



Movie Analytics and the Future of Film Finance. Are *Oscars* and Box Office Revenue Predictable?

Christophe Bruneel, Jean-Louis Guy, Dominique Haughton, Nicolas Lemerrier, Mark-David McLaughlin, Kevin Mentzer, Quentin Vialle, and Changan Zhang

C. Bruneel · J.-L. Guy
Toulouse School of Economics, Toulouse, France

Université Toulouse 1, Toulouse, France
e-mail: christophe.bruneel@gmail.com; Jean-louis.Guy@ut-capitole.fr

D. Haughton (✉)
Bentley University, Waltham, MA, USA
SAMM, Université Paris 1, Paris, France
TSE-R, Université Toulouse 1, Toulouse, France
e-mail: dhaughton@bentley.edu

N. Lemerrier
Université Toulouse 1, Toulouse, France
e-mail: nicolas.lemerrier1@hotmail.fr

M.-D. McLaughlin
Bentley University, Waltham, MA, USA
Cisco Systems, San Jose, CA, USA
e-mail: mmclaughlin@bentley.edu

K. Mentzer
Bryant University, Smithfield, RI, USA
e-mail: kmentzer@bryant.edu

Q. Vialle
Inbox, Malakoff, France
Université Toulouse, Toulouse, France
e-mail: vialle.quentinalexandre@gmail.com

C. Zhang
CTrip, Shanghai, China
e-mail: hellozca@gmail.com

1 Introduction: Movie Analytics and Film Finance

This chapter examines to which extent modern analytics techniques help us understand the success of movies. We will describe essential analytics techniques as needed here and discuss them in the context of the prediction of box office revenue and *Oscar* attribution. The work in this chapter lies in the context of the broader issue of film financing. A series of papers published by [slated.com](#) (Brown, 2015a, 2015b) makes the case that the industry has now become an attractive investment domain, and the success of the industry in terms of both raw revenue and revenue growth implies that investors must have at least some sense of how to control the risk of investments in that industry. It, however, remains true that predicting the success of a film, even with modern data mining techniques at hand, is a difficult task, as will be detailed further in the chapter (see for example El Assady et al. 2013).

This chapter addresses in detail the problem of predicting box office revenue on the basis of data available *before* the movie is released (Sect. 2). The methodology for data collection and analysis based on state-of-the-art data mining models is described. In all, this discussion draws a sobering lesson: state-of-the-art methods can identify those variables which are important for predicting box office revenue, such as the budget, whether the movie is part of a series, and the “star power” of the distributor, actors, and producer, but it remains difficult to actually predict box office revenue with decent accuracy because of the presence of very strong outliers in the dataset. The chapter then turns to a discussion of the role of “prediction markets,” that is, exchange-traded markets created for the purpose of trading the outcome of specific events, in predicting *Oscar* wins (Sect. 3). The performance of the *Intrade* market (now taken off-line for reasons explained below) in predicting the 2013 *Best Picture Award* (attributed to the movie *Argo*) is described, on the basis of data extracted from *Intrade* before it was taken off-line. Finally, the chapter discusses the role of “controversy,” as identified by a text mining of movie reviews, in *Oscar* attribution (Sect. 4). The analysis suggests that too many themes underlying the reviews may be too complex for a voting audience to rally on. On the other hand, too few may be too simple. The movie *Argo*, with six underlying themes, may very well have reached a happy medium. Hope may very well lie, both in terms of predicting box office revenue and awards such as *Oscars*, in preproduction analyses of scripts. However, such analyses, as performed, for example, by [slated.com](#), involve human judgment. Parts of this chapter are based on Haughton et al. (2015).

2 Box Office Revenue Prediction

Strong financial stakes are linked to a good box office prediction for a film, both for producers of this film and for cinema managers. The literature in this area is large, but datasets used are often unintentionally biased by their authors who exclude films with a low budget or low success (observed *ex post*) because information about such films is less reliable.

Managers of cinemas are investors, and as such, minimizing the risk of commercial failure of the films they select is a major issue. To achieve this, several methods are at their disposal. They may, for example, broadcast a film with an actor or a director who is recognized and appreciated by the public, a film adapted from a popular book, or a sequel to a successful film; or they might count on an effective advertising campaign. However, these simple strategies do not guarantee the commercial success of a film. To address this problem, predictive models have been developed for several years to identify factors that help make a movie a success.

Extensive literature exists on the subject with early attempts dating back to the 1990s and new developments each year. In general, models display relatively good predictive rates (with R-squares of more than 70% and relatively low error rates).

Most of these studies, however, have a selection bias inherent in the availability and quality of the information used to construct the models: they base their prediction only on a subsample of films with either a high production budget or, worse as far as bias is concerned, a relatively high box office revenue. The problem is that such information is not available “ex ante” for a cinema wishing to decide whether or not to accept a given film. Typical samples in the literature eliminate a considerable number of “outliers” (up to 46% of the films with total box office revenues under US\$1 million). We also note that this 46% is probably an underestimate of the proportion of film with revenues below US\$1 million since information related to certain failures was probably not included on the site www.boxofficemojo.com used in this chapter. This explains in large part the good “predictions” in the aforementioned studies (see, e.g., McKenzie (2012) for a review article). These models are however not practical for a film director wishing to identify a potentially successful movie. Determinants of success among films with a relatively high box office revenue are not necessarily the same as those for films “ignored” because of their low visibility.

In this section, we explore to what extent it is possible to predict the financial success of a film before its release, with the goal of potentially helping a cinema manager decide which films to accept to maximize profit. We consider all categories of films such as those that can be found on the US site www.boxofficemojo.com, a website that tracks box office revenue in a systematic, algorithmic way. We will focus on box office revenue for the first week after a film’s release, since most profits are realized during this first week. We compare the results of estimates using different statistical methods, more or less recent and more or less complex. Our main finding is that a precise prediction of box office receipts on the basis of prerelease information only is much more difficult than when the selection of the sample is biased by the use of “ex post” information. We will therefore try to predict box office revenue from the available “ex ante” information for the manager whatever the film. More precisely, we will only use information available before the shooting of the movie begins: a good prediction with this information only would allow the manager to position himself/herself earlier on promising films. We will test several forecasting models.

