Green-lighting Movie Scripts:

Revenue Forecasting and Risk Management

Jehoshua Eliashberg

Sam K. Hui

Z. John Zhang*

May 6, 2010

* Jehoshua Eliashberg is Sebastian S. Kresge Professor of Marketing and Professor of Operations and Information Management at the Wharton School of the University of Pennsylvania. Sam K. Hui is an Assistant Professor of Marketing at Stern School of Business, New York University. Z. John Zhang is Murrel J. Ades Professor and Professor of Marketing at the Wharton School of the University of Pennsylvania. Corresponding author: Jehoshua Eliashberg (eliashberg@wharton.upenn.edu).

Green-lighting Movie Scripts:

Revenue Forecasting and Risk Management

Abstract

Major studios and independent production firms (Indies) often have to select or "green-light" a portfolio of scripts to turn into movies. Despite the huge financial risk at stake, there is currently no risk management tool they can use to aid their decisions, even though such a tool is sorely needed. In this paper, we developed a forecasting and risk management tool, based on movies scripts, to aid movie studios and production firms in their green-lighting decisions. The methodology developed can also assist outside investors if they have access to the scripts. Building upon and extending the previous literature, we extracted three levels of textual information (genre/content, bag-of-words, and semantics) from movie scripts. We then incorporate these textual variables as predictors, together with the contemplated production budget, into a BART-QL (Bayesian Additive Regression Tree for Quasi-Linear) model to obtain the posterior predictive distributions, rather than point forecasts, of the box office revenues for the corresponding movies. We demonstrate how the predictive distributions of box office revenues can potentially be used to help movie producers intelligently select their movie production portfolios based on their risk preferences, and we describe an illustrative analysis performed for an independent production firm.

Keywords: entertainment industry, movie production, text mining, machine learning, risk management, portfolio selection.

1. Introduction

For movie studios and independent producers, "green-lighting" or deciding on a movie production portfolio is an important yet challenging task. Movie production typically involves three phases: development, production, and post-production. Green-lighting a movie represents a transition from the development to the production phase and entails enormous financial commitment. In making this critical decision, studios executives of all divisions evaluate a tentative budget that is based on shot-by-shot breakdown of the script (Epstein 2005). Usually, each of the major studios keeps a slate of around 100 to 400 films (or film ideas) at a time in development (Waterman 2005); from these potential ideas, each major studio then generates a portfolio of around 12 to 25 movies a year.[1] The production process is similar for independent production firms, albeit at a smaller scale (Marich 2005).

For both major studios and independent production firms, the main difficulty that they are facing in creating the desired movies portfolio is that very little information about the ultimate success or failure of a movie is known before production commitments need to be made. This problem with the lack of information is magnified even further by the high (and rising) cost of producing and marketing a movie. In 2007, for instance, movie studios spend around $70M to produce a movie on average (MPAA 2007). As a consequence, content providers often have to make decisions with high financial impact to their companies without knowing precisely what the potential risk-payoff structure is. Not surprisingly, industry insiders often liken green-lighting a movie to a crapshoot and the huge variation of box office revenues across movies is consistent with the random nature of green-lighting movies (Walls 2005).

[1] For instance, specified in regulatory filing in 2005, Disney released 21 films per year, while Warner Bros released 25 films per year (Marich 2005).

While plenty of academic research focuses on the movie industry, very few provide useful guidance to the green-lighting problem. There are two main reasons why current research fails to provide such an adequate decision support system. First, most researchers focus on forecasting box office revenue *after* a movie has been produced and when more tangible predictors are available. Typically, researchers forecast box office performance based on box office receipts in the early weeks (e.g., Sawhney and Eliashberg 1996), or post-production/pre-release information (e.g., Neelamegham and Chintagunta 1999; Eliashberg et al. 2000; Shugan and Swait 2000). These studies are helpful for movie distributors, but do not provide any direct guidance for producers' (studios') movie production decisions at the green-lighting stage as well as for investment decisions made by external investors.

Second, almost all academic research studies focus on producing point estimates of box office revenues, but do not give a full characterization of the predictive distributions that are crucial for managing production portfolios. Much like equity investors who not only care about the expected return but also the variance and risk of a stock portfolio, studios are not only interested in the expected upside/downside of a movie investment (i.e., point predictions), but also the uncertainty in box office revenues, quantified by their full predictive distributions.

In this paper, we propose a new methodology that tackles the two aforementioned issues together. We develop an economically meaningful tool for green-lighting decisions and risk management at this early stage of the movie production. Two features characterize this new methodology. First, we restrict ourselves to only information that is available at the point of green-lighting decisions, i.e., only movie scripts and their corresponding estimated production budgets. While there have been some previous attempts to link textual information to commercial performance of movies (e.g., Eliashberg, Hui and Zhang 2007), such work simulates

the process of adding film ideas to the development pool. This is so because these ideas are typically "pitched' to the studios as 'treatments" before the full-fledged scripts are written. To the best of our knowledge, no previous research has come as close to modeling the green-lighting practice (i.e., the transition from the development to the production phase based on the full-fledged scripts). Towards this end, we compile and analyze a large dataset, which consists of actual, full-fledged movie scripts. As a script is the very foundation of a movie, a sophisticated analysis of the textual information and hidden story structures in the script help us better predict box office revenues. We discuss it in more detail in Section 2.

Second, in our analysis, we fully characterize not only the expected return but also the uncertainty of box office revenues. In addition, our analysis also recognizes and allows for highly nonlinear interactions among different ingredients in a movie script. For both purposes, we employ a recently developed Bayesian model known as Bayesian Additive Regression Tree or BART for short (Chipman et al. 2007, 2008), and extend it to handle quasi-linear models (henceforth refer to as BART-QL) that are more appropriate for the task at hand.

After building up our model and validating its (holdout) predictive performance on actual data, we demonstrate how our model can be used to aid content providers' portfolio optimization decisions. Using the posterior prediction distributions of box office revenues, we illustrate how the mean-VaR (mean-value-at-risk) efficient frontier (e.g., Alexander and Baptista 2001) of movie portfolios can be derived. Using this efficient frontier, movie producers can manage their risk exposure based on their risk preferences and budgetary restrictions.

To summarize, Figure 1 shows a general overview of this article. The rest of this paper describes each aspect of Figure 1 in detail and is organized as follows. Section 2 describes how we assemble our dataset, including box office revenue and budget data (Section 2.1), the

genre/content variables (Section 2.2), bag-of-words variables (Section 2.3), and semantic variables (Section 2.4). Section 3 develops the BART-QL methodology that allows us to model box office revenues as a function of production budget and textual variables from scripts. Section 4 compares the predictive capabilities of our methodology vis-à-vis other benchmark models, and Section 5 discusses how our model can be used for portfolio optimization and risk management. Section 6 concludes with directions for future research.

[Insert Figure 1 about here]

2. Extracting textual information from movie scripts

Our dataset contains a sample of 200 movies released from 1995 to 2006, whose shooting scripts are available online in electronic format. Since our dataset includes only scripts that are already made into movies, we may run into the problem of "sample selection bias" or "covariate shift" (Huang et al. 2007), if the features of scripts that are made into movies are different from the features of scripts in general (including those that are not produced into movies). This is a common problem that is also shared in machine learning (Huang et al. 2007; Sugiyama et al. 2007; Zadrozny 2004), bioinformatics (Baldi and Brunak 1998), and econometrics (Heckman 1979). However, in our case, this is actually less of a problem given that the process of green-lighting a few movies out of a large number of choices has not been aided by anything other than the readers' intuitions, and that the variation in the distribution of box-office revenues is very wide.

For each movie in our dataset, we record its domestic box office revenue and its production budget from the IMDB database. From each script in our sample, we extract three layers of textual information: genre/content variables (Section 2.2), (ii) bag-of-words variables

(Section 2.3), and (iii) semantic variables (Section 2.4). The complete set of variables collected from each script is summarized in Table 1, and described in detail in the following sub-sections.

[Insert Table 1 about here]

2.1 Box office revenue and production budget

We collected the box office revenue and production budget for each movie in our sample. The histograms of box office revenue and production budget, both in absolute and in log scale, are shown in Figure 2. The corresponding summary statistics are shown in Table 3. We see that the distribution of both box office revenue and production budget appears closer to normal distribution in the log-scale, suggesting that a log-transform is appropriate.

[Insert Figure 2 about here]

Figure 3 shows a scatterplot of log-box office revenue against log-production budget. As can be seen, there is a significant positive relationship between box office and production budget, with a correlation of 0.70 ($p<.001$). A linear regression of (log-) box office revenue and (log-) production budget, as shown in Figure 3, is estimated with an intercept of 0.12 and a slope coefficient of 1.01 ($p<.001$), with a corresponding R^2 value of 0.48.

[Insert Figure 3 about here]

The strong linear relationship between (log-) box office and (log-) production budget suggests that our model should incorporate (log)-production budget as a linear effect. In Section 3, we describe how our model takes into account this relationship using a quasi-linear specification.

2.2. Genre and content variables

The highest level of textual information in movie content can be summarized by its genre and its "content" variables (Eliashberg et al. 2007). The genre of a script summarizes the theme of the movie and helps identify the size of the target segment for the movie. The "content" variables measure various aspects of the storyline of a script (e.g., premise, setting, conflict, resolution, ending).

We asked three independent readers, who are trained in film studies, to read each script and answer a questionnaire about the genre and the storyline. "Genre" is a categorical variable with eight possible categories (Drama, Romance, Thriller, Comedy, Horror, Scifi, Action, Family) that describe each movie. Note that a movie can be described by more than one category. For instance, a movie can belong to both "Drama" and "Romance" genre categories.

Readers then answer a set of 23 "content" questions about the storyline for each script. The list of questions is shown in Table I of Appendix I. These questions are simple "yes or no" questions that have been identified by experts as important to writing a successful movie script (e.g., Monaco 2000). It is very unlikely that any of the readers provided biased answers to this set of questions because of their possible familiarity with a movie, since the content variables and the box office success do not have a linear, additive relationship. For instance, a clear or important premise needs not enhance the prospect of a movie's success; instead, it may be the case that an important premise is helpful only for action movies that also have a "stronger nemesis" with "character growth". This is why, as it will become clear soon, we specify a model that allows complex nonlinear interactions among content variables. In this context, it would be impossible for anyone to know what a successful movie imply about the content variables of the

movie's script. In fact, when we plot each content variable against the box office, we do not find any discernable pattern.

As a final step to generate content variables, we averaged the three readers' responses for each question. Note that before averaging their responses, we studied the inter-rater agreement on the genre and content questions; on average, we find reasonable agreement among the three readers using Fleiss's kappa (Fleiss 1971). The details are shown in Appendix II.[2]

2.3. Bag-of-words variables

The second layer of textual information we extract comes from the actual words used in the script, captured using a "bag-of-words" representation (i.e., representing a script by a list of words with their associated frequencies). Words used in a script and their usage frequencies are the building blocks of a script and indicative of a storyline. In particular, the frequencies of key words and phrases (e.g., love, die, sex, blood) may be indicative of the overall tone and theme of the movie, beyond that captured by genre and content variables alone. Reducing a document to its bag-of-word representation has been used successfully in natural language processing applications, such as document retrieval/document classification and organization (e.g., Blei et al. 2003; Lewis 1998; Li and Jain 1998). We use the following procedure to extract bag-of-word information from each script so that we can reduce the dimensionality of these variables, while still picking up significant information. First, we use a "stemming" algorithm in a natural language processing package (Porter 1980) to reduce each word to its simplest form (e.g., "going" is reduced to "go"). We then tabulate all the unique stemmed words that occur in any

[2] We also estimated a version of our model (Section 3) where only content variables that have Kleiss's kappa higher than 0.3 are retained. The "pruned" model provides inferior holdout predictive performance compared to the model that uses all content variables. This seems to suggest that disagreements among the readers (i.e., heterogeneity) may have some predictive value.

document, and count the occurrence of each word in each document to produce a word-document matrix.

Next, we compute the "importance index" for each word; the importance index is defined as follows:

$$I_i = \left(1 - \frac{d_i}{D}\right) \times N_i \qquad [1]$$

where d_i denotes the number of scripts containing the i-th word, D denotes the total number of scripts, and N_i is the total frequency of occurrence of the i-th word across all scripts.

We then keep only the 30 most important words to allow for a stable factor analysis solution (MacCallum et al. 2001). Finally, we perform factor analysis on the word-document matrix to further reduce its dimensionality. As shown in the screeplot in Figure 4, the factor analysis shows an "elbow" between factor 2 and factor 3 (Johnson and Wichern 2007). Thus, we keep a two-factor solution, which explains 35.6% of the variance of the word-document matrix. The factor loadings of each factor on the 30 words are shown in Table 2. We extract the two factor scores for each script and use them as predictor variables in our model.

[Insert Figure 4 about here]

[Insert Table 2 about here]

2.4 Semantic variables

The third layer of textual information from scripts comes from the "semantics" information. The way a script is organized may give us some insights into how the final movie will look like; for example, a script can tell us approximately how many scenes are in the movie, how often the characters interact in interior or exterior space, etc.. It may also carry information

about the way characters speak in the movie; for instance, whether characters will give long prose, or just short dialogues, and how evenly distributed these dialogues are among the characters. To capture this semantic information, we focus on scene variables (ii) and (iii) and dialogue variables (iv)-(vi) below.

