

PREDICTING THE FINANCIAL SUCCESS OF HOLLYWOOD MOVIES USING AN INFORMATION FUSION APPROACH

Dursun DELEN and Ramesh SHARDA*

Spears School of Business, Oklahoma State University, USA

*Received: 24 July 2008; Accepted: 18 December 2009
This article was with the author 38 days for 2 revisions.*

ABSTRACT

Hollywood has often been called the land of hunches and wild guesses. The uncertainty associated with the predictability of product demand makes the movie business a risky endeavor. Therefore, predicting the box-office receipts of a particular motion picture has intrigued many scholars and industry leaders as a difficult and challenging problem. In this study, with a rather large and feature rich dataset, we explored the use of data mining methods (e.g., artificial neural networks, decision trees and support vector machines along with information fusion based ensembles) to predict the financial performance of a movie at the box-office before its theatrical release. In our prediction models, we have converted the forecasting problem into a classification problem—rather than forecasting the point estimate of box-office receipts; we classified a movie (based on its box-office receipts) into nine categories, ranging from a “flop” to a “blockbuster.” Herein we present our exciting prediction results where we compared individual models to those of the ensembles.

Keywords: Classification, movie business, box-office receipts, neural networks, decision trees, ensembles, performance measures.

BİLGİ BİRLEŞTİRME YÖNTEMİNİ KULLANARAK HOLLYWOOD SİNEMA FİLMLERİNİN FİNANSAL BAŞARISINI KESTİRMEK

ÖZET

Genellikle Hollywood (Amerika'nın sinema endüstrisinin kalbi) sezgi ve tahminler diyarı olarak adlandırılır. Ürün talebine yönelik kestirimlerdeki belirsizlik film endüstrisini risklerle dolu bir iş alanı haline getirdi. Bu yüzden bir filmin finansal başarısını kestirebilmek pek çok bilim adamı ve endüstri liderlerinin ilgisini çeken zor ve büyüleyici bir problem olarak tanımlandı. Bu çalışmada, büyük bir veri tabanını kullanarak, veri madenciliği yöntemlerinin (yapay beyin ağları, karar ağaçları, destek vektör algoritmaları ve bilgi birleştirme yöntemleri) henüz yapımına başlanmamış bir film projesinin olası finansal getirisini kestirmedeki başarı potansiyelini araştırdık. Kestirim modellerimizde, bu sıradan tahmin problemini sınıflandırma temelli bir kestirim problemine dönüştürdük; yani kesin dolar (\$) miktarını tahmin etmek yerine, bir filmin finansal başarısını dokuz sınıftan birine koyan bir kestirim modeli geliştirdik. Bu makalede kullandığımız veri madenciliği modellerinin oldukça başarılı adedilebilecek kestirim sonuçlarını ve model geliştirmede kullandığımız yöntemleri rapor ediyoruz.

Anahtar Kelimeler: Sınıflandırma, film sektörü, gişe gelirleri, sinir ağları, karar ağaçları, topluluk, performans ölçütleri

* İletişim yazarı

1. INTRODUCTION

The uncertainty associated with the predictability of product demand makes the movie business in Hollywood a risky endeavor. Therefore, predicting the box-office success of a particular motion picture has intrigued many scholars and industry leaders as a difficult and challenging problem. To some analysts, “Hollywood is the land of hunch and the wild guess” (Litman and Ahn, 1998) due largely to the difficulty and uncertainty associated with characterizing and predicting the product demand. Such unpredictability of the product demand makes the movie business one of the riskiest endeavors for investors to take in today’s competitive marketplace. In support of such observations, Mr. Jack Valenti, the long-time president and CEO of the Motion Picture Association of America, once mentioned that “... No one can tell you how a movie is going to do in the marketplace... not until the film opens in darkened theatre and sparks fly up between the screen and the audience.” Trade journals and magazines of the motion picture industry have been full of examples, statements, and experiences that support such a claim.

