

Forecasting Box-Office Receipts of Motion Pictures Using Neural Networks

Ramesh Sharda* and Dursun Delen
Department of Management Science and Information Systems
College of Business Administration
Oklahoma State University
Stillwater, OK 74078

October 2, 2002

* Corresponding Author (sharda@okstate.edu)

Forecasting Box-Office Receipts of Motion Pictures Using Neural Networks

Abstract

Forecasting box-office receipts of a particular motion picture has intrigued many scholars and industry leaders as a difficult and challenging problem. In this study, we explore the use of neural networks in forecasting the financial performance of a movie at the box-office before its theatrical release. In our model, we convert the forecasting problem into a classification problem—rather than forecasting the point estimate of box-office receipts, we classify a movie based on its box-office receipts in one of nine categories, ranging from a “flop” to a “blockbuster.” Because our model is designed to predict the financial success of a movie before its theatrical release, it can be used as a powerful decision aid by studios, distributors, and exhibitors. We present our exciting prediction results using three different performance measures: average percent success rate, improvement over random sampling, and similarity to perfect classification. Using sensitivity analysis we also present an evaluation of the decision variables and their impact on the box-office success.

Keywords: Forecasting, Classification, Motion Pictures, Box-office Receipts, Neural Networks, Performance Measures, Sensitivity Analysis

1. Introduction

Forecasting box-office receipts 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 1996] due largely to the difficulty and uncertainty associated with 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 industrial world. In support of such observations, Jack Valenti, 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” [Valenti 1978]. 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. However, most [Litman 1983, Sawhney and Eliashberg 1996] did not get sufficiently accurate results for decision support. Litman [1997] summarizes and compares some of the 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 [Litman 1997]. Thus, once the first week of box-office receipts are determined, the total box-office receipts of a particular movie can be forecasted with very 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 our study, we explore the use of neural networks in forecasting the financial performance of a movie at the box-office before its theatrical release. In our model, we convert the forecasting problem into a classification problem. Rather than forecasting the point estimate of box-office receipts, we classify a movie based on its box-office receipts in one of nine categories, ranging from a “flop” to a “blockbuster.”

Neural networks are known to be biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. For many years linear modeling has been the commonly used technique in capturing and representing functional relationships between dependent and independent variables, largely because of its well-known statistically explainable optimization strategies. In the problem scenarios where the linear approximation of a function was not valid (which was frequently the case) the models suffered accordingly. Now, such cases can easily be modeled with neural networks. Simply put, *neural networks* are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data [Haykin 1998]. Applications of neural networks have been reported in many diverse fields addressing problems in areas such as prediction, classification, and clustering. A classical reference for the fundamentals of neural networks is Rumelhart and McClelland [1986]. Many application bibliographies exist [Sharda 1994]. However, none of these include an application of the neural networks in forecasting box-office success of theatrical movies. This paper is one of the first to attempt the use of neural networks for addressing this challenging problem.

The other major contribution of this paper is the introduction of additional performance measures to assess a neural network’s forecasting ability. In addition to the most commonly used statistical performance measure, average percent success rate (a.k.a. hit rate), we borrowed some performance measures from psychology literature to better explain the value added by our neural network-based classification model. We think that

such measures should be used in determining the performance superiority of any classification technique. Accordingly, we used percent success rate, improvement over random sampling, τ (tau), and similarity to the perfect classification, R_g , to measure the predictive performance of our neural network approach. The average percent hit rate (APHR) is arguably the most intuitive measure of discrimination for predictive accuracy of classification problems. It is the ratio of total correct classifications to total number of samples, averaged for all classes in the classification problem. Intuitively, the bigger values of APHR should indicate better classification performance. One should, however, judge the magnitude of this percentage in relation to the expected percentage of correct classification if the assignment were made randomly. Therefore, the prediction accuracy should be judged with respect to the number of classes presented in the classification problem. A relatively smaller value of APHR may indicate a reasonably good performance when the number of classes is large, and vice versa. One way to measure such prediction performance is by using a proportional reduction in error statistic (τ), which measures the percent improvement over the random assignment of samples to the classes [Klecka 1980]. Another way in which we can objectively measure the accuracy of a classification algorithm is to compare its results with the perfect classification (R_g), which is simply the prior probabilities of each classes calculated from the available sample data. If the two sets of results are significantly *similar* to each other, then we can conclude that the classification algorithm under investigation is performing appropriately [Rand 1971].

The remainder of this paper is organized as follows. The next section reviews the literature on forecasting the box office success of theatrical movies. Section three gives the details of our methodology by specifically talking about the data, the neural network model, the experimental design and the performance measures used in this study. Next, the experimental results are shown and explained. 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, Litman and Kohl 1989, Sochay 1994, Litman and Ahn 1998, Neelamegham and Chintagunta 1999, Ravid 1999, Elberse and Eliashberg 2002], and (ii) Behavioral Models—those that primarily focuses 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, Zufryden 1996, De Silva 1998, Eliashberg et al. 2000]. These behavioral models usually employ a hierarchical framework where behavioral traits of consumers are combined (mostly in a sequential process) with the econometric factors in developing the forecasting models. Another classification is based on the timing of the forecast: (i) Before the Initial Release—that is forecasting the financial success of the movies before their initial theatrical release [Litman 1983, Litman and Kohl 1989, Sochay 1994, Zufryden 1996, De Silva 1998, Eliashberg et al. 2000], (ii) After the Initial Release—that is forecasting the financial success of the movies after their initial theatrical release where the first week of receipts are known [Sawhney and Eliashberg 1996, Ravid 1999, Neelamegham and Chintagunta 1999]. 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 critics, 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. Following is a chronological review of the most relevant and the most cited literature published in the field of forecasting financial success of theatrical movies.

One of the earlier and arguably the most referenced studies in predicting financial success of theatrical movies was reported by Litman [1983]. Hypothesizing on the predictability of the box-office revenues, Litman developed his theory of motion picture success as

dependent on three decision-making areas: the creative sphere (story, cast, director, production budget and MPAA rating); the scheduling and release pattern (release time, competition, number of theatres, and distributor strength); and the marketing effort (advertising budget, critics' reviews, and movie awards). Litman used a multiple regression analysis to generate, what he calls, the "revenue equation" that determines the functional relationship between the above-mentioned independent variables and the gross revenue of a movie. He used a sample of 125 movies covering the release period of 1972 to 1978. Several early computer runs were needed to eliminate those independent variables, which had no statistically significant impact on the dependent variable, gross theatrical revenues. The regression analysis on the remaining independent variables (adjusted production cost, critics' ratings, science-fiction, major distributors, Christmas release, academy award nominee, and academy award winner) generated a statistical fit (R^2) value of 0.485 indicating that nearly half of the variance of the dependent variable is explained by the independent variables included in the model. Based on these results, he claims that the ingredients of a successful movie are not as mysterious and unpredictable as claimed by industry leaders. In this study, Litman did not report the predictive performance of his regression model. Therefore, we have no base to compare our predictive performance results to those of his. In later studies, he continued to use linear regression modeling with more comprehensive explanatory variables [Litman and Kohl 1989, Litman and Ahn 1998].

