Movie Attendance Prediction

Kushal Gevaria
Department of Computer Engg
Dwarkadas J. Sanghvi COE
Mumbai, India

Rijuta Wagh
Department of Computer Engg
Dwarkadas J. Sanghvi COE
Mumbai, India

Lynette D’Mello
Department of Computer Engg
Dwarkadas J. Sanghvi COE
Mumbai, India

ABSTRACT

Posting online reviews and rating a movie is a very popular way to obtain information about movies. An online data set of reviews of the movies was taken from IMDB site. This paper uses two models to predict movie theatre capacity for the weekly released movies. The diffusion model, Sawhney and Eliashberg (1996) model predicts the capacity of movie theatre through time-to-decide and time-to-act parameters. The Hierarchical Bayes model consists of three models which are regression model, standard logit model and nested logit model and their efficiency is explained with detail. Finally, these two models are compared and their accuracy is determined.

Keywords

Attendance Prediction; Dynamic Prediction; Autoregressive Model; Diffusion Model; Sentiment Analysis; Data Mining; Linear Regression Models; Canibalization; Hierarchical Bayesian Approach

1. INTRODUCTION

Advancement in technologies leads us to use extremely large amount of data and also the creation of new data at an increasing rate everyday. With this creation of abundance of data, business and organizations focus more towards decision-making process, which helps in further progress of the use of technologies. Hundreds of new movies are launched every year which makes it difficult for consumers to evaluate the quality of the movie taking into account the lifetime of a movie running in the domestic theatre which is less than 15 days. Thus,various forcasting models are extremly important for decision-making to predict the theatre intake capacity . Forecasting movie attendance is one of the examples of decision-making process, where large amount of data as an input can be used to manipulate the attendance of the multiplex theatres everyday, thereby making the decisions of customers more and more efficient and accurate. Many a times, it’s the decision of the manager who decides that how many times a particular movie should be screened depending upon the attendance of the first day. This decision made is judgmental and one cannot rely on such decisions. If you consider the starting day to be Monday, then depending of the Monday box office revenue and attendance, one can at least get an idea of how the data can change on the next day, but finally it is just a guess.

Eliashberg in 2009 proposed a model that could improve the box office revenue by using data-driven approach [4]. Based on this data-driven approach, one can more accurately forecast movie attendances of individual showings, which would thus maximize the current capacity rate based on previous attendance. From this information, a comparison can be made between two models to calculate the capacity of multiplex theatres, which are as follows:

1. The Sawhney and Eliashberg(1996) model
2. Bayesian hierarchical model

The first model called as Sawhney and Eliashberg model uses three parameters to characterize the behavior of consumers, which allows generating forecasts based on previous data [9]. The Bayesian hierarchical model is subdivided into three types, which are linear regression model, standard logit model and nested logit model [1]. These predictions play an important role in strategic film management and budgeting. A comparison between two models is done to produce demand forecasts at different stages of the life cycle. Therefore, the goal here is to determine the potential demand and use forecast as a decision making tool for predicting the movie theatre capacity.

2. LITERATURE REVIEW

Data mining came into picture as these words were used first in 1990 in the database community. Proper knowledge extraction is what done in data mining process. Data mining feature can be used in many ways and hence can also be implemented to predict the attendance of the movie theatre. Many data mining software are available which can be used for this movie attendance prediction system. Different models helps in movie prediction, which can be used. The Bass model in 1969 was developed by Frank Bass, which is used for forecasting the new product sales as well as can be used for technology forecasting. Bass models main advantage is the explicit consideration of the word of mouth effect or imitation effect, a factor, which determines the long-term success of movie and other good experiences according to De Vany and Walls (1999) [9].

The much better model than the Bass model is the Sawhney and Eliashberg model (1996) [9]. The above model uses two factors, which are time-to-act and time-to-decide. The prediction error determined in the above two models is between 2.7% to 17.1%. A multiple regression model was developed by Litman (1983), which was able to predict the financial success of the movie. Terry, Butler, and De'Armond (2004) calculated the movie box office revenue based on action movies, children's movies, sequels, Academy Award wins, and the production budget, which helped to know about the next day attendance [8].

