Using Decision Trees to Characterize and Predict Movie Profitability on the US Market

Maria Luisa Campanario
Universidad a Distancia de Madrid
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Juan A. Lara
Universidad a Distancia de Madrid
47 PUBLICATIONS **333** CITATIONS
SEE PROFILE

David Lizcano
Universidad a Distancia de Madrid
64 PUBLICATIONS **545** CITATIONS
SEE PROFILE

Using Decision Trees to Characterize and Predict Movie Profitability on the US Market

María C. Burgos, María L. Campanario, Juan A. Lara, David Lizcano

Abstract—**The filmmaking is one of the most important branches of the entertainment industry primarily because of the huge revenues that it generates. The producer plays an essential role in filmmaking, as they provide the funding required to turn out quality blockbusters for cinemagoers. Film production is a risky business, as illustrated by the examples of films that fail to cover costs every year. In this respect, tools capable of predicting movie profitability are of potential use to producers as a decision-making tool for deciding whether or not to produce a movie project.**

In this paper we report a study using historical data on over 100 films produced in the United States (including their genre, opening month, duration, budget, etc.). Decision trees were extracted from these data in order to forecast whether or not a film will be profitable even before it is produced. Decision trees are models commonly used in the field of artificial intelligence as decision support tools.The results show that the resulting model forecasts whether or not a movie will be profitable with an accuracy of over 70%, and this model can be used as a decision support tool for film producers. The proposed approach is not designed to be used as a standalone tool; it should rather round out other forecasting methods, including producers' foresight and judgement.

The approach presented here could be equally applicable to other branches of the entertainment business, such as the music or video game industries.

Index Terms— **Movie industry, Movie profitability prediction, Data Mining, Decision trees, Decision Support Systems.**

I. Introduction

THE term entertainment refers to all the activities that provide human beings with enjoyment and amusement during their leisure time in order to temporarily evade their worries. Entertainment plays a crucial role for human beings. The philosopher Blaise Pascal opened a window on this idea when, in his 1662 work "Les Pensées", he defended that man has need of periods of diversion to avoid thinking about other more vital matters of life [1].

Beyond its social role, entertainment, which is part of the extended family of leisure, has become one of the most important sectors of the economy. In monetary terms, the entertainment industry earned US$ 1.7 trillion in 2012. With the leave of the growing video game industry, filmmaking is still one of the most important industries in the entertainment business, outperforming others like the music industry, for example. Much of this expansion is related to the consumer demand for access to quality cinema. In this scenario, film production is a key activity within the film industry. Without the people or organizations to finance filmmaking, it would be impossible to make quality movies. Film production is, however, a risky business. This is illustrated by the films that make a loss every year, profitability being the result of evaluating Equation (1).

$$P=((G\text{-}B)/B)\times 100 \quad (1)$$

where P is the profitability of a film, G represents the gross receipts and B is the amount of money invested in the activities necessary to make and distribute a film.

In the film industry, examples of movies that did not even manage to cover their costs (negative profitability) abound (see Table I). At the other end of the scale, of course, there are films that are extremely profitable. Some recent examples from 2013 are 'Despicable Me 2' (investment of 76 million vs revenue of 781 million), 'The Conjuring' (investment of 20 million vs revenue of 193 million) or 'Iron Man 3' (investment of 200 million vs revenue of 1200 million). Table II lists some other profitable films.

TABLE I
HISTORICAL EXAMPLES OF FILMS WITH NEGATIVE PROFITABILITY [1]

Title	Investment ($\times 10^6$ \$)	Gross ($\times 10^6$ \$)	Profitability
The Adventures of Pluto Nash (2002)	100	4.4	-95.6
Town & Country (2001)	90	6.7	-92.6
Heaven's Gate (1980)	44	3.5	-92.0
Cutthroat Island (1995)	98	10	-89.8
Mars Needs Moms (2011)	150	21.4	-85.7
The 13th Warrior (1999)	160	32.7	-79.6
The Alamo (2004)	107	22.4	-79.1
Final Fantasy: The Spirits Within (2001)	137	32.1	-76.6
Speed Racer (2008)	120	43.9	-63.4
Sahara (2005)	130	68.7	-47.2

TABLE II
HISTORICAL EXAMPLES OF HIGHLY PROFITABLE FILMS[1]

Title	Investment ($\times 10^6$ \$)	Gross ($\times 10^6$ \$)	Profitability
My Big Fat Greek Wedding (2002)	5	241.4	4728.0
E.T.: The Extra-Terrestrial (1982)	10.5	359.2	3321.0
Star Wars (1977)	11	307.3	2693.6
Grease (1978)	6	160	2566.7
Home Alone (1990)	18	285.8	1487.8
Pretty Woman (1990)	14	178.4	1174.3
The Passion Of The Christ (2004)	30	370.3	1134.3
Ghost (1990)	22	217.6	889.1
Slumdog Millionaire (2008)	15	141.3	842.0
American Beauty (1999)	15	130.1	767.3

Manuscript received October 22, 2014; revised November 18, 2014.