Eliashberg and Sawhney [1994] looked at the prediction problem from a consumer behavior perspective. They developed a conceptual framework to model the dynamics of hedonic consumption experiences in terms of the determinants of enjoyment, and applied the framework to predict individual differences in the enjoyment of a movie. Through formalization and application of the conceptual framework, they demonstrated the usefulness of the modeling methodology in predicting individual differences in enjoyment of movie experiences, upon which valuable managerial implications can be drawn, such as segmentation and targeting for a given movie. Though some empirical test results are presented for enjoyment prediction of a specific movie, they also did not report on the predictive performance of their model for a large sample size of movies.

Zufryden [1996] proposed a hierarchical behavioral model to evaluate financial success of new movies. This three-step linear model structure sequentially links planned advertising expenditures for a new film introduction to awareness, intention to see the film, and projected ticket sales at the box office. He used a collection of 63 movies (from the French movie market covering a six-month period—from January through June 1993) to develop and empirically evaluate his model. He differentiates his study from those of others by stressing that he had utilized the hierarchical three-step prediction model where he incorporated consumer traits with distribution intensity, advertising expenditures, and film characteristics. Two points worth mentioning in regard to his study are that: (1) he did not provide predictive accuracy of his model on a test dataset, and (2) advertising and production budgets, which were the cornerstone of his study are mostly proprietary and therefore, very hard to obtain.

Sawhney and Eliashberg [1996] also developed a model, which they call adaptive forecasting schema, for forecasting the gross box-office revenues of new motion pictures based on their early box-office data. Drawing upon a queuing theory framework, they stochastically conceptualized the customer's movie adoption process in two consecutive steps—the *time to decide* to see the new movie, and the *time to act* on the adoption decision. They used box-office revenues reported for 111 movies released theatrically in 1992. One hundred and one movies were used for building the regression model and the remaining 10 movies were used to test the forecasting performance of the model. They adapted many of the same variables mentioned in the previous studies including genre, special effects, sequel, sexual content, MPAA rating, star value, and critics reviews. Given the large set of explanatory variables, they had to use three separate step-wise regressions to reduce the number of independent variables to a manageable size. They reported a gross box-office receipts predictive power (R^2) of 0.419 ($p < 0.00001$) for those of the reduced set of movie attributes. Their model produces increasingly more accurate prediction results as the new early box-office data becomes available. More specifically, their model generated mean absolute percent error of 71.1%, 51.6%, 13.2%,

7.2% and 1.8%, for using no data, one week of data, two weeks of data, three weeks of data and all available data, respectively.

De Silva [1998] studied consumer behavior to understand movie theatre attendance habits as a function of viewer's demographic characteristics and other movie related variables. In his regression models he consistently found five variables, director, advertising, reviews, age, and marital status, as being significant contributors for explaining attendance. Among all the demographic variables included, age and marital status were found to be the only ones to attain significance. His regression models for predicting theatre attendance generated statistical fit values (R^2) between 0.1871 and 0.2069. That is, with all the variables included into the models, only up to 20 percent of the variation in attendance could be explained. In this study, no attempt was made to report the predictive performance of the consumer behavior based regression model, therefore, we have no base to compare our predictive performance results to those of his.

Neelamegham and Chintagunta [1999] developed a Bayesian modeling framework to predict early viewership for new movies in both domestic and international markets. They used such factors as the number of screens, distribution strategy, and movie attributes such as MPAA ratings, genre, and presence and absence of movie stars and high profile directors to model the viewership influence. Their hierarchical Bayesian prediction model provides viewership forecasts at different stages of the new movie release process. Thus, forecasts can be obtained under a number of information availability scenarios, starting with just information from historical databases containing data on previous new movie launches, to more and more information about the movie after its release in domestic and international markets. As more information becomes available, the forecasting model allows for combining historical information with data on the performance of the new movie and thereby making more accurate forecasts. They used a sample of 35 movies released in domestic markets (as well as 13 other international markets) between 1994 and 1996. They reported a mean absolute percent error of 44.85% as pre-release viewership forecasting performance for the domestic markets when they combine their Bayesian model with a rule of thumb. Their model

performs significantly better for the international markets (mean absolute percent error as low as 21% in some international markets) as the actual performance data from the domestic market becomes available for the forecasting model. It is not certain from the study whether this performance measure is based on a holdout sample or based on the 35 movies used in developing the model. Since our model is addressing a classification problem and their model is addressing a viewership point estimate forecasting problem, making an objective comparison of the two models predictive performance would be difficult.

Ravid [1999] has studied the role of movie stars and other explanatory variables in the success of motion pictures. He used a combination of means comparisons and linear regression models on a random sample of 200 movies released between 1991 and 1993 in the U.S. His reported results indicate that based on the means comparisons the super stars have a high positive correlation with the movie revenues, whereas linear regression did not specifically indicated the same level of significance for super stars. He has also reported that the production budget, family oriented stories, MPAA ratings, and initial distribution volumes all have found significant predictors for box office revenues. He also did not report on the predictive accuracy of his model.

Eliashberg et al. [2000] expand on their previous work (Eliashberg and Sawhney 1994, Sawhney and Eliashberg 1996) by modeling the behavioral representation of the consumer adoption process for movies in six consecutive stages (as opposed to two in their previous work). According to their Markov chain based macro-flow model, which they call MOVIE MOD, at any point in time with respect to the movie under study, a consumer can be found in one of the following behavioral stages: undecided, considerer, rejecter, positive spreader, negative spreader, and inactive. The progression of consumers through the behavioral stages depends on a set of movie-specific and behavioral factors. Because of the cumulative nature of their modeling framework, they managed to include (in several stages of the model execution) the word-of-mouth influence, publicity and reviews of critics, and promotion and distribution strategies. They claim that their model can be used as a decision aid for marketing managers in allocating promotional budgets.

Anita and Eliashberg [2002] developed a dynamic and iterative framework for predicting box-office revenues of theatrical movies. According to them, in order to gain a thorough understanding of the drivers of motion picture performance, one should consider the determinants of both revenues and screens (i.e., the drivers of both audiences and exhibitors) as separate but interrelated modeling components. Accordingly, they used simultaneous equation models to capture the essence of both revenues and screens. Since some of the variables are concurrently used in both equations, such as number of screens, an iterative solution approach is employed. However, they did not report any experimental results of their modeling framework.

We differentiate our study from the others as follows. First, there is no reported study on using neural networks (NNs) to predict box office receipts of new motion pictures. Our study seems to be the first attempt of its kind in this problem domain. Second, most of the above mentioned studies did not report on the predictive performance of their models. We not only report on the most commonly used predictive performance measure (i.e., percent correct classification), but also borrowed a couple of other relevant performance measures from psychology literature. Another difference of our study comes from its longitudinal nature. We based our study on a comprehensive four-consecutive-years of data that covers movies released between 1997 and 2000. Our study also compares the difference between those of individual years and the combined data set of all four years.

3. Method

The initial launch of our study was in 1998. For each year since 1997, the available data has been collected, organized, and analyzed on year-by-year bases. An early performance was reported in Sharda et al. [2000]. Now, enough analysis results exist to report on this project.