Stimpert, Laux, Marino, and Gleason (2008) mainly focused on the factors that can affect the outcome of a successful movie [11]. Factors that can affect the movie attendance are as follows:

- Release date
- Presence of dramatic top talent
- The Motion Picture Association of America (MPAA) rating of a film

- Genre
- Promotion Efforts
- Production Budget
- Country of Origin
- Word of mouth advertising
- Internet Presence
- Critics Reviews

Hand (2002) and Judge (2012) investigates a time-series analysis of attendance, which is useful for the analysis of the industry as a whole but not for the managerial decisions at the theatre level. This model is not useful to predict for just a particular movie and it's success but the entire film industry itself. Finally, it's the Sawhney and Eliashberg model, which is more efficient as this model predicts both the dynamic and the total behavior of movie attendance at the individual level [3].

Hierarchical Bayes models are hierarchical models analyzed using Bayesian methods. Bayesian methods are based on the assumption that probability is operationalized as a degree of belief, and not a frequency as is done in classical, or frequentist, statistics. Most researchers in marketing have been trained to think about statistics in terms of frequencies [2]. When computing a sample mean or test statistic, for example, many of us think of multiple realizations of a dataset that could lead to variability of the statistic. Even though the statistic is fixed for the data under investigation, we admit the possibility that other realization of the data could have been obtained. Assuming that the model under investigation is true, we compute the expected variability of the statistic. Limitations of Sawhney and Eliashberg model (1996) are overcomed by Linear regression models in hierarchical bayes structure [9].

3. IMPLEMENTATION OF MODELS

3.1 The Sawhney and Eliashberg(1996) model

The Sawhney and Eliashberg model is used to determine whether a person goes to watch a movie, which depends on various factors that influence their decision. Awareness, together with other movie characteristics such as genre, determines customers' intention to watch the movie [6]. This type of model allows for the analysis and prediction of the life cycle of individual movies. It considers a behavioral model in which the consumer goes through two stages in making his or her decision to attend a movie. The first stage is the time-to-decide and the second stage is the time-to-act [9].

In the time-to-decide stage, the consumer takes time to decide his likeliness to watch a movie. In this stage, the person receives influential information about a new movie like movie media advertising, prices and promotions, featured movie stars, which can determine his decision. Thus, an individual's media exposure and his ability to be persuaded by word of mouth can help determine time-to-decide parameter. The second stage, time-to-act immediately follows the first stage because individuals act after receiving motivational information. Different set of factors influence the time-to-act parameter such as movie-going habits, free time and the willingness to go to a suitable theatre (distance, theatre comfort, screening quality, parking lots and others). The time-to-decide and time-to-act are two independent processes, as the time the person takes to act to his or her previous decision doesn't depend on the time an individual takes to hear or see influential information.

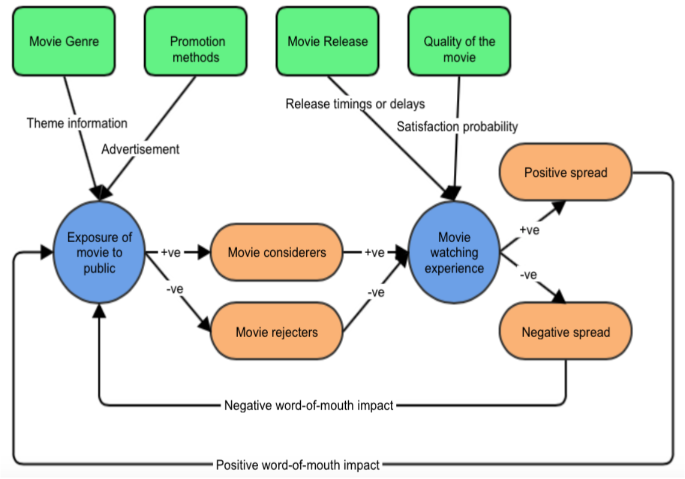

Figure 1 Decision support system proposed by Eliashberg

The probability that the time-to-decide process takes place before time t for any member of the population is exponentially distributed with a parameter λ. The expected time taken to get movie exposure for any individual is 1/ λ. When λ is extremely high, the average time taken to decide whether to watch a movie or not tends toward zero. λ is expected to be particularly large for blockbuster movies where the intensity of information in the marketplace through publicity and media advertisement is extremely high. In this case, people will take quick decisions on whether to watch a movie or not.