María C. Burgos, María L. Camapanrio, Juan A. Lara (phone +34630524530; e-mail: juanalfonso.lara@udima.es) and David Lizcano are with the Open University of Madrid, UDIMA, Ctra. De la Coruña, km 38.500 – Vía de Servicio, 15 - 28400, Collado Villalba, Madrid, Spain.

[1] Data sourced from http://www.boxofficemojo.com/

Based on the above examples, it is easy to fathom that one of the most important decisions to be taken by film producers is whether or not to agree to produce a film project that they have been offered. This decision is unquestionably as important as it is complex.

A film's profitability depends on many financial, social, commercial and technical factors. In most cases, these issues are out of the reach of producers and will not, therefore, be addressed in this paper. At a time of economic recession, for example, film industry revenue tends to drop considerably, thereby compromising the profitability of a film as receipts are smaller.

In the research reported here, however, we show that in a sizeable percentage of cases it is possible to forecast whether or not a film will be profitable by examining its key features (duration, genre, budget, etc.), even before starting production. We reached this conclusion after conducting a study of over 100 movies produced in 2012 in the United States, which, together with India, is the largest film producer in the world. The study suggests that it is possible to predict, with an accuracy of over 70%, whether or not a particular film will be profitable based on the key features of the movies.

As illustrated throughout the paper, this proposal is more accurate than most other reported research. Apart from its strong predictive power, our research makes a major contribution to the field because, as far as we know, it is the only proposal capable of directly identifying the key issues influencing movie profitability and their relative weight.

In the study we have used data mining techniques, which is a branch of computer science responsible for analysing large quantities of data in search of useful and interesting knowledge. In particular, we have used decision trees, a predictive model widely used in the field of artificial intelligence. The knowledge extracted from this research can be used to develop decision support mechanisms for film producers. A decision support system does not make decisions; rather it is an additional mechanism to help producers decide whether or not to produce a movie. The methods described here are equally extendible to other entertainment industries, like the video game or music business.

The remainder of the paper is as follows: Section II describes other work related to this research. Section III presents the data used and the data conditioning tasks. Section IV then describes the methods used. Section V reports the results and their application. Finally, Section VI outlines the conclusions and future lines of research.

II. RELATED WORK

Large quantities of data are generated and stored in almost all walks of life nowadays. The entertainment industry is no exception, since, as outlined in the introduction, huge quantities of data are generated regarding, for example, film openings and receipts every year. Some aspects related to film industry activity are likely to benefit from the analysis of these huge quantities of data.

The analysis of large volumes of data in order to discover knowledge poses a major challenge in the field of computer science. The extraction of useful, implicit and previously unknown knowledge from large volumes of data is a process called knowledge discovery in databases (KDD). KDD extends from the understanding and preparation of the data to the interpretation and use of the data processing results [2].

Data mining is a stage within the KDD process during which different techniques can be applied to solve a wide range of problems. The problems addressed in data mining include:

- **Classification**. Classification techniques are used to identify to which of the predefined classes a new individual belongs. To do this, a classification model is built from a set of training individuals with some known attributes, including the class to which they belong. The classification model will determine the class of a new unclassified individual from the known value of its attributes. Prominent classification techniques are decision trees (used in this paper and addressed in more depth in Section IV) [3, 4, 5], neural networks [6, 7] and Bayesian classifiers [8].
- **Regression**. The aim of regression techniques is to predict the value (unknown) of an attribute of a particular individual from the values (known) of other attributes of that individual. There are two major regression techniques depending on whether or not the generated regression model is linear.
- **Association rules**. Association techniques aim to find rules that show the relationships between different variables of database records. A common example of this problem type is to identify products that are often purchased together. The *Apriori* algorithm is the best known association technique [9].
- **Clustering**. Clustering techniques aim to divide objects into groups (called clusters) depending on their characteristics and/or behaviour. There are different types of clustering techniques, the most prominent being hierarchical clustering, partitioning clustering, density-based clustering and grid-based clustering [10, 11].

