JOURNAL OF INFORMATION SYSTEMS ENGINEERING & MANAGEMENT

Journal of Information Systems Engineering & Management, 2018, 3(3), 22
ISSN: 2468-4376

lectito

Forecasting Movie Box Office Profitability

Marta Galvão [1]*, Roberto Henriques [1]

[1] *NOVA IMS, Universidade NOVA de Lisboa*, Lisboa, PORTUGAL

***Corresponding Author:** martagalvao28@gmail.com

Citation: Galvão, M. and Henriques, R. (2018). Forecasting Movie Box Office Profitability. *Journal of Information Systems Engineering & Management*, 3(3), 22. https://doi.org/10.20897/jisem/2658

Published: July 16, 2018

ABSTRACT

This study intends to estimate the profit of a movie through the construction of a predictive model that uses several Data Mining techniques, namely neural networks, regression and decision trees. The model will allow obtaining the prediction of box office revenue. Three different dependent variable approaches were used (interval, categorical and binary) aiming to study the difference and predictive influence that each one has on the results. Two metrics were used to determine the most accurate predictions: the misclassification error for the categorical models and the average squared error for the continuous one. In this study, the best predictive results were obtained through the use of multi-layer perceptron. Regarding the representative distinction between the dependent variable, the multiclass model presents a much higher error rate comparing to the remaining ones, which is explained with the increase of the number of classes to predict.

Keywords: data mining, movie profitability, box office profit, predictive analysis, neural networks, decision tree, regression

INTRODUCTION

The losses of an only slightly lucrative movie can contribute to the partial or even total downfall of the financial status of a movie studio. Data regarding 2012 allow to establishing a clearer vision of the importance of the profitability of movie production, assuming that only 10% of the released movies were responsible for more than 68.8% of the total box office revenue of that year (Ghiassi et al., 2015). This field is characterized by being one of the riskier to investors due to unpredictability. This is a multimillionaire industry that has generated more than 11 billion dollars in revenue in 2016 and a growth of more than 6 billion dollars in the last 11 years (Nash Information Services, 2016). Due to this uncertainty, it is imperative to know the cinematographic features that influence more the audiences of a movie, and their impacts assessed so that the risk associated with movie production is minimized. In this paper, we propose to model the movie profitability and compare the results using three different dependent variables: 1) an interval variable with the value of profitability for each movie; 2) a categorical variable with the profitability values transformed into 9 classes and; 3) a binary variable indicating if the movie resulted in profit or deficit. Ultimately, creating three different predictive models will allow us to understand better the phenomena and improve the accuracy.

LITERATURE REVIEW

Allied to the great computational development that we see nowadays, an urgent need for obtaining insights through big volumes of raw data emerged. This growth of the available data is the result of our society's digital transformation and the rapid evolution of powerful tools for its storage (Han et al., 2012). Through data mining

analysis, it is possible to discover patterns and tendencies that rapidly will become an essential pillar for this industry to be able to succeed today (García et al., 2007). The success of this industry is directly related to the box office revenues that reflects on the profit of the cinematographic distributor as the main indicator of the commercial success of a movie (Wallace et al., 1993).

Litman (1983) was the great pioneer of this field of research through the elaboration of a multiple regression model capable of predicting the financial success of a movie. Delen and Sharda (2006) developed the same study applying a different methodology: neural networks. In this approach, the predictive problem was converted in a classification problem, the division of the movies being made in nine distinct categories according to their box office revenue, having the success rate only reached 36.9%. This value was improved in 2009 by the same authors that conducted another investigation where they tested several prediction models besides neural networks, including decision trees and a more comprehensive sampling of movies. Recent studies in the area were able to significantly increase the percentage of success of the cinematographic forecast, using the main techniques used in the project work (Fazzion et al., 2013; Hunter et al., 2016; Im and Nguyen, 2011; McKenzie, 2013; Song and Han, 2013). The need to explore this theme motivated the existence of new studies that allowed an exponential increase in the prediction of the financial success of the movies, with the greater percentage being of 94.1% (Ghiassi et al., 2015).