In our study, we used 588 movies released between 1997 and 2000. The sample data was drawn (purchased) from ShowBiz Data Inc. [ShowBiz 2002]. The dependent variable in

our study box-office gross revenues, not including auxiliary revenues such as video rentals, international market revenues, toy and soundtrack sales, etc. Another important difference between our study and previous efforts is that we convert the forecasting problem into a classification problem. Rather than forecasting the exact amount of the dependent variable (box-office receipts), 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 in NN literature as “discretization.” In our study, we discretized the dependent variable into nine classes using the following breakpoints

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

We used eight different types of independent variables. Our choice of independent variables is based on previous studies conducted in the field. Each independent categorical variable is then converted into a binary representation, which created a number of pseudo variables increasing the independent variable count from 8 to 43. In the process, all pseudo variables of a categorical variable are given the value of 0 except the one that holds true for the current record, which is given the value of 1.

Neural networks require the data to be in numerical format. In the cases where a variable is represented with categorical (a.k.a. symbolic) values, the *binary encoding* should be utilized [Principe 2000]. Binary encoding is a common practice in neural network modeling for translating categorical variables in to numerical (binary) columns (assuming that the data is presented in a way that the columns represent variables and rows represent exemplars). A categorical variable, say *marital status* would have values like *married*, *single* and *divorced*. When a column of categorical data is translated using binary encoding, one column is created for each unique category type. For the example of marital status, we should create three columns, each representing the values that the categorical variable can take. Within a translated column, a “1” signifies the existence of the corresponding category and a “0” signifies the existence of any other symbol.

We used most of the same variables mentioned in previous studies [Litman 1983, Litman and Kohl 1989, Sochay 1994, Sawhney and Eliashberg 1996, Neelamegham and Chintagunta 1999, Ravid 1999, Elberse and Eliashberg 2002]. Following is a short description of these variables with their respective representation schemas.

Month of Release: Most studies found the time of release of a movie as a significant contributor to box-office receipts [Krider and Weinberg 1998, Radas and Shugan 1995]. Radas and Shugan (1995) developed a model of seasonality to capture the effect of the timing of the release of new motion pictures on their financial outcomes. We use 12 binary variables to represent the month of release of a movie in the input vector. The release month of the movie gets the value of 1 and the rest of the related variables get the value of 0 in accordance with binary encoding.

MPAA Rating: Another commonly used variable in predicting the financial success of a movie is the rating assigned by the Motion Picture Association of America (MPAA). There are five possible rating categories, each represented with a binary variable: G, PG, PG-13, R, and NR. These ratings help assess the degree of sexual content, violence, and adult language for any movie before its theatrical release. Though earlier studies that used regression based forecasting models showed no significant contribution of MPAA rating to predict box-office receipts [Litman 1982, Litman and Kohl 1989, Sochay 1994], others indicated a high level of significance [Ravid 1999]. Litman [1997] and Ravid [1999] indicated that some ratings, (e.g., G and PG) get bigger audiences than others when all other influencing factors are kept equal, due largely to the fact that these films appeal to a larger potential crowd. For completeness sake, in this study we decided to include MPAA rating by using five binary variables to represent the MPAA ratings.

Competition: Movies are not released in a vacuum. Each movie competes for the same pool of entertainment dollars against movies released at the same time or released some time ago and carried over to the present time. Therefore, the financial success of a movie is expected to be highly correlated with the current competitive forces in the marketplace. Some of the previous research studies on the topic of forecasting motion picture box-office results also included a parameter representing the competitive forces [Litman and Kohl 1989, Sochay 1994]. These forces are expected to negatively influence the success

of a movie. In our study we represent the competition using three binary variables (high competition, medium competition, and low competition).

Star value: As others before us, we included another independent variable to measure the financial success of a movie depending upon the presence or absence of any box office superstars in the cast (e.g., actors, actresses, and directors). Ravid [1999] found conflicting results with respect to the contribution of superstars to the financial success of a movie. A superstar actor/actress can be defined as one who contributes significantly to the up-front sale of the movie regardless of the script, the costars, and the director. The value of a star/superstar is determined by averaging his/her recent past history of movie-making prices. We used three independent binary variables to represent the degree of star value of actors/actresses in our model: A+/A (high) star value, B (medium) star value, and C (insignificant) star value. We did not use a variable to signify the importance (or star value) of the director due to the fact that most earlier studies that included star value of the movie director did not find it as being a significant contributor [Litman 1982, Litman and Kohl 1989, Sochay 1994], and recent studies did not even include it into their prediction models [Sawhney and Eliashberg 1996, Neelamegham and Chintagunta 1999].

Content category (Genre): One commonly used, yet rarely found to be significant contributor, is the content category variable [Litman 1983, Litman and Kohl 1989, Sochay 1994]. In our study, we followed the tradition and included the content category as one of the independent variables in our model. Specifically, we placed each film into one of content categories using nine binary independent variables. The categories include Sci-Fi, Historic Epic Drama, Modern Drama, Thriller, Horror, Comedy, Cartoon, Action, and Documentary.

Technical effects: We also used a variable to capture the technical merit of a movie. We used three levels of technical effect categories, which are represented by using three binary independent variables. Movies with high technical content and special effects, such as animations and science fiction movies, are given the highest variable or rating. Movies with moderate special effects are given a medium rating, and movies with little or no special effects are given the low technical affect ratings.

Sequel: Similar to other studies [De Silva 1997, Ravid 1999], we also included a binary variable to specify whether a movie is a sequel (value of 1) or not (value of 0).

Number of screens: Previous research efforts showed close correlations between a movies' financial success and the number of screens on which the movie is shown during its initial launch [Neelamegham and Chintagunta 1999, John and Ritz 1991]. In our model, we capture the number of screens a movie is scheduled to be shown at its opening with six categorical binary variables representing six intervals of number of screens: less than 500, 500 - 1000, 1000 - 1500, 1500 - 2000, 2000 – 2800, and more than 2800.

Above mentioned and briefly defined decision variables are summarized in Table 1. In total, 43 decision variables (representing 8 categories) are used in this study.

Table 1. Summary of variables

Variable Category	Input or Output	No. of Variables	Possible Values
Month of release	Input	12	January thru December
Rating of the movie	Input	5	G, PG, PG-13, R, NR
Competition	Input	3	High, Medium, Low
Star value	Input	3	A+/A, B, C
Genre	Input	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Technical effect	Input	3	High, Medium, Low
Sequel	Input	1	Yes or No
Number of screens	Output	6	<500, 500-1000, 1000-1500, 1500-2000, 2000-2800, >2800

A neural network model was developed by using a commercial software product, Neuralware Professional Plus [Neuralware 2002]. We used multi-layer perceptron (MLP) neural network architecture with two hidden layers, and assigned 18 and 16 processing elements (PE) to them, respectively. In both hidden layers Tangent Hyperbolic (Tanh) transfer functions were utilized. The learning algorithm was based on back-propagation. Given the description of the variables above, we had 43 input neurons and 9 output

neurons. Figure 1 depicts an MLP neural network architecture with two hidden layers having 4 and 3 PEs, respectively.