(i) Number of words in title (NTITLE)

(ii) Total number of scenes in the script (NSCENE)

(iii) Percentage of interior scenes (INTPREC)

(iv) Total number of dialogues (NDIAG)

(v) Average length of dialogues (AVGDIAGLEN)

(vi) The "concentration index" of dialogues (DIAGCONC)

To obtain the concentration index of dialogues (variable vi), we use the Herfindahl-Hirschman index (Hirschman 1964) of the share of dialogue that each character has in the script. Let $s_i \in [0,1]$ be the share of the dialogue by character i ($\sum_i s_i = 1$). The Herfindahl-Hirshman concentration index of dialogue is defined as:

$$DIAGCONC = \sum_i s_i^2 \qquad [2]$$

These variables say something about the theme and pace of a storyline and also about how the story is being told.

To sum up, the summary statistics of each variable in our dataset are shown in Table 3. We now proceed to discuss our statistical learning procedure, which uses production budget and the textual information described in this section to generate the predictive distributions of box office revenues.

[Insert Table 3 about here]

3. Bayesian additive regression tree for quasi-linear model (BART-QL)

After we extract textual information from the scripts, we use them as covariates, together with production budget, in a predictive model of box office revenue. Section 3.1 describes the BART-QL model; in particular, it discusses how the methodology allows for a flexible semi-parametric modeling of box office revenue while allowing for interactions among predictors. Such flexibility and interactions are two pre-conditions for predicting a movie's success. Section 3.2 outlines our computational procedure used to sample from the posterior distribution of our model parameters.

3.1. BART-QL model

Our BART-QL model specification is a semi-parametric model that is comprised of two parts. As shown in Section 2.1, (log-) production budget is linearly related to (log-) box office revenue. That is, the magnitude of box office revenue is controlled, to a moderate extent, by the size of the production budget. We add to it the other textual information extracted from the script. Formally,

$$\log(y_i) = \alpha + \beta \log(z_i) + f(\vec{x}_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad [3]^3$$

where y_i denotes the box office revenue of the *i*-th movie; z_i denotes the production budget of the *i*-th movie, and $\vec{x}_i$ denotes the vector of textual covariates for the *i*-th movie; ε_i is a random error term that is assumed to be independent and normally distributed with mean zero and variance σ^2. α and β are two free parameters (estimated in our model) for intercept and slope,

[3] Note that if the data exhibits heteroskedasticity, Equation [3] can be generalized to $\varepsilon_i \sim N(0, \sigma_i^2)$ to handle the unequal variances across observations. In our empirical analyses we found that this was not the case and the homoskedasticity assumption was a reasonable approximation for our dataset.

respectively; thus, the expression $\alpha + \beta \log(z_i)$ captures the linear effect of (log-) production budget on (log-) box office revenue in the first component of our model. In the second component of our model, the excess/under residual in box office revenue is captured by other covariates based on the textual information extracted from the scripts. Given the highly nonlinear interactions among predictors, a flexible tree-based specification is used to model the effect of textual covariates. The function $f(\vec{x}_i)$ (described next) denotes the effect of the other textual covariates.

We model $f(\vec{x}_i)$ using a flexible BART (Bayesian Additive Regression Tree) specification proposed by Chipman et al. (2007, 2008). The BART specification can be explicitly expressed as:

$$f(\vec{x}) = g(\vec{x}; T_1, M_1) + g(\vec{x}; T_2, M_2) + \cdots + g(\vec{x}; T_N, M_N) \qquad [4]$$

where each $g(\vec{x}; T, M)$ denotes a tree specification; *T* denotes a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and *M* denotes a set of parameter values associated with each of the terminal nodes of *T* (See Chipman et al. 2007, 2008 for details). Thus, the specification in Equation [4] can be viewed as a sum-of-trees model.

The BART specification used in Equation [4] has been widely applied as a flexible modeling/predictive method in various settings. For instance, Zhang et al. (2007) applied a spatially-adjusted version of BART for data fusion and used it to analyze data on health status and income. Zhang and Hardle (2008) used BART to model default risk and predict corporate insolvency. Kourtellos et al. (2007) exploited the flexible nature of BART modeling to test whether the relationship between aid and economic growth is nonlinear. In biology, Zhou and Liu (2008) and Liu and Zhou (2007) compared different machine learning methods in an application designed to predict how and where transcription factors interact with DNA. They

found that BART outperforms other methods (stepwise linear regression, multivariate adaptive regression splines, neural networks, support vector machines) both in terms of accuracy and robustness. In computer science, Abu-Nimeh et al. (2008) applied BART for automatic detection of phishing emails, and found that BART outperforms logistic regression, random forests, support vector machines, CART, and neural networks.

Putting together Equation [3] and [4], our BART-QL model can be formally expressed by Equation [5] as follows:

$$\log(y_i) = \alpha + \beta \log(z_i) + \sum_{n=1}^{N} g(\vec{x}_i; T_n, M_n) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad [5]$$

Towards our goal of predicting box office revenue from scripts, the BART-QL formulation specified by Equation [5] has a number of key advantages. First, given that the sum-of-tree model is essentially an additive model with multivariate components, it can easily handle both additive effects (through the summation of trees), as well as interaction effects (which can be captured by each individual tree) of varying orders (Chipman et al. 2007). The specification in Equation [4] can thus be viewed as a generalization of both the multiple linear regression model and the single-tree model (e.g., Chipman et al. 1998); hence it allows for more flexibility, which is crucial for the modeling of box office revenue based on textual variables, where interactions are highly significant. As will be shown in Section 4, this additional flexibility allows us to capture the variations in box office revenue more accurately than alternative models.

Second, unlike other machine learning methods (e.g., neural networks, support vector machine, lasso, MARS; see Hastie et al. 2001) which are essentially "black-box" prediction methods that generate only point estimates, the BART-QL model is a fully specified Bayesian model with a proper prior distribution and a likelihood function. Through the fully specified Bayesian model, we can directly obtain not only point predictions of box office revenues, but

also their posterior *predictive distributions*. As we will show in Section 5, these predictive distributions can be used to help manage risk and optimize movie production portfolios.

Third, because of the sum-of-trees specification, results from BART-QL can easily be interpreted to obtain substantive insights. After obtaining the posterior distribution of the model parameters, one can generate partial dependence plots (Friedman 2001), which summarizes the marginal effect of a predictor on the response,[4] to study the relationship between each textual variable and box office revenue. In Section 4.2, we will study the estimates from the BART-QL model in detail.

3.2. Prior specification and computational procedure

To complete our Bayesian model, we need to specify a set of prior distributions on our model parameters. For the parameters of the BART model, we use the default "regularization priors" used in Chipman et al. (2007). The general idea is that these priors penalize larger trees and thus allow each tree to make a small contribution to overall fit, thereby delivering a strong predictive performance in the spirit of "boosting" (i.e., using many weak learners to improve estimation; see Hastie et al. 2001). For more technical details, readers are referred to Chipman et al. (2007, 2008). For the regression parameters, we apply conjugate, weakly information priors on α and β; specifically,

$$\alpha, \beta \sim N(0, 100^2). \qquad [8]$$

[4] More precisely, let $f(x) = f(x_s, x_c)$ where the set of predictors x has been partitioned into the predictors of interest, x_s, and the complement $x_c = x \setminus x_s$, the partial dependence function is defined as $f(x_s) = \frac{1}{n}\sum_{i=1}^{n} f(x_s, x_{ic})$ where x_{ic} is the i-the observation of x_c in the data. (For details, refer to Friedman (2001) and Chipman et al. (2008)).

Next, we outline the MCMC procedure we used to obtain the posterior distribution of the BART-QL model parameters, and hence the posterior predictive distributions of the movies' box office revenues.

We initialize α_0, β_0 by running a simple linear regression of log(*y*) on log(*z*) on the training sample; then, with this as a starting point we start the MCMC iterations. We use a two-stage Gibbs sampler (Casella and George 1992), which consists of two main steps: (I) draw $f(.)$ and σ^2 conditional on α and β, and (II) draw α and β conditional on $f(.)$ and σ^2.

More specifically, in the (*t+1*)-th iteration, we take two steps. In the first step, we draw $f(.)$ and σ^2 conditional on α and β. To do that, we first subtract the linear terms, $\alpha_t + \beta_t \log(z)$, from the left side. Thus, we have $\log(y) - \alpha_t - \beta_t \log(z) = f(x) + \varepsilon$. Then, we draw the tree parameters of BART (i.e., *T* and *M*) using the MCMC backfitting procedure proposed by Chipman et al. (2007), which is implemented in the BART package in R. We store the resulting draw as $f_{t+1}(.)$ and σ^2_{t+1}. For the second step, we draw α and β conditional on $f(.)$ and σ^2. This is done by first subtracting the $f(.)$ from the left-hand side. That is, $\log(y) - f_{t+1}(x) = \alpha + \beta \log(z) + \varepsilon$. Then, given conjugate prior on α and β, the full conditional distribution of α and β of standard conjugate form; we can use a Gibbs sampler to sample from their full conditional distribution. The resulting draws of α and β are stored as $\alpha_{t+1}, \beta_{t+1}$.

We then iterate these two steps for 2000 iterations, drop the first 1000 iterations as "burn-in" iterations (Gelman et al. 2003), and store the last 1000 iterations as a sample from the posterior distributions of $(\alpha, \beta, f(.), \sigma^2)$.

4. Empirical results

In this section, we study the performance of our BART-QL approach in predicting box office revenues, based on the rich information available in full-fledged scripts. We begin by examining the in-sample fit of our model, and next assess and discuss the out-of-sample predictive performance of our approach in comparison to other methodologies, both in terms of point estimates and predictive distributions in Section 4.1. In Section 4.2, we interpret the coefficients and parameters of our model. Section 4.3 describes how our model results are used to compute the risk-adjusted profit of each movie as well as the relationship between risk adjusted gross profit, the genre, MPAA ratings, and the studio.

4.1. Model validation (in-sample) and out-of-sample prediction

We first investigate the in-sample fit (validation sample of size 200) of our model. We obtain the posterior distribution of our model parameters using the MCMC procedure described in Section 3.2, and plot the actual (log-) box office revenue versus the corresponding model-fitted values in Figure 5.

[Insert Figure 5 about here]

Given the flexibility of our BART-QL approach, our model should be able to describe the data very well. As can be seen, our model provides an excellent within-sample fit of the box office revenues. The R^2 value is 0.663, which is much higher than the R^2 value (0.480) obtained by a simple linear regression with only (log-) production budget as the independent variable, as shown in Section 2.1. This indicates that the textual variables extracted from the movie scripts do provide additional information about box office revenues, above and beyond the information provided by production budget alone.

Next, we study the out-of-sample predictive ability of our model, using ten-fold cross validation (Hastie et al. 2001), which provide an unbiased assessment of out-of-sample mean squared prediction error. To perform ten-fold cross validation, we randomly divide the 200 movies in our dataset into ten subgroups, each group with 20 movies. Our model then cycles through the data ten times; in each pass, a different subgroup of movies is used as the holdout sample, and the rest is used as training data to estimate the model.

Using the above procedure, we compare the out-of-sample prediction of our BART-QL model with all variables (i.e., genre/content, words, and semantics) versus six reduced models that use only a subset of the variables (i.e., (i) content and words, (ii) content and semantics, (iii) words and semantics, (iv) content only, (v) words only, (vi) semantics only), as well as a regression model that uses only production budget as predictor variables. We compare the accuracy of the point estimates across model using MSE (Mean Square Error) and MAE (Mean Absolute Error) criteria, and we compare the accuracy of the predictive densities using predictive log-likelihood (e.g., Bjornstad 1990). Predictive log-likelihood refers to the log-likelihood of the observed box office revenue under the posterior predictive distribution. The results are summarized in Table 4 below; the predicted box office revenues under the full data-based model (along with 95% posterior intervals[5]) are plotted against the actual box office revenues in Figure 6. As can be seen, most of the actual box office revenues are covered by the 95% posterior intervals.

[Insert Table 4 about here]

[Insert Figure 6 about here]

[5] We plot the 95% posterior highest probability density (HPD) intervals (Chen and Shao 1999), defined as the narrowest interval that covers 95% posterior probability; it is appropriate to consider HPD intervals given that the predictive distribution is not symmetric.

From Table 4, we find that the all variables-based model (content, word, and semantics variables) outperforms all of the seven alternative (reduced) models, both in terms of point estimates and predictive distributions. Our full model has the lowest MSE and MAE compared to all seven reduced models. In addition, the predictive log-likelihood of the full model, -267.27, is much higher than that of the regression model with production budget as covariates (-284.80); this corresponds to a Bayes factor of 17.53, which indicates very strong posterior evidence in support of our model (Berger 1985). Thus, our model not only provides more accurate point predictions of box office revenue, but also more accurate predictive densities, which is crucial for making portfolio optimization and risk management decisions.

Furthermore, we compare the out-of-sample predictive performance of the BART-QL model (with all variables) against other statistical learning methodologies, again using ten-fold cross validation. We include the following list of benchmark models for comparison: (i) the Bag-CART model used in Eliashberg et al. (2007), (ii) the naïve projection method used in Eliashberg et al. (2007) as a benchmark, (iii) linear regression with only production budget as a covariate, (iv) linear regression with all variables, and (v) stepwise regression. The results are shown in Table 5 below. As can be seen, the proposed BART-QL model outperforms all the benchmark methods.

[Insert Table 5 about here]

4.2 Parameter interpretation

Having validated both the in-sample fit and out-of-sample predictive ability of our model, we now move on to interpret the model parameters.

First, we examine the posterior estimates that correspond to the intercept (α) and slope estimates (β) for the (log-) production budget in Equation [5]. The posterior mean of α is -2.60,

with a 95% confidence interval of (-4.12, 0.94). The posterior mean of β is 0.908, with a 95% posterior interval of (0.719, 1.113). This confirms, as expected, that there is a strong positive correlation between (log-) box office revenue and (log-) production budget. Further, since the posterior mean of beta is less than 1, the box office revenue exhibits a diminishing return to scale with respect to production budget, which provides support for the need to incorporate and estimate β in our model framework.