The data to be analysed may have to be cleaned and prepared before the above techniques can be applied. There are many different data cleaning and preprocessing tasks. Two of the most prominent, which are used in this research, are [12]:

- **Feature creation**. This task creates a new data attribute, normally calculated as a function of other existing attributes. For example, if we have the *Gross_Monthly_Salary* attribute and the *Number_of_Pay_Periods* attribute, a new *Gross_Annual_Salary* attribute can be constructed using the following function:

Gross_Annual_Salary
*=Gross_Monthly_Salary*Number_of_Pay_Periods*

- **Discretization**. This task transforms a quantitative attribute into an ordinal qualitative attribute. For example, a person's height in centimetres can be discretized into the intervals tall (≥180 cm), medium (from 150 cm to 180 cm) and short (≤150 cm).

According to the literature review that we conducted, different research approaches have been taken to analyse

TABLE III
FRAGMENT OF THE RAW DATA TABLE USED AS A STARTING POINT FOR THE STUDY

Title	Duration	Genre	Restricted	RealEvents	Remake	Month	Budget	Gross
Life of Pi	126	Adventure	N	N	N	November	120000000	609000000
Journey 2: The Mysterious Island	94	Adventure	N	N	N	February	79000000	335300000
Ice Age: Continental Drift	94	Animation	N	N	N	July	95000000	877200000
Madagascar 3: Europe's Most Wanted	93	Animation	N	N	N	June	145000000	746900000
Dr. Seuss' The Lorax	95	Animation	N	N	N	March	70000000	348800000
Hotel Transylvania	91	Animation	N	N	N	September	85000000	358400000
Brave	100	Animation	N	N	N	June	185000000	539000000
Wreck-It Ralph	108	Animation	N	N	N	November	165000000	471200000
The Pirates! Band of Misfits	88	Animation	N	N	N	April	55000000	123100000
…								

data from the film industry for knowledge discovery has taken different approaches. The most prominent includes research described by Simonoff and Sparrow [13], using regression techniques to predict revenue from movies. More recent work has addressed similar research to the investigation reported in this paper, like, for example, research by Im [14] using a linear gradient descent algorithm to predict whether or not films will be profitable, with a mean accuracy of 72.4%. Another noteworthy study [15] uses neural networks to predict film success in terms of profitability. In this case, the results report a mean accuracy of 72.5%. The results section will discuss these figures compared with the proposal introduced here.

Other proposals in the field of movie profitability prediction are substantially different from the line of research described here and tend to use other unstructured or semi-structured information items. Research reported in [16, 17], based on the analysis of film screenplays using knowledge-based and natural language processing techniques among others, is a prominent example. Approaches like this are, however, based on resources that are mostly not freely available and which are computationally expensive to process as they require an exhaustive analysis.

III. DATA

For the research presented here, we used data from 104 films that opened in 2012 in the United States. We selected 2012 because it was the most recent year for which all the required data were available at the time of writing.

Table III shows a fragment of the raw data. We find that the data compiled about each film included its title, duration (minutes), genre, whether or not it was rated as restricted, whether or not the film is based on a true story, whether or not it is a remake, the opening month, its budget ($) and its gross ($).

This research considered all the variables that in principle provided relevant information and were freely available. Some of the other variables considered initially were omitted after they were found to be of absolutely no relevance. We have endeavoured at all times to use freely available information, thereby leaving the door open for the scientific community to reproduce our proposal in other fields using other data sets covering other time periods. This is a valuable feature of any scientific proposal and is especially important in the data analysis field.

Section III.A details the data collection process, whereas Section III.B details the preprocessing tasks required to condition the data for the construction of decision trees.

A. Data Collection

The data in Table I were retrieved from the information tables posted on the Box Office Mojo website (http://www.boxofficemojo.com/), which publishes information on films, including revenue. This Amazon.com-owned website receives over one million visits a month and has been operational for 15 years.

The information on the duration, genre, opening month, budget and grosses was gathered more or less automatically from the web site. We selected 104 films that opened in the United States in 2012 about which the web site contained all the necessary information. The other 2012 films, which had some missing attribute values, were not taken into account for this research.