DATA AND USED VARIABLES

The majority of variables used in this paper were adopted in previous studies in its continuous attempt to improve the sensitivity of the predictive methods (Ghiassi et al., 2015; Sharda and Delen, 2006). It is considered of the utmost importance to present a brief description of each one of the independent variables integrated into this analysis, putting in context their respective importance according to the authors that were considered to be the most relevant. The variables selected considered not only the most relevant studies and with a higher success rate (sequel, actor, director, budget, nominations, genre, season, MPAA, Oscars), but also other variables not so common (reviews, duration, awards, spectators).

The dataset used in this paper was collected through the Opus Data, a data service that belongs to Nash Information Services. Subsequently, this dataset was enhanced with information from one of the largest virtual archives of current times, the Internet Movie Database (IMDB). The final dataset used includes 1920 movies from 2000 to 2016 and is related to the global box office revenues, which includes the domestic market and the international one. Some of the variables used were extracted not from the main IMDB database, but from other digital platforms: (1) the variable "spectators" was obtained through a rating created by IDMB, that is calculated through the weighted average of all the votes made by the users of the site; (2) the variable "critics" was obtained through the site Metacritic, that gathers all the evaluations of the most important critics of movies of the world in one value that will illustrate the global decision that characterizes a certain production and (3) the variables "actor" and "director" were classified according to the list of the most lucrative actors and filmmakers of all time (Mojo, 2016a) .

Independent Variables

Sequel

Sequels nowadays present an ever more relevant strategy in the introduction of new movies in the cinematographic culture of spectators through the capitalization of the success of previous productions, which is proven by some studies that concluded that sequels contribute for an increase in box office revenue (Ravid, 1999; Simonoff and Sparrow, 2000).

MPAA

Motion Picture Association of America (MPAA) is the entity that is responsible for the movies classification per age in America. Its organization is based on five categories: G (General Audiences), PG (Parental Guidance Suggested), PG-13 (Parents Strongly Cautioned), R (Restricted) e NC-17 (Adults Only). A study conducted by Sawhney and Eliashberg (1996) reveals that movies with an MPAA R classification will obtain lesser box office revenue than the remaining ones, a conclusion is also taken by other authors such as Sochay (1994).

Genre

Many studies prove the influence of cinematographic genre in box office revenue (Simonoff and Sparrow, 2000; Vany and Walls, 2002; W. David Walls, 2005). Certain authors defend that the only relevant genre that shows a positive relationship is the science fiction one (Litman, 1983), as others state through their studies that action movies and thrillers are the most significant and popular for the spectators when they decide which movie to watch.

Table 1. Independent variables used in the model

Independent Variable	Type	Values
Sequel	Binary	0,1
MPAA	Nominal	G, PG, PG-13, R, NC-17
Genre	Nominal	Action, Animation, Adventure, Biography, Comedy, Crime, Documentary, Drama, Familiar, Fantasy, Scientific Fiction, History, Horror, Mystery, Musical, Thriller, Western
Budget	Interval	10350000 – 425000000
Oscars	Interval	0 – 11
Awards	Interval	0 – 234
Nominations	Interval	0 – 549
Directors	Binary	0,1
Actors	Binary	0,1
Season	Nominal	Winter, Summer, Easter, Normal
Spectators	Interval	0 – 9
Critics	Interval	9 - 98

Budget

The budget of a movie is strongly correlated with the predictability of its box office revenue (Deniz and Hasbrouck, 2012). Litman (1983) argues that big budgets represent bigger quality and more popularity in box offices. It also argues that still associated to the budget of the production of a movie, are the excessive salaries of the stars, significant production delays or inefficient management, which inflate the final values that do not translate in its entirety to the quality of the production.