Next, we turn our attention to the estimates from the textual information part of the model. As we stated earlier, a main advantage of the BART-QL approach, in addition to excellent predictive performance, is its interpretability. We focus on interpreting the relative importance and marginal effects (Chipman et al. 2008) of the textual variables. For interpretation purposes, we fix the estimates of α and β at their posterior means and focusing on the posterior distribution of the trees.

We examine the relative importance of each textual variable by looking at the number of times it is used in the trees across all iterations. If a variable is included in a higher percentage of iterations, it is more important in predicting box office revenue (Chipman et al. 2008). The relative importance of the top 10 textual variables, based on the percentage of iterations (out of 1000 iterations) that use each variable, is shown in Table 6. The results indicate that the most important textual variable is GENRE_ACT (Genre: Action), which are used in all of the iterations, followed by BUILD (Conflict build up) and MULCONF (Multidimensional conflict), which are used in 82.4% and 72.4% of the iterations, respectively. However, this does not mean that these three variables are the most important among the textual variables in contributing to a movie's box office success. What it does mean is that they are the three most relevant variables in predicting a movie's box office performance, good or bad. We also see that the top 10

variables include variables from each of the three types (content, word, semantics), as indicated in Table 6, providing additional evidence that all three types of variables are important in prediction box office revenues accurately.

[Insert Table 6 about here]

We study the effects of these top 10 variables in more detail by looking at their marginal effects on box office, through the partial dependence plot (Chipman et al. 2008), shown in Figure 7.

[Insert Figure 7 about here]

Note that when interpreting partial dependence plots, one should keep in mind that these plots only describes the *marginal* effect of a variable, without taking into account interactions among variables. If a variable exerts its effect on box office mainly through interactions, the partial dependence plot may not provide much insight into the effect of such variable. With this important caveat in mind, Figure 7 offers some interesting insights on the marginal effects of our textual variables. First, we see that movies that belong to the action genre have, on average, lower box office revenues (see first row, left panel). Second, movies with longer titles tend to do better in the box office (see second row, left panel). Third, the content variable "familiar setting" has an important marginal effect: movies that take place in familiar setting tend to have higher box office revenues than ones that do not. For the other variables, the partial dependence plots are mostly flat, indicating that their effect is mainly through the interactions among different variables.

4.3 Risk adjusted profit analysis

As we mentioned in Section 3, the BART-QL approach allows us to not only obtain point estimates on box office performance, but also the entire predictive distribution of box office revenue associated with the script under consideration. This enables us to compute a measure of the excess return (or Risk Premium) per unit of "risk" (defined below) in each script. One of the most commonly used metric of risk-adjusted return is the Risk-Adjusted Return on Capital (RAROC), defined as the excess return divided by the value-at-risk (Jorion 2006). The larger the RAROC metric, the more attractive is the risky asset. For a major studio, the RAROC metric associated with any given script can be written as follows:

$$\text{RAROC} = \frac{R}{VaR}, \qquad [9]$$

where R denotes the studio's (actual) return from a movie, calculated as R = 0.55 * box office – production budget, where the 0.55 is the studio's share of the box office revenues net of the exhibitor's share of the revenues (Eliashberg et al. 2007). VaR denotes the value-at-risk of producing a movie (Holton 2003; Jorion 2006; McNeil et al. 2005). It is a metric that is widely used in economics and finance to measure risk (e.g., Jorion 2006). It goes beyond the first- and second- moments of the predictive distribution to quantify financial risk, which is appropriate because the distributions of box office revenues are highly skewed and non-Gaussian. The VaR (with confidence level α) of a portfolio P (denoted as $\text{VaR}_\alpha(\text{P})$) is defined by the smallest number l, such that the probability that the loss L exceeds l is not larger than (1-α) (see McNeil et al. 2005). Formally, the value at risk of a portfolio P (with confidence level α) is defined by the following equation:

$$VaR_\alpha(P) = \inf\{l \in R : \text{Prob}(L > l) \le 1 - \alpha\} \qquad [10]$$

Conceptually, having a portfolio with $\text{VaR}_\alpha(\text{P}) = V$ means that the manager is (1-α)% confident that his/her loss from the portfolio will be lower than V. Usually, α is set to be 0.05 or

0.01, which corresponds to confidence level of 95% and 99%, respectively (Pearson 2002). Throughout this paper, we set $\alpha = 0.05$.

Using the predictive distributions of box office from our model, we obtain an estimate for VaR ($\alpha = 0.05$) for each movie in our dataset. We use the calculated RAROC metrics for different evaluations of the movies in our dataset. In the analysis below, we divide our movies into different subsets based on (i) genre, (ii) MPAA rating, and (iii) production studio, and study the general relationship between these subsets and their associated RAROC metrics.

Figure 8 shows the RAROC metrics for movies in different genres. Across the eight different genres, "family" movies have the highest median RAROC metric of 0.0726, presumably because those movies appeal to a large pool of audiences of any age, and also movie going is a favorite family activity. The genre with second highest metric is "comedy" movies with a median of -0.0404. At the other end of the spectrum, "horror" movies have the lowest metrics with median of -0.550.

[Insert Figure 8 about here]

Next, we look at the distribution of the RAROC metric for movies with different MPAA ratings. We note that 58% of the movies in our dataset are R movies, which is consistent with the population proportion of R movies made over the last 10 years (MPAA 2007). Thus, we divide our movies into two groups: R movies, and PG-13/PG/G movies, and study whether there are any differences between the two groups. As we can see in the boxplot in Figure 9, non-R movies (PG-13/PG/G) has a higher median RAROC of -0.226 compared to R-rated movies (-0.432); this is consistent with the findings in De Vany and Walls (2002).

[Insert Figure 9 about here]

We take a step further and divide movies into different groups depending on the studio that produced the movie. We compare the eight major studios (Warner Bros, Walt Disney Motion Pictures Group, Sony Pictures Entertainment, Fox Entertainment Group, NBC Universal, Paramount Motion Pictures Group, MGM, and Dreamworks) with respect to the RAROC metrics of their movies (in our sample) that they made. The results, shown in Figure 10, offer some interesting insights into how studios compared in terms of the metric. Dreamworks has the highest median (0.529), followed by NBC Universal (-0.100). The studios with the lowest median RAROC metrics are Paramount Motion Pictures Group (-0.703) and Warner Bros (-0.453).

[Insert Figure 10 about here]

Thus, to sum up, our analyses suggest that the attractiveness of the movies in terms of their risk/return differs across genre, MPAA rating, and the production studio. Family movies and non-R rated movies are, in general, more attractive and some studios seem to be better at identifying and producing movies that have higher RAROC metric. In the next section, we discuss how studios can improve their movie production portfolio decisions using our methodology.

5. Portfolio optimization and risk management

A key property of our proposed methodology is that it generates the predictive distributions of box office revenues based on the scripts. In Section 5.1 and Section 5.2, we return to the risk management problem we stated in the introduction, and demonstrate how our model can be used by producers (Section 5.1) and investors (Section 5.2) to optimize their production portfolios and adequately manage financial risk, subject to various constraints.

5.1. Optimization of production portfolio from a producer's perspective

In this subsection we illustrate how a producer can optimally choose a movie production portfolio. For the sake of illustration, we assume that the producer (e.g., an Indie or a major studio) is considering a set of ten movie scripts (randomly selected for illustrative purpose), shown in Table 7.[6] The goal is to invest and produce a subset of them. Further, we assume that the cost c_i, of turning script i into a movie, is known and fixed in advance.[7] As discussed earlier, we assume that the studio' share of the box office revenue is 55%.

[Insert Table 7 about here]

Using the BART-QL model, we obtain the posterior predictive distribution of the box office revenue for each movie. We let the (untransformed) box office revenue of the i-th movie (a random variable) be y_i, and the predictive distribution of y_i be $f_i(y_i)$, i.e., $y_i \sim f_i(y_i)$. From the predictive distribution of y_i, we can compute the expected return $E(R_i) = 0.55E(y_i) - c_i$ for each movie in the consideration set. The title, production cost, expected return, and the value-at-risk for each movie are listed in Table 7.

Given a portfolio $P = \{p_1, p_2, p_3, ..., p_{10}\}$, where $p_i = 1$ if movie i is chosen and $p_i = 0$ otherwise, we can easily compute the expected return of portfolio P by summing over the expected return from each movies in the portfolio. That is,

$$\Pi(P) = \sum_{i \in P} E[Ri]_i = \sum_{i \in P} (0.55E(y_i) - c_i). \qquad [11]$$

[6] We have held various discussions with independent producers and verified that a slate of ten movie scripts is roughly the average size of their consideration set. For much larger consideration sets, other optimization tools such as genetic algorithm (Tsao and Liu 2006) are required.

[7] Discussions we had with several movie makers suggest that it is a common practice to have a contemplated budget for a movie, even before any casting and other productions-related decisions have been made. If there are any uncertainties in production costs, it is straightforward to incorporate that uncertainty into the predictive distributions of profit.

The value-at-risk of each portfolio is then computed by simulating the box office revenues of the selected movies using their predictive distributions (details available upon request). We then exhaustively solve for the expected return and value-at-risk (VaR, $\alpha = 0.05$) for each of the $2^{10} - 1 = 1023$ possible movie portfolios, hence derive and plot the "mean-VaR efficient frontier" (Alexander and Baptista 2001) of movie portfolios that are not "dominated" (i.e., have a lower expected profit and a higher value-at-risk) by another portfolio. The efficient frontier is shown as solid points in Figure 11. The small gray points denote all other possible portfolios that can be chosen, but they are not on the efficient frontier and hence inferior.

[Insert Figure 11 about here]

One way to assess the practical value of our forecasting methodology is to compare the "optimal" movie portfolios identified using our methodology (i.e., the portfolios that are on the efficient frontier) with those identified using the predictive distributions generated by a simple linear regression of (log-) box office vs. (log-) budget. If the two sets of "optimal" movie portfolios are exactly the same (or very similar), the value of our approach is minimal because even without our approach, producers would have arrived at the same set of movie production decisions.

Our result in Figure 11 shows that this is not the case. The trianglar points on Figure 11 denote the portfolios that are on the efficient frontier if a box-office/budget-regression model is used as a forecasting tool. As can be seen, some of the portfolios that are recommended by the simple regression approach do not lie on the efficient frontier but in fact below it. For instance, the portfolio that is circled in Figure 11 can be improved by moving it towards the upper-left direction, resulting in a movie portfolio that has a higher expected return and, at the same time, a lower value-at-risk ($\alpha = 0.05$). More generally, since for the same expected return, the portfolio

generated by our forecasting methodology involves less risk than the one generated by the box-office/budget regression model, the former is said to be second-order stochastically dominating the latter, and it implies that all risk-averse expected-utility maximizers prefer it (Bawa 1975).

So far, for the efficient frontier in Figure 11 we have assumed that the producer does not have a strict budgetary constraint in mind, which would be the case where financing is easy to obtain and thus the producer is only concerned about the expected return versus value-at-risk tradeoff. In reality, however, the producer may have limits on the maximum amount of capital he/she can access, which causes some movie portfolios to be infeasible, if the total cost exceeds the amount of accessible capital. Starting from Figure 11, the efficient frontier under budgetary constraints can easily be solved by simply excluding the portfolios on Figure 11 whose total costs exceed the budget constraints. The result is shown in Figure 12; we demonstrate how to address such a problem by solving for the efficient frontiers with budgetary restrictions of varying from $400M (upper left panel), $300M (upper right panel), $200M (lower left panel), and $100M (lower right panel). As can be seen, our tool can be easily tailored to fit the specific needs and budgetary restrictions for producers.

[Insert Figure 12 about here]

5.2. Optimization of capital allocation from an investor's perspective

In the previous section, we study a producer's portfolio choice as a binary 0-1 problem; i.e., the producer has to either pay the full cost of a movie to produce it, or not produce the movie at all. More recently, Indies, hedge fund managers, and external financiers (and soon movie fans) are able to invest in movie productions or trade shares of movies that are still in production (Plambeck 2010).Typically, such investors provide some capital to partially fund x% of a

movie's production cost (in return of x% of the movie's box office revenue).[8] Having an opportunity to collaborate with one such entity, we discuss and demonstrate below how it can use our model to derive the efficient frontier for its capital allocation decision.

The company we collaborated with is an independent movie production firm, headed by two producers with solid track record who decided to make independent movies, and provided us with two actual, full-fledge scripts that they are planning to produce. They have limited access to capital at this point and thus seeking external investment from financiers. For confidentiality purpose, we henceforth refer to these two actual scripts as "script A" and "script B." To produce script A and script B it was estimated that their costs would be $15M and $50M, respectively. Following the methodology proposed and described in this paper, we extracted the textual variables (as shown in Table 1) from the two scripts, and applied BART-QL model to obtain the predictive distributions of box office revenues; the predictive distributions of the two movies are shown in Figure 13. The mean (median) predictive box office revenues for script A and script B are $19.4M ($11.9M) and $87.2M ($54.3M), respectively. The 90% posterior highest probability density (HPD) intervals (Chen and Shao 1999) are ($0.6M, $39.1M) and ($2.8M, $182.9M) for script A and script B, respectively.