The other attributes for each film were added manually from information published on the FilmAffinity web site (http://www.filmaffinity.com), a web site set up in 2002 which contains an exhaustive database of films opening all over the world.

B. Data Preprocessing

A series of data preprocessing tasks were performed on the original data table (Table III).

The first preprocessing task was to calculate profitability from the *Budget* and *Gross* attributes, as specified in Equation (1). The calculated profitability was discretized into two intervals: *POSITIVE* (profitability > 0) and *NEGATIVE* (profitability <= 0), resulting in the new variable *ProfitabilityBin*.

As decision trees are models especially designed to work with discrete data, the Duration and Budget attributes were also discretized according to the rules shown in Table IV, resulting in two new attributes: *DurationD* and *BudgetD*.

TABLE IV
DETAILS ON THE DISCRETIZATION OF INDEPENDENT VARIABLES

Attribute	Discretized value	Condition
DurationD	SHORT	Duration <= 90
	MEDIUM	90 < Duration <= 120
	LONG	Duration > 120
BudgetD		
	LOW	Budget <= 20000000
	MEDIUM	20000000 < Budget <= 80000000
	HIGH	80000000 < Budget <= 150000000
	VERY HIGH	Budget > 150000000

The above preprocessing tasks were applied to produce

TABLE V
FRAGMENT OF THE CLEAN DATA TABLE USED IN THE STUDY

Title	DurationD	Month	Genre	BudgetD	Restricted	RealEvents	Remake	ProfitabilityBin
Life of Pi	LONG	November	Adventure	HIGH	N	N	N	POSITIVE
Journey 2: The Mysterious Island	MEDIUM	February	Adventure	MEDIUM	N	N	N	POSITIVE
Ice Age: Continental Drift	MEDIUM	July	Animation	HIGH	N	N	N	POSITIVE
Madagascar 3: Europe's Most Wanted	MEDIUM	June	Animation	HIGH	N	N	N	POSITIVE
Dr. Seuss' The Lorax	MEDIUM	March	Animation	MEDIUM	N	N	N	POSITIVE
Hotel Transylvania	MEDIUM	September	Animation	HIGH	N	N	N	POSITIVE
Brave	MEDIUM	June	Animation	VERY HIGH	N	N	N	POSITIVE
Wreck-It Ralph	MEDIUM	November	Animation	VERY HIGH	N	N	N	POSITIVE
The Pirates! Band of Misfits	SHORT	April	Animation	MEDIUM	N	N	N	POSITIVE
…								

TABLE VI
SUMMARY OF ATTRIBUTES USED IN THE STUDY

	Attribute	Meaning	Domain
Independent variables			
	DurationD	Film duration	{SHORT, MEDIUM, LONG}
	Month	Opening month	{JANUARY, FEBRUARY, …, DECEMBER}
	Genre	Film genre	{ADVENTURE, ACTION, SCI-FI, …}
	BudgetD	Film budget	{LOW, MEDIUM, HIGH, VERY HIGH}
	Restricted	Whether or not the film is rated as *Restricted*	{Y(es), N(o)}
	RealEvents	Whether or not the film is based on a true story	{Y(es), N(o)}
	Remake	Whether or not the film is a remake of an earlier film	{Y(es), N(o)}
Dependent variable			
	ProfitabilityBin	Whether or not the film is profitable	{POSITIVE, NEGATIVE}

the final minable data table, an extract of which is shown in Table V. In this table, *ProfitabilityBin* is the variable to be predicted or explained (dependent variable) and the other variables, save the title which is unimportant for the analysis, are the predictor or explanatory variables (independent variables).

Table VI summarizes all the attributes considered in the analysis, their meaning and their values.

One last point to be considered is the number of cases covered by the 104 films in each of the two classes considered for the dependent variable. Table VII illustrates this question.

TABLE VII
DEPENDENT VARIABLE

ProfitabilityBin		
POSITIVE (*Profitability* > 0)	89	85.58%
NEGATIVE (*Profitability* <= 0)	15	14.42%
Total	104	100.00%

IV. APPLIED METHODS

Classification is a predictive data mining task. One of the alternatives for performing this task is to use decision trees.

Decision trees are tree-shaped structures that are used as predictive models in many different areas [18]. To do this, the value of the known attributes of the object is used to move down through the tree (each tree node contains a condition on those known attribute values, which determines the branch to be taken) to a leaf node. The leaf node specifies the class within which the object has been classified.