Oscars, Awards, and Nominations

The cinematographic awards, including the academy ones, represent an important and prestigious stepping-stone for worldwide cinema. It is important to understand that the monetary distinction of these awards is not limited to the winners, being their nomination sufficient to generate bigger revenues than the others (Deuchert et al., 2005).

Stars (Directors and Actors)

The phenomenon "star" that an actor reaches by its popularity in Hollywood seems to favor productions. Nowadays, according to some authors, it is an important variable for their profitability (Prag and Casavant, 1994; Wallace et al., 1993). The associated cost to become a "star" in the cast is quite high; however, it can bring very favorable revenues in the long term. It is estimated that the average impact of an important actor is an increase in 6,5 million dollars in the final revenue (Walls, 2009).

Season

The movies premiere date is a variable that influences box office revenue since spectator's affluence to the movie theaters increases significantly on holidays or festive seasons. Several authors proved their direct relationship with cinematographic revenue o (Basuroy et al., 2003; Chang and Ki, 2005; Simonoff and Sparrow, 2000).

Spectators and Critics

The evaluations received before the visualization of a movie are always important to the spectator, either given by other spectators or attributed by professional critics. For the cinematographic industry, this investment consists of added value regarding marketing since it makes it possible to predict the success or non-success of a movie in the long term (Chakravarty et al., 2010; Duan et al., 2008).

Dependent Variable

The dependent variable used in this paper was "profit" (box office revenue). This variable is used in the study in three distinct ways: (1) the first one will be represented in its original form, interval, being the objective to determine the exact value of the profit of a certain production; (2) the second one consists of its binary form, where it will take the value "1" if its box office revenue is equal or higher than the double of the budget or "0" otherwise and (3) lastly, the profit will be based on the methodology of Sharda and Delen (2006, 2009) in which the interval variable is transformed in a discrete one, more precisely in nine distinct classes (the more reduced class - class 1 - represents the movie that were a failure in movie theaters, while successes are exhibited by the bigger class – class 9). The set of data that was used in this project was extracted through the Opus Data, a data service that belongs to Nash Information Services, a powerful platform that puts the effort in the supply.

DIMENSIONALITY REDUCTION

The model's complexity can hinder the comprehension of the forecast, so it is necessary to reduce the study's number of variables, removing from the model the variables that are less relevant and that have the least influence on the dependent variable (profit) to obtain a less complex analysis. Some authors claim that for the variables to be highly relevant, the values of the variables should vary according to the different classes (Gennari et al., 1989). In this way, it is possible to avoid redundancy and reduce the dimensionality of the predictive model. Three distinct techniques of variable elimination were used, allowing for the model to provide more intuitive predictions:

Chi-squared and coefficient of determination

Regarding the discrete dependent variable, the chi-squared (X^2) is a numerical test that measures the expected distribution's deviation, being the study's variables independent from their class value, i.e., the influence of the independent variables on the dependent variable (box office revenue) is evaluated (Ikran and Cherukuri, 2016). When dealing with continuous dependent variables, the chi-squared test cannot be used, and the coefficient of determination (R2) has to be used, which is defined as a statistical measure of the percentage of adjustment of the variable in relation to the linear model.

Spearman and Pearson's correlation matrix

It is fundamental to analyze the correlation matrix, avoiding including variables that are highly correlated in the analysis. The Spearman's correlation coefficient, a non-parametric statistic of classification, is the adequate statistic for ordinal data unlike the Pearson's correlation method, a parametric statistic classification, only applicable for interval data (Chok, 2010). Two variables are classified as highly correlated if their correlation coefficient is equal or higher than 0.75 (Marôco, 2014).

Coefficient of regression

It is possible to reject variables if their regression coefficient is not significant for the model, using p-values. The acceptable value for this measure is 0.05, the usual significance level, meaning that all variables with a p-value higher than this will be considered insignificant and unessential for the model.

