ISSN 2349-4506
Impact Factor: 2.265

Global Journal of Engineering Science and Research Management

A KERNEL BASED APPROACH: USING MOVIE SCRIPT FOR ASSESSING BOX OFFICE PERFORMANCE

Mr.K.R. Dabhade[*1] Ms. S.S. Ponde[2]
[*1]Computer Science Department. D.I.E.M.S.
[2]Asst. Prof. Computer Science Department, D.I.E.M.S.

KEYWORDS: Semantic variable, green lightning (key words).

ABSTRACT
A method to predict performance of box office of movie at time of Green lightning, when only budget and script is available. The level of extraction of textual features from screen writing domain knowledge like Genre & content, natural language processing techniques ie. bag of words and human input as Semantic variables. Textual variable which defines an distance metrics of scripts which used as input for kernel based approach for assessing box office performance.

INTRODUCTION
Movie studios have to choose among thousands of scripts to decide which ones to turn into movies. Despite the large amount of money is invested in movies this process known as "green-lighting" in the movie industry is largely a guesswork based on experts' experience and intuitions[1][3]. In this proposed system [5] a new approach to help studios evaluate scripts which will then lead to more profitable prior decisions. It combines screenwriting domain knowledge, natural language processing techniques, and statistical learning methods to forecast a movie's return-on-investment [4] based only on textual information available in movie scripts [5]. It tests the model in a holdout decision task to show that this model is able to improve a studio's gross return-on-investment significantly. While deciding which scripts to turn into movies (i.e. "green-lighting") movie studios and film makers need to assess the box performance of a movie based only on its script and allocated production budget as most post-production drivers of box office performance e.g., actor, actress, director, MPAA rating are unknown at the point of green-lighting when financial commitments have to be made. Usually movie producers rely on a "comps"-based approach [3] to assess the box office potential of a new script. Specifically, they identify around past movies which are "similar" to the script and we use the box office performance of those movies as benchmarks for the revenue potential .So question is "similarity" between movies scripts should be measured. As instance one should focus on theme the actual words/language used or structure of the scenes and dialogues? The goal is to answer above questions and in develop a decision which helps studios make decisions based green-lighting. We can develop a method based on text mining and the kernel approach that identifies the "comps" of a new script based on its content and textual features, and hence assesses its revenue potential.

The research contribution is in three parts. First that collects and analyzes actual movie scripts. Second show that the kernel approach outperforms both regression and tree-based methods in the context of assessing box office performance. Third the estimated "feature weights" provide some insights about which textual features require particular attention when identifying useful "comps" for a new script. The next section describes an overview of the script data set & how we extract textual information from script and section 3 describes the kernel-based approach and how can we estimate the obtained feature weights. In next section we compare our method with other benchmark methods and present a hypothetical portfolios selection scenario this proposed method can gives lower mean square error.

TEXTUAL FEATURES FROM MOVIE SCRIPTS
Data is comprised of more than 300 movies script which are available online we than record the U.S box office revenue and production budget from IMDB i.e Internet Movie Database.

Genre and Content Variable:
The textual information in movie script can summarize by the "content" variable and genere of scripts summarize by overall theme of movie so genre of script we considered eight genres and the content describes the variable which give detail about script like ending of story is happy or sad?

We considered eight genre based on category of movie as follows

Romance(ROM),Thriller(THR),Drama(DRA),Comedy(COM)Horror(HOR),Family(FAM),Action(ACT) and Sci-fic (SCI). The set of few questions is provided regarding storyline of each script based on genre which questions are simply yes & no type which have been identified by script writing experts.

Semantic Variables

This textual information captures from the scripts of movie an semantic variables is used and it provides a preview that how the final movie will look. The script is organized into interior/exterior scenes whereas each scene is comprised characters dialogue. The semantic variable is second layer of textual information where structure of an script is captured and final preview is provided about the script.

Here we define two level.

(i)At scene level- Here we can obtain total no of scenes in movie & the way how an character interact with co-actor.

(ii) Dialogue level- Here We can obtain the manner how character communicates all information is carried from script.

i) Number of scenes (NSCENE).
ii) Interior scenes percentage (INTPREC).
iii) Number of dialogues (NDIAG).
iv) Average of dialogues length (AVGDIAGLEN).
v) The "concentration index" of dialogues (DIAGCONC).

We use HH index to compute the concentration index of dialogues. The value of HH index is between 0 & 1.The higher index indicates concentration of a few characters in a dialogues.