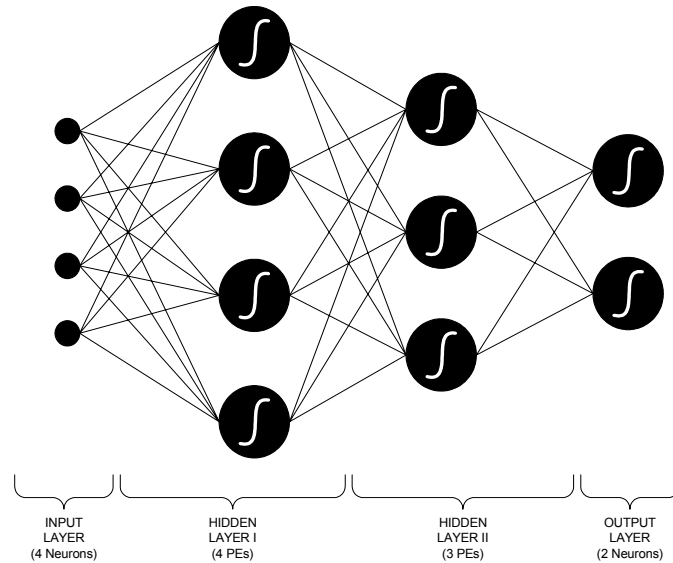


Figure 1. Graphical representation of an MLP neural network architecture

Many researchers over the recent past have studied the performance of neural networks in predicting a variety of classification problems in a wide range of different business settings. Many of these studies however, were based on a single experiment and/or the method of selecting the training and testing samples was unclear. We believe that because of the stochastic nature of the neural network training, more rigorous and more statistically sound experimental design methods are necessary to develop the objective performance measures of neural networks. As opposed to using a single neural network experiment to base our results upon, we chose to follow a more statistically sound experimental design methodology. In our model development effort, we treated each of the four years as separate data sources. For each year's data we created 20 data sets (training and testing files), by using a resampling technique. These data sets were then used to train 20 neural networks. Each training set included 100 movies and each testing data set included another 20 movies. We made sure that the training and testing data sets did not overlap (did not share the same movie) for the same neural network model. The first 100 data points (training set) were introduced to the network in the training phase, and then the 20 data points (which were not used in the training data set) were introduced

to the trained neural net for testing purposes. This process was repeated 20 times for each year. Our classification results are based on the statistical measures of those twenty neural network test runs. After developing the models and obtaining the performance measures for each of the four years, we combined all four years into a single data set and repeated the same experimental design yet one more time. The prediction results of the combined data set model were comparable to those of the ones obtained from the individual years' neural network models. This result indicates that the pattern of the independent variables considered in this study was not significantly different from one year's data to the next, at least for the duration considered.

We used three different performance measures to evaluate the predictive performance of our neural network models in this problem domain. First, we used the most commonly applied statistical performance measure, average percent hit rate (APHR), to calculate the rate at which the movies are assigned to their correct classes. Then, we used improvement over random sampling, τ (tau), to calculate the percent improvement of our classification over random assignment to the classes. Finally, we used the similarity to the perfect classification, R_g , measure to show the closeness of our prediction results to those of the perfect classification.

In classification problems, the average percent hit rate indicates the rate at which the testing data samples are classified into the correct classes. In our case, we have two different hit rates: the exact (bingo) hit rate (only count the correct classifications to the exact same class) and the within 1 class (W1C) hit rate. The hit rate measures the average accurate classification rate of the neural network prediction and the desired output.

Algebraically, APHR can be formulated as follows:

$$APHR = Average \left(\frac{\text{Number of samples correctly classified}}{\text{Total number of samples}} \right)$$

$$APHR_{Bingo} = \frac{1}{g} \sum_{i=1}^g \frac{p_i}{n_i} \qquad APHR_{W1C} = \frac{1}{g} \sum_{i=1}^g \frac{p_{i-1} + p_i + p_{i+1}}{n_i}$$

where, g is the total number of classes, n_i is the total number of samples in class i , and p_i is the total number of samples correctly classified in class i .

As a direct indicator of predictive accuracy, average percent correct classification is the most intuitive measure of discrimination. One should, however, judge the magnitude of this percentage in relation to the expected percentage of correct classification if assignment were made randomly. If we have two groups, we can expect to get 50% of the predictions right by pure random assignment. With four groups, our expected accuracy is only 25%. Should the classification process yield only 60% between two groups, the improvement is rather small. With four groups, however, 60% is a considerable improvement, because we would expect only 25% to be correct by chance. A proportional reduction in error statistic, τ (tau), which would give a standardized measure of improvement regardless of the number of groups is given in equation (1) [Klecka 1980].

$$\tau = \frac{n_c - \sum_{i=1}^g p_i n_i}{n - \sum_{i=1}^g p_i n_i}$$

where, n is the total number of cases in the experiment, n_c is the number of cases correctly classified, p_i is the prior probability of group i , n_i is the total number of cases in group i , and g is the total number of groups in the classification problem. The term involving the summation is the number of cases that would be correctly classified on the basis of random assignment to groups in proportion to their prior probabilities. The maximum value of τ is 1.0, and it occurs when there are no errors in prediction. A value of zero indicates no improvement over random assignment.

Another way we can objectively measure the accuracy of a classification algorithm is to compare its results with the perfect classification, which includes the prior probabilities of each classes calculated from the available data sample. If the two sets of results are significantly *similar* to each other, then we can conclude that the classification algorithm under investigation is performing appropriately. This method has been widely used in

evaluating the performance of clustering algorithms by comparing the results of one clustering scheme to another [Rand 1971]. We adopt this measure for our use by treating the known classification (actual performance) as one clustering technique and the predicted classification (by the neural network model) as the second clustering technique. In the core of this method lies the basic unit of comparison between the two classifications based on how pairs of points are classified. If the elements of a point-pair are placed together in a class in each of the two classifications, or if they are assigned to different classes in both classifications, this represents a similarity between the classifications, as opposed to the case in which the elements of the point-pair are placed together in different classes in both classifications. The measure of the similarity between the two classifications (calculated classification and the perfect classification) is denoted by R_g and can be calculated as follows:

$$\begin{aligned}
 R_g &= \frac{\text{Total number of pair objects that are classified together that fall in different classes in both classifications}}{\text{Total number of pair objects}} \\
 &= \text{Probability that two objects are treated alike in both classifications} \\
 R_g &= \left[T_g - \frac{1}{2} P_g - \frac{1}{2} Q_g + \binom{n}{2} \right] / \binom{n}{2} \\
 T_g &= \sum_{i=1}^g \sum_{j=1}^g m_{ij} - n \\
 P_g &= \sum_{i=1}^g m_i^2 - n \\
 Q_g &= \sum_{j=1}^g m_j^2 - n
 \end{aligned}$$

where m_{ij} is the number of points simultaneously classified in the i th class of one classification and the j th class of the other classification. The total number of classes in each of the two classifications is denoted by g , and n is the total number of samples. The value of R_g ranges between 0 and 1, denoting the strength at which the two classifications (calculated and perfect) can be treated similarly.