METHODOLOGY

The methodology used in this project is called SEMMA (Sample, Explore, Modify, Model, Assess), which is a Data Mining process developed by the SAS Institute (REF to SEMMA). One of the main decisions on the design of a predictive model is the learning method to be used. In this paper, neural networks, decision trees, and multiple regression were used. The main goal of this work is to obtain a model that is realistic and precise in its behavioral predictions of future observations, having historical observations as a basis.

Neural Networks

A neural network is an innovative predictive tool that has been applied in some scientific areas, whose goal is to train models to be capable of answering to several kinds of problems (Craven and Shavlik, 1997). In this project, we used the Multi-Layer (MLP) neural network, one of the most well-known and successful architectures used in predictive and classification problems. MLP neural network uses a backpropagation algorithm, typical in this kind of artificial intelligence, and which is characterized by its decreasing gradient throughout the whole network, capable of minimizing the average squared error of the model's output (Henriques and Hajek, 2017). Each variable of the initial layer is trained and synthesized by the hidden layers, and the values obtained in the output layer are compared with the real ones, concluding the process with the calculation of the error value that the network obtained. One of the reasons for the popularity of neural networks is their flexibility, that is, the capacity for modeling a large number of functions (Williams, 1998) to generalize the model and classify new data.

Decision Tree

The Decision tree, that in this study uses the CART algorithm, represents an effective predictive tool with some specific characteristics: the use of a graph that is identical to a tree, where decisions and consequences were represented and where a certain strategy was chosen to maximize the forecast decision trees.

Regression

Multiple Regression has been one of the most popular tools in predictive models that account for the relationship between multiple independent variables and one dependent variable. It is mostly used for continuous numerical predictions, although it also can deal with discrete tendency identification (Han et al., 2012). As

Table 2. Average squared error of Interval Model in the three methods

			NN	RG	DT
WITH OUTLIERS	70/15/15	R	0.2707	0.3298	0.4018
		PC	0.2707	0.3226	0.4957
		r^2	0.2859	0.3241	0.3961
	70/30	R	0.2950	0.3763	0.3711
		PC	0.2972	0.3741	0.4627
		r^2	0.3092	0.3725	0.3593
WITHOUT OUTLIERS	70/15/15	R	0.3180	0.3215	0.3992
		PC	0.2694	0.3258	0.3036
		r^2	**0.2361**	0.3264	0.3034
	70/30	R	0.2437	0.3278	0.3268
		PC	0.2676	0.3314	0.3310
		r^2	0.3030	0.3352	0.3301

Table 3. Misclassification error of binary model in the three methods

			NN	RG	DT
WITH OUTLIERS	70/15/15	R	0.1035	0.1241	0.1069
		ρ	0.1104	0.1	0.1
		X^2	0.1104	0.1207	0.1
	70/30	R	0.0951	0.0986	0.1021
		ρ	0.0917	0.0900	0.0865
		X^2	0.0882	0.0986	0.0900
WITHOUT OUTLIERS	70/15/15	R	**0.0669**	0.0775	0.0810
		ρ	0.0880	0.0986	0.0845
		X^2	0.0845	0.0986	0.0845
	70/30	R	0.0671	0.0830	0.0989
		ρ	0.0972	0.0901	0.1025
		X^2	0.1060	0.0919	0.1025

regression is one of the most used statistical techniques, it is capable of modeling the relationship between the intervenient variables (Cerrito, 2008).

RESULTS

To improve the prediction accuracy, different factors such as the data partition schema (training, validation, and test or training and test), the outliers' treatment (with outliers and without outliers) and the reduction of dimensionality (chi- squared/ coefficient of determination, Spearman's/Pearson's correlation matrix or regression coefficient) were tested. For each combination of the previously presented factors, three predictive models were tested: MLP, multiple regression and decision tree.

