A random forests approach to predict movie success

Ryan Tsoi
Department of Biomedical Engineering
Duke University
ryan.tsoi@duke.edu

Carolyn Zhang
Department of Biomedical Engineering
Duke University
carolyn.zhang@duke.edu

Tushar Krishnan
Department of Biomedical Engineering
Duke University
tushar.krishnan@duke.edu

Xiaoshuang Yin
Department of Electrical and Computer Engineering
Duke University
xiaoshuang.yin@duke.edu

Abstract

The entertainment sector makes up a significant portion of the world's economy with new movies being released on a weekly basis. However, predicting the success of these movies can be difficult: on one side, are movies that reach the top of the box office despite a lesser known cast and a limited budget. In contrast, are lackluster movies with popular actors and actresses with a substantially larger budget. As a result, a predictive metric is required to evaluate the potential success of newly-released movies. We develop a model based on random forests to construct a predictive measure of movie success, represented by rating and profit. Our model can be used to inform theaters of the potential earnings for each movie, actors of which movies to participate in, and production companies on who to cast. This is an improvement upon current rating-based systems, which are biased towards a small subset of the population, as well as previous attempts to predict movie success based on similar parameters.

1 Introduction

Global cinematography is a multi-billion dollar industry. While there are many metrics to measure the success of a movie, few exist to predict movie success before the release date. Although movie data is widely available from as far back as 1890, creating a predictive metric is statistically challenging. This stems from the sheer volume of movies, most of which have one or several unreported parameters. We are motivated to develop a predictive model as it may benefit filmmaking and spur similar analyses into other facets of the entertainment industry. For example, movies would be led by the best directors and cast the best actors and actresses (which we will refer to as actors from this point), theoretically improving overall movie quality.

Generating a predictive measure based on random forests requires the construction of decision trees using a random subset of the pre-processed data. Once a forest has been trained, the model is evaluated based on its predictive accuracy of a test set. We devise predictive models for the success of movies based on five sets of features that include admissions, budget, earnings, ratings, and a subset of actors, directors, and production companies. We represent success by either IMDb rating (on a scale of 1-10) or profit. Furthermore, we evaluate the predictive power of new models for varying number of variables. We demonstrate that this model improves on previous attempts to predict movie success from IMDb data.

1.1 Previous work

Predicting movie success using the IMDb dataset and other similar databases have been previously attempted using a variety of statistical techniques including regression and machine learning [1-3]. One study by Demir et. al. attempts to use Google search data to predict movie ratings. The authors select a small subset of 400 movies and categorized ratings into high rated movies ($\geq$ 6) and low rated movies ($<$ 6) [1]. Using logistic regression on Google trends data, the authors achieve a prediction success rate of 64.13%.

Their best model, which uses l1-SVM, a binary classification machine learning technique, achieves a prediction success rate of 72.25%. In another study by Saraee et. al., IMDb ratings are generalized into 4 categories (1-2.4, 2.5-4.9, 5-7.4, and 7.5-10) and the query is constrained to movies with a budget in US dollars [2]. Here, the prediction success rate is calculated to be 57.1% from a list of 26 rating predictions of, at the time, unreleased 2004 and 2005 movies [1, 2]. With Newton, stochastic Newton ,and matrix factorization techniques, Singh et. al. demonstrate test error rates of 33% or a predictive success of 67% on three-factor models [3].