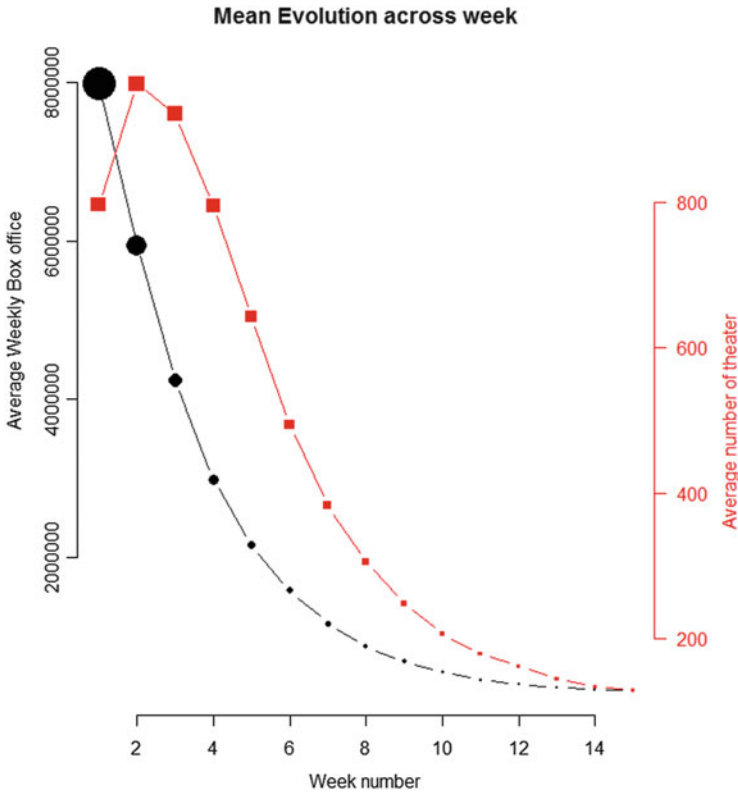


Fig. 1 Average weekly box office performance and average number of movie theaters showing a movie. Source: Computed by the authors

2.1 Methodology

Our forecasts will focus on the box office revenue obtained during the first week after the release of the movie. This is highly relevant to cinema managers since, as illustrated in Fig. 1, the bulk of the revenue from a film is obtained during the first week (40% on average), reflecting the delay in other cinemas deciding to show the movie.

Concerning our predictions, we want to put ourselves in a “real situation,” faced by a realistic cinema manager, with the information available to him/her. We will therefore predict the revenues of the films released after the end of 2011 (2166 films), from models built on data available “ex ante”: on the basis of films released before January 2012. More specifically, we consider all those films whose first week of broadcasting was over by January 1, 2012. Different time series learning samples will be tested to construct our models:

- The 5874 films released between 2000 and 2012
- The 3392 films released between 2006 and 2012

- The 2193 films released between 2008 and 2012
- The 1090 films released between 2010 and 2012
- The 566 films released in 2012

This temporal distinction offers us the possibility of testing what is most appropriate for predicting box office revenues and of guiding the choice of the most suitable reference sample. This helps control for an unobserved economic conjuncture. We will refer to this process as the choice of an “optimal temporal horizon.” We will construct the following predictive models:

- Linear regression
- Decision tree
- Random forest
- Conditional forest
- Gradient boosting

We will also use the so-called stacking method to optimize our results. These different models will finally be applied to the *ex ante* data on the films released after January 2012, estimated as if we were at the end of 2012. Therefore, it is clear that, in contrast to previous studies, we will not exclude any film. We, finally, compare our predictions with the actual revenues realized by films released in 2013, 2014, and 2015.

In summary, our objective is to identify the optimal time span a manager must observe as a test period to construct his/her estimation model and the modeling technique with the highest predictive power.

2.2 Data

2.2.1 Data Collection: Web-Scraping

The database used in this chapter contains more or less detailed information (total, weekly or even daily) on the box office revenue of 15,459 films in the United States. It also contains “classical” details about the genre of the film, its production budget, distributor, cast members, etc. This information was gathered via web-scraping on the American reference site: www.boxofficemojo.com.

The data collection method (fully implemented via R) proceeded as follows:

- Collection of all the links to the pages referencing the films
- Creation of a “scraping function” to apply to these links

This function must be general enough to be able to retrieve the information available on any type of page, which implies extensive investigative work on the site (discovery of patterns in the coding of pages) and management/anticipation of potential errors. Among other things, the function must be able to determine the granularity of the available information (daily, weekly, etc.). To speed up data

collection, and to avoid overloading the server of the query site, we aim for the function to visit the smallest number of links possible. This function returns the results of each page as a frame, with one row for each “period” of box office revenue (e.g., 1 week). Finally, this function is applied to all links gathered in the first stage. Once data collection was completed, a dataset of 107,760 weekly box office revenues was obtained for 11,544 different films. We recall that we focus on first-week revenues in this study.

2.3 Data Processing: Creation of Variables

We now possess a very rich database, but several data transformations were necessary in order to be able to make use of it.

2.3.1 Box Office Revenue Deflation

We choose to measure box office revenues in financial terms rather than in terms of audience (number of viewers) since we are focusing on the financial return of the films. To make the intertemporal comparison possible, we deflated box office revenue by the monthly CPI of the weeks in which they were launched. When that information was not available, we simply deflated the global box office revenue by the CPI of their launch year. Table 1 displays the ten highest revenue films, in both nominal and real terms.

2.3.2 Production Budget

The film production budget is only available for a little less than a quarter of the movies. To retain this 25% of available information, we will split the production budget into two “subvariables”:

- A first binary variable indicating whether or not the information is available.
- A 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 is equal to the production budget otherwise.

This “dichotomy” is the best way to maintain this variable which is positively correlated with the box office, hereby avoiding a non-negligible bias.

2.3.3 Experience, Quality, and Star Power

The decision to “consume” a film is special because the good is consumed once (with a few exceptions, since a few individuals will see the same film multiple times in a cinema). The individual therefore bases his/her choice of viewing on an implicit estimate of the quality of the film. Watching trailers, reading opinions by experts, or observing the enthusiasm of other better-informed individuals (measured on social networks, e.g., see Mestyán, Yasserli, & Kertész, 2013) can influence this estimate. The observation of these different variables makes it possible to estimate in a very fine way the box office revenue and even its dynamic

Table 1 Ten films with the highest box office revenue: nominal and actual (basis, January 2010)

Rang	Top films (nominal)	Year	BO nominal (in \$ millions)	Top films (real)	Year	BO real (in \$ millions)
1	Avatar	2009	750	Star Wars 4	1977	1099
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

Source: Computed by the authors

evolution over time. Without any variables related to the viewing of the film (“after-launch” variables), information on casting (actors, producers, directors) are our only means of obtaining an indication of quality, for three main reasons:

1. The reputation of the actors, their “star power,” attracts consumers because it sends, a priori, a positive signal on the quality of the film: an actor who has acted “well” in good films in the past will have a good reputation and will attract more people than less well-known actors, all things being equal.
2. The experience of the director is another indicator of ex ante quality for spectators: the more experienced a director is, the more he/she is likely to attract viewers.
3. The “quality” (judged by specialists) of past performances of actors suggests that quality will be present in the new film.

