth effects on short-term and long-term product judgments. *Journal of business research*, *32*(3), 213-223.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton: CRC press.

Brown. C. (2015a). *Key considerations in film finance*. www.slated.com.

Brown, C. (2015b). *Filmed entertainment as an attractive asset class*. www.slated.com.

Bruneel, C., Guy, J.-L., Haughton, D., Lemercier, N., McLaughlin, M.-D., Metzner, K., Vialle, Q., & Zhang, C. (2018). Movie Analytics and the Future of Film Finance. Are *Oscars* and Box Office Revenue Predictable?. In: Murschetz P. C., Teichmann R., & Karmasin M. (eds.), *Handbook of State Aid for Film. Media Business and Innovation* (pp. 581-578). Cham: Springer. DOI: 10.1007/978-3-319-71716-6_30

Chakraborty, P., Rahman, M. Z., & Rahman, S. (2006). Movie Success Prediction using Historical and Current Data Mining. *International Journal of Computer Applications, 178*(47), 1-5.

Debande O. (2018). Film Finance: The Role of Private Investors in the European Film Market. In: Murschetz P., Teichmann R., Karmasin M. (eds.), *Handbook of State Aid for Film. Media Business and Innovation* (pp. 51-66). Springer, Cham. DOI: 10.1007/978-3-319-71716-6_4

Delen, D., & Sharda, R. (2012, October). Forecasting financial success of hollywood movies a comparative analysis of machine learning methods. In *9th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2012* (pp. 653-656).

Divakaran, P. K. P., Palmer, A., Alsted Søndergaard, H., & Matkovskyy, R. (2017). Pre-launch prediction of market performance for short lifecycle products using online community data. *Journal of Interactive Marketing, 38*, 12–28.

El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., & Keim, D. A. (2013). *Visual analytics for the prediction of movie rating and box office performance*. IEEE VAST Challenge USB Proceedings.

Erikson, R. S., & Wlezien, C. (2008). Are political markets really superior to polls as election predictors? *Public Opinion Quarterly*, *72*(2), 190-215.

Film Industry Statistics (2018). *Film Industry - Statistics & Facts*. https://www.statista.com/topics/964/film/

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. icml, *96*, 148-156.

Gaustad, T. (2019). How streaming services make cinema more important: Lessons from Norway. *Nordic Journal of Media Studies*, *1*(1), 67-84.

Gold, M., McClarren, R., & Gaughan, C. (2013). The lessons Oscar taught us. Data science and media & entertainment. *Big Data*, *1*(2), 105-109.

Gomery, D. (2004). The Economics of Hollywood: Money and Media. In: Alison, A. et al (eds.), *Media economics: theory and practice*, 3rd ed. (pp. 193-207). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Haughton D., McLaughlin M.-D., Mentzer, K., & Zhang C. (2015). Can We Predict Oscars from Twitter and Movie Review Data?. In: Bruneel C., Guy, J.-L., Haughton, D., Lemercier. N., McLaughlin, M.-D., Mentzer, K., Vialle, Q., & Zhang, C. (eds.), *Movie Analytics*. A Hollywood Introduction to Big Data. Cham: Springer. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-319-09426-7_6

Hennig-Thurau, T., & Houston, M. B. (2019). *Entertainment Science. Data Analytics and Practical Theory for Movies, Games, Books, and Music*. Cham: Springer International.

Hennig-Thurau, T., Houston, M. B., & Walsh, G. (2007). Determinants of motion picture box office and profitability: an interrelationship approach. *Review of Managerial Science*, *1*(1), 65-92.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning. A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651-674.

Kolli, S., & Khajeheian, D. (2020). How Actors of Social Networks Affect Differently on the Others? Addressing the Critique of Equal Importance on Actor-Network Theory by Use of Social Network Analysis. In: Williams I. (ed.), *Contemporary Applications of Actor Network Theory* (pp. 211-230). Palgrave Macmillan, Singapore. DOI: 10.1007/978-981-15-7066-7_12

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, *33*(3), 874-903.

Leamer, E. E. (2008). *What's a recession, anyway?* (No. w14221). National Bureau of Economic Research. http://www.nber.org/papers/w14221.

Leonhardt, D. (2015). Oscars 2015. An Excellent Night for Prediction Markets. *The New York Times*, February 23.

Litman, B. R. (1983). Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, *16*(4), 159-175.

Litman, B. R., & Kohl, L. S. (1989). Predicting financial success of motion pictures: The 80s experience. *Journal of Media Economics*, *2*(2), 35-50.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Massey, A. (2020). Could AI and Data Analytics Deliver Blockbuster Movies? https://www.esri.com/about/newsroom/publications/wherenext/movie-greenlighting-with-ai

McKenzie, J. (2012). The economics of movies. A literature survey. *Journal of Economic Surveys*, *26*(1), 42-70.