4. Results

The results presented in this paper are not directly comparable to those of earlier studies because our model predicts a class of performance in which the movie should belong, rather than an actual dollar figure. A general sense of accuracy is, however, still relevant. Recall that the model aims to categorize a film in one of the nine categories. Accuracy is measured in two ways. The first metric is the percent correct classification rate. This metric is a reasonably conservable measure for accuracy in this scenario. In reality, the movie studios might be glad to predict within one (or maybe two categories) on either side. Table 2 presents our neural network results. For each of the four years, 20 neural network model results are presented along with their average, standard deviation, and median. The first column denotes the number assigned to each of the 20 neural network models trained and tested for the given years' data. The second column denotes the percent hit rate for the same category (Bingo) and the third column shows the percent hit rate within one category (W1C). The summary statistics for each year of each metric is also given in the last three rows of each year's table. As the summary statistics given in Table 1 indicate, our neural network model predicted the exact category (Bingo) of a movie in the average of 30% of the time and predicted within one category (1Away) more than 72% of the time. We also wanted to know if the misclassification errors are biased towards upper or lower classes of the output variable. The statistical study conducted using the combined prediction results of all neural network models indicates that the misclassification errors for upper and lower classes of the output variable are not significantly different from each other.

Table 2: Percent correct classification performance measures for each of the four years and all four years combined

Runs	Year 1997		Year 1998		Year 1999		Year 2000		All Four Years	
	Bingo	1 Away	Bingo	1 Away	Bingo	1 Away	Bingo	1 Away	Bingo	1 Away
1	25%	70%	50%	80%	25%	55%	30%	70%	41%	79%
2	15%	60%	25%	55%	40%	70%	30%	75%	30%	76%
3	35%	90%	60%	90%	40%	70%	30%	60%	29%	75%
4	10%	55%	30%	70%	25%	65%	45%	90%	34%	77%
5	30%	75%	40%	85%	45%	85%	40%	80%	24%	64%
6	25%	55%	25%	70%	20%	70%	30%	85%	29%	72%
7	30%	65%	30%	75%	20%	70%	50%	70%	31%	74%
8	25%	75%	40%	80%	35%	80%	35%	80%	36%	69%
9	30%	60%	25%	65%	30%	65%	25%	65%	34%	67%
10	25%	70%	30%	65%	50%	85%	30%	85%	33%	79%
11	25%	70%	50%	80%	25%	60%	20%	70%	27%	74%
12	15%	90%	30%	65%	30%	75%	20%	95%	30%	74%
13	20%	65%	25%	55%	30%	70%	40%	60%	30%	69%
14	15%	75%	40%	65%	40%	80%	35%	70%	28%	71%
15	25%	75%	20%	70%	10%	65%	5%	65%	32%	68%
16	30%	90%	30%	75%	20%	65%	15%	70%	29%	72%
17	25%	70%	40%	75%	25%	85%	35%	85%	31%	76%
18	25%	75%	20%	65%	25%	80%	30%	75%	26%	68%
19	30%	75%	30%	75%	40%	80%	40%	75%	27%	73%
20	20%	65%	10%	75%	25%	65%	20%	80%	27%	74%
Mean	24.00%	71.25%	32.50%	71.75%	30.00%	72.00%	30.25%	75.25%	30.40%	72.55%
StdDev	6.41%	10.37%	11.87%	9.07%	10.00%	8.80%	10.70%	9.66%	3.89%	4.07%
Median	25%	70%	30%	73%	28%	70%	30%	75%	30%	74%

Figure 2 presents the summary statistics of all four years, both individually and combined, using standard bar charts. Notice that the mean, standard deviation, and the median calculated for each of the four years are not significantly different from the ones calculated for All Four Years. This result indicates that the predictable behavior of the domain does not change over time for the variables included into the model and for the time period used in this study.

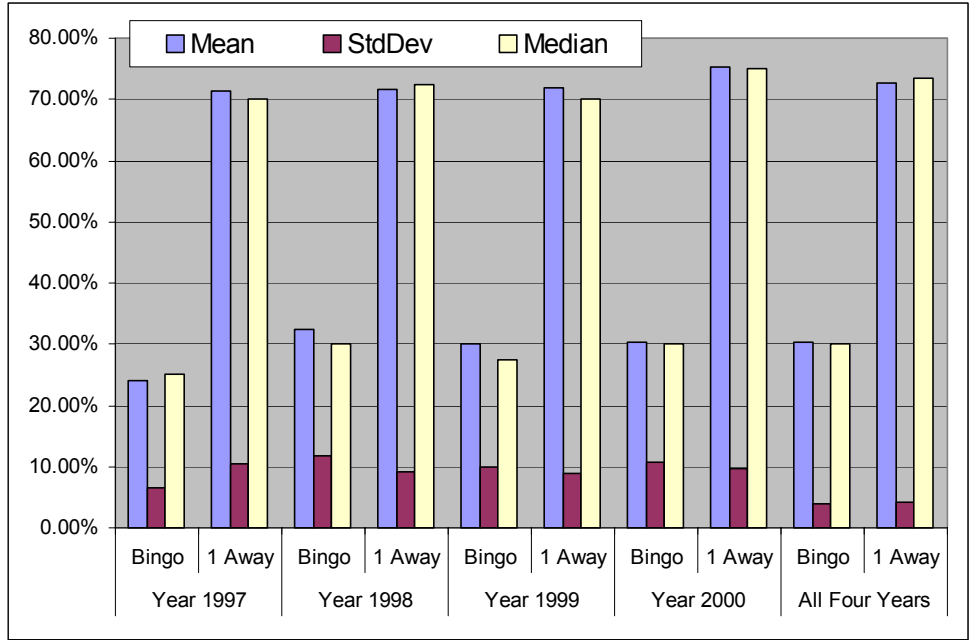


Figure 2. Graphical representation of the hit rate summary statistics

In addition to the performance measures presented in Table 2, we also calculated the measures of improvement from the random guessing τ (tau), and degree of similarity to the perfect classification (R_g). The results are presented in Table 3. Our results show a significant improvement over random guessing with an average of 22% based on the calculated τ values for all four years. The degree of similarity to the perfect classification, R_g , performance measure shows even better results with an average of 77% based on the calculated R_g values for all four years.