For example, an Oscar can be perceived as a guarantee of quality for an actor. It is therefore necessary to create a variable summarizing the star power and the experience of the actors, etc. involved in a film. We use the following approach: for a given film, the star power of each actor in the film is estimated by the sum of the box office revenues of the films in which he/she played a major role in the past. The total star power of the actors of the film is then obtained by summing these individual star powers. The star power of the directors, the producers, and the distributors are computed in the same way.

For a given film, the experience of each actor is simply estimated by the number of films in which he/she played a major role in the past. We obtain the total experience of the actors of a film by summing up the individual experiences. The same applies to directors, producers, and studio distributors. For a given film, the recognized reward-based quality of an actor or a director is estimated by the number of nominations he/she had for an *Oscar* in the past. For a producer or distributor, this quality is captured by the number of films in which he/she has intervened and that has been nominated for a *Best Film Oscar*. The total reward-based quality of a given new film is simply estimated by the sum of the individual reward-based qualities of the actors, distributors, producers, and directors.¹

2.4 Descriptive Statistics

2.4.1 Box Office Revenue

As illustrated in Fig. 2, the distribution of the logarithm of box office revenue in the first week follows a bimodal distribution, with a main mode representing movies with a relatively low (<US\$100,000) box office revenue in the first week. The second mode represents movies with an average to high box office revenue. It can be seen here that the box office revenue is clearly not normally distributed, and it is in this bimodality that the problem lies. The usual literature which implicitly excludes low-performing films, actually the majority of films, concentrates only on the “second mode” of the distribution.

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. In this chapter we will work directly with this bimodal dependent variable.

2.4.2 Production Budget

Table 2 gives summary statistics for the availability and the deflated level of budgets.

¹Note that all these variables also give us an estimate of the production budget, even when it is not available. They therefore have a useful role to play in our predictions and have the advantage of being available very early in the shooting of the movie and can very easily be obtained since the movies themselves use the information to advertise.

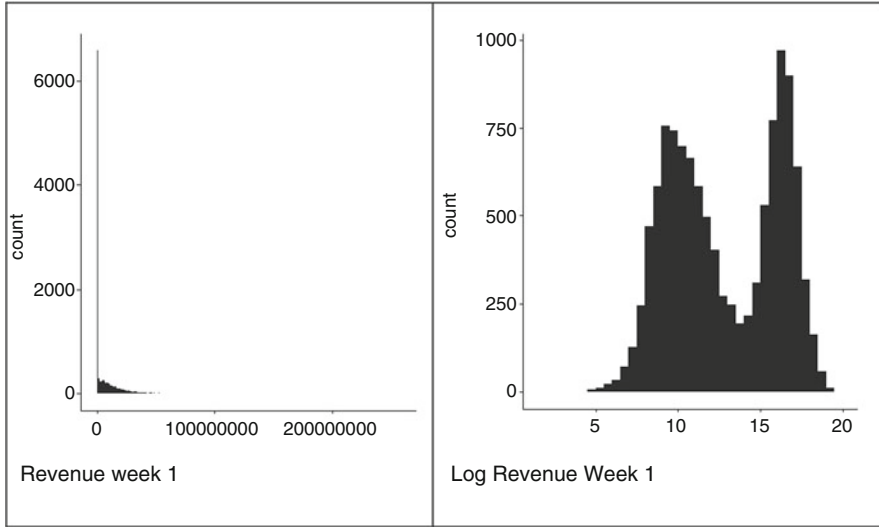


Fig. 2 Distribution of operating revenue in the first week. Source: Computed by the authors

Table 2 Description: production budget

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

Source: Computed by the authors

2.4.3 Star Power: Actors, Directors, Producers, and Distributors (Table 3)

2.4.4 Other Explanatory Variables and Controls

The other variables that the manager has ex ante to make his/her predictions are genre and MPAA rating. We also know if the film is a remake, if it is a book adaptation, if it belongs to a series of films, if it is a prequel, if it is broadcast in a foreign language (non-English), and if it received a golden palm at *Cannes* before arriving on US territory. We also add a dummy indicating whether the film has had a limited release before its wide release and a variable with the duration of this limited release. One may also consider the day of the week when the movie was launched. Finally, seasonal effects are captured by the month of release of the film.

All of these variables are described in Tables 4, 5, 6, and 7.

Table 3 Description: star power

Statistic	Median	Mean	Standard deviation
<i>Actors</i>			
Star power of actors in the film	4,274,693	972,238,048	1,888,323,059
Sum of the 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
<i>Director</i>			
Star power of the director	0	79,763,117	303,180,322
Number of previous films directed by the director	0	1.37	3.50
Number of nominations to Oscars for the director	0	0.07	0.42
<i>Distributor</i>			
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
<i>Producers</i>			
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

Source: Computed by the authors

2.5 Models

We now describe the models we will use to predict box office revenue in terms of ex ante available predictors.

2.5.1 Classification and Regression Tree

A classification and regression tree (CART) decision tree is an improved nonlinear and entirely nonparametric statistical learning technique introduced by Breiman, Friedman, Stone, and Olshen (1984) which allows to classify or predict a dependent variable from independent variables. A decision tree is built intuitively and is easily interpretable, thanks to its graphic appearance. The construction of the tree proceeds as follows: at each node, the algorithm splits the dataset into two subsets, using any possible predictor and any cutoff point for continuous predictors, in such a way that the two subsets are as homogeneous as possible with respect to the dependent variable. This technique has the advantage of being nonparametric, thus

Table 4 Description: genres of films (a movie may have several genres)

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

Source: Computed by the authors

Table 5 MPAA rating: proportions

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

Source: Computed by the authors

Table 6 Description of other controls

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

Source: Computed by the authors

Table 7 Control variables: seasonality and day of release

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
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

Source: Computed by the authors

not postulating any a priori assumption on the distribution of the data, being robust to outliers, and supporting all types of variables. In addition, the CART algorithm handles missing values in an effective manner. When the learning sample is large as is the case here for most reference periods, the CART algorithm has properties which 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 periods 2010–2012 and 2011–2012).

2.6 Random Forests

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 majority vote on the resulting CART trees (hence the name of *random forests*).

2.7 Conditional Forests

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

2.8 Gradient Boosting

The gradient boosting algorithm introduced by Freund and Schapire (1996) is a prediction method that minimizes several types of loss function with respect to a prediction function. This method can adapt to any type of data even when the number of variables exceeds the number of observations and gives very good results.