2 Motivations

The movie industry has been rapidly growing over the past few years. From 2010 to 2015, global box office revenues grew from $31.6 billion to $36.4 billion [4]. This was fueled by growing international viewership increasing revenue from $21.0 billion to $26.0 billion [4]. During this period, the number of digital screens increased worldwide from 35,063 to 127,689 [4]. Increasingly, people are turning to online sources like Rotten Tomatoes and IMDb for movie recommendations, with unique monthly visitors exceeding 39 million globally [5, 6]. These websites allow any user to post anonymous reviews for movies along with a rating between 1 and 10 stars, which are averaged to produce a net score representing a movie's popularity. These sites also include reviews from movie critics from other sources to provide a separate rating based solely on critics. Despite multiple resources to evaluate movie success, few metrics exist to predict the quality of upcoming releases. Critic reviews are typically a movie's first evaluation, but these are subjective and few watchers actively research this information. Companies attempt to attract audiences by advertising famous casts or directors, but many failed movies feature well-known casts. While online reviews are a comprehensive summary of a movie's quality, this information is not available immediately after its release.

As a result, we propose a model with the ability to predict movie quality based on budget, cast, director, and production company. We choose these variables due to their availability prior to a movie's official release. In addition, a predictive tool of movie success would not only benefit viewers. For example, companies could cast actors that generate the most profit. Moreover, actors that receive a percentage of the movie's profit could selectively participate in movies with the potential to generate the most revenue. Finally, theaters can use this tool to pick which movies to show. An accurate predictive model for movies may also motivate statistical studies in other aspects of the entertainment sector. For example, one could imagine a similar analysis for television series to determine the number of seasons to fund.

2.1 The data

The IMDb data is obtained from a subset of the open-source plain text data files found on the website. This database contains a variety of information on different movies including details of both movie success and the filming process. While the full IMDb database contains information on all movies released, we have access to only a small portion of the data. The movies contained in each categorical text file seem to be chosen randomly by IMDb, making them probabilistic samples. However, IMDb's data collection method is unknown?some monetary values represent estimates, while others appear to be exact. While we assume that these lists act as representative samples of the movie population, this data may be subject to biases, stemming from a number of causes. These include the location of the movie, data availability, and IMDb's relationship with the production company. While this is beyond our scope, analysis of this subset in comparison to the full database must be conducted before drawing conclusions about predictive accuracy. We use the portion of movies from the final processed data which contains every parameter. This is necessary to construct models with

the capability to accurately predict the success of upcoming movies. As a result, we process the data such that movies missing specific parameters are removed, and the remaining movies represent the population of all movies.

2.2 Unique challenges posed by IMDb data

In order to compare movies directly, the text files were converted into tables through the processing of strings based on observed patterns (Supplementary A,B). With over 40,000,000 total lines of text corresponding to more than 300,000 possible movies, parsing these files poses a significant challenge. The number of potential variables must be drastically reduced to generate an interpretable model without compromising predictive power. Therefore, our analysis focuses on movies released between 1980 and 2014. Moreover, we reduce our scope to a set of monetary variables corresponding to revenue and nominal variables representing a subset of actors, directors, and production companies.

Multiple features of the data reduce the final list of movies and complicate the analysis. First, most movies do not all desired parameters, a common occurrence among less well-known movies. Data not found in the text files is retrieved separately using the IMDbpy package and integrated into the tables. Second, many foreign (non-American) movies contain unique characters or accents that can not be parsed. Third, each file follows different schemes and contains entries with ill-defined formats. In these cases movies are excluded from the final dataset. Finally, monetary data is not inflation adjusted and reported in multiple currencies, these are adjusted to a baseline of 2015 US dollars (USD). Additional data pre-processing steps are discussed in the supplementary.

3 Regression

We explore various linear regression models for a simple predictive model of movie success. Regression models estimate the conditional expectation of the response variable given input variables [7]. As a result, predictive performance is measured using the deviance from the true value [8]. For the IMDb data, we construct a least squares linear regression of the form [9]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Using a training set, the coefficients β are calculated to minimize the residual sum of squares. From the model, a subset of the most influential predictors are retained. However, a reduced model may exhibit high variance due to discrete removal of variables. As a result, shrinkage methods like LASSO are used to impose a penalty on the regression coefficients [8]. In this study, we explore both shrinkage methods but will solely discuss the LASSO, which minimizes the sum of squared errors with a bounded sum for the absolute values of coefficients [10]. Model quality is evaluated by looking at information criteria, which measure goodness of fit and correct for overfitting [8].