Bag-of-Words Variables

The bag of words is third layer of textual information by using natural language processing technique. The words used in scripts and frequencies of their usage are backbone of story-line.

We can extract bag of words through scripts using the following steps.
(i)We then eliminate all punctuation as stop words and a Standard English names.
(ii)A stemming algorithm is used for reducing words to simplest form.
After the eliminating stemming & stop words even though there are thousands of unique words appeared in one or more scripts. Hence we compute an "importance index" for each word.

$$I_i = \left(1 - \frac{d_i}{D}\right) \times N_i \quad \ldots\ldots (1)$$

Above formula is used to measure importance index where d_i denotes no of scripts which contains i_{th} Word. And N_i is total frequency occurrence of i_{th} word. We keep few 100 words as important words and finally we perform LDA to further reduce dimensional of the words document matrix. Based on singular-value decomposition (SVD) it provide us to index each script by a set of "scores".

Summary and Potential Data Limitations

Summary statistics for each variable in data set is taken. All textual variables and the (log-) production budget is considered and used as predictors in a kernel-based approach which forecast box office performance.

A KERNEL-BASED APPROACH TO FORECAST BOX OFFICE PERFORMANCE

The kernel-based method utilizes a distance metric to Identify the "similarity" between a new observation and each observation in the training database. The kernel- based approach is free of functional form this allow flexibility to capture complex relationship between features in textual script and box office performance. So we feel that kernel based approach is appropriate & correct relationship between textual variable of scripts & box office. Another approach of kernel based is it is business friendly as we can directly communicate to studio manager.

Textual Variable	MAX	Mean	SD	Min
GENRE_DRA	1.00	0.55	0.44	0.0
GENRE_ROM	1.00	0.41	0.33	0.0
GENRE_COM	1.00	0.28	0.36	0.0
GENRE_HOR	1.00	0.25	0.41	0.0

Figure2.1Table Summary statistic of variables

KERNEL BASED APPROACH

With use of following notations scripts in the training sample are indexed by i 1 .. .N. Each script is comprised of J distinct "features" and is denoted as X long with a "response" variable y_i.We define the response variable for each movie by its (transformed) return of investment (ROI). Specifically:

Yi = log (BOX OFFICE i/ BUDGET) ---------- (2)

We specify (transformed) ROI as the response variable in the kernel based method because such specification confers several statistical advantages. First the distribution of y_i is much closer to normality than box office revenues which has a heavy right tail. The Notation based y_i is response variable we define it for each movie y_i is much closer to normality than box office revenue the features we consider here are the textual variables extracted from each script along with its production budget. The distance metric between two observations is defined, based on (weighted) Euclidean distance as follows:

$$d(\vec{x_i}, \vec{x_l}) = \sqrt{\sum_{i=1}^{j} v_j^2 (x_{ij} - x_{lj})^2} \text{ ----------(1)}$$

is a vector of "feature weights".As shown that the conceptual argument above we set the value of θ by appealing to studios' domain knowledge. The studio managers typically look at no more than 10 "comps" when making a green-lighting decision. Therefore, we select θ such that any "comp" beyond the 10th will receive minimal weight this is achieved by setting θ so that on average the 10th comp receives a weight that is proportional to the density of a standard normal distribution at two standard deviations from the mode, Hence the 11th or further comps have weights that are negligible.

Featured Weight calibration ($\vec{v}$):

The calibrated featured weight $\vec{v}$ as a starting point a reasonable "default choice" is to put equal weight on every variable i.e V_j =1 . We refer it as Kernel-I approach. We will evaluate its predictive performance verses kernel-II Approach that involves features weight. The proposed approach is based on cross validation to calibrated features weight $\vec{v}$ for kernel – II approach. We define Leave one out mean squared error, LOOMSE a key component of our objective function. We let i=1…n(n=265)index the scripts in training sample & let $\hat{z}_i(\theta, \kappa, \vec{v})$ be the predicted value of the log box office revenue of i^{th} script when all except the i^{th} script are used as training data. Z_i denotes actual log box office revenue for i^{th} script.

$$\text{LOOMSE}(\theta, \kappa, \vec{v}) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i(\theta, \kappa, \vec{v}))^2 \text{----------(2)}$$