[Insert Figure 13 about here]

We presented the predictive results to the management of the production firm and they were in agreement with their relative estimates. Again, for confidentiality, we cannot disclose the budgetary constraint they face and we will analyze their capital allocation problem assuming that they are planning to invest a total of $10M into both movies. Unlike the integer problem discussed in Section 5.1, the $10M can be allocated to produce both script A and B and the balance will be raised from external investors. Figure 14 plots the expected return vs. value-at-

[8] We make this assumption for illustrative purpose. In practice, other types of arrangements are possible.

risk (VaR, $\alpha = 0.05$) for each of the different capital allocations, in increments of \$0.5M. As can be seen, the minimum VaR portfolio is at the point {\$4.5M, \$5.5M} (i.e., investing \$4.5M in the script A and \$5.5M in script B). Figure 14 also shows that portfolios that lie below the point {\$4.5M, \$5.5M} are dominated by the minimum VaR portfolio; i.e., those portfolios have higher risk yet lower expected return than the {\$4.5M, \$5.5M} portfolio. The efficient frontier is, therefore, the arc between {\$4.5M, \$5.5M} and {\$0M, \$10M}. Thus, from a risk-return standpoint, our analysis suggests that the production firm should allocate at least \$5.5M to script B to ensure that they choose a capital allocation that is on the efficient frontier.

[Insert Figure 14 about here]

So far, we have assumed that the box office performances of different movies in the portfolio of interest are *a priori* independent from each other. In reality, however, box office performance of a movie slate can be dependent across movies. Box office performance can be positively correlated (e.g., if a movie that uses 3D effects did well in the box office, one may believe that another movies that also uses 3D effect will perform well too); or they could be negatively correlated (e.g., if two movies are "too similar", consumers may not want to see the other one after they see the first one). It would therefore be useful to conduct sensitivity analysis with respect to the independence assumption in order to test how the portfolio decision would change if the performance of script A and script B are likely to be correlated.

Here, we briefly demonstrate how correlations can be introduced to solve for the efficient frontier in the two-movie capital allocation problem. Correlations between the box office returns of the two movies can be incorporated using copulas (Nelson 1999). A copula is a mathematical device that allows us to form a multivariate distribution (with a certain dependence structure), from the marginal distributions of several random variables. Here, we use a Gaussian copula

(e.g., Anderson and Sidenius 2004) that has been widely used in the finance literature to introduce correlations into the performance of the two scripts, and solve for the efficient frontier given different correlation coefficients. The details of our simulation procedure using copulas are described in Appendix III.

Using copulas, we can experiment how the minimum VaR portfolio (and hence the efficient frontier) will change with different degrees of dependence between the movies, captured by the correlation coefficient ρ. Specifically, we experimented with three values of ρ: -.5 (negatively correlated), 0.0 (independent, as assumed in our previous analysis), and .5 (positively correlated). Figure 15 shows the simulated bivariate distributions, generated using the simulation procedure in Appendix III, of the box office revenues of the two scripts under different values of ρ. (The plots are shown in log- scale to highlight the dependence). As can be seen, the resulting simulations capture the potential correlations across the two scripts.

[Insert Figure 15 about here]

Figure 16 shows how the efficient frontiers vary as a function of the correlation coefficient. As may be expected, value-at-risk is reduced if the box office revenues have a negative correlation; the minimum VaR portfolio {\$4.5M, \$5.5M} is the same as the independent case. When box office revenues are positively correlated ($\rho = +0.5$), the effect of diversification is attenuated, and hence value-at-risks of the production portfolios increase. The minimum VaR portfolio becomes {\$3.5M, \$6.5M}; i.e., more resources should be allocated to the script with the higher expected return, given that diversification benefits are reduced due to the positive correlation.

[Insert Figure 16 about here]

Together, the illustrative scenario in Section 5.1 and the quasi-implementation described in Section 5.2 above serve mainly to demonstrate the *kinds* of problems that our methodology can provide valuable insights for movie producers. By taking into account their own risk preferences, budgetary constraints, and their particular financial objectives, movie producers can use our predictive distributions to derive their own optimal portfolios effectively.

6. Conclusion

In this paper, we developed a methodology to forecast the entire predictive distribution of box office revenue, based only on the textual information from movie scripts and the production budget of a movie. We extracted three layers of textual information from full-fledged scripts: genre/content, words, and semantics, and used them as predictors in a BART-QL model (Bayesian Additive Regression Tree for Quasi-Linear model), a semi-parametric statistical learning technique recently developed in the statistics literature (Chipman et al. 2008). Being a fully specified Bayesian model, with BART-QL we can obtain not only accurate point predictions but also the predictive distributions of box office revenues. We compared our approach to other benchmark models, and found that our approach has the most accurate predictive performance.

Most importantly, our model's capability to generate predictive distribution of the box office revenue not only allows a studio to assess the risk associated with a point forecast, but also opens new doors for a studio to optimize its portfolio choice and manage its risk exposures. Based on our interactions with industry executives, forecasting and risk management are the two capabilities that are sorely needed in the movie industry in order to transform it from an intuition and experience-based decision making into a more science-based decision making. In this paper, we have shown that a science-based approach can pay off handsomely. We demonstrated,

through two illustrative portfolio optimization problems, how the predictive distributions from our model can be used to aid studio's risk management and movie production decisions. To the best of our knowledge, our paper is the first to introduce such tools that enable the movie industry to acquire those two capabilities.

By applying this new methodology to a database of 200 movies, we have also generated some interesting insights about the factors that are conducive to a movie's success. Our analysis shows that for the movies compiled in our dataset, a higher movie budget tend to increase the movie's box office, but at a diminishing rate. Thus, throwing money at a script does not always generate a blockbuster movie. How a movie will turn out at the box office will critically depend, among other things, on what the genre of the movie is, how the main conflict in the movie is built up, and how different conflicts are structured. Finally, our model suggests that family movies and comedies tend to generate highest risk-adjusted gross profits, whereas horror movies tend to perform the worst; non-R rated movies generate higher risk-adjusted gross profits than R-rated movies. This means that a movie production firm with different portfolio of movies can deliver different risk adjusted gross profits. For that reason, the portfolio choice is an important management decision.

For the future research, we can extend our model to rationalize the movie budgeting process. At this point, a studio sets its movie budget through many processes of negotiations among diverse stakeholders. For that reason, the budget for a movie is taken as given in our model. However, it is conceivable that budgeting can be done in the framework of our model based on the box office potential of a script. We will leave this exploration to a future study.

Variable name	Description
Dependent variable	
LNBOX	(log) Box office revenue
Budget variable	
LNBUDGET	(log) Production budget
Genre and content variables	
GENRE	Categorical variable describing the genre of the movie. A movie may belong to any number of the following categories: • GENRE_DRA: Drama • GENRE_ROM: Romance • GENRE_THR: Thriller • GENRE_COM: Comedy • GENRE_HOR: Horror • GENRE_SCI: Sci-fi • GENRE_ACT: Action • GENRE_FAM: Family
Other content variables	For the other content variables, please refer to Table I
Word variables	
WF1	Factor score 1 for bag-of-words variables
WF2	Factor score 2 for bag-of-words variables
Semantic variables	
NTITLE	Number of words in title
NSCENE	Total number of scenes
INTPREC	Percentage of interior scenes
NDIAG	Number of dialogues
AVGDIAGLEN	Average length of dialogues
DIAGCONC	Concentration index of dialogues

Table 1. Summary description of variables extracted from each script.

Word	Loadings 1 (WF1)	Loadings 2 (WF2)
Man	0.19	0.04
Fuck	0.38	-0.07
Sword	-0.35	0.13
Phone	0.31	-0.09
Ship	-0.25	0.74
Gun	0.23	0.04
Car	0.52	-0.15
Girl	0.24	-0.07
Office	0.16	-0.07
Boat	0.13	0.30
Corridor	-0.32	0.09
Plane	0.02	0.27
Truck	0.34	0.00
Hotel	0.11	-0.04
Dad	0.20	-0.10
Mom	0.24	-0.10
Beach	-0.03	0.13
Woman	0.00	0.01
Police	0.34	-0.07
Tunnel	-0.16	-0.07
Van	0.25	-0.08
Deck	-0.07	0.89
Japanese	-0.14	-0.03
Gonna	0.60	0.03
Head	-0.06	-0.01
Kid	0.35	-0.10
Fucking	0.30	-0.05
Monitor	-0.09	-0.05
TV	0.15	-0.11
Chamber	-0.29	-0.02

Table 2. Factor loadings of the two-factor solution on bag-of-word variables.

Variable	Mean	SD	Min	Max
LNBOX	3.72	1.33	-3.66	6.05
LNBUDGET	3.58	0.92	1.10	5.42
GENRE_DRA	0.53	0.43	0.00	1.00
GENRE_ROM	0.28	0.34	0.00	1.00
GENRE_THR	0.24	0.36	0.00	1.00
GENRE_COM	0.23	0.37	0.00	1.00
GENRE_HOR	0.13	0.31	0.00	1.00
GENRE_SCI	0.17	0.32	0.00	1.00
GENRE_ACT	0.45	0.44	0.00	1.00
GENRE_FAM	0.06	0.20	0.00	1.00
CLRPREM	0.93	0.20	0.00	1.00
IMPPREM	0.73	0.26	0.00	1.00
FAMSET	0.76	0.31	0.00	1.00
EAREXP	0.92	0.20	0.00	1.00
COAVOID	0.90	0.22	0.00	1.00
INTCON	0.87	0.23	0.00	1.00
SURP	0.96	0.14	0.33	1.00
ANTICI	0.89	0.21	0.33	1.00
FLHBACK	0.45	0.43	0.00	1.00
CLRMOT	0.90	0.25	0.00	1.00
MULDIM	0.86	0.26	0.00	1.00
HEROW	0.54	0.33	0.00	1.00
STRNEM	0.58	0.46	0.00	1.00
SYMHERO	0.98	0.11	0.33	1.00
LOGIC	0.99	0.07	0.33	1.00
CHARGROW	0.66	0.30	0.00	1.00
IMP	0.93	0.20	0.00	1.00
MULCONF	0.81	0.29	0.00	1.00
INTENSITY	0.98	0.09	0.33	1.00
BUILD	0.85	0.26	0.00	1.00
LOCKIN	0.89	0.23	0.00	1.00
RESOLUT	0.59	0.39	0.00	1.00
BELIEVE	0.92	0.19	0.00	1.00
SURPEND	0.52	0.41	0.00	1.00
WF1	0.00	0.85	-2.13	2.62
WF2	0.00	0.91	-0.80	7.33
NTITLE	2.73	1.83	1.00	10.00
NSCENE	157.03	58.50	33.00	354.00
INTPREC	0.63	0.15	0.03	1.00
NDIAG	832.71	207.60	356.00	1541.00
AVGDIAGLEN	10.45	1.91	6.57	18.52
DIAGCONC	0.16	0.06	0.04	0.41

Table 3. Summary statistics of all variables.

Data subset	Mean Sq. Error	Mean Abs. Error	Predictive Log-Likelihood
All variables	0.8698	0.6815	-267.2729
Word x Semantics	0.9226	0.6848	-282.9272
Content x Semantics	0.8774	0.6843	-270.4336
Content x Word	0.9257	0.7125	-273.0631
Content Only	0.9232	0.7138	-272.6515
Word Only	0.9614	0.7210	-284.1746
Semantics Only	0.9246	0.6857	-282.2958
Regression on Budget	0.9406	0.7091	-284.7976

Table 4. Out-of-sample predictive performance of our model (all variables vs. subsets of variables).

Methodology	Mean Sq. Error	Mean Abs. Error
BART-QL	0.8698	0.6815
Bag-CART (Eliashberg et al. 2007)	0.9464	0.7170
Naïve Projection (benchmark used in Eliashberg et al. 2007)	0.9270	0.6988
Linear Regression with production budget only	0.9406	0.7091
Linear Regression with all variables	1.0708	0.7868
Stepwise Regression	1.0440	0.7790

Table 5. Holdout performance of BART-QL model versus other methodologies.

Variable	Type	Relative Importance
GENRE_ACT	Content	100.0%
BUILD	Content	84.4%
MULCONF	Content	72.4%
NTITLE	Semantics	65.8%
GENRE_COM	Content	49.5%
CHARGROW	Content	33.3%
GENRE_ROM	Content	31.8%
WF1	Word	29.7%
FAMSET	Content	26.9%
AVGDIAGLEN	Semantics	24.3%

Table 6. Relative importance of the top 10 textual variables.

Title	Cost ($M)	Expected box office($M)	Expected gross profit ($M)	Value-at-risk ($M)
Boogie Nights	15.0	35.3	4.4	11.9
Bruce Almighty	81.0	168.5	11.7	65.8
Enemy of the State	90.0	160.3	-1.8	75.4
Harry Potter and the Chamber of Secrets	100.0	303.5	66.9	74.8
I am Sam	22.0	47.7	4.2	17.6
I Know What You Did Last Summer	17.0	64.9	18.7	12.5
Jay and Silent Bob Strike Back	22.0	58.8	10.4	17.9
Kate and Leopold	48.0	91.1	2.1	39.8
Panic Room	48.0	92.1	2.6	39.3
Thirteen Ghost	42.0	73.7	-1.5	36.2

Table 7. Ten movies used to illustrate the portfolio optimization problems.