4 Random Forest

Random forests integrate multiple machine learning methods to construct highly accurate prediction models. This is accomplished through the construction of multiple classification or regression trees, which have the capacity to be arbitrarily enlarged to increase the generalization accuracy [7, 11, 12].

4.1 Regression trees

Regression trees is a machine-learning method for constructing prediction models. This method is advantageous over typical regression such as linear regression, which produces global models that can be inaccurate when the variables interact in a nonlinear manner [13]. These models are obtained by recursively partitioning the data until simple models can be fit within each partition [14]. Prediction trees use a tree representation of this recursive structure, where each leaf represents a partition containing a simple model applying only to that partition. In classical regression trees, the local models are a constant estimate of Y such as the sample mean of the dependent variable in that cell is [9]. This results in fast, easily interpretable predictions with the ability to fit nonlinear behavior. Once the trees are fixed, the models are fixed as well, meaning that model optimization requires finding a good tree.

Table 1: Random forest algorithm [16]

Step	DESCRIPTION
1.	Select n bootstrap samples out of original data
2.	For each of n samples, grow an unpruned regression tree
2a.	At each node, randomly sample m of the variables
2b.	Choose the best split
3.	Predict on new data by aggregating predictions from n trees
4.	Estimate the error rate of OOS predictions

4.2 Random forest algorithm

Random forests grow many classification or regression trees. Each tree is grown by selecting n samples at random with replacement from the original data, where the sample represents the training set. A subset m of M variables is then selected at random to calculate the best split to use at that node. At each split, a new sample of m variables are chosen from a single parameter [7] Each of these trees is grown without pruning. An aggregation of the resulting trees are then used to estimate the error rate of OOS predictions [7].

Random forests provide many advantages over traditional regression methods. They are adept at handling large databases with many input variables without variable deletion. Resulting models can also estimate the importance of each variable in the regression with increasing mean squared error (MSE in regression or misclassification rate in classification) [12, 15]. While the final prediction models are significantly more accurate compared to linear models, they may be difficult to interpret in the context of the data. Despite this, we investigate the use of random forests as a method for the construction of a more accurate predictive model.

5 Application to the IMDb data

We apply regression and random trees to the processed set of 1795 movies released between 1980 and 2014 from IMDb's database. We consider numerous features that include admissions, budget, gross revenue, opening weekend revenue, weekend gross revenue, ratings, and a subset of actors, directors, and production companies. After evaluating the fit of standard regression methods, we construct a predictive model using random forests. To measure movie success, we use two evaluation metrics: the IMDb movie rating, a measure of popularity, and the overall profit, which represents commercial success.

Evaluation methods We evaluate predictive accuracy by comparing the predicted and true values in the test set. For ratings, the model makes a correct prediction if its value falls within $\pm$ 1 rating point of the "rue rating. For profit, a prediction is correct if its value is within $\pm$ 25% of the true profit. In both cases, prediction accuracy is the percentage of correct predictions within the defined ranges.

5.1 Linear and Lasso regression

Simple linear regression produces models with weak prediction accuracy that fit poorly to the data. Looking at the coefficient of determination (R^2), the model predicting movie profit using average profits of the cast members has the highest R^2 value at less than 0.7. In addition, none of the six regression models produce an average prediction accuracy greater than 32% with the worst model having an accuracy of 8.57% (Figure 1).

5.2 Random forests

We use random forests to construct two models for the prediction of movie success, defined as either rating or profit. One model predicts the rating, and a second predicts profit. For each model, the dataset is randomly divided into a training and test set. Of the total 1795 movies, 1595 are randomly

sampled into the training set and the remaining 200 form the test set. Each random forest model is created using the training dataset with the number of trees set to 1000 and the number of variables randomly sampled as candidates at each split equal to $\frac{1}{3}$ of the total predictors.