Table 3: τ and R_g performance measures for each of the four years and all four years

	Year 1997		Year 1998		Year 1999		Year 2000		All Four Years	
Runs	Tau	Rg	Tau	Rg	Tau	Rg	Tau	Rg	Tau	Rg
1	0.155	0.653	0.458	0.774	0.157	0.795	0.231	0.742	0.339	0.789
2	0.034	0.763	0.160	0.737	0.333	0.726	0.224	0.805	0.212	0.747
3	0.251	0.626	0.559	0.758	0.328	0.758	0.200	0.789	0.207	0.783
4	-0.017	0.716	0.222	0.753	0.160	0.742	0.397	0.742	0.260	0.778
5	0.193	0.721	0.346	0.753	0.385	0.816	0.341	0.763	0.148	0.785
6	0.153	0.679	0.169	0.747	0.109	0.789	0.233	0.726	0.193	0.770
7	0.186	0.674	0.227	0.837	0.101	0.758	0.454	0.658	0.216	0.763
8	0.169	0.716	0.350	0.732	0.286	0.637	0.270	0.774	0.287	0.783
9	0.198	0.732	0.148	0.742	0.213	0.774	0.185	0.747	0.264	0.781
10	0.143	0.684	0.213	0.774	0.440	0.774	0.218	0.779	0.248	0.786
11	0.130	0.705	0.451	0.800	0.164	0.768	0.096	0.753	0.176	0.744
12	0.026	0.721	0.224	0.737	0.209	0.774	0.126	0.742	0.221	0.761
13	0.064	0.732	0.143	0.737	0.231	0.784	0.322	0.758	0.216	0.764
14	-0.003	0.611	0.331	0.821	0.326	0.826	0.292	0.758	0.187	0.779
15	0.138	0.811	0.111	0.732	-0.014	0.805	-0.038	0.695	0.239	0.780
16	0.207	0.800	0.227	0.768	0.086	0.753	0.058	0.726	0.205	0.788
17	0.125	0.653	0.346	0.768	0.189	0.653	0.268	0.732	0.224	0.764
18	0.167	0.679	0.093	0.795	0.145	0.774	0.231	0.800	0.164	0.736
19	0.209	0.800	0.202	0.684	0.326	0.763	0.328	0.758	0.186	0.786
20	0.080	0.695	-0.011	0.805	0.145	0.716	0.109	0.711	0.188	0.770
Mean	0.130	0.708	0.248	0.763	0.216	0.759	0.227	0.748	0.219	0.772
StdDev	0.076	0.055	0.139	0.036	0.114	0.048	0.118	0.035	0.045	0.016
Median	0.148	0.711	0.223	0.755	0.199	0.771	0.231	0.750	0.214	0.778

As we did for the percent hit rate summary statistics, in Figure 3 we present the summary statistics for τ , and R_g for all four years, both individually and combined, using standard bar charts. Again, notice that the mean, standard deviation, and median calculated for τ and R_g for each of the four years are not significantly different from the ones calculated for all four years combined. This result, again, reinforces our previous conclusion that the predictable behavior of the problem domain does not change over time for the variables included into the model and for the time period used in this study.

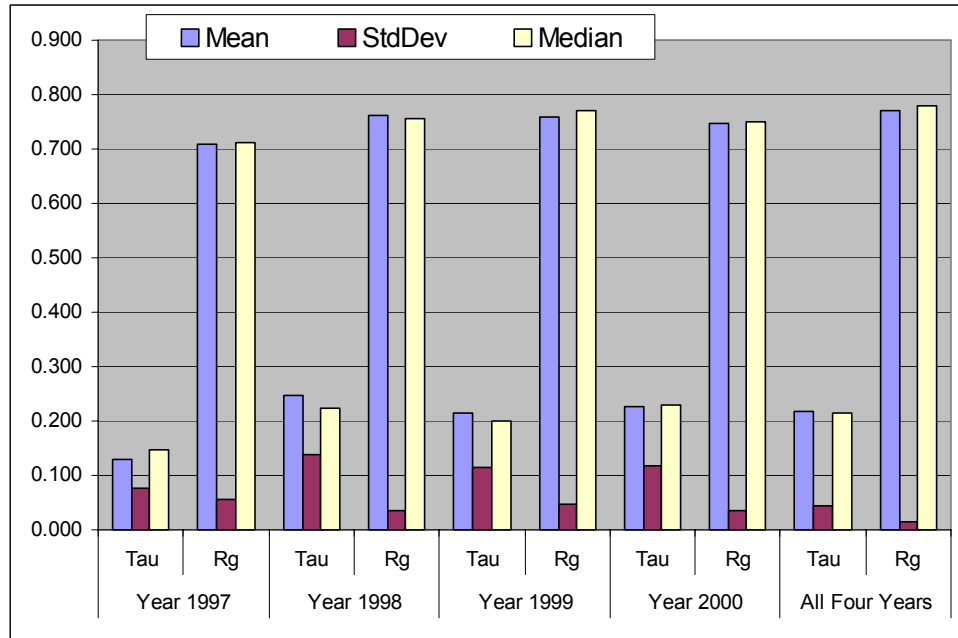


Figure 3. Graphical representation of the τ and R_g summary statistics

In addition to the visual aid provided via the bar charts, we also present the results of pair two sample t-tests in Table 4. Specifically, Table 4 lists the two-tail P statistics (at the 95 percent confidence level) in comparing the results of the four individual years with the results of all four years combined. As the results indicate, with the exception of the year 1997, the difference between the performance results of *all four years combined* and the *individual years* came out to be not significant.

Table 4: Two-tail P statistics for comparing the performance results of individual years and the all four years combined.

Year	Performance Measures			
	Bingo	1Away	Tau	Rg
1997	0.0017*	0.6192	0.0003*	0.0001*
1998	0.4348	0.7207	0.3522	0.3504
1999	0.8759	0.8170	0.9164	0.2573
2000	0.9533	0.2111	0.7775	0.0167*
* indicates significance at $p < 0.05$ level				

Sensitivity analysis on neural network output

Sensitivity analysis is a method for extracting the cause and effect relationship between the inputs and outputs of a neural network model. It is a commonly used method in neural network studies for identifying the degree at which each input channel (independent variables) contributes to the identification of each output channel (dependent variables). In the process of performing sensitivity analysis, the neural network learning is disabled so that the network weights are not affected. The basic idea is that the inputs to the network are shifted (perturbed) slightly and the corresponding change in the output is reported as a percentage (Principe et. al 2000). The first input is varied between its mean plus (or minus) a user-defined number of standard deviations while all other inputs are fixed at their respective means. The network output is computed for a user-defined number of steps above and below the mean. This process is repeated for each input. As an outcome of the process, a report (usually a column plot) is generated, which summarizes the variation of each output with respect to the variation in each input.

Our sensitivity analysis results are based on the combined data set of all four years. After the training, the network weights are frozen (testing stage) and the cause and effect relationship between the independent variables and the dependent variables are investigated as per the above-mentioned procedure. The results are summarized and presented as a column plot in Figure 4. The x-axis represents the input variables and the y-axis represents the percent change on the output variables, while the input variables (one at a time) are perturbed gradually around their mean with the magnitude of \pm one standard deviation.