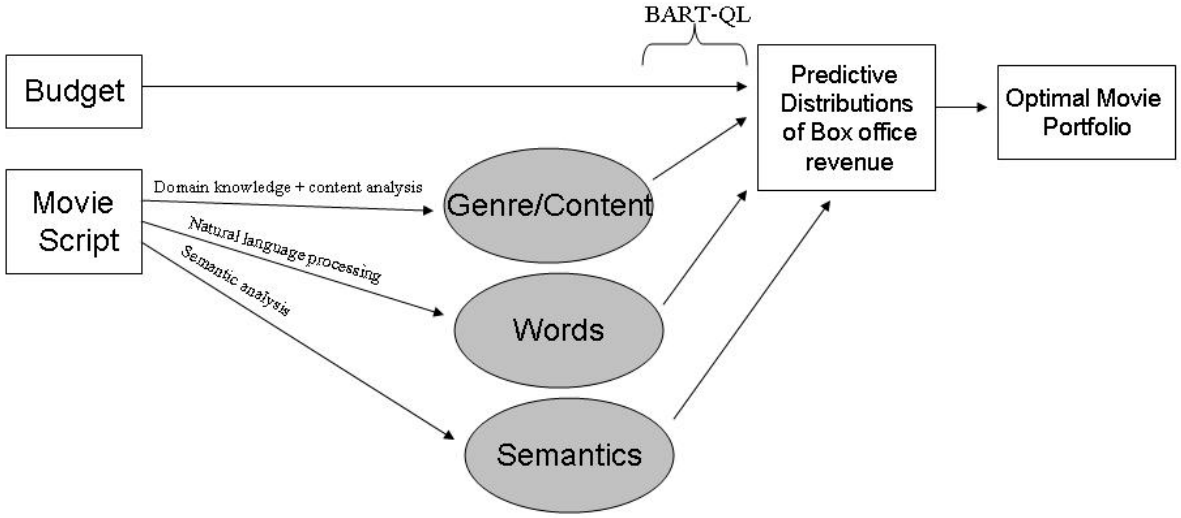

Figure 1. Overview of our approach.

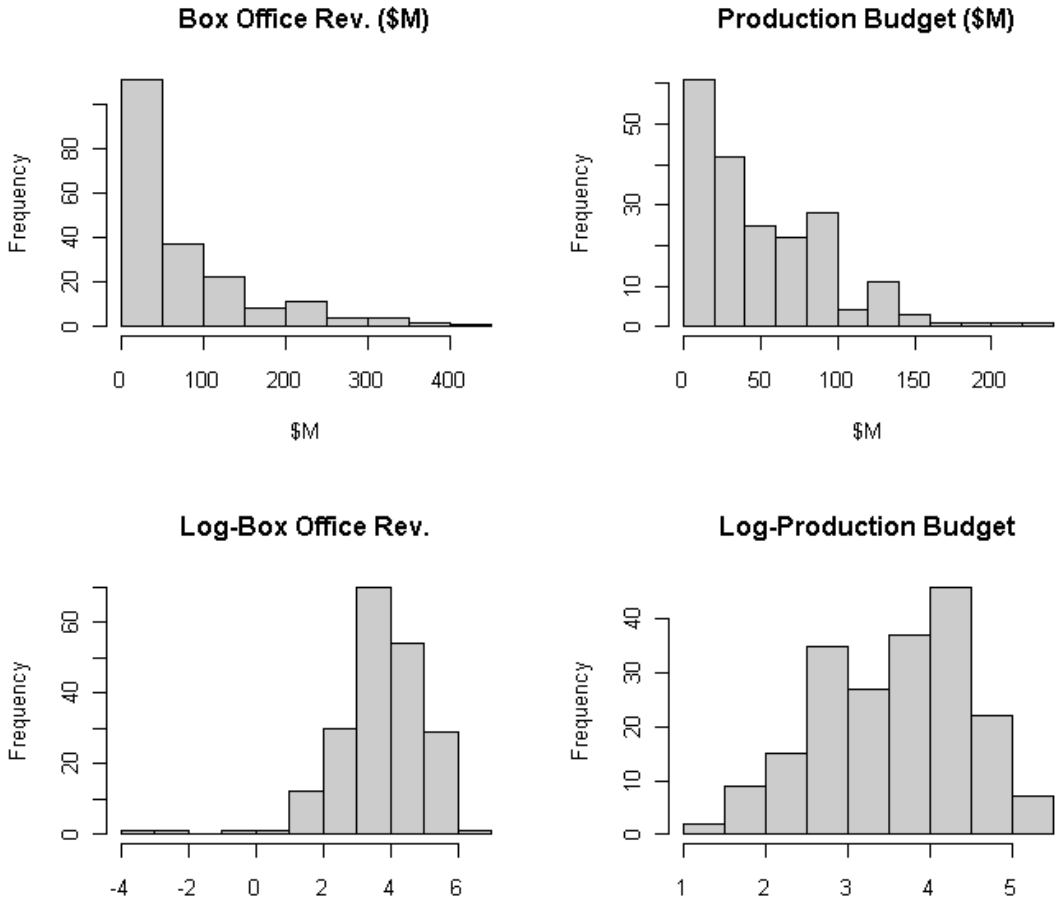

Figure 2. Histograms of box office revenue and production budget.

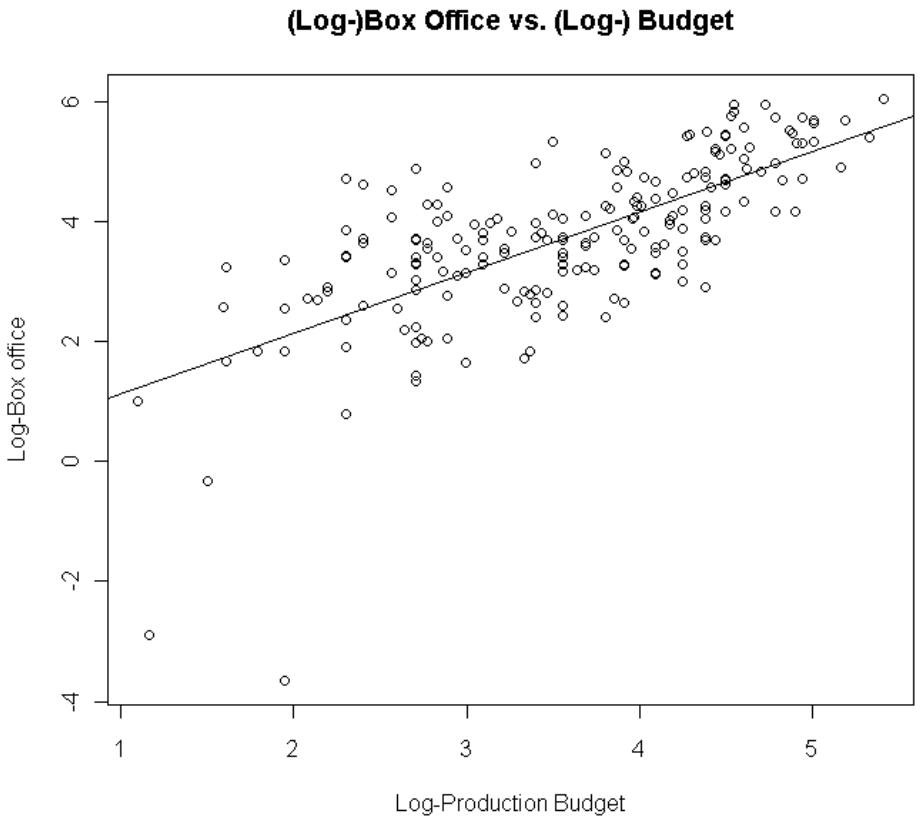

Figure 3. Scatteplot of (log-) box office revenue vs. production budget. The solid line shows a regression line estimated using simple linear regression.

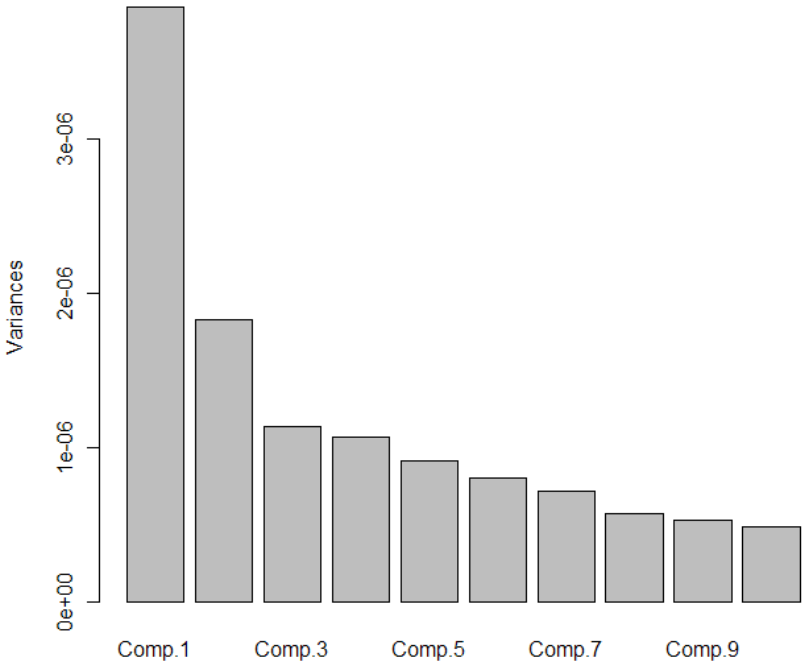

Figure 4. Screeplot from the factor analysis of the bag-of-word variables.

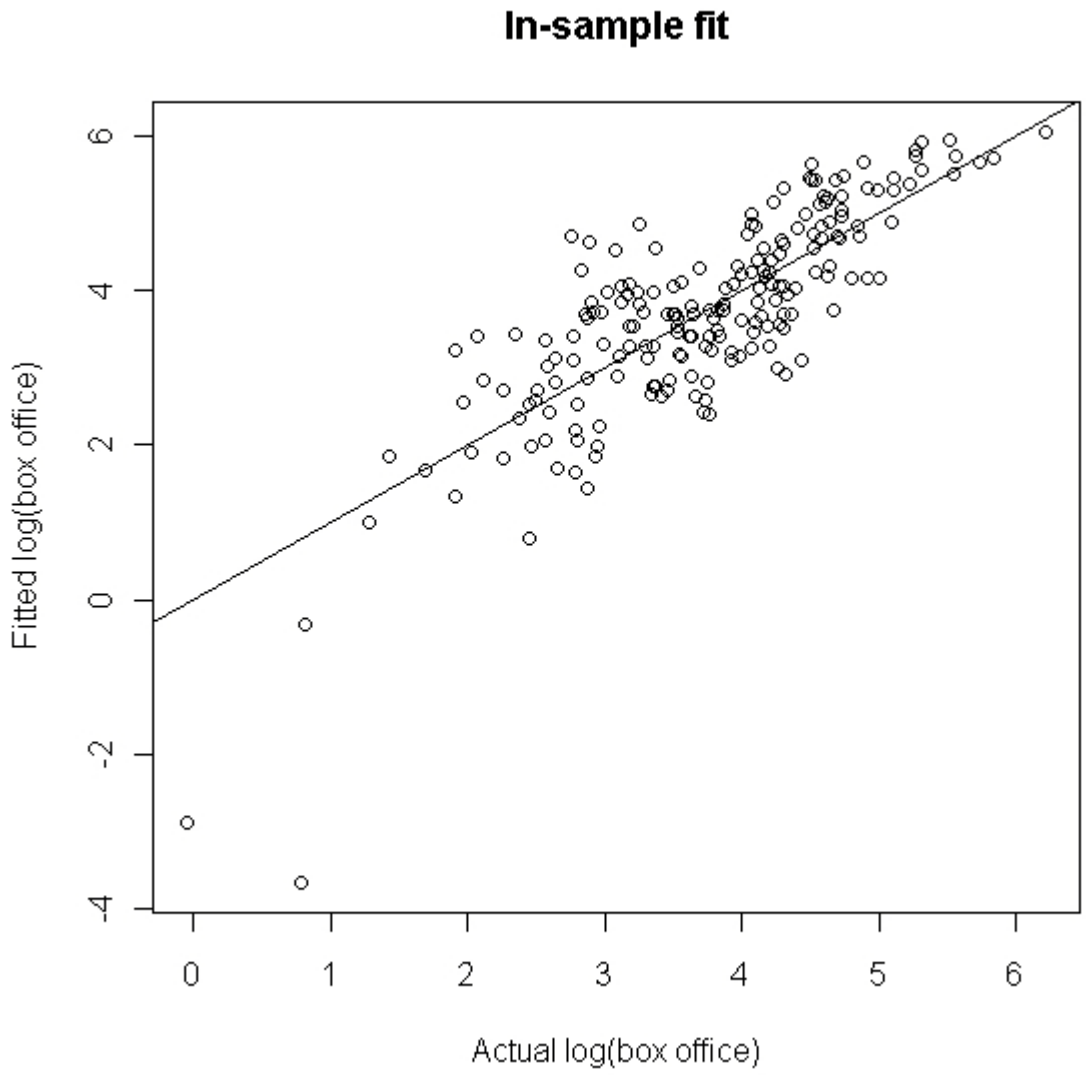

Figure 5. Actual log-box office revenue versus model-fitted log-box office revenue.