Once the individual has made a decision, he or she is assumed to wait for a certain time period before actually adopting (seeing) the movie (i.e., the time-to-act). We assume that time-to-act is exponentially distributed with a stationary parameter γ. We recognize that the stationary assumption maybe somewhat limited in realism, but the stationary time-to-act parameter-based model is appealing from a forecasting perspective because of its parsimony. The parameter γ, is assumed to be an explicit time-varying function of screening intensity [6].

Screening intensity is the number of movie theatres playing a certain movie that changes with time [8]. Thus, changes in time change the time-to-act parameter γ. Therefore, screening intensity can be related to γ by calculating the total number of theatres in the market playing a certain movie. There are various movie release strategies, which determine the screening intensity, which affects γ. One of them is wide-mass release of movies through word-of-mouth, which occupies more theatres in the market and thus influences audience to watch a movie.

Theoretically, the expected time-to-act can approach zero as the time-to-act parameter γ approaches infinity. In practice, however, an intensive screening strategy can reduce the expected time-to-act to a minimum nonzero value, since we would always expect a finite time to act after making a decision to see a movie for reasons that are not related to the movie itself, (e.g., consumers free time availability). Hence, we would expect the parameter γ to be bounded by a finite value. The final expression for the cumulative distribution function of new adopters by time t is

$$Z(t) = \frac{1}{\lambda - \gamma}[(\lambda - \gamma) + \gamma e^{-\lambda t} - \lambda e^{-\gamma t}] \qquad (1)$$

This is the dynamic adoption probability of an individual customer. However, we are ultimately interested in calculating the expression for the cumulative number of adopters as a function of time. Since the unordered points of time at which each of the individual consumers makes the decision to adopt are randomly and identically distributed, the distribution of the cumulative number of adopters by time, N (t), can be written as

$$N\ (t) \sim Binomial\ (N,p)\ where\ p = Z\ (t) \qquad (2)$$

The expected value of the cumulative number of adopters is then the expected value of the above binomial distribution

$$E\ [N\ (t)]\ =\ N.Z\ (t) \qquad (3)$$

Substituting for Z (t) from equation (1), the expression for expected number of cumulative adopters at time t becomes

$$E[N(t)] = \frac{N}{\lambda - \gamma}[(\lambda - \gamma) + \gamma e^{-\lambda t} - \lambda e^{-\gamma t}] \qquad (4)$$

Where N>0 is the potential population size, N (t)>0 is the cumulative number of adopters by time t, E [N (t)]>0 is the expected number of cumulative adopters by time t which represents both the time to decide and the time to act, λ>0 is the time-to-decide parameter and γ>0 is the time-to-act parameter.

3.2 Bayesian hierarchical model

Bayesian hierarchical model overcomes the limitations of the Sawhney and Eliashberg model and it provides three submodels which are linear regression model, standard logit model and nested logit model. The linear regression model predicts the movie attendance by using the parameters in Bayesian approach. It does not capture any substituion effects because it assumes each movie showing is independent of the other movies running at the same time. Adding the extra movie would not affect the the admissions of the existing movie and hence the total demand of watching the movie wil always increase [5].

Linear regresion model makes the assumption of log relation between the number of tickets sold, the predictor variable and the response variable. The formula used is as follows:

$$\ln(S_{jkhd}) = \propto_{jk} X_{jkhd} + \omega Y_d + \epsilon_{jkhd} \qquad (1)$$

where,

$$\propto_{jk} = [\theta_{jk}, \lambda_{jk}, \beta_{jk_{hour}}, \beta_{jk_{hour^2}}, \beta_{jk_{hour^3}}, \beta_{jk_{hour^4}}]'$$

$$\omega = [\omega_{dw}, \omega_{hw}]'$$

$$\epsilon_{jkhd} \sim N(0, \sigma^2)$$

S_{jkhd} is the response variable where k is the version of movie j in hour h of the date d. The predictor variable is decomposed of two components which are X_{jkhd} and Y_d. The first component includes version k of movie j, the age of the movie, and the starting time of the movie showings. The latter component includes the dummy variables of day of weeks and holidays.