McKenzie, J. (2013). Predicting box office with and without markets: Do internet users know anything? *Information Economics and Policy*, *25*(2), 70-80.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, *8*(8), e71226. https://doi.org/10.1371/journal.pone.0071226.

MPA. (2019). Theme report 2019. Retrieve from: https://www.motionpictures.org/research-docs/2019-theme-report

O'Leary, S., & Sheehan, K. (2008). *Building buzz to beat the big boys: word-of-mouth marketing for small businesses*. Westport, Conn.: Praeger Publishers.

Ravid, S. A. (2018), The Economics of Film Financing: An Introduction. In: Murschetz P. C., Teichmann R., & Karmasin M. (eds.), *Handbook of State Aid for Film. Media Business and Innovation* (pp. 39-49). Springer, Cham. DOI: 10.1007/978-3-319-71716-6_3

Rothschild, D., & Wolfers, J. (2008). *Market manipulation muddies election outlook*. http://online.wsj.com/article/SB122283114935193363.html.

Saxon, I. (2010). *Intrade Prediction Market Accuracy and Efficiency. An Analysis of the 2004 and 2008 Democratic Presidential Nomination Contests*. University of Nottingham. Dissertation.

Smit, E., & Pangarker, N. A. (2013). The determinants of box office performance in the film industry revisited. *South African Journal of Business Management*, *44*(3), 47-58.

Ulin, C. J. (2010). *The business of media distribution: Monetizing film, TV and video content in an online world*. Burlington, MA: Focal Press.

Valenti, J. (1978). *Motion pictures and their impact on society in the year 2001*. Midwest Research Institute.

Vogel, H. L. (2014). *Entertainment industry economics*, 9th ed., New York: Cambridge University Press.

Yfantidou, I., Riskos, K., & Tsourvakas, G. (2017). Advertising message strategy analysis for award-winning digital ads. *International Journal of Technology Marketing*, *12*(4), 340-355.

Zhang, Z., & Li, X. (2010). *Controversy in Marketing. Mining Sentiments in Social Media*. Proceedings of the 43rd Hawaii International Conference on Systems Sciences.

Biography:

Christophe Alain Bruneel is a first year PhD student in economics at the Toulouse School of Economics. His main research topics are econometrics and theoretical search models applied to the real estate market.

Jean-Louis Guy is an affiliated faculty member of the Toulouse School of Economics. He is the director of the magisterium program and oversees numerous case studies.

Dominique Haughton is a professor of mathematical sciences and global studies at Bentley University and an affiliated researcher at Paris 1 and Toulouse 1 Universities. She is a fellow of the American Statistical Association with research interests in global analytics, music analytics, business analytics, data mining, and applied statistics.

Nicolas Lemercier is a graduate of the Toulouse School of Economics in the magisterium in statistics and economics; he is a research analyst in a statistics department in a banking sector.

Mark-David McLaughlin is a PhD student at Bentley University with research interests in social policy and qualitative and quantitative social research. He is also a security incident manager at Cisco Systems.

Kevin Mentzer is an assistant professor in the Department of Information Systems and Analytics at Bryant University. He has 20 years of professional experience with more than half that time serving as a consultant for start-up organizations, assisting them with sourcing strategies and IT development and deployment. His research interests are social networks applied to policy issues.

Quentin Vialle is a graduate of the Toulouse School of Economics in the magisterium in statistics and economics; he works as a data scientist in Paris area.

Changan Zhang is a business analytics consultant and data scientist. He holds a PhD in business with specialization in business analytics at Bentley University. He is the senior business intelligence manager, advanced data engineer, and scientist at Ctrip.com. His specialties are predictive modeling, data mining, Bayesian analysis, social network analysis, and big data analysis.

Paul Clemens Murschetz, PhD, is researcher in media economics and media management associated to University of Digital Sciences (UDS) Berlin. He lectures media management and issues of the digital business at various higher education institutions in Austria and Germany. He has published widely in journals such as the International Journal on Media Management, the European Journal of Communication, and ICA's Communication Yearbook and is a frequent conference speaker (German Communication Association, World Media Economics Conference, European Media Management Association).

Barira Bakhtawar, is a PhD scholar at Lahore College for Women University, Lahore and a researcher in media economics and media management. She lectures media studies and media management at various universities in Lahore Pakistan. She is currently working on her doctoral dissertation and various research projects in the field of Media Management and Economics. She has been working as a radio broadcaster for the past ten years in a leading commercial FM station of Pakistan.

Other Information:

Conflicts of Interest: The authors declare no conflict of interest.