Interval Profit

The model with the interval dependent variable, represented in its original form, was tested. **Table 2** represents the average squared error of the three methods used in this study combined with some factors like data partition - 70/30 or 70/15/15 -, outliers' treatment - with outliers and without outliers - and method of dimensionality reduction - Coefficient of determination (**r2**), Pearson's correlation matrix (PC) or regression coefficient (R). The neural network with four neurons in the hidden layer was the one that best translates the required prediction with an average squared error of 0.2361 and a maximum absolute error of 3.0232, value obtained through a sample without outliers with a partition in three distinct sets (set of training, validation, and test), and recurring to the coefficient of determination for dimensionality reduction.

Binary Profit

The binary form of the dependent variable will take the value "1" if its box office revenue is equal or higher than the double of the budget or "0" otherwise. With a misclassification error extremely small (0.0669), and with a ROC index quite high (0.979), obtained through the neural network with three neurons in the hidden layer, resorting to the treatment of outliers, with a partition of three distinct sets (set of training, validation and test), and to the regression coefficient for the dimensionality reduction, it is proven that the number of classes by which the dependent variable is categorized makes all the difference for the success of the predictive model.

Table 4. Confusion Matrix of the binary model

		Actual Class		
		0	1	
Prediction Outcome	0	35	4	39
	1	3	59	61
		37	63	100%

Table 5. Categorical profit based on the methodology of Sharda and Delen

Class	Profit Range (in millions)
1	<1
2	[1,10[
3	[10,20[
4	[20,40[
5	[40,65[
6	[65,100[
7	[100,150[
8	[150,200[
9	>200

Table 6. Misclassification error of categorical model in the three methods

			NN	RG	DT
WITH OUTLIERS	**70/15/15**	**R**	**0.6**	0.6509	0.6712
		ρ	0.6081	0.6350	0.6103
		X^2	0.6475	0.6610	0.6644
	70/30	**R**	0.6501	0.6604	0.6621
		ρ	0.6106	0.6604	0.6587
		X^2	0.6535	0.6570	0.6621
WITHOUT OUTLIERS	**70/15/15**	**R**	0.6338	0.6444	0.6796
		ρ	0.6197	0.6585	0.6796
		X^2	0.6373	0.6655	0.6620
	70/30	**R**	0.6418	0.6472	0.6614
		ρ	0.6153	0.6525	0.6614
		X^2	0.6454	0.6472	0.6507

Of the 295 observations of the test set, only 7% of them were misclassified, a value that is quite favorable for this model's accreditation and evaluation.

Categorical Profit

The categorical profit is based on the methodology of Sharda and Delen (2006, 2009) in which the continuous variable is discretized according to the criteria shown in **Table 5**.

Table 6 presents the misclassification error of the three methods used in this study – neural networks (MLP), multiple regression (MR) and decision trees (DT) – for different data partitions (70/30 or 70/15/15 for training/validation/testing), outliers' treatment (with or without outliers), and applying different methods for dimensionality reduction (R, ϱ or X2).The neural network with three neurons in the hidden layers was the one that presented a smaller misclassification error (0.6) as well as a quite high ROC index (0.933) and, consequently, it was the one that presented a better predictive power when used in a test set. The error value, despite being the smallest among all the used methodologies is also, in representative terms, very high for the model to be considered favorable in response to the main question of this study, which is given to the big number of classes of the dependent variable.

Of the 295 observations of the test set, only 39% were correctly classified, which translates well the question described above that relates the weak percentage of success in the classification with the high number of classes to be predicted

Therefore the prediction of the sample was extended, being that for a successful percentage is not only necessary to get the real class right, but also in the two inferior and superior classes: that is, if the model classifies a movie that presents a monetary class 6 with the class 4 it means that the learning succeeded for this observation since, despite not having predicted the real class, a class that stands two digits below the one that was predicted, which for this study translates into a predictive success. The results of the test set prediction were analyzed to