The demand expansion and cannibalization effects are not mentioned properly in linear regression model. This problems are overcomed by the **standard logit model.** The choice model is an example of standard logit model and also a good candidate for the above given shortcomings [5]. Choice alternatives are given to the customers where a customer have an option of waching a movie howing at specific time or an outside option of not watching the movie at all. Choice alternatives are represented by a ultility function. The ultility function for choice j for an individual n, U_{jn} consists of two main components which are systematic component (V_{jn}) and random component (ε_{jn}). The formula is as follows:

$$U_{jn} = V_{jn} + \epsilon_{jn} \qquad (2.1)$$

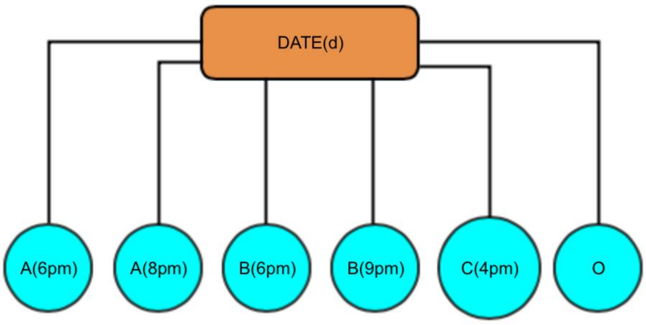

Figure 2 Example of standard logit structure on a given day d

In the above diagram, on a given day d, there are choices of movie and hour combinations as well as not watching movie, which is the outside option. Standard logit model fails to provide intuitive substitution pattern where clash of the movies at same timing may result into watching that movies at equal probability. The standard logit model, P_g, of the probability of choosing choice alternative g in the set containing all movie showings plus the outside option on date d, C_d, is defined as

$$P_g = \frac{\exp\ (\alpha_g X_g + \omega Y)}{\sum_{g \epsilon C_d} \exp\ (\alpha_g X_g + \omega Y)} \qquad (2.2)$$

As the standard logit model has unintuitive substation pattern due to IIA property, **nested logit model** is what said to be the further progress. Nested logit model uses grouping or one can say nesting similar alternatives together within the nest. Nested logit model is used when the choices are further divided into the smaller subsets consisting of elements, which are relatively homogeneous [1].

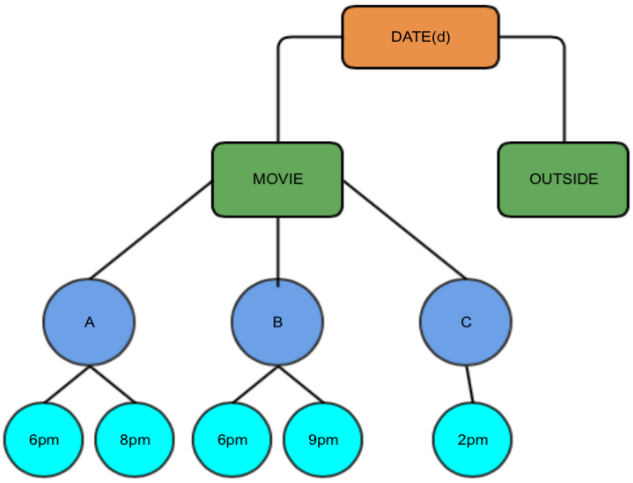

Figure 3 Example of nested logit structure on a given d

On a given date d, watching a movie or doing other activity are two options, which are further sub divided into more option like first, different types of movies and second, different timings of movies. Probability function by nested logit model is given as follows:

$$\begin{aligned} P_g &= P(hour\ h | movie\ j, version\ k, date\ d) \qquad (3) \\ &\times\ P(movie\ j, version\ k | any\ movie, date\ d) \\ &\times\ P(any\ movie | date\ d \end{aligned}$$

4. RESULTS AND DISCUSSIONS

A data set of reviews of 30 movies was taken from IMDB out of which 15 movies were used for training both the methods and remaining 15 movies were used for testing both of the methods. Based on the negative and positive impact factor of the movies accuracy of the given two methods was calculated [10].