2.9 Results

We now present results obtained by each of these methods for predicting box office revenue (Table 8).

In general, whatever the estimation method used, the R-square tends to increase with the number of observations serving as reference. The more films we have to construct our model, the better we can explain 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). Thus, it would appear that our optimal time range of films to be considered for estimation is 6 years (with a preference for the 2006–2012 range), if we want 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.

To obtain higher coefficients of determination, we would need other variables that measure expectations of the quality of the film itself, which we do not have (and which are not so easily obtainable by cinema managers). As far as the *root mean square error* (RMSE) is concerned, it can be seen that it is very large, at US\$10 million (January 2010 basis), and the average error rate is also very high (from 596.99% maximum to 22.30% minimum). This is due to the presence of

Table 8 Estimation results for box office revenue in the first week

Reference period	Method	R2 (adjusted)	Root mean square error	Average relative error	Root median square error	Median relative error
2000–2012	OLS	0.74	10,901,374.00	253.91	1,494,037.00	26.99
2000–2012	Decision tree	0.72	11,265,019.00	143.44	743,540.80	21.77
2000–2012	Random forest	0.78	9,986,626.00	36.91	188,348.30	1.56
2000–2012	Conditional forest	0.77	10,218,146.00	43.90	218,966.50	6.86
2000–2012	Gradient boosting	0.76	10,414,030.00	35.92	153,258.20	1.00
2006–2012	OLS	0.74	10,909,403.00	264.01	1,276,029.00	26.29
2006–2012	Decision tree	0.71	11,517,285.00	189.99	943,572.10	28.13
2006–2012	Random forest	0.79	9,867,366.00	27.79	180,777.40	1.41
2006–2012	Conditional forest	0.76	10,507,541.00	47.14	216,554.30	7.97
2006–2012	Gradient boosting	0.76	10,531,481.00	38.15	175,234.30	1.10
2008–2012	OLS	0.73	10,958,782.00	309.64	1,349,485.00	29.10
2008–2012	Decision tree	0.66	12,396,390.00	244.44	1,119,748.00	33.23
2008–2012	Random forest	0.76	10,378,329.00	37.97	168,820.90	1.43
2008–2012	Conditional forest	0.74	10,813,495.00	41.55	190,900.50	6.88
2008–2012	Gradient boosting	0.76	10,486,929.00	42.28	156,116.70	1.23
2010–2012	OLS	0.72	11,344,506.00	382.79	1,749,335.00	28.29
2010–2012	Decision tree	0.31	17,701,598.00	80.72	375,038.40	11.47
2010–2012	Random forest	0.74	10,769,061.00	22.30	113,640.10	1.31
2010–2012	Conditional forest	0.71	11,478,915.00	27.28	126,784.80	4.33

2010–2012	Gradient boosting	0.74	10,835,496.00	34.14	161,450.40	1.73
2011–2012	OLS	0.63	13,014,399.00	596.99	2,007,622.00	40.80
2011–2012	Decision tree	0.58	13,845,035.00	66.04	336,170.90	9.77
2011–2012	Random forest	0.70	11,709,300.00	66.24	156,481.00	1.48
2011–2012	Conditional forest	0.63	12,921,583.00	32.81	156,024.20	5.24
2011–2012	Gradient boosting	0.69	11,891,466.00	31.08	153,910.30	2.01

Source: Computed by the authors

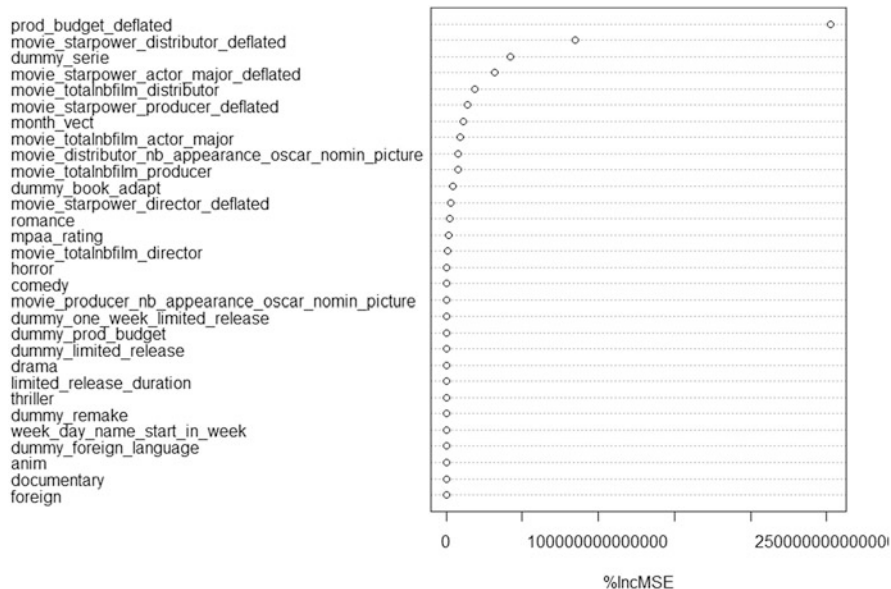


Fig. 3 Importance of variables. Source: Computed by the authors

extreme values and is the reason why we are looking at the root median square error and the median error rate to remedy this issue.

These indicators of predictive power are much more favorable than the previous ones since the *root median square error* fluctuates between 110,000 and US\$2,000,000 regardless of the technique used. These indicators clearly identify random forests and gradient boosting as the best models.

Gradient boosting seems to dominate when it has the longest range of data available (2000–2012 or 2006–2012) with a median error rate ranging between 100 and 110%. These rates, however, remain very large. This implies that even the most advanced predictive techniques, with *ex ante* information only, fail to correctly discriminate between films that will achieve a low box office revenue and those that will achieve such a medium or large revenue. Stronger predictive models would need data from social networks, for example, quite a bit more difficult to extract than the data utilized in this chapter. We can however identify which variables are more important than others in predicting box office revenue, as shown in Fig. 3.

From the graph of the importance of variables in our “best” prediction model (random, based on the period 2006–2012), we find that the production budget is the most important variable. We also find a few binary variables that are of importance, notably belonging to a series of films or being adapted from a book. Seasonality is also an important variable.