5.2.1 Prediction of movie ratings

We transform the binary data corresponding to cast, directors, and production companies from discrete outcomes into continuous variables. Specifically, we set 0s to the overall average movie rating within our database. We then set the 1s corresponding to each actor or director to the average movie rating of all the movies in IMDb's text database that person has worked on. By doing this, we minimize the potential bias against movies that contain few big name actors or directors from our list of variables. This includes smaller movies lacking famous celebrities, or movies with stars that we do not account for in our models. We first construct a full rating model using random forests in which ratings is the response variable while all remaining variables are input parameters. This model results in a prediction accuracy rate of 85% (Figure 1).

We then construct a reduced model with fewer parameters based on the available information one week before a movie's release. While it may be statistically more accurate to reference the variable importance plot, influential variables may be correlative instead of predictive. In this reduced rating model, the rating remains our response variable, but we remove gross revenue, weekend gross revenue, and admissions. These predictors are excluded because they are collected after a movie is no longer in theaters, making them more appropriate as response variables. As a result, this reduced model can be interpreted as a true predictive model of movie success before its official release. Compared to the full rating model, the prediction accuracy decreases to 71% (Figure 1).

We then perform 5 fold cross validation on this second model and analyze the impact of the number of parameters on the predictive accuracy. On both the full model and the reduced model, the CV error reduces significantly and stabilizes at around 7 variables. Additionally, as the number of the most important variables included in the reduced model increases from 7 to 30, the predictive success rate remains the same with no noticeable trend at $\sim 70\%$ (Supplementary Section 8.6).

5.2.2 Prediction of profit

We utilize random forests to predict movie profits, which we define as:

$$\text{Profit} = \text{Gross Revenue} - \text{Budget}$$

Similar to the rating prediction, we modify the binary variables into continuous variables for the same reasons. We perform similar transformations as the rating data, except replacing average ratings with average profits. We also construct an additional model that uses average gross revenue as the predictor instead of average profits (Supplementary F). Again, we start with a full profits model that includes all parameters excluding gross revenue, which would simply reproduce our equation for profits above. This results in a prediction accuracy of 27.5%. Selecting similar variables as the reduced rating model, we construct a reduced profits model, whose predictive parameters are the budget, opening weekend revenues, ratings, cast, directors, and production houses. This reduced model decreases the prediction accuracy to 18.5% in comparison to the full model (Figure 1). While this does not have strong predictive power, to our knowledge this is the first prediction model for movie profits based on IMDb's database.

We perform 5 fold cross validation on this second model and analyze the impact of the number of different variables on the predictive accuracy. On both the full model and the reduced model, the CV error reduces significantly and stabilizes at around 7 variables (Supplementary Section 8.6).

6 Discussion

Random forests allow us to generate the best possible prediction models. Specifically, not only can we predict the success of a past movie based on both monetary and categorical data, but we can also accurately predict the success of new or upcoming movies with minimal information. In the former, our full rating model improves more than 10% in prediction accuracy compared to previous studies that also estimated movie ratings based on IMDb data [2, 3]. In the latter, our reduced rating