Table 7. Confusion Matrix of multiclass model

		Actual Class									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Prediction Outcome	**1**	0	0	0	0	0	0	0	0	0	1
	2	0	1	0	2	1	1	0	0	0	6
	3	0	0	0	4	1	1	1	0	0	8
	4	0	1	0	8	1	2	1	0	1	14
	5	0	0	0	6	2	3	1	0	1	14
	6	0	0	0	3	1	5	1	0	4	15
	7	0	0	0	1	1	4	3	0	3	12
	8	0	0	0	0	0	3	2	0	3	8
	9	0	0	0	1	0	1	1	0	20	22
		0	3	1	25	8	22	8	0	32	**100%**

Table 8. Comparison of different approaches to the multiclass model

Success Criterion	**Misclassification**
Predict the real class	0.6
Predict the real class with a margin of error of two classes (lower / upper)	0.15

Table 9. Errors of the three predictive models

Dependent Variable	**Misclassification**	**Average Squared Error**	**Predictive Methodology**
Multi-class	0.6		NN 3
Binary	0.0669		NN 3
Interval		0.2361	NN 4

readjust the misclassification error based on the new approach described above: it was observed a clear decrease in that error of approximately 45%.

CONCLUSION

This project had as a main goal to develop a model able of predicting the box office financial success of a certain set of movies through specific variables and historical data. It was possible to conclude that the percentage of success of the cinematographic revenue prediction is quite different based on the typology of the dependent variable used in the study.

The empirical model demonstrated good statistical results when the dependent variable was binary and interval. However, in what regards the multiclass prediction, the results were very far from reality, negatively influencing the model. In **Table 9**, the final classification errors for each one of the models and their respective predictive tools are represented.

It should also be noted that the high misclassification error of the multiclass variable presents a percentage of success of about 40%, less than 12.6% comparing to the rate obtained by Sharda e Delen in 2009.

The binary variable presents a quite favorable misclassification error since one of the studies with the biggest percentage of success resorting to neural networks and the same binary typology (Rhee and Zulkernine, 2016), obtained a positive performance of 88.8%, about 4.5 percentage points below the result of this study. On the other hand, the average squared error reached in the continuous model obtained very favorable results and overcame, significantly, the one obtained by a similar study that used the same typology (Hunter et al., 2016), and that was not able to reach error values inferior to 0.427.

Regarding the used methodologies, neural networks was the one that presented a better predictive power, with a clear advantage in all of the three developed models. This is a result in line with the literature review made in this study, as neural networks frequently presented the highest percentages of success (Ghiassi et al., 2015; Kaur and Nidhi, 2013; Rhee and Zulkernine, 2016). In what concerns the influence of the variables present in the study it was possible to conclude that some of them are not relevant to a bigger box office revenue, being for that reason, removed. The variable "actor" in the particular case of multiclass and interval profit did not present a great explanation value comparatively to the remaining ones, a situation that occurred simultaneously with the binary dependent variable regarding the variable "Oscars". On the other hand, there is a group of variables that explain the model quite well and that contribute in a clear way to its predictive success: "budget", "director" and "sequel".

FUTURE WORK

Some limitations in the predictive model that prevented it from reaching smaller error rates were identified. One of these, already described, is the use of a dependent variable characterized by nine classes. This choice was justified by the same use of this in one of the main studies performed in the area.
(Sharda and Delen, 2006, 2009), where the highest percentage of success reached 56.1%, a value quite higher when comparing to the one obtained with this study (40%). This significant difference can be justified by the use of a bit more diversified sample and with more observations, where the choice of the variables was also different and more selective, which represents another one of the limitations encountered: the high number of variables little significative to the model.