In decision trees, the nodes represent the test on an attribute, the branches represent the value of the test performed in the node from which they branch off and the leaf nodes represent the class labels.

There are many decision tree building algorithms. Some of the best known are Id3, C4.5, C5.0, CHAID and CART, CART being the algorithm used in this research [19]. Each algorithm has certain particularities, although they all adhere to a similar iterative procedure:

1. Assign all the elements of the training set to the tree root.

2. Divide the classification tree according to a particular heuristic.

3. Repeat step 2 until the leaf nodes are reached.

4. Finally, prune the tree if necessary to remove branches that represent noise.

Normally, the heuristic used to build the tree (step 2) involves selecting the attribute that provides the biggest *information gain* at each node.

In order to explain this concept, suppose that we have two classes, P and N, and a set of examples S that contains p elements of the class P and n elements of the class N. In that case, the *amount of information* required to decide whether any object of S belongs to P or N is defined as specified in Equation (2).

$$I(p,n) = -\frac{p}{p+n}log_2\frac{p}{p+n} - \frac{n}{p+n}log_2\frac{n}{p+n} \quad (2)$$

Suppose also that we use an attribute A in a particular tree node, and the set S is divided into subsets $\{S_1, S_2,\dots, S_v\}$. In that case, if S_i contains p_i examples of P and n_i examples of N, then the *entropy*, or the information necessary to classify objects into either of the subtrees S_i, is calculated using Equation (3).

$$E(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i,n_i) \quad (3)$$

Finally, the *information gain* in the event of using attribute A is given by Equation (4). This value measures the discriminatory capability of the attribute in question considering the different problem *classes*.

$$Gain(A) = I(p, n) - E(A) \qquad (4)$$

For example, Figure 1 shows a decision tree for predicting whether new company customers will (class *Y*) or will not (class *N*) buy a particular product from their data (*Age*, *Student Status*, *Income Level*). We find that subjects of medium age (central branch) will buy; younger subjects (left branch) will only buy if they are students; finally, older subjects (right branch) will only buy if their income is excellent.

For this research we have used Salford Systems' CART (http://www.salfordsystems.com).

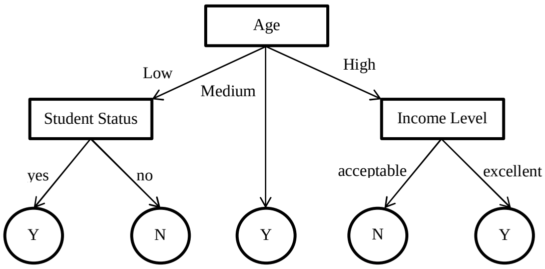

Fig. 1. Example of a decision tree to decide whether (*Y*) or not (*N*) customers will buy a product based on their data (*Age*, *Student Status* and *Income Level*).

V. Results

We used techniques based on predictive data mining models, in particular decision trees, in an attempt to explain the behaviour of the dependent variable. To do this, we used CART (a tool which implements the algorithm of the same name) to build a decision tree from the prepared data. The result was a classification tree with five intermediate and six terminal nodes, as shown in Figure 2.

Analysing the tree from the root to the leaves, the tree determines that films opening in April, December, July and May are usually more profitable. The profitability of films of the *Animation*, *Documentary*, *Fantasy*, *Horror*, *Romance* and *Western* genres is not usually positive in the rest of the year. Analysing the other genres, profitability is usually good where the budget is *HIGH* or *VERY HIGH*. The profitability of *MEDIUM* or *LOW* budget movies tends to be negative unless they are rated as *Restricted*. On the other hand, the profitability of films rated Restricted is usually positive in the months of *August*, *February*, *January*, *March*, *November* and *October*.

The resulting tree model was validated using the cross validation technique, setting aside 10% of the data for testing and building the model with the remaining 90%. This process was repeated 10 times, and the results are shown in Table VIII, which shows that the overall predictive accuracy was 72.66 %. We had to source the test set from the database of historical films containing the respective information. We could not use films that have not yet been premiered or are still being shown (because the data are inconclusive). Note, however, that the resulting model would be applied to films that have not yet opened in order to predict whether or not they will be profitable.