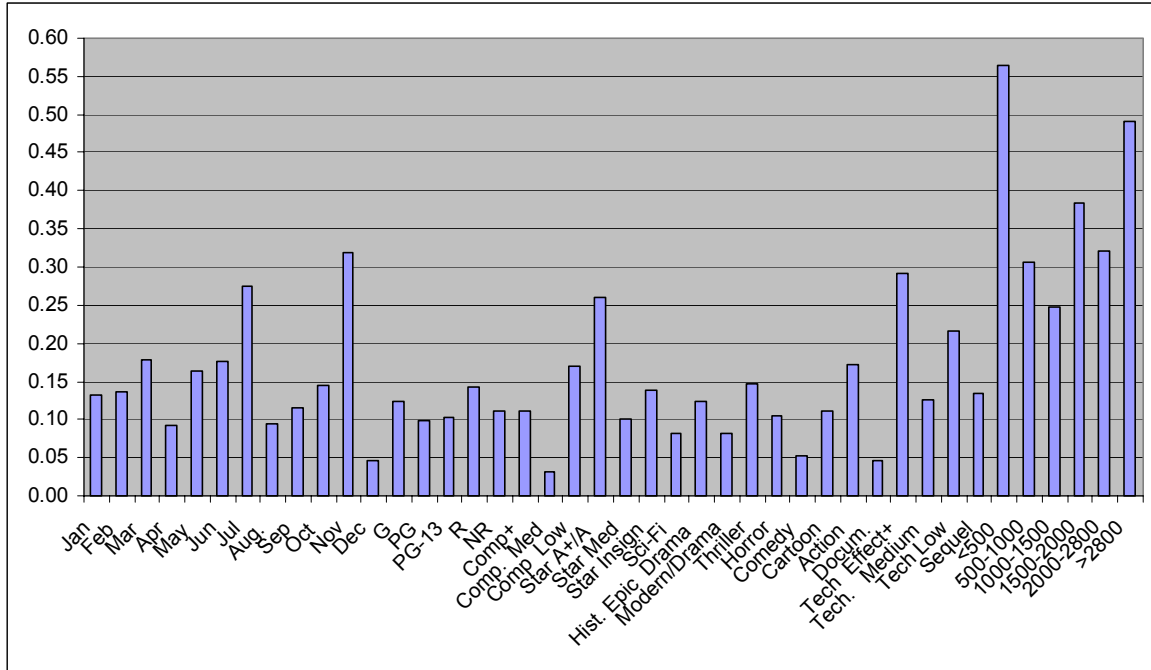


Figure 4. Sensitivity analysis results for all independent and dependent variables

Figure 4 shows the interaction between 43 input variables (represented in the x-axis) and the combined percent sensitivity value of the output variable. Though it gives a general idea as to which independent variables play a major role in (contributes to) determining the dependent variable, it is possible to see the individual pair relationships between the input and the nine categorical output variables. It would be interesting to see if any of the input variables has a significantly different effect on different classes of the output variable. For instance, can we answer the question of “Does the star value have a bigger effect on the higher box-office movies than it does on the lower box-office movies?” Therefore, in Figure 5, we split the same histogram into nine separate histograms, each of which is specific to an output variable.

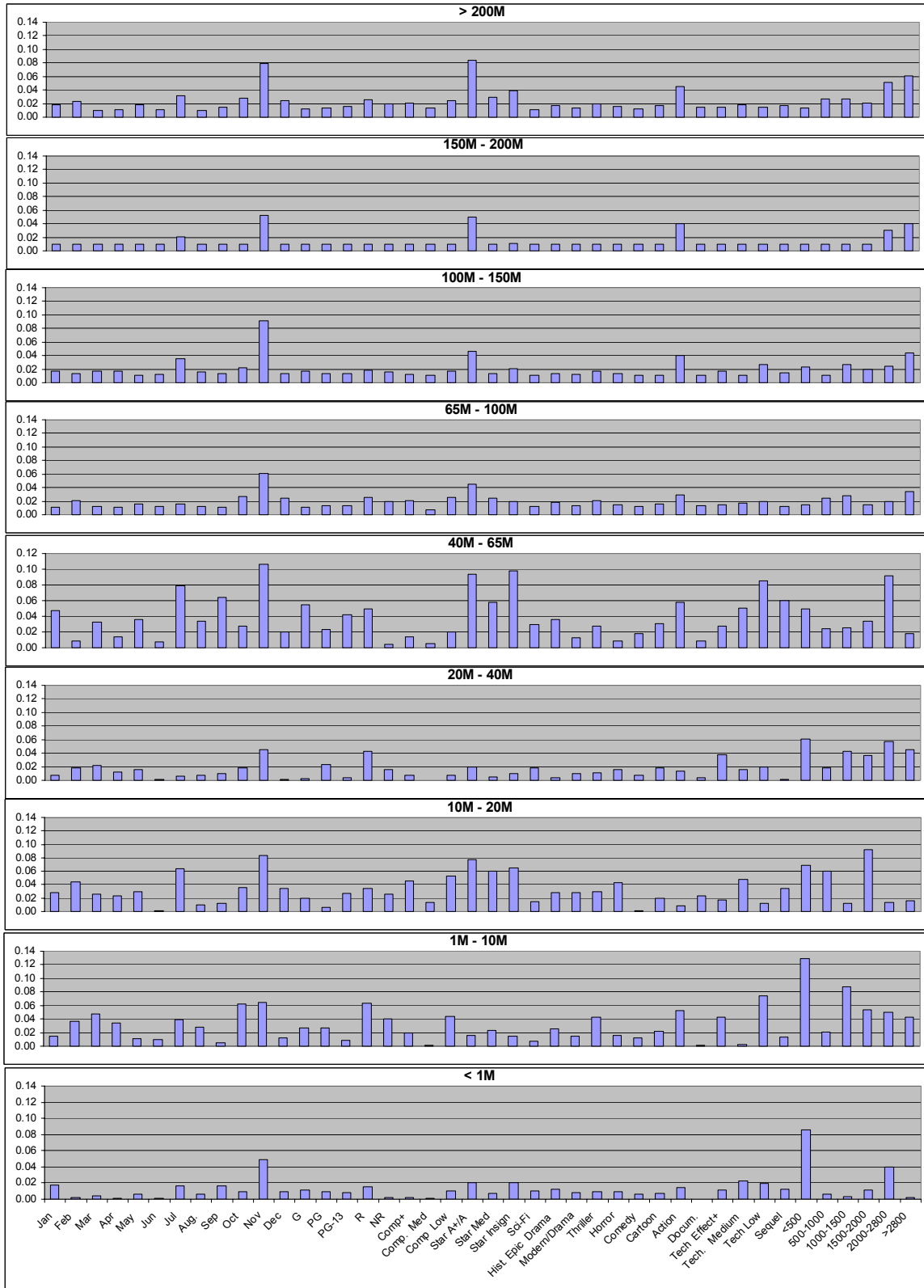


Figure 5. Sensitivity analysis results for each dependent variable

The sensitivity analysis shows that the most important independent variable is the number of screens on which the movie is to be released. This finding is consistent with the findings of the study conducted by Neelamegham and Chintagunta [1999] and Elberse and Eliashberg [2002]. The secondary importance goes to several other independent variables: month of release (especially the months of July and November stand out), start value (especially A+/A star category), and a high level of technical effects (high level of technical effects). Some of the independent variables such as Sequel, MAPE Ratings, and Genre were not among the major contributors.

5. Discussion and Conclusion

Though our numbers are not directly comparable to those of earlier studies (because our model classifies a movie into a category rather than predict a point estimate of the box office receipts), our results are quite good. The accuracy of prerelease forecasting reported by Sawhney and Eliashberg [1996] had a mean absolute error of over 71%. Our results show that we can predict the box-office receipts of a motion picture with 75% accuracy within one category before its theatrical release. Because our model is designed to predict the financial success of a movie before the theatrical release, it can be used as a powerful decision aid by studios, distributors, and exhibitors. In any case, our results are very intriguing, and once more, prove the continued value of neural networks in addressing difficult forecasting problems. The other specific contributions of our study are as follows:

1. Prediction accuracy, Generalization to other media products, Sensitivity analysis for decision makers: Beyond the attractive accuracy of our results in predicting box-office receipts, these neural network models could also be adapted 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 neural network model in order to better understand the implications of different parameters on the end result, box-office receipts. During this alternative experimentation process, the manager 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 in the financial success of a film, or a television program.