Despite the difficulty associated with the unpredictable nature of the problem domain, several researchers have attempted to develop models for forecasting the financial success of motion pictures, primarily using statistics-based forecasting approaches. Most analysts have tried to predict the total box-office receipt of motion pictures after a movie’s initial theatrical release with some level of success (Sawhney and Eliashberg, 1994). Litman and Ahn (1998) summarized and compared some of these major studies on predicting financial success of motion pictures. Yet, these previous studies leave us with an unsatisfied need for a more accurate forecasting method, especially prior to a movie’s theatrical release. Most studies indicate that box-office receipts tend to tail-off after the opening week. Research shows that 25 percent of total revenue of a motion picture comes from the first two weeks of receipts. Thus, once the first week of box-office receipts are determined, the total box-office receipts of a particular movie can be

forecasted with relatively high accuracy (Sawhney and Eliashberg, 1996). Therefore, the accurate estimate of the box-office receipts of motion pictures before its theatrical release is the most difficult and the most critical to the industry.

In this study, with a rather large and feature rich dataset, we explored the use of data mining methods (e.g., artificial neural networks, decision trees and support vector machines along with information fusion based ensembles) to predict the financial performance of a movie at the box-office before its theatrical release. In our prediction models, we have converted the forecasting problem into a classification problem—rather than forecasting the point estimate of box-office receipts; we classified a movie (based on its box-office receipts) into nine categories, ranging from a “flop” to a “blockbuster.” Herein we present our exciting prediction results where we compared individual models to those of the ensembles. The remainder of this paper is organized as follows. The next section briefly reviews the literature on forecasting the box office success of theatrical movies. Section three provides the details of our methodology by specifically talking about the data, the predictions models, the experimental design and the performance measures used in this study. Next, the prediction results of all data mining models are presented and compared to each other. Finally, the last section of the paper discusses the overall contribution of this study, along with its limitations and further research directions.

2. LITERATURE REVIEW

Literature on forecasting financial success of new motion pictures can be classified based on the type of forecasting model employed: (i) Econometric/Quantitative Models—those that explore factors that influence the box office receipts of newly released movies (Litman 1983, Sochay 1994, Litman and Ahn 1998, and Ravid 1999, Eliashberg et al. 2007) - and (ii) Behavioral Models—those that primarily focus on the individual’s decision-making process with respect to selecting a specific movie from a vast array of entertainment alternatives (Eliashberg and

Sawhney 1994, Sawhney and Eliashberg 1996). These behavioral models usually employ a hierarchical framework, where behavioral traits of consumers are combined with the econometric factors in developing the forecasting models. Another classification is based on the timing of the forecast (Sharda and Delen, 2006; Olson and Delen, 2008): (i) Before the Initial Release—that is, forecasting the financial success of the movies before their initial theatrical release, and (ii) After the Initial Release—that is forecasting the financial success of the movies after their initial theatrical release, when the first week of receipts are known. Forecasting models that fall into the category of “after the initial release” tend to generate more accurate forecasting results due to the fact that those models have more explanatory variables including box-office receipts from the first week of viewership, movie critic reviews, and word-of-mouth effects. Our study falls into the category of quantitative models for model type classification, and into the category of before the initial release in timing of the forecast classification.

From the standpoint of a DSS type of implementation in the motion picture industry, the only notable study comes from Eliashberg et al. (2001). They developed a marketing management support system (called SilverScreener) that helps theatre managers to choose among alternative movies to show in their theatres. Their system is based on a multi-phase mathematical attendance-forecasting model. The model was developed using the historical data from Pathé theatres in Netherlands, along with managerial judgment and theatre-specific factors. Their limited tests on the same theatre indicated that the overall financial success (total revenues) of the theatre increased compared to two other theatres in the same time period. They also reported improved managerial attitudes towards the system once the system started to generate better decisions resulting in higher revenues. However, in contrast to our study, this system was not meant for decision support in pre-production stage.

In a recent study, Eliashberg et al. (2007) proposed a text-mining-based modeling approach to help

studios decide on which movie scripts to pursue for production. Their approach combines screenwriting domain knowledge and natural-language processing techniques to predict a movie’s financial success based only on textual information available in movie scripts. They formulated the problem as a binary classification where the outcome is whether the movie will make profit (1) or not (0). They tested their text-mining-based classification model on a small holdout sample of movies to show that their model is able to predict the outcome better than the random chance (with roughly 60% prediction accuracy).