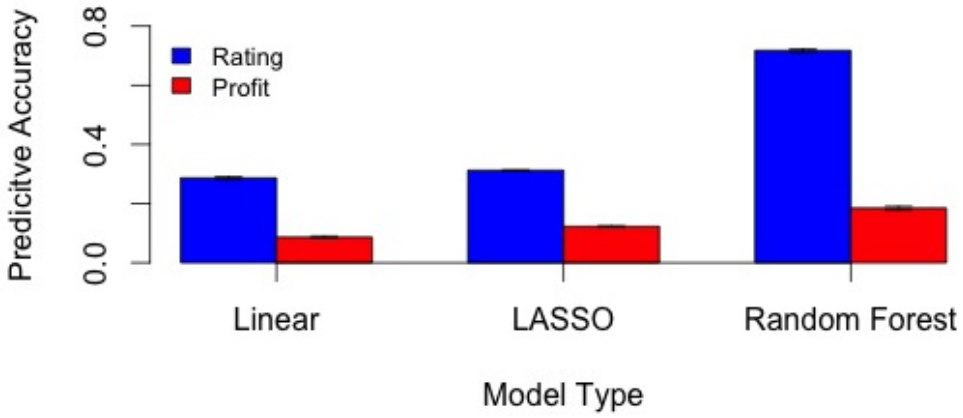


Figure 1: A series of bar plots comparing prediction accuracy between multiple models. Random forest reduced models performs the best in comparison to common regression methods.

model, which can be applied to unreleased movies, determines the correct rating 71% of the time. We also demonstrate, to our knowledge, the first two models that can predict movie profits, though with poor accuracy. These represent first steps towards developing full, predictive models of box office success.

Due to limitations attributed to the runtime and memory requirements of data parsing in R, our study was limited to a small subset of the database. Many of the observations in the full dataset were unable to be analyzed as a result of the immense size of the original IMDb dataset and the slow speed of syntactic analysis in R. Compounded with constraints in the open-source dataset, our models were constructed using a very small final processed dataset representing less than 1% of IMDb's total database. More observations and variables are necessary to improve predictive power of both models, though it will increase the required computational time and may complicate final models. This would require using a different language such as C/C++ to parse the text more efficiently as well as more rigorous variable selection.

Moving forward, optimizing the model may be as simple as obtaining the official movie database from IMDb. This would avoid the need for parsing, which removed numerous movie entries due to inconsistent formatting. With complete datasets, we can again use random forests to construct models with even better predictive power. Such models are promising in the context of the film performance but also in other facets of the entertainment industry. Some examples include determining the existence of a "golden age" for movies, gender or age preferences for specific movie genres, and the most influential people in the entertainment industry. Additionally, we can use movies previously searched by users from other databases to recommend movies based on a combination of our metric and other methods.

7 References

[1] Demir D, Kaproalova O, & Lai H. *Predicting IMDB movie ratings using Google Trends*. 2012

[2] Saraee M, White S, & Eccleston J. *A data mining approach to analysis and prediction of movie ratings*. The Fifth International Conference on Data Mining, Text Mining and their Business Applications. 2004;15-17.

[3] Singh, Ajit P.; Gordon, & Geoffrey J. *Relational learning via collective matrix factorization*. 2008;:650.

[4] (2014). *Theatrical Market Statistics 2014*. Retrieved from Motion Picture Association of America website: http://www.mpaa.org/wp-content/uploads/2015/03/MPAA-Theatrical-Market-Statistics-2014. Accessed November 25, 2015.

[5] Available at: https://www.quantcast.com/rottentomatoes.com. Accessed October 30, 2015.

[6] Available at: https://www.quantcast.com/imdb.com. Accessed October 30, 2015.

[7]James, Gareth, Daniela Witten, Trevor Hastie, & Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013. Print.

[8] Steorts, R. C. *Lecture 15: Regularization and Model Selection Retrieved from Duke University*. 2015.

[9] Hastie, Trevor, Robert Tibshirani, & J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009. Print.

[10] Tibshirani, Robert. *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological) 58.1 (1996): 267-88. JSTOR. Web. 25 Nov. 2015.

[11] Tin Kam Ho. 1995. *Random decision forests*. In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1 (ICDAR '95), Vol. 1. IEEE Computer Society, Washington, DC, USA, 278-.

[12] Breiman, Leo. *Random Forests*. Machine Learning. 2001;45(1):5.

[13] E. Alpaydin. *Introduction to Machine Learning*. 2nd ed. Boston: MIT Press; 2010

[14] Loh WY. *Fifty years of classification and regression trees (with discussion)*. International Statistical Review 2014; 34:329-370.