2. More statistically sound experimental design procedure for NN training: In contrast to most previous neural network studies, we used a more statistically sound experimental design procedure in order to objectively measure the predictive performance of neural networks which are simply stochastic simulation techniques prone to generating less than optimal results. We randomly resampled from the same data set to generate a statistically significant number of training and testing data for neural network modeling. We based our results on the summary statistics obtained from those multiple neural network runs.

3. More comprehensive performance measures for classification problems: We presented several performance measures, which can be used in evaluating the superiority of classification techniques by using our neural networks prediction results. Some of these measures have been used in the psychology literature in evaluating the performance of classical statistical methods, but underused in neural network applications. We think that these measures along with the classical performance indicators such as mean absolute percent error, give a better indication of the superiority of a classification method and provide the decision maker or modeler with additional insight into the strengths and weaknesses of the method being used.

4. Limitations / Future directions: The accuracy of the neural network model 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 trade secrets and are not publicly released. Just like many other stochastic modeling techniques, neural networks starts from a random set of weights, by utilizing the architectural parameters such as learning algorithm, learning rate, number of PEs in the hidden layers, etc., it adjusts those weights to create a map between the input and the output vectors. Correct choice of those architectural parameters plays a great role in developing better neural network models. There is no close form solution to what those architectural parameters should be for a given data set of a given problem domain. Modelers use their experiences, hunches, and rule of thumb, along with trial and error procedures to better configure those architectural parameters. Lately, researchers are developing hybrid architectures where they apply genetic algorithms and other intelligent search techniques to optimize

the architectural parameters of neural networks. Reported results are promising. Application of such hybrid architecture can improve the results we have obtained in this study. In addition to MLP, some other neural network architectures can be used to improve the accuracy of the neural network models.

From an application perspective, once developed to a production system, such a neural network model can be made available (via a web server or an ASP) to industrial decision makers, where individual users can plug in their own movie parameters to forecast the potential success of a motion picture before its theatrical release. A neural network model can be designed in a way such that it can calibrate its weights (continuous self learning) by taking into account new samples (movies that are released and determined box-office receipts) as they become available.

Acknowledgements

Several graduate assistants have participated in this project as the first author attempted to continue the study for several years. These include, in chronological order, Edith Meany, Niketu Mithani, and Prajeeb Kumar. Their contributions for data collection and analysis are much appreciated.

References

- De Silva, I. (1998). "Consumer Selection of Motion Pictures" appeared in *The Motion Picture Mega-Industry* by B. Litman. Allyn & Bacon Publishing, Inc.: Boston, MA.
- Elberse, A. and J. Eliashberg (2002). "The Drivers of Motion Picture Performance: The Need to Consider Dynamics, Endogeneity and Simultaneity," to appear in the *Proceedings of the Business and Economic Scholars Workshop in Motion Picture Industry Studies*, Florida Atlantic University, pp. 1-15.
- Eliashberg, J. and M.S. Sawhney (1994). "Modeling Goes to Hollywood: Predicting Individual Differences in Movie Enjoyment," *Management Science*, Vol. 40, No. 9, pp. 1151-1173.

- Eliashberg, J., J.J. Junker, M.S. Sawhney, and B. Wierenga (2000). "MOVIEMOD: An Implementable Decision Support System for Prerelease Market Evaluation of Motion Pictures," *Marketing Science*, Vol. 19, No. 3, pp. 226-243.
- Hykin, Simon S. (1998). *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall.
- Klecka, W.R., (1980). *Discriminant Analysis*. Sage Publishing: Beverly Hills, CA. pp. 49-51.
- Krider, R.E. and C.B. Weinberg (1998). "Competitive Dynamics and the Introduction of New Products: The Motion Picture Timing Game," *Journal of Marketing Research*, Vol. 35, No. 1, pp. 1-15.
- Litman B.R. and H. Ahn (1998). "Predicting Financial Success of Motion Pictures" appeared in *The Motion Picture Mega-Industry* by B.R. Litman. Allyn & Bacon Publishing, Inc.: Boston, MA.
- Litman, B.R. and A. Kohl (1989). "Predicting financial success of motion pictures: The 80's experience," *The Journal of Media Economics*, Vol. 2, No. 1, pp. 35-50.
- Litman, B.R. (1983). "Predicting Success of Theatrical Movies: An Empirical Study," *Journal of Popular Culture*, Vol. 16, No. 9, pp. 159-175.
- Neelamegham, R and P. Chintagunta (1999). "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," *Marketing Science*, Vol. 18, No. 2, pp. 115-136.
- Neuralware, Inc. (2002). The World Wide Web address is www.neuralware.com, Carnegie, PA.
- Principe, J.C., N.R. Euliano and W.C. Lefebvre (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. New York: John Wiley and Sons.
- Radas, S. and S.M. Shugan, (1998). "Seasonal Marketing and Timing New Product Introductions," *Journal of Marketing Research*, Vol. 35, No. 3, pp. 296-315.
- Rand, W.M. (1971). "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, Vol. 66, No. 336, pp. 846-850.
- Ravid, S.A. (1999). "Information, Blockbusters, and Stars: A Study of the Film Industry," *Journal of Business*, Vol. 72, No. 4, pp. 463-492.

- Rumelhart, D.E., and J.L. McClelland (1986). *Parallel Distributed Processing*, Vol. 1, MIT Press.
- Sawhney, M.S. and J. Eliashberg (1996). "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," *Marketing Science*, Vol. 15, No. 2, pp. 113-131.
- Sharda, R. (1994). "Neural Networks for the OR/MS Analyst: An Application Bibliography," *Interfaces*, Vol. 24, No. 2, pp. 116-130.
- Sharda, R. and R. Wilson (1996). "Neural Network Experiments in Business-Failure Forecasting: Predictive Performance Measurement Issues," *International Journal of Computational Intelligence and Organizations*, Vol. 1, No. 2, pp. 107-117.
- Sharda, R., H. Amato and E. Meany (2000). "Forecasting Gate Receipts Using Neural Networks and Rough Sets," the *Proceedings of the International DSI Conference*, Athens, Greece, pp 1-5.
- ShowBiz Data, Inc. (2002). The World Wide Web address is www.showbizdata.com, Los Angeles, CA.
- Sochay, S. (1994). "Predicting the performance of motion pictures," *The Journal of Media Economics*, Vol. 7, No. 4, pp. 1-20.
- Valenti, J. (1978). "Motion Pictures and Their Impact on Society in the Year 2000," speech given at the Midwest Research Institute, Kansas City, April 25, pp. 7.
- Zufryden, F.S. (1996). "Linking Advertising to Box Office Performance of New Film Releases: A Marketing Planning Model," *Journal of Advertising Research*, Vol. July-August, pp. 29-41.