Predicting the future is a difficult task, especially when the predicted phenomenon is complex and can only be characterized with partial information (Karakan and Koç, 2008). In movie industry, most still believe that the financial success is hidden in the artistic nature of the product. However, given the significant investment in the movie industry, the need for effective and efficient decision support tools and models are inevitable. A general lack of published research that describes successful implementation of these decision models suggests challenges as well as opportunities for researchers in developing practical predictive analytics for this domain. This paper describes a research effort that to be a small step towards filling such gap.

3. METHODOLOGY

3.1 Data and Variable Descriptions

In our study, we used 2632 movies released between 1998 and 2006. The sample data was drawn (and partially purchased) from public and commercial movie databases. The dependent variable in our study is the box-office gross revenues. In this study, following the advice of the industry experts in Hollywood, we classify a movie based on its box-office receipts in one of nine categories, ranging from a “flop” to a “blockbuster.” This process of converting a continuous variable in a limited number of classes is commonly called “discretization” in data mining literature. Many prefer using discrete values as opposed to continuous ones in developing prediction models because discrete

values are closer to a knowledge-level representation. In this study, we discretized the dependent variable into nine classes using the following, somewhat arbitrary, breakpoints (following neither equal-length nor equal-frequency based binning approaches). These ranges are determined based on the information and in-depth advice obtained from several industry experts (decision makers) in Hollywood.

review section), and partially based on the feedback we have obtained from the movie industry experts. As per our experiences with the professionals in Hollywood (including studio executives, producers, investment firm representatives and consultants), we deduced that these variables are known (or roughly estimated) by the decision makers before the start of a movie production process. In essence, these are

Class No.	1	2	3	4	5	6	7	8	9
Range	< 1	> 1	> 10	> 20	> 40	> 65	> 100	> 150	> 200
(in Millions)	(Flop)	< 10	< 20	< 40	< 65	< 100	< 150	< 200	(Blockbuster)

We used a rich set of independent variables. Our choice of independent variables was mostly based on previous studies conducted in the field (see literature

the parameters that they would be most interested in “optimizing”. For instance, the *number of screens* is an aggregate measure of a production (or distribution)

Table 1. Summary of Independent Variables

Independent Variable Name	Definition	Range of Possible Values
MPAA Rating	The rating assigned by the Motion Picture Association of America (MPAA).	G, PG, PG-13, R, NR
Competition	Indicates the level at which each movie competes for the same pool of entertainment dollars against movies released at the same time.	High, Medium, Low
Star value	Signifies the presence of any box office superstars in the cast. A superstar actor/actress can be defined as one who contributes significantly to the up-front sale of the movie.	High, Medium, Low
Genre	Specifies the content category the movie belongs to. Unlike the other categorical variables, a movie can be classified in more than one content category at the same time (e.g., action as well as comedy). Therefore, each content category is represented with a separate binary variable.	Sci-Fi, Epic Drama, Modern Drama, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	Signifies the level of technical content and special effects (animations, sound, visual effects) used in the movie.	High, Medium, Low
Sequel	Specifies whether a movie is a sequel (value of 1) or not (value of 0).	Yes, No
Number of screens	Indicates the number of screens on which the movie is planned to be shown during its initial launch.	A positive integer between 1 and 3876

studio’s marketing and advertising efforts. Prior to the production, the decision makers have a rough idea of what these numbers are going to be for a specific movie project. In fact, it is one of the decision variables that they would like to optimize by conducting sensitivity analysis on its values in order to find the best possible value, because there is a significant cost to commit too many screens and not have enough viewership. The variables and their brief descriptions are given in Table 1.

3.2 Prediction Models

In this section we provide a brief description of each of the prediction models (including the ensemble methods) we used in this study.

Neural Networks: Neural networks are commonly known as biologically inspired analytical techniques, capable of modeling extremely complex

non-linear functions (Hykin, 1998). We used a popular neural network architecture called Multi-Layer Perceptron (MLP) with back-propagation algorithm. MLP is essentially the collection of nonlinear neurons organized and connected to each other in a feed-forward multi-layer structure. MLP is known to be a strong function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied NN architecture. Hornik et al. (1990) empirically showed that given the right size and the structure, MLP is capable of learning arbitrarily complex nonlinear functions to an arbitrary accuracy level. Figure 1 shows the graphical representation of the MLP used in this study.