Initially, twelve independent variables were introduced and were subjected to a dimensionality reduction resorting to distinct techniques that sometimes ended up rejecting seven variables. If a more comprehensive evaluation of the relationship between the variables used in the study and the cinematographic revenue were previously done, their use would have been avoided, and others with more predictive value would have been included. As a reference and suggestion for future works, a predictive model that accounts for the needs and weakness of this study is proposed. Three dependent variable approaches were tested, being the binary one the most successful. However, in practical and monetary terms for the cinematographic studies, to adopt a more precise interval predictive model will give a higher competitive advantage in the current market. This way, the model using a continuous dependent variable, where the choice of the variables should be based on the ones that presented more value to the corresponding model of this study is recommended ("budget", "nominations", "awards", "director" and "sequel). Also, adding other interesting ones like the budget and profit obtained with the movie's marketing, the mouth-to-mouth opinion through social networks, exploring Text Mining models, and direct competitors (movies that were launched in the same month and characterized by the same genre) an improve the prediction.

REFERENCES

Basuroy, S., Chatterjee, S. and Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, 67(4), 103–117. https://doi.org/10.1509/jmkg.67.4.103.18692

Cerrito, P. B. (2008). The Difference between Predictive Modeling and Regression, 1–19.

Chakravarty, A., Liu, Y. and Mazumdar, T. (2010). The Differential Effects of Online Word-of-Mouth and Critics' Reviews on Pre-release Movie Evaluation. *Forthcoming at Journal of Interactive Marketing*. https://doi.org/10.1016/j.intmar.2010.04.001

Chang, B.-H. and Ki, E.-J. (2005). Devising a Practical Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property. *Journal of Media Economics*, 18(4), 247–269. https://doi.org/10.1207/s15327736me1804

Chok, N. (2010). Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. *Graduate School of Public Health*, 1–53. https://doi.org/10.1017/CBO9781107415324.004

Craven, M. W. and Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, 13(2–3), 211–229. https://doi.org/10.1016/S0167-739X(97)00022-8

Deniz, B. and Hasbrouck, R. B. (2012). What Determines Box Office Success of a Movie in the United States? *Proceedings for the Northeast Region Decision Sciences Institute*, (757), 447.

Deuchert, E., Adjamah, K. and Pauly, F. (2005). For Oscar Glory or Oscar Money? *Journal of Cultural Economics*, 29(3), 159–176. https://doi.org/10.1007/s10824-005-3338-6

Duan, W., Gu, B. and Whinston, A. (2008). The Dynamics of Online Word-of-Mouth and Product Sales – An Empirical Investigation of the Movie Industry. *Forthcoming at Journal of Retailing*. https://doi.org/10.1016/j.jretai.2008.04.005

Eliashberg, J. and Shugan, S. M. (1997). Film critics : Influencers or predictors ? *Journal of Marketing*, 61(2), 68. https://doi.org/10.2307/1251831

Galvão, M. and Henriques, R. (2018). Forecasting model of a movie's profitability. *In Information Systems and Technologies (CISTI), 2018 13th Iberian Conference on (pp. 1-6). IEE*. https://doi.org/10.23919/CISTI.2018.8399184

García, E., Ventura, S. and Romero, C. (2007). Data mining in course management systems : Moodle case study and tutorial. *Computers & Education*. https://doi.org/10.1016/j.compedu.2007.05.016

Gennari, J. H., Langley, P. and Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence*, 40(1–3), 11–61. https://doi.org/10.1016/0004-3702(89)90046-5

Ghiassi, M., Lio, D. and Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, 42(6), 3176–3193. https://doi.org/10.1016/j.eswa.2014.11.022

Hajek, P. and Henriques, R. (2017). Modelling innovation performance of European regions using multi-output neural networks. *PLoS ONE,* 12(10): e0185755. https://doi.org/10.1371/journal. pone.0185755

Han, J., Kamber, M. and Pei, J. (2012). *Data Mining: Concepts and Techniques*. *Journal of Chemical Information and Modeling* (Vol. 3). https://doi.org/10.1017/CBO9781107415324.004

Hunter, S., Smith, S. and Singh, S. (2016). Predicting Box Office from the Screenplay: An Empirical Model. *Journal of Screenwriting*, 7(2). https://doi.org/10.1386/josc.7.2.135

Ikran, S. T. and Cherukuri, A. K. (2016). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University - Computer and Information Sciences*, 1319–1578. https://doi.org/10.1016/j.jksuci.2015.12.004

Im, D. and Nguyen, M. T. (2011). Predicting Box-Office Success of Movies in the U. S. Market, 1–5.