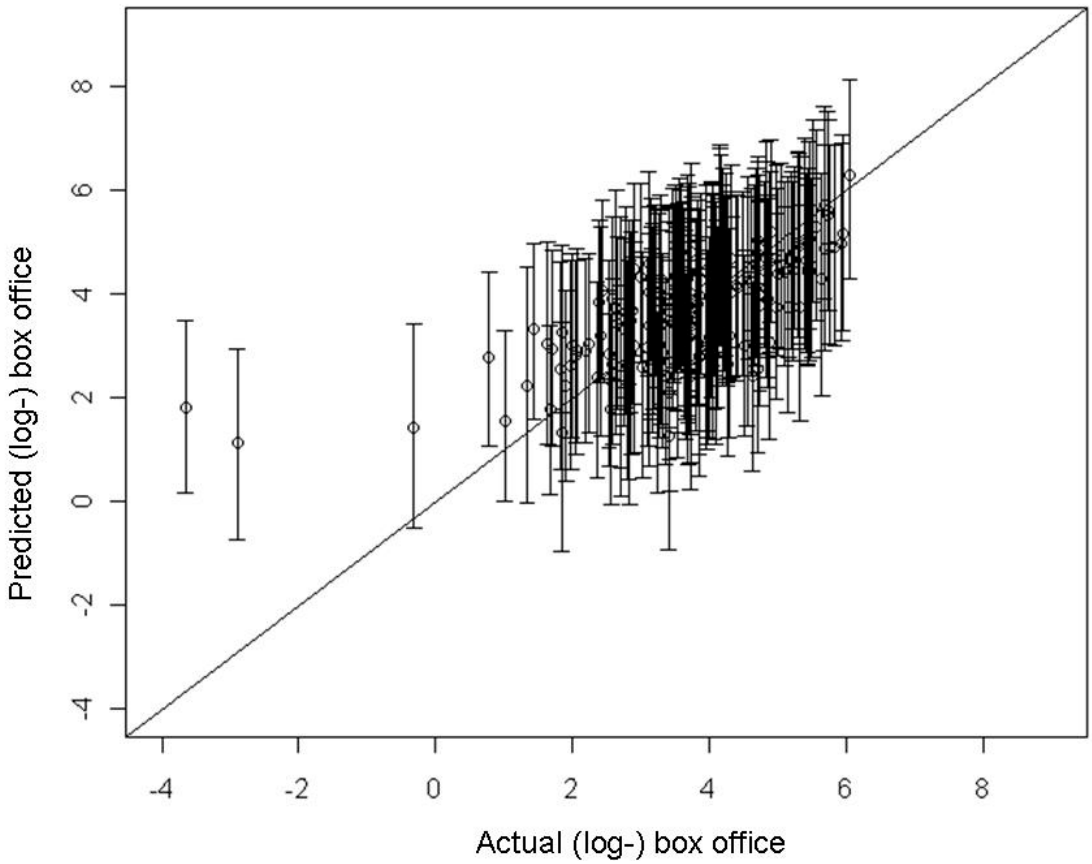

Figure 6. (Out-of-sample) predicted box office revenues vs. actual box office revenues; the 95% posterior highest probability density (HPD) intervals are also plotted.

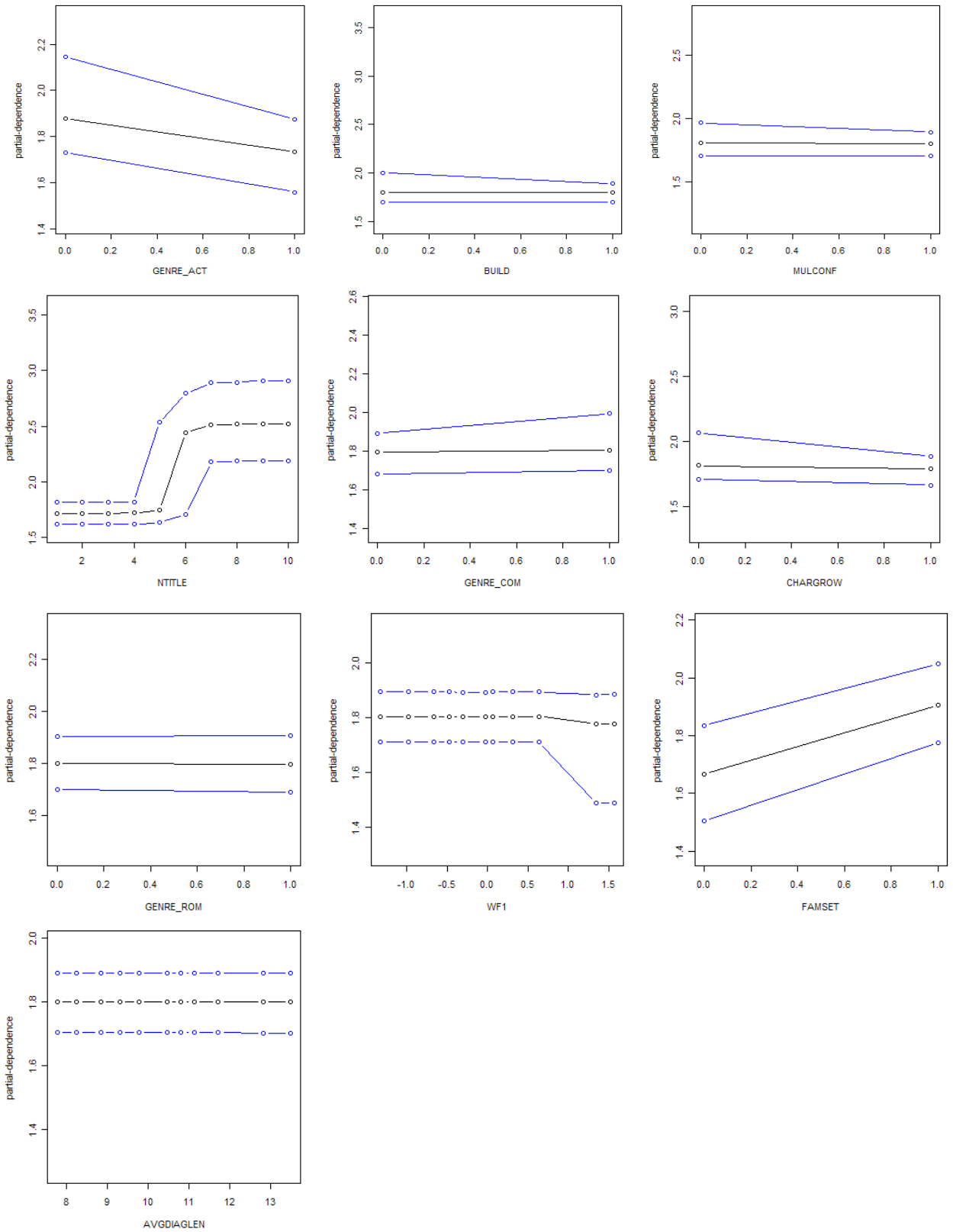

Figure 7. Partial dependence plots for the top 10 textual variables (in relative importance).

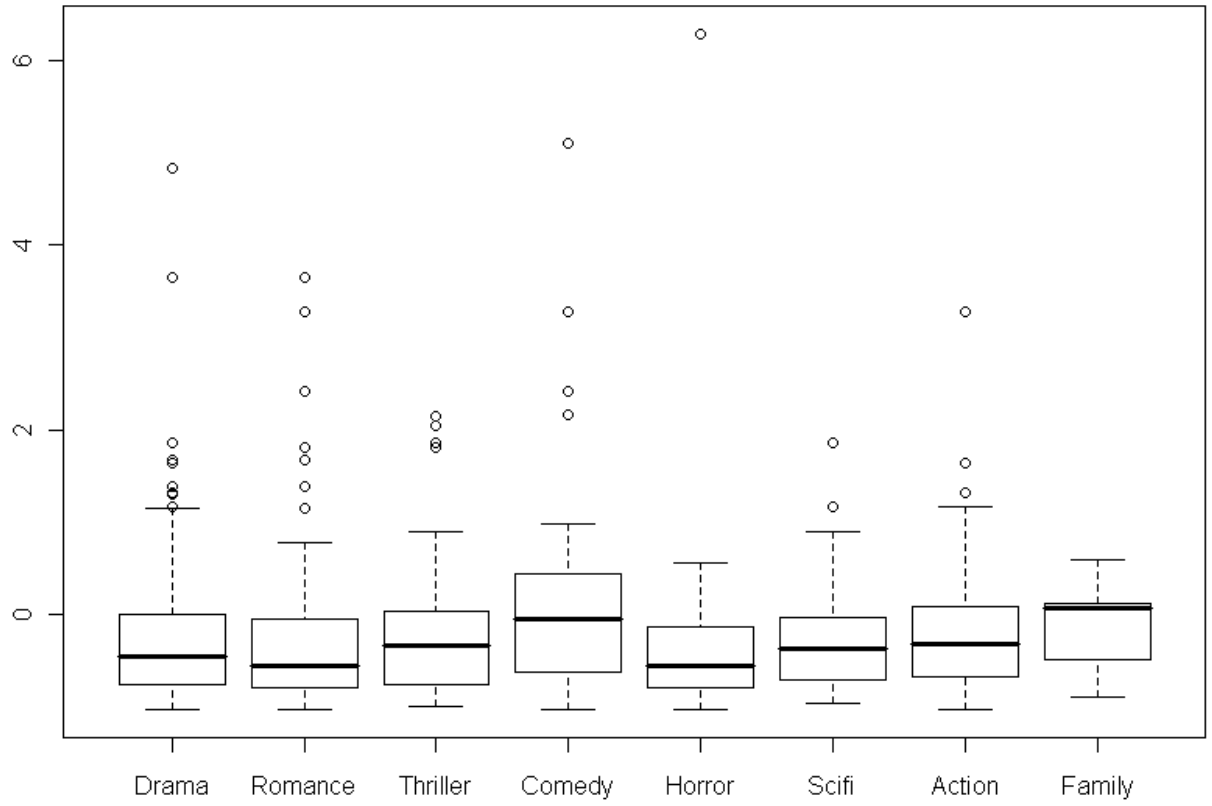

Figure 8. Boxplots of the RAROC metrics for movies grouped by their genres.

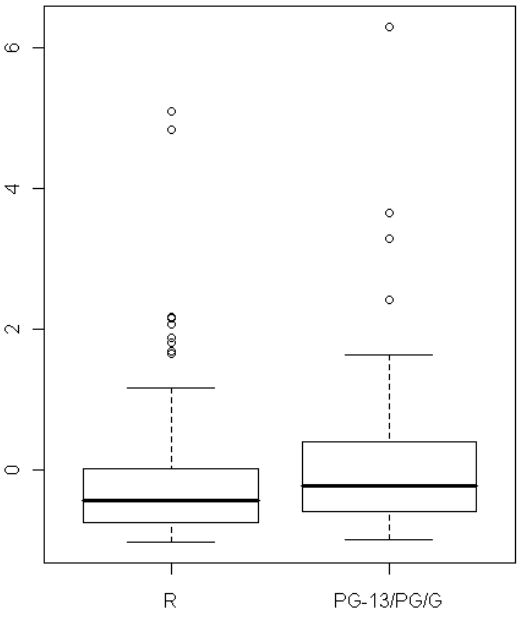

Figure 9. Boxplots of the RAROC metrics for movies grouped by MPAA ratings.

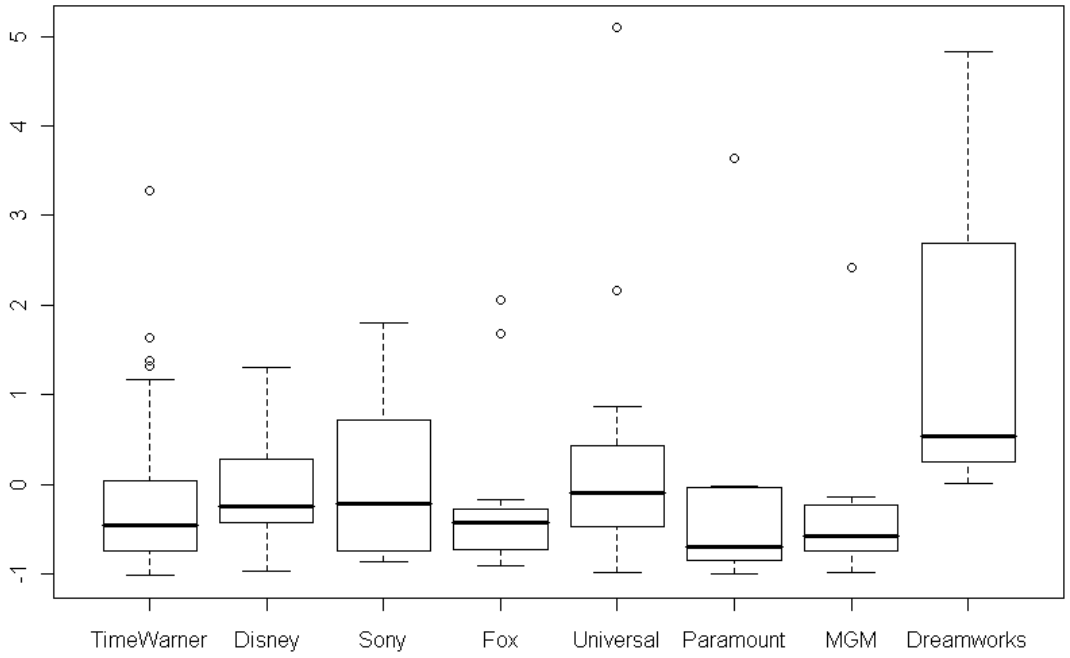

Figure 10. Boxplots of the RAROC metrics for movies grouped by production studio.