3 Predicting *Oscars* from “Prediction Markets”

The *Intrade* market (Intrade.com) was an online predictive betting exchange operated by *Intrade* The Prediction Market Limited. It allowed members to purchase or sell contracts on whether a future event will occur. Popular topics included upcoming elections, movie and music awards, and financial predictions of stock market indexes. *Intrade* did not participate in the buying or selling of contracts directly but instead had a flat monthly fee structure for members regardless of the participation level of that member. Trading was done on a per-unit basis with each unit paying US\$10.00 if the event occurs and US\$0 if the event does not occur. The contracts traded on a 100-point scale with 100 points representing the full US\$10.00 value. For example, a contract might have stated “Mitt Romney will win the U.S. presidential election in 2012,” and the contract might have traded at 25 points. Therefore, a member would purchase this contract for the value of US\$2.50, and if Mitt Romney was elected, then the member would receive US\$10.00. If Mitt Romney was not elected, then the member would lose the US\$2.50 to the person who sold the contract.

Intrade received significant media exposure during the 2012 presidential elections with the accurate prediction of nearly all US state electoral contests, but the exposure was overshadowed later in the year with the filing of a civil suit on unregulated trading by the US *Commodity Futures Trading Commission*. On December 23, 2012, *Intrade* ceased allowing US members from participating, resulting in a significant drop in overall participation, and on March 10, 2013, *Intrade* ceased all trading. The prediction market *Paddy Power* (www.paddypower.com) typically hosts best picture bets in *Academy Award* competitions, as does *Bet Victor* (www.betvictor.com). But unlike for *Intrade*, there is no convenient way to get historical pricing information out of *Paddy Power* or *Bet Victor*, and rapid changes in pricing may be difficult to track without employing some form of screen scraping. Researchers in the United States may also be shut out from even looking at international betting sites (such as www.WilliamHill.com).

Scholars have used *Intrade* data to investigate issues such as participation in the Euro currency (Shambaugh, 2012), the probability of a US recession (Leamer, 2008), elections (Saxon, 2010; Rothschild & Wolfers, 2008; Erikson & Wlezien, 2008), and entertainment awards such as the *Grammy Awards* and *Oscars* (Gold, McClarren, & Gaughan, 2013). Prediction market estimates of the probability of a win are considered to be very accurate, at least for events such as *Oscar* wins. Prediction markets were reportedly successful again in predicting the 2015 *Academy Awards* (Leonhardt, 2015). Figure 4 displays the price per contract for each of the nominees winning the *Oscar* for the 2013 Best Picture Award. We can see that up until December, there was no clear front-runner. Then beginning in December, the film *Lincoln* emerged as the clear favorite. However, in late January, the film *Argo* began to gain on *Lincoln*, surpassing *Lincoln* and holding onto that position until the end.

In analyzing the average contract price for each movie, we see that the five serious contenders for the *Best Picture Award* were in alphabetical order *Argo*, *Les Misérables*, *Lincoln*, *Silver Linings Playbook*, and *Zero Dark Thirty*. What

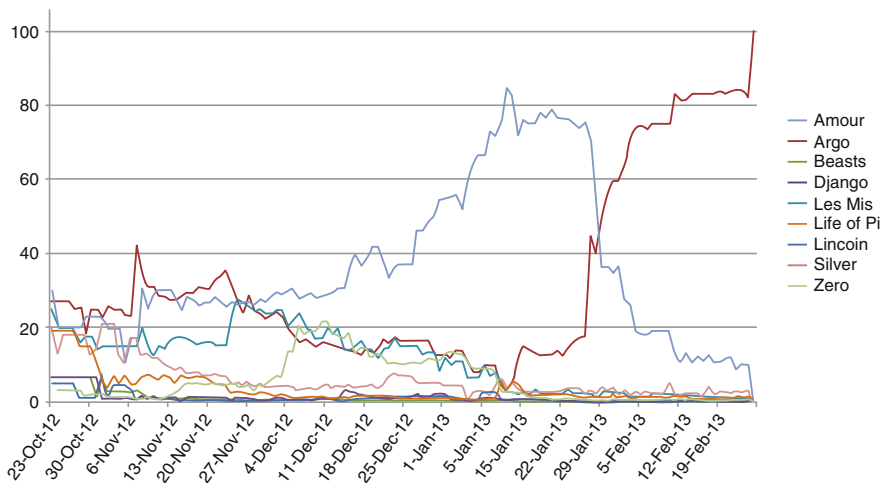


Fig. 4 Close Intrade contract prices for each nominated movie to win the 2013 *Best Picture Award*. Source: Extracted by the authors

happened over this time period that could have contributed to the perceptions of which film would win the *Award*? Specifically, what occurred around the time frame of late January that caused such a dramatic change? On January 26, 2013, a *Los Angeles Times* headline read “The Gold Standard; now for real insight into *Oscars* – by the guilds.” The article reported that the guilds’ awards, beginning with PGAs, had been fairly reliable predictors.

The “PGAs” denote the *Producer Guild of America Awards* (PGAs) which were announced that evening. The headline coming out of the PGAs that evening was that *Argo* won the top prize of the night, the *Zanuck Award for Outstanding Producer of Theatrical Motion Pictures*.

This awards ceremony was followed the next evening with the 19th *Annual Screen Actors Guild* (SAG) awards. Their top award is the *Outstanding Performance* by a Cast in a Motion Picture which was awarded to *Argo*. So can we simply use those two awards ceremonies to predict the *Oscar’s Best Picture Award*? Although we focused only on the 2012 movies, we can take a quick look at the winners over the past decade. Over those 10 years, the PGA and SAG have awarded the same picture five times, and four of those times, the *Oscars* have followed suit and awarded the same film. In the other 5 years where the PGA and SAG have awarded different films, the *Oscars* selected one of the two 4 of those 5 years. Only in 2004 did the PGA, the SAG, and the *Oscars* each give the top award to a different film (Table 9).

Table 9 PGA awards, SAG awards, and *Oscars*

	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 Predicting Oscars from Movie Review Data

In this section, we focus attention on whether text reviews of movies which are nominated for a *Best Picture Award* carry any sign of the likelihood of a movie winning the Award. We suggest that a measure of how controversial the movie is perceived to be, the value of which could be extracted by a text analysis of the reviews, is a potential predictor of a win, aside from other predictors identified in the past work.

4.1 IMDb Review Data


4.1.1 IMDb Review

In terms of text mining the opinions of movie watchers, IMDb user reviews have several advantages compared to tweets. First, most user reviews on IMDb are much longer than tweets (which are constrained to a maximum of 140 characters). Therefore, a review can contain richer and more complex thoughts than a tweet. Second, some review writers on IMDb are prolific authors, while the quality of tweets is not guaranteed at all. One can filter out reviews by non-prolific authors by choosing the “Prolific Author” filter on the IMDb review page (Fig. 5). Third, IMDb review readers can vote up or down to a review, as in “2 out of 12 people found the following review useful” in the middle part of Fig. 5. We can use it to measure the quality of a review. However, IMDb does not provide any API or structured database for downloading movie reviews. Therefore, we need to crawl the raw HTML webpage to extract review data.