TABLE VIII
CLASSIFICATION ACCURACY OF THE DECISION TREE MODEL

Actual Class	Predicted Class			
	Positive	Negative	Total	%Correct
Positive	70	19	89	78.65%
Negative	5	10	15	66.67%
Total	75	29		
Average				72.66%

Although other similar proposals operate predominately with quantitative data and are not immediately evaluable with our data (and vice versa), we can compare our proposal against previous models developed by other authors in terms of accuracy. In this respect, our proposal slightly outperforms models developed by Im [14] (accuracy of 72.66% vs 72.4%). Our results come close to the results reported by Sharda and Denle [15], using neural networks (accuracy of 72.66% vs 75.2%). Sharda and Denle also reported accuracy figures for other approaches (regression, discriminant analysis and other decision trees), all of which are worse than for our proposal. Table IX summarizes the figures of the comparison.

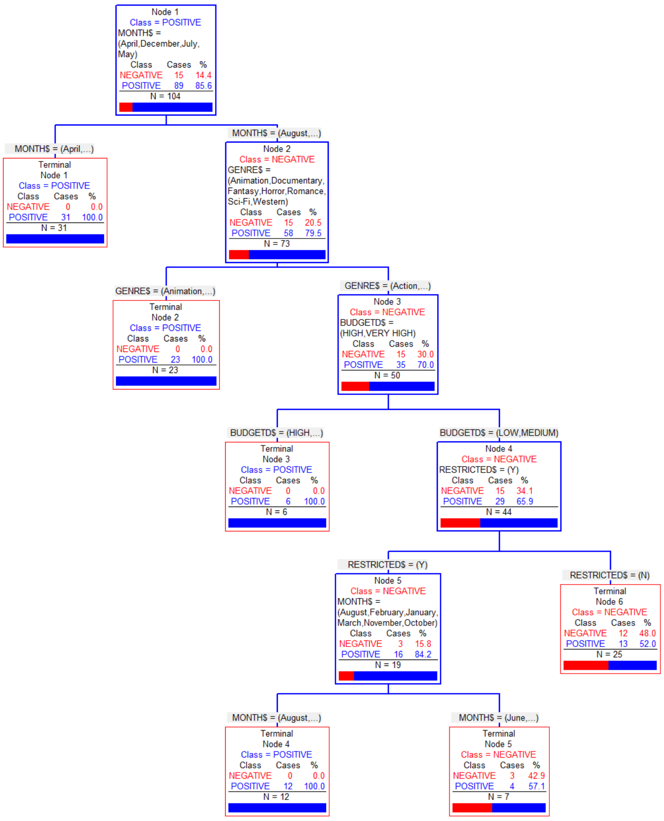

Fig. 2. Decision tree built from movie data.

On top of the relatively good results compared with other proposals, our proposal has, unlike other models based on neural networks for example, the advantage of being easily interpretable because it is founded on the CART algorithm.

Additionally, our study identifies the importance of each

TABLE IX
COMPARISON OF THE ACCURACY OF DIFFERENT METHODS FOR PREDICTING PROFITABILITY FROM FILM INDUSTRY DATA

	(Im, 2011)	(Sharda and Denle, 2006)			
Our proposal	*Linear gradient*	*Neural Network*	*Discriminant Analysis*	*Regression Trees*	*Regression*
72.66%	72.4%	75.2%	67.9%	71.1%	69.6%

predictor variable. In particular, the CART tool outputs an ordered list specifying importance, as shown in Table 10. The best predictor variable (in this case *Month*) always receives a score of 100, whereas the other predictor variables receive a relative score depending on their predictive capability. Table X shows that, apart from opening month, film genre and budget are very important. The other variables appear to have a much smaller share in explaining the dependent variable.

Let us pick a film at random in order to illustrate the usefulness of the resulting decision tree, for example, ‘Man of Steel’. This is an action film, which is neither a remake nor restricted, whose duration is *LONG* and whose budget is *VERY HIGH*. This film opened in June 2013. All this information was known (or decidable) before the film was produced. Therefore, the resulting decision tree could be applied to predict whether or not the film would be profitable, which would be of interest to potential producers. Traversing the tree using these data from the root node down, the first branch on the right (*June*) should be chosen. At the next node, the right branch (*Action* genre) should be taken. At the third node, the left branch should be chosen (*VERY HIGH* budget), leading to the leaf node labelled *POSITIVE*. According to the model, the profitability of this film would be positive, as it really was (it earned 291 million dollars in the three months that it was running compared with the 225 million dollars that it cost to produce).