[15] Genuer, Robin, Jean-Michel Poggi, & Christine Tuleau. *Random Forests: some methodological insights*. arXiv preprint arXiv:0811.3619 (2008).

[16] A. Liaw & M. Wiener (2002). *Classification and Regression by randomForest*. R News 2(3), 18–22.

[17] (2014). *World Economic Outlook Database April 2015: Inflation, average consumer prices*. Retreived from Intermational Monetary Fund website: http://www.imf.org/external/index.html. Accessed November 15, 2015.

[18] Anderberg MR. *Cluster Analysis for Applications, Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Academic Press; 2014.

8 Supplementary

8.1 Parsing text files (full dataset) into tables

The conversion of .list files into table form is conducted in R with the code found in projectParsingFull.Rmd. One complication is that not all of the potential parameters we use are found in the open source dataset, which limits the available information. Here, the text format of each data file is converted into a tabular form through the processing of strings based on observed patterns with each individual file having differing schemes. For example, the business.list file, which contains all information pertaining to financial data needs to be separated into individual components. While various parameters are expected for each movie due to the documentation from IMDb, this is not necessarily true. Most movies only contain a subset of this information (further explanation in Supplementary D). With respect to the copyright holder information, a single company could be found under many names. To simplify this case, all copyright holders of a common pattern are classified under a single name. With mergers between companies, we group companies currently under a single parent company (Full list in Supplementary B). The treatment of ASCII characters presents an additional hurdle. Because the required information cannot be parsed for in the presence of ASCII characters, all observations with these are discarded. This biases the final dataset against foreign movies, especially those from countries with languages that incorporate such characters.

8.2 Converting nominal into binary tables

The conversion of .list files into table form is conducted in R as explained in Supplementary A. Due to the extensive size of the IMDb dataset, we use a small subset of actors, actresses, production companies, and directors to use as binary parameters (listed below). If a particular movie was produced by a company, it had a 1 for that parameter and 0 if it did not. The codes in predictors_table_directors.Rmd, predictors_table_actress.Rmd, predictors_table_actors.Rmd, and copyright_parse_businesslist.Rmd are used to retrieve all information pertaining to a known list of actors, actresses, directors, and production companies (see excel file).

8.3 Using IMDb python package to retrieve ratings from IMDb

Few movies in the open-source rating list intersect with the movies from the business list. Therefore, we us IMDbpy, an open source python package for the IMDb database, to retrieve ratings data.

8.4 Cleaning business data and currency conversion

The parsing script (Supplementary 8.1, 8.2) converts the business text file into multiple sparse tables: admissions, gross revenue, budget, weekend gross, and opening weekend revenue. This is processed (cleaning.R) such that empty rows and rows lacking a movie name or movie year are removed. To constrain the scope of our analysis to movies, entries corresponding to TV shows are removed. We create a unique key (combination of movie name and year) to distinguish between re-releases or sequels sharing the same name. Using this key, we confine the business data to movies containing each variable.

A significant challenge here is due to the complexity of monetary values. These are typically reported in the movie's release year, and not adjusted for inflation (in various non-USD currencies). For comparison, a baseline is set to 2015 USD. Using data from the International Monetary Fund, inflation values are based on country CPIs from 1980 to 2014 [17]. The 2015 USD exchange rate for each currency from XE"s live currency converter is used to then convert all currencies to USD. One obstacle in the conversion involves discontinued currencies, a common occurrence among the European countries (use of Euro began in 1992). As a result, it is difficult to determine if XE's data corresponds to a current exchange rate or the last known rate. To address this, we assume that yearly inflation is independent of currency and that given exchange rates represent current exchange rates.

All monetary values are consolidated into a single value for each parameter (moneyprocessing.R). While gross revenue is reported in a cumulative manner, the weekend gross revenue is not and is summed for each movie. Further complications arise due to inconsistencies in reported locations. Aside from individual country data, some movies contain worldwide data (with or without USA). In the case that worldwide data exists, that value is used rather than summing up individual locations. And when worldwide (excluding USA) data is available along with US data, the two are summed.