4.1.2 XPath and R XML Library

In this chapter, *XPath* is used to mechanically navigate through elements and attributes in an XML document, such as all IMDb reviews on one webpage. It is easy to read and easy to reuse and is supported by most programming languages and software packages such as *Python* or *R*. *XPath* expressions such as in Table 10 are

IMDb > [Argo \(2012\)](#) > Reviews & Ratings - IMDb



Watch It

at Amazon

or

Buy it at Amazon

More at IMDb Pro

Discuss in Boards

Add to Watchlist

Update Data

Reviews & Ratings for **Argo** [More at IMDbPro >>](#)

Filter: Prolific Authors Hide Spoilers:

Interleaved...


Reviews from users who have written at least 100 reviews, most prolific authors first.

Reviews from users who have written at least 100 reviews, most prolific authors first.

Page 1 of 23: [\[1\]](#) [\[2\]](#) [\[3\]](#) [\[4\]](#) [\[5\]](#) [\[6\]](#) [\[7\]](#) [\[8\]](#) [\[9\]](#) [\[10\]](#) [\[11\]](#) ▶

[Index](#) 222 matching reviews (656 reviews in total)

2 out of 12 people found the following review useful:



Oddly, I couldn't find anything to dislike about this film!

★ ★ ★ ★ ★

Author: [planktonrules](#) from Bradenton, Florida

18 February 2013

I will readily admit that I am a very critical person when it comes to movies. After all, a norm counting) reviews to IMDb! However, "Argo" is an unusual film because I honestly can't think Really...it's THAT good! The film is about a joint effort by the Canadian government and the (Iran during their revolution in 1979-80. It seems that most of the Americans in the US emba: of folks escaped and sought shelter in the Canadian embassy. What happened next? See t

It's odd. In light of the film's greatest strength, how could the Oscar folks NOT have nominat many ways the film was wonderfully directed. Although I know the fate of the six refugees, I

Fig. 5 “Argo” IMDb reviews including prolific authors only. Source: Extracted by the authors

Table 10 Path expressions for *XPath* (http://www.w3schools.com/xpath/xpath_syntax.asp)

Expression	Description
<i>nodename</i>	Selects all nodes with the name <i>nodename</i>
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

Source: Extracted by the authors

not difficult to learn. Handy *XPath* tutorials are available at <http://www.w3schools.com/xpath/>.

We can get the *XPath* of the part of a webpage we are interested in by simply using the Google Chrome web browser. First, we open an IMDb review webpage in Chrome, choose the part we want to crawl, right-click on it, and from the pop-up menu select “Inspect Element.” Once that option is selected, we see two windows on the bottom side of the browser, and the part chosen earlier is highlighted in the left-down side HTML code area. We right-click on the highlighted area, and select “Copy XPath” from the pop-up menu and then obtain the raw *XPath* expression for the IMDb review, such as “//*[@id="tn15content"]/p[1]/text()” (Fig. 6).

The screenshot shows a web browser displaying a movie review for "Under the Dome". The review text is highlighted in blue, and the browser's developer tools are open, showing the HTML and XPath for the selected text.

Review Text:

*** This review may contain spoilers ***

I understand it may flatter US patriotism, or recall memories to those who remember the events and I don't even dispute Affleck's directorial and acting skills. However, this is a completely superfluous, empty and desperately predictable movie. The historical inaccuracy has been pointed out by several other reviews: no, things didn't happen that way, the Canadians deserve much more credit in that operation than this portrayal ever shows. Notwithstanding the role of the US in sustaining a puppet dictatorship during the Shah and actively interfering in a sovereign country's domestic politics for decades prior to the events. But this is only a secondary concern. Historical accuracy is not the most important factor for a fiction, even when it's based on actual events. What I dispute is how incredibly shallow and predictable the storytelling is: cliché anonymous US CIA antihero agent with issues at home goes to a dangerous place, saves innocent lives, takes risks against orders, comes out victorious to reunite with his family. Who on Earth cares, seriously? And no, the fact that it's based on historical events - and therefore you can't argue with history - is not an answer precisely because the script takes so many liberties with the events. I don't care about the liberties taken with history but I care about the ability to portray convincingly the complexity of human emotions and relationships. There is none here. And make no mistake, a fictitious 2 min car chase at an airport is the closest you'll get to see some emotions (ie. anguish at being killed by the revolutionary guards). The characters come out of a cardboard factory, they have zero critical self-reflection about their own role in interfering with a foreign country's domestic affairs, total solidarity with each other and pure love for their partners. This is a Disney version of human psyche, a dishonest and partial historical account and a debauchery of time, energy and money, ill spent. Affleck is an able actor and I hope will grow more convincing in his future efforts as a director, but what really baffles me is not the mediocrity of this film, it's the uncritical enthusiasm of so many for it.

Developer Tools (Styles):

```

element.style {
  #tn15content p {
    line-height: 131%;
  }
  p {
    margin: 0.5em 0 0.75em;
    padding: 0;
  }
  user-agent stylesheet {
    display: block;
    -webkit-margin-before: 1em;
    -webkit-margin-after: 1em;
  }

```

Developer Tools (XPath):

```

//div[@id='tn15content']

```

Fig. 6 Obtaining the XPath of review text using Google Chrome. Source: Extracted by the authors

```

library(XML)

#Crawling IMDB
doc <-htmlParse("http://www.imdb.com/title/tt2013293/reviews?count=76&start=0")

#Get Review Quality and Score, and Review
xpath_quality<-xpathSApply(doc, "//*[@id=\"tn15content\"]//div//small[1]",xmlValue)
xpath_score<-xpathSApply(doc, "//*[@id=\"tn15content\"]//div//img[last()]", xmlGetAttr,
"alt")
xpath_text<-xpathSApply(doc, "//*[@id=\"tn15content\"]//p[not(b)]",xmlValue)
xpath_text1 <- gsub("\n", " ",xpath_text[1.length(xpath_text)-1])
xpath_text2 <- gsub("\r", " ",xpath_text1)

# Combine lists to matrix
table<-cbind(xpath_score,xpath_quality,xpath_text2)


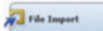
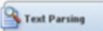
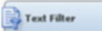
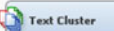

# Write matrix to file
write.table(table, file = "Your_file_path.txt",sep="\t")

```

Fig. 7 R code to extract IMDb reviews. Source: Extracted by the authors

In this chapter, we use R to handle the process of crawling, transforming, and loading IMDb reviews. To handle the *XPath* in R, we need to first install the “XML” package. After installation, we can run the R code in Fig. 7 to crawl and parse movie reviews. The result looks like that in Fig. 8.