Kaur, A. and Nidhi, A. P. (2013). Predicting Movie Success Using Neural Network. *International Journal of Science and Research (IJSR)*, 2(9), 69–71.

Litman, B. R. (1983). Predicting Succes of Theatrical Movies: An Empirical Study. *Journal of Popular Culture*, 159–175. https://doi.org/10.1111/j.0022-3840.1983.1604_159.x

Litman, B. R. and Kohl, L. S. (1989). Predicting Financial Success of Motion Pictures: The '80s Experience. *Journal of Media Economics*, *2*(2), 35–50. https://doi.org/10.1080/08997768909358184

Marôco, J. (2014). *Análise Estatística com o SPSS Statistics*. (ReportNumber, Ed.).

Mojo, B. O. (2016a). People index by gross. Available at: http://www.boxofficemojo.com/people/?view=Actor&sort=sumgross&p=.htm

Mojo, B. O. (2016b). Yearly Box Office. Available at: http://www.boxofficemojo.com/yearly/?view2=domestic&view=release date&p=.htm

Nash Information Services, L. (2016). Annual Ticket Sales. Available at: http://www.the-numbers.com/market/

Neelamegham, R. and Chintagunta, P. (1999). A Bayesian Model to Forecast New Product Performance in Domestic and International Markets. *Marketing Science*, 18(2), 115–136. https://doi.org/10.2307/193212

Prag, J. and Casavant, J. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. *Journal of Cultural Economics*, 18(3), 217–235. https://doi.org/10.1007/BF01080227

Ravid, S. (1999). Information, Blockbusters and Stars: A Study of the Film Industry. *Journal of Business*, 72. https://doi.org/10.1086/209624

Rhee, T. and Zulkernine, F. (2016). Predicting Movie Box Office Gross: A Neural Network Approach. *15th IEEE International Conference on Machine Learning and Applications*, 665–670. https://doi.org/10.1109/ICMLA.2016.138

Sawhney, M. S. and Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science*, 15(2), 113. https://doi.org/10.1287/mksc.15.2.113

Sharda, R. and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243–254. https://doi.org/10.1016/j.eswa.2005.07.018

Sharda, R. and Delen, D. (2009). Predicting the financial success of Hollywood movies using an information fusio approach, 30–38.

Simonoff, J. and Sparrow, I. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance-Berlin Then New*, 13(3), 40. https://doi.org/10.1080/09332480.2000.10542216

Sochay, S. (1994). Predicting the Performance of Motion Pictures. *Journal of Media Economics*, *7*(4), 1–20. https://doi.org/10.1207/s15327736me0704_1

Vany, A. and Walls, D. (2002). Movie stars, big budgets, and wide releases. Empirical analysis of the blockbuster strategy. *Latin American Meeting of the Econometric Society*.

Wallace, W., Seigerman, A. and Holbrook, M. (1993). The role of actors and actresses in the success of films: How much is a movie star worth? *Journal of Cultural Economics*, 17(1), 1–27. https://doi.org/10.1007/BF00820765

Walls, W. D. (2005). Modeling Movie Success When "Nobody Knows Anything": Conditional Stable-Distribution Analysis Of Film Returns. *Journal of Cultural Economics*, 29(3), 177–190. https://doi.org/10.1007/s10824-005-1156-5

Walls, W. D. (2009). Screen wars, star wars, and sequels. *Empirical Economics*, 37(2), 447–461. https://doi.org/10.1007/s00181-008-0240-z

Williams, C. K. I. (1998). Prediction with Gaussian processes: from linear regression to linear prediction and beyond. *Learning and Inference in Graphical Models*.