Decision trees: As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy (Briemen et al, 1984). In doing so,

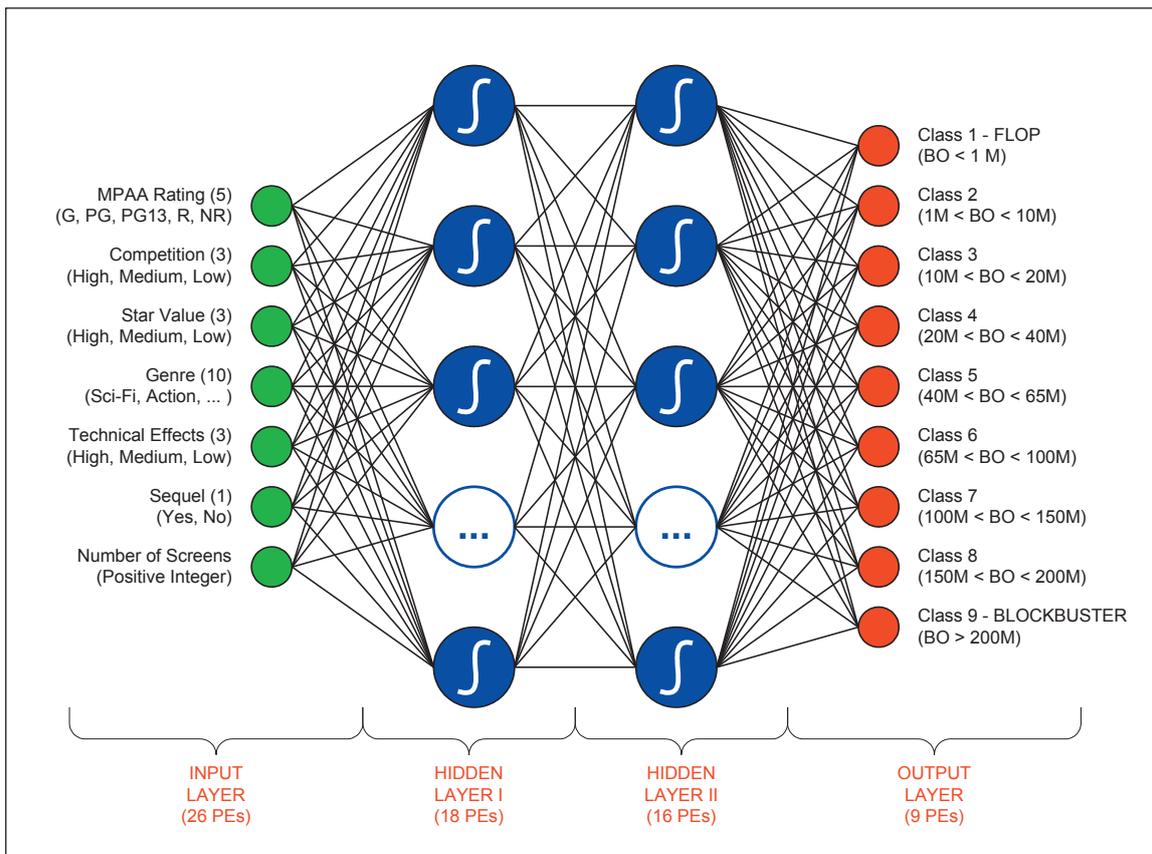


Figure 1. Graphical Representation of our MLP Neural Network Model

different mathematical algorithms (e.g., information gain, Gini index, etc.) are used to identify a variable and the corresponding threshold for the variable that splits the pool of observations into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. Popular decision tree algorithms include ID3, C4.5, C5, CART, and CHAID.

Support Vector Machines (SVM): Support vector machines (SVMs) belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. The mapping function in SVMs can be either a classification function (used to categorize the data, as is the case in this study) or a regression function (used to estimate the numerical value of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be (Cristianini, 2000).

Even though SVMs and ANNs are often assumed to belong to the same family of methods, there are some significant differences between the two. For instance, the development of ANNs followed a heuristic path, with applications and extensive experimentation preceding the theory. In contrast, the development of SVMs involved sound theory first, then implementation and experiments. A significant advantage of SVMs is that whilst ANNs can suffer from multiple local minima, the solution to an SVM is global and unique (Drucker et al., 1997). Two more advantages of SVMs are that they have a simple geometric interpretation and give a sparse solution.