4.1.3 Text Mining Using SAS Enterprise Miner

In the next step, we handle the textual dataset using SAS Text Miner, which is a plug-in for the SAS Enterprise Miner environment. The Enterprise Miner interface is displayed in Fig. 9, after we have created a New Project and New Diagram. We can then create a SAS dataset from the IMDb review documents, using the Text Import node  or File Import node . The Text Parsing node  in SAS Text Miner decomposes the documents into detailed terms or phrases, and the Text Filter node  automatically detects misspelling in the data and transforms the quantitative representation into a compact and informative format. The Text Cluster node  clusters documents into disjoint sets of documents, and the Text Topic node  creates topics for each document, where one document can be associated with more than one topic (Fig. 10).

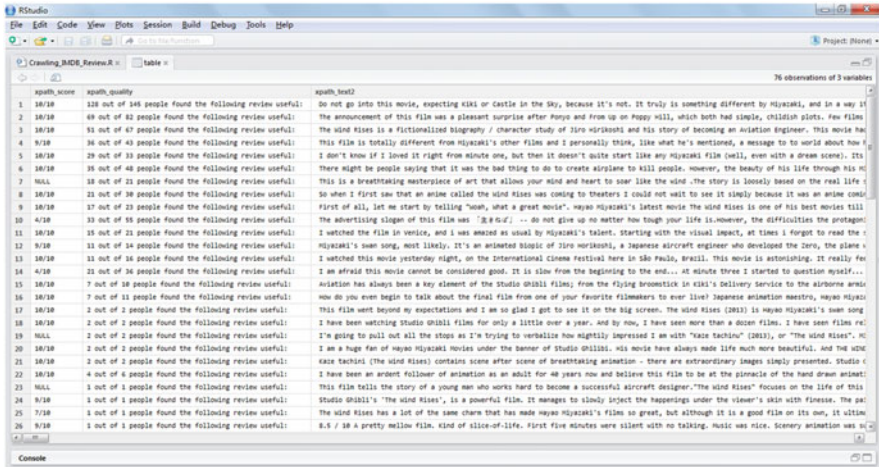


Fig. 8 Extracted IMDb reviews. Source: Extracted by the authors

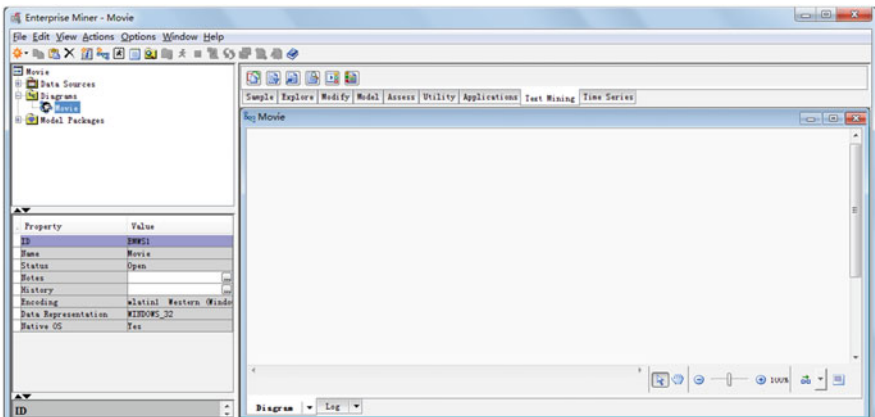


Fig. 9 SAS Enterprise Miner diagram. Source: Extracted by the authors

4.1.4 Review Themes and Oscar Chances

In this section, we discuss how a text mining of the IMDb pre-Oscars reviews gives an idea of the numbers of different themes which are perceived by reviewers for each movie and potentially yields a preliminary measure of perceived controversy. The question is then of how much “controversy” is optimal for Oscar winning purposes.

Measures of controversy and how they are used in marketing are discussed in Zhang and Li (2010). A quote from this article is very pertinent to our discussion. “From a persuasion point of view, our belief is that a convincing argument is not

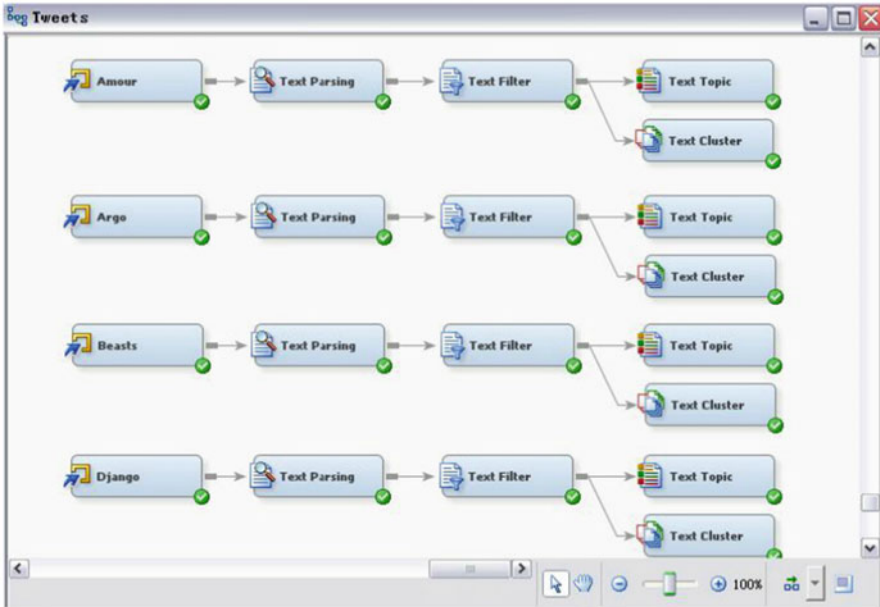


Fig. 10 SAS Enterprise Miner diagram for movie tweets. Source: Extracted by the authors

necessarily a mono-color picture, but instead a meaningful “bag” of positive and negative reflections” (p. 2).

This point of view may very well apply to movie chances for *Oscars* and other measures of success such as profit. Zhang and Li (2010) state that one possible quantitative measure of controversy is simply the standard deviation of the numerical ratings, and we adopt that point of view here as well. 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 value decomposition* (SVD) techniques are used to reduce the matrix without losing too much information, and a 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.

As an illustration, Table 11 displays the results of this clustering exercise for *Argo*. For example, the main theme in cluster 3 is clearly related to perceived *Oscar* chances for the movie, director, and leading actor (Ben Affleck), and the main theme in cluster 4 is about the thrilling aspects of the movie.