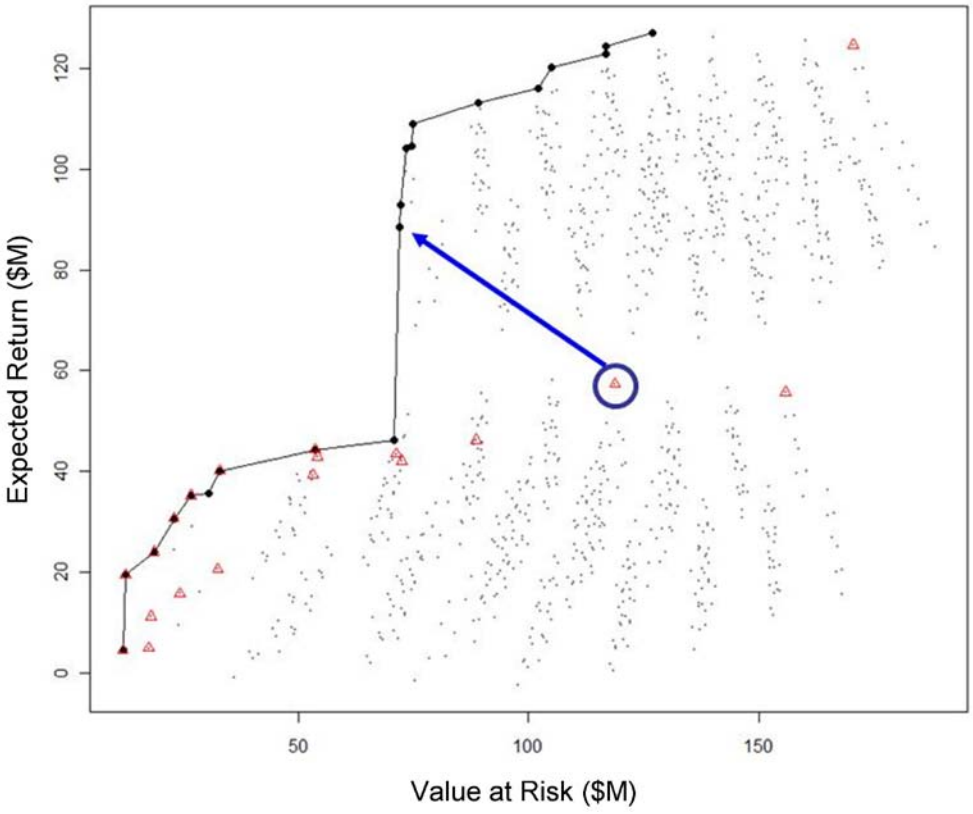

Figure 11. Mean-VaR efficient frontier of movie portfolios.

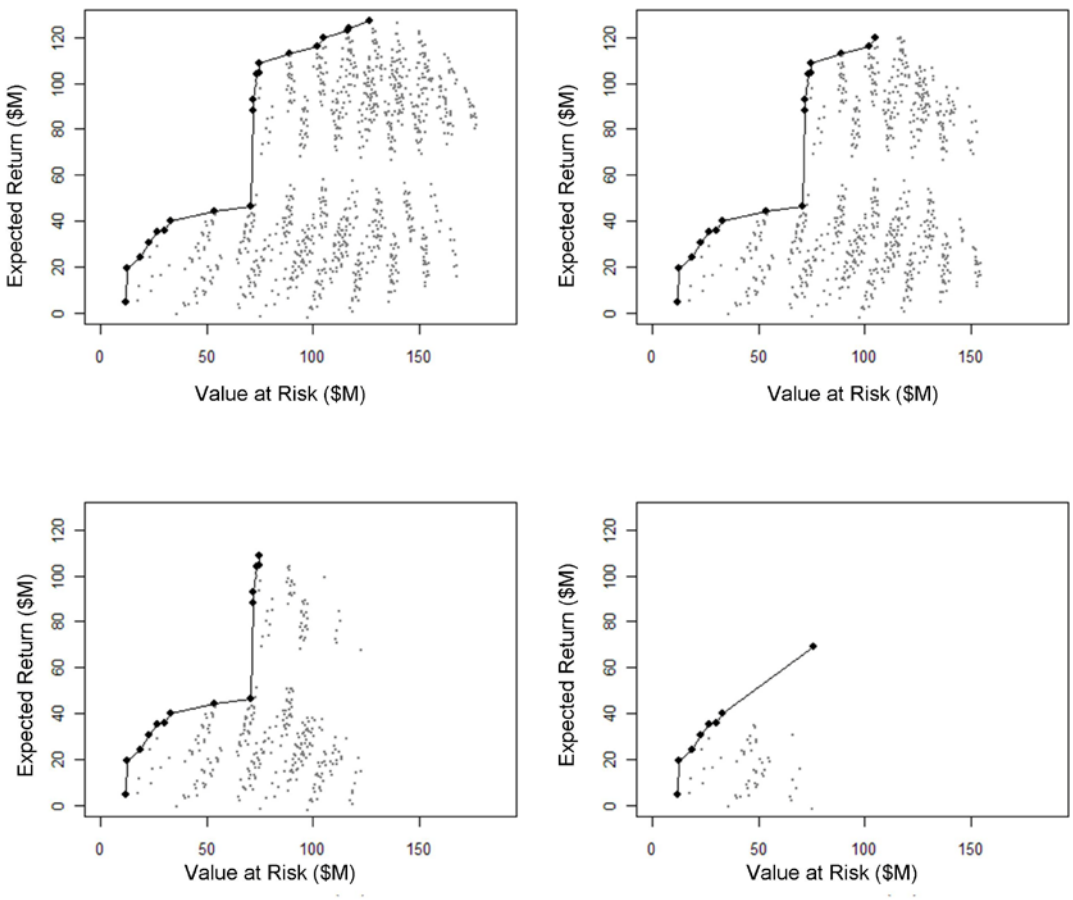

Figure 12. Mean-VaR efficient frontier under budgetary restrictions of $400M (upper left panel), $300M (upper right panel), $200M (lower left panel), and $100M (lower right panel).

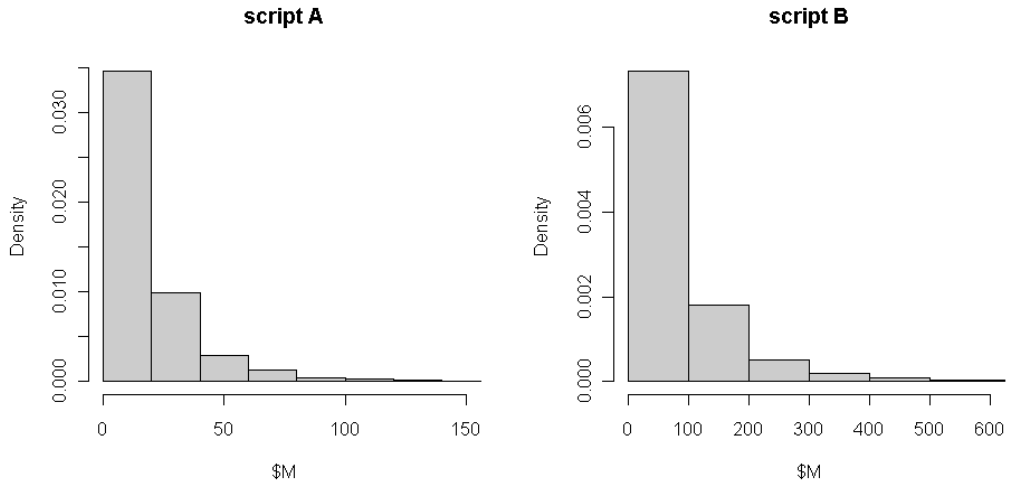

Figure 13. Predictive distribution of box office revenues for script A and script B, respectively.

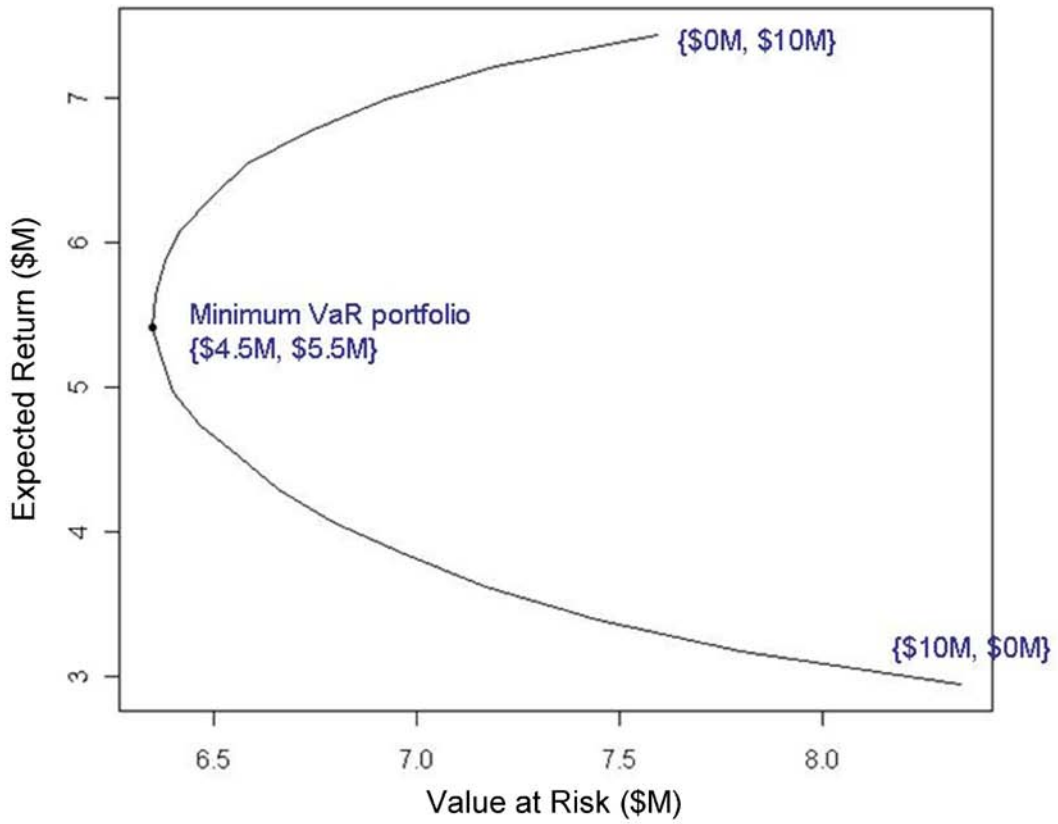

Figure 14. Expected return vs. value-at-risk for different capital allocation.

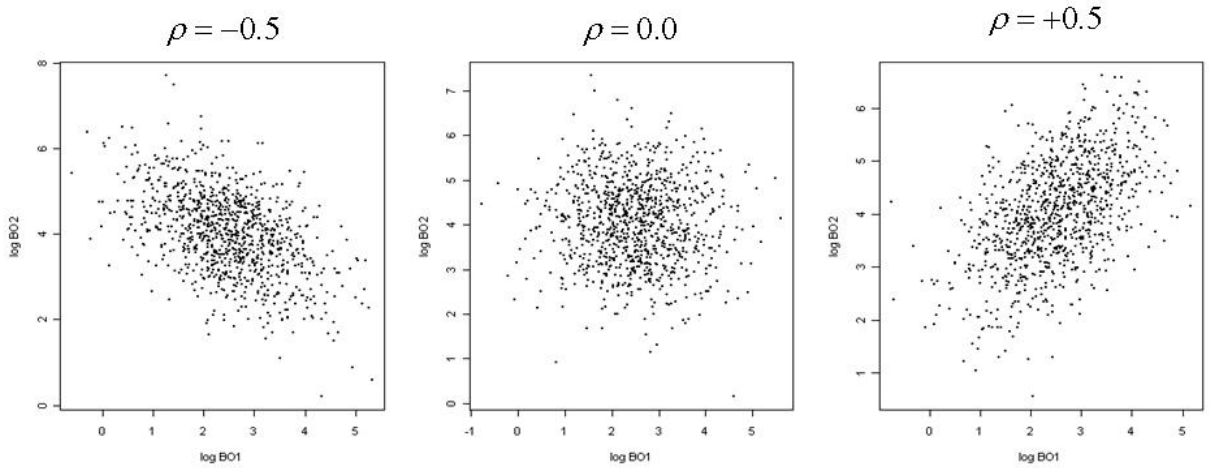

Figure 15. Simulated bi-variate distribution of the box office revenues for the two scripts under different values of ρ, using a Gaussian Copula (see Appendix III). Log-box office revenue of script A is plotted on the x-axis, while log- box office revenue of script B is plotted on the y-axis.