Unlike ANNs, the computational complexity of SVMs does not depend on the dimensionality of the input space. ANNs use empirical risk minimization, whilst SVMs use structural risk minimization. The reason that SVMs often outperform ANNs in practice is that they are less prone to over fitting.

Ensembles / Bagging (Random Forest): A random forest is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. In essence, a random forest consists of a collection (ensemble) of deceptively simple decision trees, each capable of producing a response when presented with a set of predictor values. Random forest has shown to run very efficiently on large datasets with large number of variables. The algorithm for inducing a random forest was first developed by Breiman (2001).

Ensembles / Boosted Trees: The general idea of boosted trees is to compute a sequence of very simple trees, where each successive tree is built for the prediction residuals of the preceding tree. In essence, it learns from the previous tree, in order to construct the succeeding one so that the misclassification of cases is minimized. Detailed technical descriptions of this methods can be found in Hastie et al (2001).

Ensembles / Information Fusion: Information fusion is the process of “intelligently” combining the information (predictions in this case) provided by two or more information sources (i.e., prediction models). While there is an ongoing debate about the sophistication level of the fusion methods, there is a general consensus that fusion (combining predictions) produces more accurate and more robust prediction results (Elder, 2003).

3.4 Performance Metrics

We used percent success rate to measure the predictive performance of the data mining methods. The percent success rate (a.k.a. average percent hit rate or APHR) is arguably the most intuitive measure for predictive accuracy. It is the ratio of total correct classifications to total number of samples, averaged for all classes in the classification problem. In our case, we have two different success rates: bingo (which

measures the exact classification into the same class and the within one class) and 1-Away (which includes the neighboring classes as success). Algebraically, APHR can be formulated as follows:

$$APHR_{Bingo} = \frac{1}{n} \sum_{i=1}^g p_i \tag{1}$$

$$APHR_{1-Away} = \frac{1}{n} \left((p_1 + p_2) + \sum_{i=2}^{g-1} p_{i-1} + p_i + p_{i+1} + (p_{g-1} + p_g) \right) \tag{2}$$

where, g is the total number of classes ($= 9$), n is the total number of samples ($= 2632$), and p_i is the total number of samples classified as class i .

4. RESULTS

Table 2 shows the prediction results of all three data mining methods as well as the results of the three different types of ensembles. These results are obtained from the holdout sample (the sample movies that are not used in the training or validation procedures): the movies released in 2006. That is, all of the models are built using the movies from 1998-to-2005

fuse the reader, these training results (which has no significance on judging the true classification accuracy of the models) are omitted from Table 2.

The first performance measure is the percent cor-

rect classification rate, which we have called “bingo”. We also report the 1-Away correct classification rate. As can be seen SVM performed the best among the individual prediction models, followed by ANN, and worst of the three was found to be CART decision tree algorithm. In general, the ensemble models performed better than the individual predictions models, of which fusion algorithm performed the best. What is probably more important to decision makers, and standing out in the results table, is the significantly low standard deviation obtained from the ensembles compared to the individual models.

Table 2. Classification Accuracy

Performance Measure	Prediction Models and Methods					
	SVM	ANN	C&RT	Random Forest	Boosted Tree	FUSION (Average)
Count (Bingo)	192	182	140	189	187	194
Count (1-Away)	104	120	126	121	104	120
Accuracy (% Bingo)	55.49%	52.60%	40.46%	54.62%	54.05%	56.07%
Accuracy (% 1-Away)	85.55%	87.28%	76.88%	89.60%	84.10%	90.75%

dataset and these models are then tested on the movies from 2006. In Table 2 only the test dataset results are reported. The training dataset results were slightly better than the test results, but since machine learning methods are prone to overfitting (a phenomenon where the prediction results for the training dataset appear to be significantly higher than the test results due to overtraining/overfitting), in order not to con-

5. CONCLUSION

To the best of our knowledge these prediction results are better than any reported in the published literature for this problem domain. Beyond the attractive accuracy of our prediction results of box-office receipts, models could also be used to forecast the success rates of other media products. The particular parameters used within the model of a movie or

other media products could be altered using the already trained prediction models in order to better understand the impact of different parameters on the end results. During this experimentation process, the decision maker of a given entertainment firm could find out, with a fairly high accuracy level, how much a specific actor, a specific release date, or the addition of more technical effects, could mean to the financial success of a film, or a television program. The deployment of these prediction models is achieved using a Web based decision support system architecture Delen et al. (2007).