In the case of *Argo*, the text analysis yielded six clusters. Reviews of *Amour*, a movie by a controversial director (Haneke) on a very complex theme, related to

Table 11 Clusters and main terms for *Argo* reviews

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

Source: Computed by the authors

death and euthanasia, yielded 23 clusters. Aside from whether these extracted themes are expressing positive or negative sentiments (and the approach to measuring perceived controversy in Zhang and Li (2010) does use the number of positive and negative sentiments), it is reasonable to surmise that the number of issues such a complex movie rises may be simply too large for a group to rally on.

Our study, based on just nine movies, is still preliminary; we suggest that a very interesting future direction would involve looking at measures of complexity for a much larger number of movies and investigating how correlated such measure would be to, for example, the standard deviation of ratings.

Figure 11 displays a scatter plot of the standard deviation of ratings against the number of clusters extracted by the text analysis for the nine movies.

With a few caveats, first that the standard deviation of ratings is fairly small for all nine movies, and that *Zero* and *Amour* act as outliers, we can see that the standard deviation of ratings shows a propensity to increase with the number of clusters. It is interesting to note that the five serious contenders for the *Best Picture Award* as perceived by the *Intrade* market (*Argo*, *Les Misérables*, *Lincoln*, *Silver Linings Playbook*, and *Zero Dark Thirty*) tend to yield a moderate number of clusters, in other words tend to raise a number of issues which a group can potentially rally on (see Fig. 11 and Table 12).

Further investigations of controversy indices in movie reviews and their role on measures of success would be very interesting. Controversy is highly correlated with word-of-mouth (WOM) activity and WOM marketing. WOM is the process of information exchange, involving in particular recommendations about products and services, between two people in an informal way (O’Leary & Sheehan, 2008). WOM communication could have a strong influence on consumer short-term and long-term purchasing behavior, influencing both short-term and long-term judgments (Bone, 1995). Another advantage of WOM is that cost of WOM marketing is low, for both online and off-line channels.

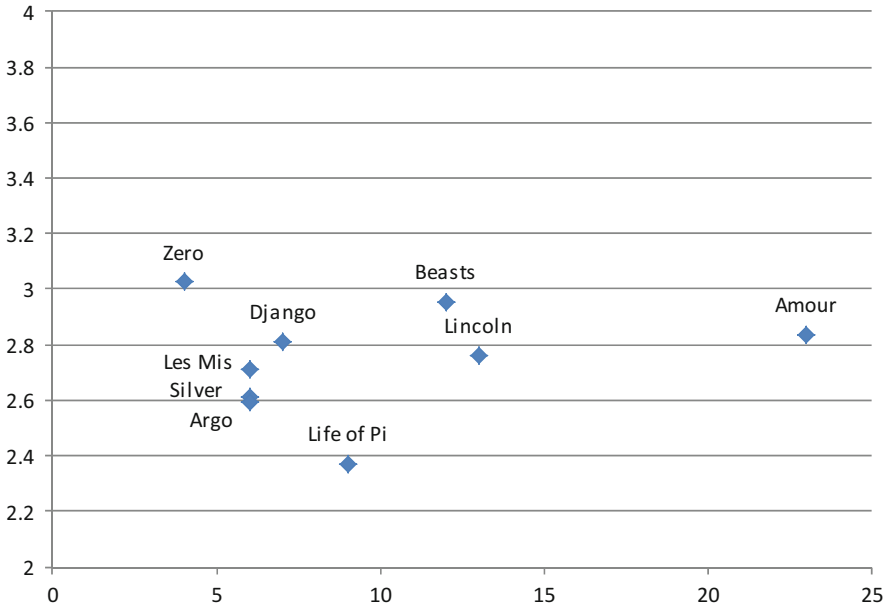


Fig. 11 Standard deviation of ratings and number of clusters for each nominated movie to win the 2013 *Best Picture Award*. Source: Computed by the authors

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

	Mean rating	Number of themes (of clusters)	Mean <i>Intrade</i> close	Last <i>Intrade</i> close	Standard deviation 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

Source: Computed by the authors

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 to brand coming from negative WOM or mixed WOM (so-called controversy)? Some

research indicates the possibility that controversy arising from consumers' opinion might have a positive impact. Liu (2011) finds that movie box office revenue is correlated with the volume of WOM activity but *not* correlated with the percentage of negative critical reviews. Zhang and Li (2010) observe that controversy can attract market attention and potentially yield strong sales. On the other hand, consider the role of controversy in the 2014 *Best Animated Feature Award*, as pertains to Hayao Miyazaki's *The Wind Rises*. *The Wind Rises* lost the award in large part because of controversy surrounding the theme in the movie. We suggest that getting a better handle on what constitutes controversy, relying on progress in text mining techniques, is likely to illuminate problems which are to date difficult to apprehend.

To conclude, this chapter has discussed a number of approaches to predicting box office revenue using variables available before the release of the movie and has also presented a number of correlates of *Oscar* awards. It is clear that data analysis, coupled with strong human judgment, is likely to be the key combination to investor's risk control. 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

- 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, FL: 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
- El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., et al. (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.
- Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. *ICML*, 96, 148–156.
- Gold, M., McClarren, R., & Gaughan, C. (2013). The lessons Oscar taught us. Data science and media & entertainment. *Big Data*, 1(2), 105–109.
- Haughton, D., McLaughlin, M.-D., Mentzer, K., & Zhang, C. (2015). *Movie analytics. A Hollywood introduction to big data*. Heidelberg: Springer.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning. A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674.
- Leamer, E. E. (2008). *What's a recession, anyway?* (NBER working paper). National Bureau of Economic Research. <http://www.nber.org/papers/w14221>
- Leonhardt, D. (2015, February 23). Oscars 2015. An excellent night for prediction markets. *The New York Times*.
- Liu, B. (2011). *Web data mining: exploring hyperlinks, contents, and usage data*. Berlin: Springer.
- McKenzie, J. (2012). The economics of movies. A literature survey. *Journal of Economic Surveys*, 26(1), 42–70.

- 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>
- O'Leary, S., & Sheehan, K. (2008). *Building buzz to beat the big boys: word-of-mouth marketing for small businesses*. Westport, CT: Praeger.
- 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*. Dissertation, University of Nottingham.
- Shambaugh, J. C. (2012). The Euro's three crises. *Brookings Papers on Economic Activity*, 43, 157–231.
- Zhang, Z., & Li, X. (2010). Controversy in marketing. Mining sentiments in social media. In *Proceedings of the 43rd Hawaii international conference on systems sciences*.

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 Houghton 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 Lemerrier 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.