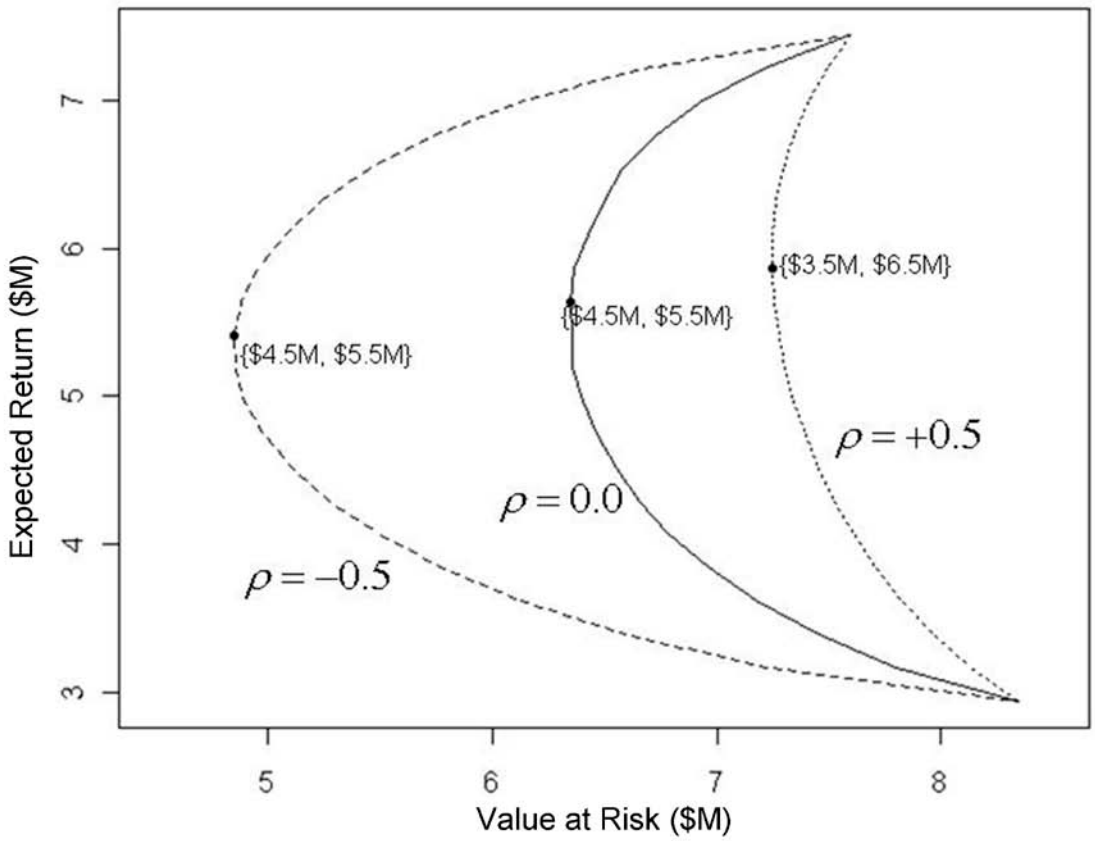

Figure 16. Efficient frontier based on different values of ρ. Solid line: $\rho = 0$; broken line: $\rho = -0.5$; dotted line: $\rho = +0.5$.

Appendix

I. List of storyline questions (Eliashberg et al. 2007)

The storyline questions are listed in Table I below.

Variable	Description
CLRPREM	Clear Premise: The story has a clear premise
IMPPREM	Important Premise: The story has a premise that is important to audiences.
FAMSET	Familiar Setting: The setting of the story is familiar to audiences.
EAREXP	Early Exposition: Information about characters comes very early in the story.
COAVOID	Coincidence Avoidance: The story follows a logical and causal relationship; coincidences are avoided.
INTCON	Inter-Connected: Each scene description advances the plot and is closely connected to the central conflict.
SURP	Surprise: The story contains elements of surprise, but is logical within context and within its own rules.
ANTICI	The story keeps readers trying to anticipate what would happen next.
FLHBACK	The story contains flashback sequences.
CLRMOT	Clear Motivation: The hero of the story has a clear outer motivation (what he/she wants to achieve by the end of the movie).
MULDIM	Multi-dimensional Hero: Many dimensions of the hero are explored.
HEROW	Hero Weakness: Hero has an inherent weakness.
STRNEM	Strong Nemesis: There is a strong nemesis in the story.
SYMHERO	Sympathetic Hero: The hero attracts your sympathy.
LOGIC	Logical Characters: The actions of the main characters are logical considering their characteristics. They sometimes hold surprises but are believable.
CHARGROW	Character Growth: Hero changes because of the conflict in the story.
IMP	Important Conflict: The story has a very clear conflict that involves high emotional stakes.
MULCONF	Multi-Dimensional Conflict: The central conflict has multiple dimensions.
INTENSITY	Intensity of Conflict: Parties to the central conflict have strong convictions in what they do.
BUILD	Conflict Build-up: The hero faces a series of hurdles. Each successive hurdle is greater and more provocative than the previous ones.
LOCKIN	Conflict Lock-in: The hero is locked into the conflict very early in the movie.
RESOLUT	Unambiguous Resolution: Conflict is unambiguously resolved through confrontation between the hero and the nemesis at the end.
BELIEVE	Believable Ending: The ending is believable.
SURPEND	Surprise Ending: The ending carries surprise and is unexpected.

Table I. List of storyline questions.

II. Inter-rater agreement on storyline questions

The genre/content questions are rated on a 0/1 scale (yes/no). Thus, we measure inter-rater agreement using coefficient kappa as defined in Fleiss (1971), which measures the agreement between three or more judges beyond that expected purely by chance. The Fleiss's kappa across our three readers in each question is shown in the Table II below.

Variable	Fleiss's κ	Variable	Fleiss's κ
GENRE_DRA	0.60	FLHBACK	0.61
GENRE_ROM	0.35	CLRMOT	0.52
GENRE_THR	0.57	MULDIM	0.31
GENRE_COM	0.64	HEROW	0.15
GENRE_HOR	0.74	STRNEM	0.79
GENRE_SCI	0.60	SYMHERO	0.41
GENRE_ACT	0.66	LOGIC	0.33
GENRE_FAM	0.58	CHARGROW	0.10
CLRPREM	0.34	IMP	0.35
IMPPREM	0.01	MULCONF	0.32
FAMSET	0.30	INTENSITY	0.17
EAREXP	0.33	BUILD	0.28
COAVOID	0.35	LOCKIN	0.34
INTCON	0.20	RESOLUT	0.45
SURP	0.28	BELIEVE	0.25
ANTICI	0.20	SURPEND	0.51

Table II. Inter-rater agreement measured using kappa coefficient Fleiss (1971).

On average, the three readers show reasonable agreement on the set of genre/content questions, with an overall average kappa of around 0.4, indicating "moderate agreement" among raters (Landis and Koch 1977).

III. Simulation of correlated box office revenues using a Gaussian copula

We use the simulation procedure in Nelson (1999).

(a) From our BART-QL model, we obtain the predictive distribution of y_1 and y_2 (the box office revenue of script A and script B, respectively), and hence the estimated CDF, $\hat{F}_{y_1}(\cdot)$ and $\hat{F}_{y_2}(\cdot)$.

(b) For each simulation, we simulate two random variates $(u_1^{(t)}, u_2^{(t)}) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

(c) Take $y_1^{(t)} = F_{y_1}^{-1}(\Phi(u_1^{(t)}))$ and $y_2^{(t)} = F_{y_2}^{-1}(\Phi(u_2^{(t)}))$.

Steps (b) and (c) are repeated to attain the number of simulated draws needed. For details about copula and related simulated techniques, readers can refer to Nelson (1999).

Reference

Abu-Nimeh, S., D. Nappa, X. Wang, and S. Nair (2008), "Detecting Phishing Emails Via Bayesian Additive Regression Trees," *Technical Report (South Methodist University)*.

Alexander, Gordon J., and Alexandre M. Baptista (2001), "Economic Implications of Using a Mean-VaR Model for Portfolio Selection: A Comparison with Mean-Variance Analysis," *Journal of Economic Dynamics and Control*, 26 (7/8), 1159-1193.

Anderson, Leif, and Jakob Sidenius (2004), "Extensions to the Gaussian Copula: Random Recovery and Random Factor Loadings," *Journal of Credit Risk*, 1(1), 29-70.

Baldi, P., and S. Brunak (1998), *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge.

Bawa, Vijay S. (1975), "Optimal Rules for Ordering Uncertain Prospects" *Journal of Financial Economics*, 95-121.

Berger, James O. (1985), *Statistical Decision Theory and Bayesian Analysis, 2nd Edition*, Springer.

Bjornstad, Jan F. (1990), "Predictive Likelihood: A Review," *Statistical Science*, 5(1), 242-265.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993-1022.

Casella, George, and Edward I. George (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46 (3), 167-174.

Chen, M-H., and Q-M. Shao (1999), "Monte Carlo Estimation of Bayesian Credible and HPD Intervals," *Journal of Computational and Graphical Statistics*, 8(1), 69-92.

Chipman, H. E. George, and R. McCulloch (1998), "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935-948.

Chipman, H., E. George, and R. McCulloch (2007), "Bayesian Ensemble Learning," in *Advances in Neural Information Processing Systems*, 19, Eds. E. Scholkopf, J. Platt, and T. Hoffman, Cambridge, MA: MIT Press.

Chipman, H., E. Geroge, and R. McCulloch (2008), "BART: Bayesian Additive Regresion Trees," *Working paper*.

De Vany, Arthur, and W. David Walls (2002), "Does Hollywood Make Too Many R-Rated Movies? Risk, Stochastic Dominance, and the Illusion of Expectation," *Journal of Business*, 75(3), 425-451.

Eliashberg, Jehoshua, Sam K. Hui, and Z. John Zhang (2007), "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts," *Management Science*, 53(6), 881-893.

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

Epstein, Edward J. (2005), *The Big Picture: The New Logic of Money and Power in Hollywood*, Random House, NY

Fleiss, J. L. (1971), "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, 76 (5), 378-382.

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189-1232.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2003), *Bayesian Data Analysis, 2nd Edition*, Chapman & Hall.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001), *The Elements of Statistical Learning*, Springer.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153-162.

Hirschman, Albert O. (1964), "The Paternity of an Index," *The American Economic Review*, 54(5), 761.

Holton, Glyn (2003), *Value-at-Risk: Theory and Practice*, Academic Press.

Huang, J. A. Smola, A Gretton, K. M. Borgwardt, and B. Scholkopf (2007), "Correcting Sample Selection Bias by Unlabeled Data," in B. Scholkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, 19, MIT Press, Cambridge, MA.

Johnson, Richard A., and Dean W. Wichern (2007), *Applied Multivariate Statistical Analysis, 6th Edition*, Prentice Hall.

Jorion, Philippe (2006), *Value at Risk: The New Benchmark for Managing Financial Risk, 3rd Edition*, McGraw-Hill.

Kourtellos, Andros, Chih Ming Tan, and Xiaobo Zhang (2007), "Is the Relationship between Aid and Economic Growth Nonlinear?" *Journal of Macroeconomics*, 29, 515-540.

Landis, J. R., and G. G. Koch (1977), "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 33, 159-174.

Lewis, Davd (1998), "Naïve (Bayes) at Forty: The Independence Assumption in Information Retriveval," *Proceeding of ECML-98, 10th European Conference on Machine Learning*, Heidelberg, DE, 4-15.

Li, Y. H., and A. K. Jain (1998), "Classification of Text Documents," *The Computer Journal*, 41 (8), 537-546.

Liu, J. S., and Q. Zhou (2007), "Predictive Modeling Approaches for Studying Protein-DNA Binding," *Proceedings of ICCM 2007*.

MacCallum, R. C., K. F. Widaman, K. J. Preacher, and S. Hong (2001), "Sample Size in Factor Analysis: The Role of Model Error," *Multivariate Behavioral Research*, 36, 611-637.

Marich, Robert (2005), *Marketing to Moviegoers: A Handbook of Strategies Used by Major Studios and Independents*, Elsevier Inc. Burlington, MA

McNeil, Alexander, Rudiger Frey, and Paul Embrechts (2005), *Quantitative Risk Management Concepts Techniques and Tools*, Princeton University Press.

Monaco, James (2000), *How to Read a Film*, Oxford University Press, NY

MPAA (2007), "Motion Picture Association of America Theatrical Market Statistics 2007", available at http://www.mpaa.org/2007-US-Theatrical-Market-Statistics-Report.pdf.

Neelamegham, R., P. Chintagunta (1999), "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," *Marketing Science*, 18 (2), 115-136.

Nelson, R. B. (1999), *An Introduction to Copulas*, Springer, New York.

Pearson, Neil (2002), *Risk Budgeting: Portfolio Problem Solving with Value-at-Risk*, John Wiley & Sons.

Plambeck, Joseph (2010), "A Place to Bet Real Money on Movies," *New York Times*, March 10, 2010.

Porter, M.F. (1980), "An Algorithm for Suffix Stripping," *Program*, 14(3), 130-137.

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

Sharpe, W. F. (1994), "The Sharpe Ratio," *Journal of Portfolio Management*, 21 (1), 49-58.

Shugan, S. M., and J. Swait (2000), "Enabling Movie Design and Cumulative Box-Office Predictions," *American Research Foundation Entertainment Conference Proceedings*, CA: Beverly Hills.

Sugiyama, M., M. Krauledat, and K.-R. Muller (2007), "Covariate Shift Adaptation by Importance Weighted Cross Validation," *Journal of Machine Learning Research*, 8, 985-1005.

Tsao, Chueh-Yung, and Chao-Kung Liu (2006), "Incorporating Value-at-Risk in Portfolio Selection: An Evolutionary Approach," *Working Paper*.

Walls, W. D. (2005), "Modelling Heavy Tails and Skewness in Film Returns," *Applied Financial Economics*, 15, 1181-1188.

Ward, Lisa S., and David H. Lee (2002), "Practical Application of the Risk-Adjusted Return on Capital Framework," *Casualty Actuarial Society Forum*, 2002 (Summer), 79-126.

Waterman, David (2005), *Hollywood's Road to Riches*, Harvard University Press, Cambridge, MA

Zadrozny, B. (2004), "Learning and Evaluating Classifiers under Sample Selection Bias," *Proceedings of the Twenty-First International Conference on Machine Learning*, ACM Press, New York, NY.

Zhang, Junni L., and Wolfgana Hardle (2008), "The Bayesian Additive Classification Tree Applied to Credit Risk Modelling," *Working Paper*.

Zhang, Song, Ya-Chen Tina Shih, and Peter Muller (2007), "A Spatially-adjusted Bayesian Additive Regression Tree Model to Merge Two Datasets," *Bayesian Analysis*, 2 (3), 611-634.

Zhou, Qing, and Jun S. Liu (2008), "Extracting Sequence Features to Predict Protein-DNA Interactions: A Comparative Study," *Nucleic Acids Research*, 36 (12), 4137-4148.