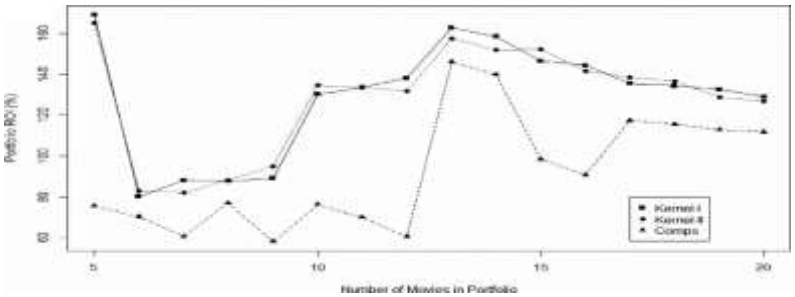
Portfolio Selection

Now we demonstrate the potential economic significance of our proposed method & we conduct a hypothetical portfolio selection exercise so that we can compares the performance of the comps-based approach with our proposed Kernel-I/II methods. We consider the following portfolio selection setting. Suppose we would like to pick r scripts to form a movie portfolio.

First, based on the predicted box office revenue and the given production budget, we compute the predicted ROI of each of the 35 scripts in the holdout sample. Then scripts in the holdout sample are ranked based on predicted ROI, and the r scripts that have the highest predicted ROI are selected. We vary from 5 to 20 and compare the ROIs of the overall portfolios which are selected by the comps based method Kernel-I and the Kernel-II method, respectively. The results are shown in Fig. 2. While there is a lot of variability in portfolio ROIs ((total box office – budget)/budget) across all methods, portfolios selected by Kernel-I and Kernel-II approaches consistently provide higher portfolio returns compared to those selected by the comps-based method. when r = 10 movies scripts are selected to form a portfolio, the selections by Kernel-I and Kernel-II method yield portfolio ROIs of 130.3 percent (Box office = $1184.7M; Budget = $514.5M) and 134.6 percent (Box office = $1236.3M; Budget = $527.0M), respectively, while the selection by the comps based method yields a ROI of 76.4 percent (Box office =$307.8M; Budget = $174.5M).8 Across different values of r(from 5 to 20), the median ROI of portfolios selected by Kernel-I and Kernel-II is around 134.0 and 134.1 percent (respectively), while the median ROI of portfolios selected by comps-based method is only around 83.9 percent. Thus it is clear that the improvement in prediction accuracy afforded by the Kernel-I/II methods is also economically significant.

	Kernel1	**Kernel-II**
Bag of Words	0.4300	0.4021
Semantics	0.444	0.4219
Original result	0.4096	0.3822

Table 3: Holdout Predictive Performance (in terms of MSE) for Kernel I and Kernel-II



CONCLUSION

The paper consist a methodology which is depend on the kernel-based approach to predict the box office potential of movie scripts at the point of green-lighting with lowest mean square error by which it can possible to access box office performance using movie scripts.

REFERENCES

1. J. Eliashberg, S.K. Hui, and Z. John Zhang, “From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts,” Management Science, vol. 53, no. 6, pp. 881-893, 2007 .
2. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data:
3. *Mining Online Reviews for Predicting Sales Performance in the Movie Domain*:
4. Online Review Mining For Forecasting Sales:
5. H. Chipman, E. Geroge, and R. McCulloch, “BART: Bayesian Additive Regresion Trees,” The Annals of Applied Statistics, vol. 4, no. 1, pp. 266-298, 2010.
6. S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.Harshman, “Indexing by Latent Semantic Analysis,” J. Am. Soc. for Information Science, vol. 41, no. 6, pp. 391-407, 1990.
7. H. Mukerjee, “Nearest Neighbor Regression with Heavy-Tailed Errors,” The Annals of Statistics, vol. 21, no. 2, pp. 681-693, 1993.
8. J. Eliashberg, S.K. Hui, and Z. John Zhang, “From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts,”Management Science, vol. 53, no. 6, pp. 881-893, 2007.
9. J. Eliashberg, C. Weinberg, and S. Hui, “Decision Models for the Movie Industry,” Handbook of Marketing Decision Models,pp. 437-468, Springer, 2008.
10. E.J. Epstein, The Big Picture: The New Logic of Money and Power in Hollywood. Random House, 2005.

ISSN 2349-4506
Impact Factor: 2.265

Global Journal of Engineering Science and Research Management

11. S. Field, Screenplay: The Foundations of Screenwriting. third ed., DellPublishing, 1994.