The accuracy of the data mining models presented in this study can be improved by adding some of the other determinant variables such as production budget and advertising budget, which are known to be industry secrets and are not publicly released. Another method to improve the predictive accuracy of a system is through more sophisticated ensemble models (combining multiple classifiers into a single predictive model by considering their historical accuracy levels).

From the usability perspective, once developed into a production system, such a prediction system can be made available (via a Web-based application server) to industrial decision makers (Hollywood managers), where individual users can plug in their own parameters to forecast the potential success of a movie before its theatrical release. Using the sensitivity analysis feature of the system, decision makers can also “optimize/improve” the parameter values. Such a user friendly Web-based software system is under development (and close to completion and deployment) at the IRIS labs at Oklahoma State University under the guidance of this paper’s authors. Continuing algorithmic work on this prediction system is primarily dedicated to design the system in a way that it can calibrate its parameters (continuous self learning) by taking into account new samples (movies that are released and determined box-office receipts) as they become available.

REFERENCES

- Breiman, L. 2001. Random Forests. *Machine Learning* 45 (1), 5-32.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*. Wadsworth & Brooks Books & Software. Monterey, CA.
- Cristianini, N. and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press: London.
- Delen, D., Sharda, R. and Kumar, P. 2007. Movie Forecast Guru: A Web-based DSS for Hollywood Managers, *Decision Support Systems* 43(4), 1151-1170.
- Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A. and Vapnik, V. 1997. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems*, MIT Press, 9, 155-161.
- Elder, J.F. 2003. The Generalization Paradox of Ensembles. *Journal of Computational and Graphical Statistics*, 12(4), 853-864.
- Eliashberg, J., Hui, S.K. and Zhang, Z.J. 2007. From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts, *Management Science*, 53(6), 881-893.
- Eliashberg, J., Swami, S., Weinberg, C.B. and Wierenga, B. 2001. Implementing and Evaluating SilverScreener: A Marketing Management Support System for Movie Exhibitors, *Interfaces*, 31(3), 108-127.
- Eliashberg, J. and Sawhney, M.S. 1994. Modeling Goes to Hollywood: Predicting Individual Differences in Movie Enjoyment, *Management Science* 40(9), 1151-1173.
- Hastie, T., Tibshirani, R. and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Publishing, New York, NY.
- Hykin, S. 1998. *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall.
- Hornik, K., M. Stinchcombe and H. White, 1990, Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Network, *Neural Networks*, 3, 359-366.
- Karakan, G. and Koç, T. 2008. Technology Forecasting Methods and an Application to White Goods Sector. *Industrial Engineering Journal (Endüstri Mühendisliği Dergisi)*, 20(1), 20-38.
- Litman B.R. and Ahn, H. 1998. Predicting Financial Success of Motion Pictures. In *The Motion Picture Mega-Industry* by B. Litman. Allyn & Bacon Publishing, Boston, MA.
- Litman, B.R. 1983. Predicting Success of Theatrical Movies: An Empirical Study, *Journal of Popular Culture*, 16(9), 159-175.
- Olson, D. and Delen, D. (2008). *Advanced Data Mining Techniques*. New York, NY: Springer.
- Ravid, S.A. 1999. Information, Blockbusters, and Stars: A Study of the Film Industry, *Journal of Business*, 72(4), 463-492.
- Sawhney, M.S. and Eliashberg, J. 1996. A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures, *Marketing Science*, 15(2), 113-131.
- Sharda, R. and Delen, D. 2006. Predicting Box-office Success of Motion Pictures With Neural Networks. *Expert Systems with Applications* 30, 243-254.
- Sochay, S. 1994. Predicting the performance of motion pictures, *The Journal of Media Economics* 7(4), 1-20.