4.1 The Sawhney and Eliashberg(1996) model

The data set collected from imdb site consists of information such as (1) when the showing started, (2) how long ago the movies were first released, (3) whether it is played during holidays, (4) what day of week is the showing on and (5) the number of tickets sold. Furthermore, the characteristics of movies such as genre and age restriction are contained in the data set. Tabular representation is as follows:

Movies	Release date	Played on holidays	Genre
Terminator 2	24/8/1991	Yes	Action
Robin Hood	12/5/2010	No	Drama
Die Hard 2	4/7/1990	Yes	Action
Ghost	13/7/1990	No	Horror
Exorcist 3	17/8/1990	Yes	Horror

The time-to-act (γ) and time-to-decide (λ) parameters of 5 movies are as follows:

Movies	No. Of Weeks	N (Total demand)	γ	λ
Terminator 2	24	188	0.55	42.32
Robin Hood	20	166	0.31	57.49
Die Hard 2	15	119	0.43	2.91
Ghost	20	155	0.35	0.34
Exorcist 3	6	28	0.96	0.96

The coefficients in the regression for γ suggest that R-rated movies have higher time-to-act parameters, while movies with major stars and movies receiving positive critic reviews are associated with lower time-to-act parameter that will influence viewers to decide faster. Depending on these parameters coefficient of determination (R^2) calculated is 0.76.

Thus, the Sawhney and Eliashberg model provides an accuracy of approximately 76%.

4.2 Bayesian hierarchical model

In the three Hierarchical Bayes models, the actual movie ticket sales and predicted median of movie ticket sales are compared. The co-efficient of determination (R^2) determines whether the predictions are good or bad.

Model	Linear	Standard Logit	Nested Logit
Existing movies	0.67	0.63	0.70
New movies	0.66	0.66	0.29
All movie	0.71	0.72	0.59

R^2 of the given three hierarchical based models for the provided 30 movies are compared which shows that linear regression model and standard logit model are much better that nested logit model for all movies while nested logit model shows better result if we consider only for the existing movies.

Finally, after testing 15 movies, the best suggested model is standard logit hierarchical model, which shows an accuracy of approximately 72%. Thus the predictions of attendance of 11 movies out of 15 were correct.

5. CONCLUSION

The two different models are compared based on the review data selected from imdb website. The Sawhney and Eliashberg model proves out to be better than the three Hierarchical Bayes models, as its accuracy is 76%, while the accuracy of Hierarchical Bayes model is 72%. The predictions derived from these models are a useful tool for the movie industry and for theatres when deciding the capacity of movie theatres. More methods are available to predict the attendance of the movie theatre, which can be utilized and thus could provide more efficiency. Future implementation of this particular article would be addition of more attributes to get a better precision.

5. REFERENCES

[1] Dellarocas, C., Zhang, X.,& Awad, N.F.(2007). Online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive Marketin.g, 21 (4),23-45.

[2] Hand, C. (2002). The distribution and predictability of cinema admissions. Journal of Cultural Economics, 26, 53-64.

[3] Hand, C., &Judge, G. (2012). Searching for the picture:Forecasting UK cinema admissions using Google Trends data. Applied Economics Letters, 19, 1051-1055.

[4] J. Eliashberg, Q. Hegie, J. Ho, D. Huisman, S. J. Miller, S. Swami, C. B. Weinberg, and B. Wierenga. Demand-driven scheduling of movies in multiplex. Intern. J. of Research in Marketing, 26:75–88, 2009.

[5] A. Gelman, J. Carlin, H. Stern, and D. Rubin. Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton, 2003.

[6] Eliashberg, J., J.-J. Jonker, M.S. Sawhney, B. Wierenga. 2000. MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures. Marketing Science 19(3) 226–243.

[7] Liu, Y. 2006. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. Journal of Marketing 70(3) 74–89.

[8] Pringle, L., R.D. Wilson, E.J. Brody. 1982. NEWS: A Decision Analysis Model for New Product Analysis and Forecasting. Marketing Science 1(1) 1–30. Ravid, S.A. 1999. Information, Blockbusters, and Stars: A Study of the Film Industry. Journal of Business 72(4) 463–492.

[9] Sawhney, M.S., J. Eliashberg. 1996. A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. Marketing Science 15(2) 113–131.

[10] Zufryden, F.S. 2000. New Film Website Promotion and Box-Office Performance. Journal of Advertising Research 40(1) 55–64.