TABLE X
IMPORTANCE OF VARIABLES TO PREDICT PROFITABILITY

Rank	Variable	Score
1	*Month*	100.00
2	*Genre*	98.72
3	*BudgetD*	70.48
4	*Restricted*	16.80
5	*DurationD*	4.04
6	*Remake*	1.96

VI. CONCLUSIONS AND FUTURE WORK

Filmmaking is a major branch of the entertainment industry, in which producers play a key role, as it is they who finance blockbusters. Investment in a film is, however, no guarantee of its profitability. The many cases throughout cinema history of movies that have not managed to cover expenses stand as proof of this point. In this respect, film production is a complex issue that carries a sizeable financial risk. In this scenario, tools capable of predicting whether or not a film will be profitable can be of a lot of help to producers. Such decision support tools can round out the producers’ foresight and help them to opt for projects that have more chances of making a profit.

In this paper, we have conducted a study using historical data of US movies. We used these data to build a decision tree to predict whether or not a movie will be profitable. The resulting model was evaluated on a real set of test data, and accuracy was found to be 72.6%. This model is as or more accurate than other previous approaches and has the advantage of being easily interpretable by experts who can also discover how important each analysed feature is.

The proposed approach is equally applicable to other entertainment industries and can be used with other attributes depending on the case. In fact, one of the intended future lines of research is to apply decision trees in other fields. To do this, we would have to identify the features of the elements to be analysed (for example, video games), their production expenses and revenue, and run a similar study to try to identify predictive patterns of positive profitability.

Most of the data had to be collected manually, which was a tiresome process, due partly to the fact that there is no film data repository to encourage other researchers to examine this domain. A centralized and organized repository of open film industry data would be a major advance in this sector.

As the last line of future research, we suggest the possibility of extending the historical film database considered in this research to other years. This should produce more representative models.

ACKNOWLEDGMENT

The authors would like to thank Rachel Elliott for translating this manuscript and Marisa Gómez for her help in collecting data**.**

REFERENCES

[1] Pascal, B., Pensées, Baltimore-Penguin Books, 1662.

[2] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., From Data Mining To Knowledge Discovery: An Overview. In Advances In Knowledge Discovery And Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., pp. 1-34, 1996.

[3] Breiman, L., Friedman, J. H., Olshen, R., Stone, C. J.: *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California, 1984.

[4] Quinlan, J. R.: *Induction of Decision Trees*, Machine Learning, 1(1), pp. 81-106, 1986.

[5] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[6] Haykin, S.: *Neural Networks*, Prentice-Hall, 1999.

[7] Freeman, J. A., Shapura, D. M.: *Neural Networks Algorithms*, Addison-Wesley, 1991.

[8] Phyu, T. N., Survey of Classification Techniques in Data Mining. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, 2009.

[9] Agrawal, R., Srikant, R., Fast algorithms for mining association rules, International Conference on Very Large Databases, pp. 487-499, Santiago, Chile, 1994.

[10] MacQueen, J. B.: *Some Methods for classification and Analysis of Multivariate Observations*. Proceeding of the 5-th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1, pp. 281-297, 1987.

[11] Kaufman, L, y Rousseeuw, P. J.: *Clustering by means of medoids*. Y. Dodge Ed., Statistical Data Analysis based on the L1 Norm, pp. 405-416, 1987.

[12] Lara, J. A., Manual de Minería de Datos, Ed. Udima, 2013.

[13] Simonoff, J. S., Sparrow, I. R., Predicting movie grosses: winners and losers, blockbusters and sleepers. Chance, 13(3), 2000.

[14] Im, D., Nguyen, M. T., Predicting box-office success of movies in the U.S. market, CS 229, Univ. Stanford, 2011.

[15] Sharda, R., Delen, D., Predicting box-office success of motion pictures with neural networks, Expert Systems with Applications, 30, pp. 243-254, 2006.

[16] Eliashberg, J., Hui, S. K., Zhang, Z. J., From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts, Management Science, 53(6), pp. 881-893, 2007.

[17] Eliashberg, J., Hui, S. K., Zhang, Z. J., Assessing Box Office Performance Using Movie Scripts: A Kernel-based Approach, IEEE Transactions on Knowledge and Data Engineering, 2014.

[18] Huo, X., Kim, S. B., Tsui, K.-L., & Wang, S. A frontier-based tree pruning algorithm (FBP). INFORMS Journal on Computing, 18, 494–505, 2006.

[19] Hastie, T., Tibshirani, R., & Friedman, J. The element of statistical learning. New York, NY: Springer, 2001.