8.5 Regression analysis code

The penalty term is calculated from the 1se, which is defined as the largest λ with an average out of sample (OOS) deviance no larger than 1 standard deviation from the mean, as well as the minimum λ, which has the minimum OOS deviation [8]. This creates a balance between false discoveries and predictive performance [8]. Here, Cross Validation (CV), an algorithm for model selection, uses a subset of the original dataset to train the model (training set) and measures the error rate in the remaining data points (test set) [8]. All analyses can be found in Regressions.Rmd.

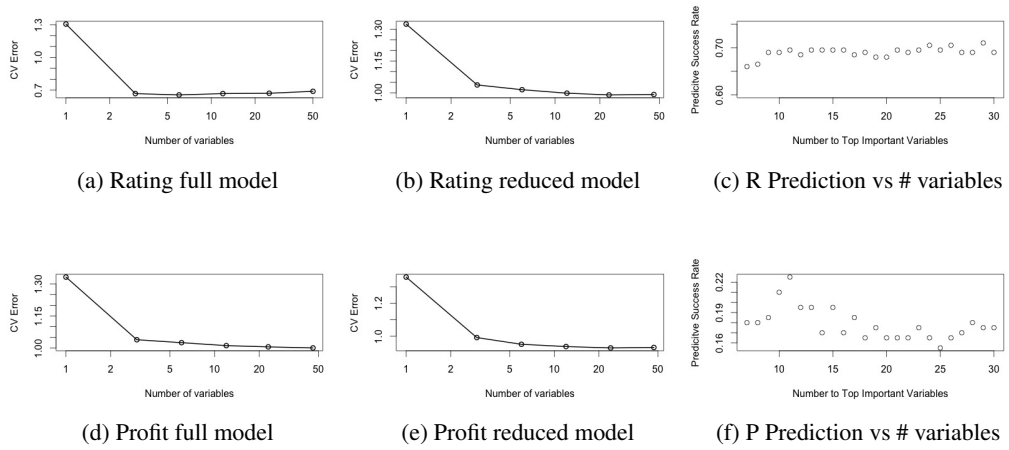
8.6 Random forest analysis code

We also construct a profit-gross revenue model that uses average gross revenue to transform the binary data instead of average profits. Again, we start with a full model that includes all parameters excluding gross revenue. This results in a prediction accuracy of 25.5%. Choosing similar variables as the reduced rating model, we also construct a reduced gross revenue model, whose predictive parameters are the budget, opening weekend revenues, ratings, cast, directors, and production houses. This reduced model maintains the same prediction accuracy of 18.5% in comparison to the full model.

8.7 Exploratory data analysis

8.7.1 Clustering analysis

To fully examine the data, we use cluster analysis, a general process which includes a variety of techniques with the goal of identifying structures within a complex dataset by grouping objects into clusters [18]. In this manner, the movies are clustered according to either the information from the business file or the binary classification with respect to the actors, actresses, directors, and production house. The result with single and complete linkage methods, two hierarchical clustering algorithms, are difficult to read and interpret. Therefore, no further analysis is conducted using clustering analysis.



(a) Rating full model (b) Rating reduced model (c) R Prediction vs # variables

(d) Profit full model (e) Profit reduced model (f) P Prediction vs # variables

Figure 2: An analysis of the impact of the number of variables on both full and reduced models. Plots (c) and (f) correspond to reduced models and measure change in predictive accuracy.

8.7.2 Univariate analysis

Observing the density plots, we see that most of the monetary variables are centered close to 0. However, since money cannot be negative, the distribution appears to be skewed to the right. Notably, none of the data appears to be bimodal. Thus, the averages can be used to transform the binomial data into continuous values.

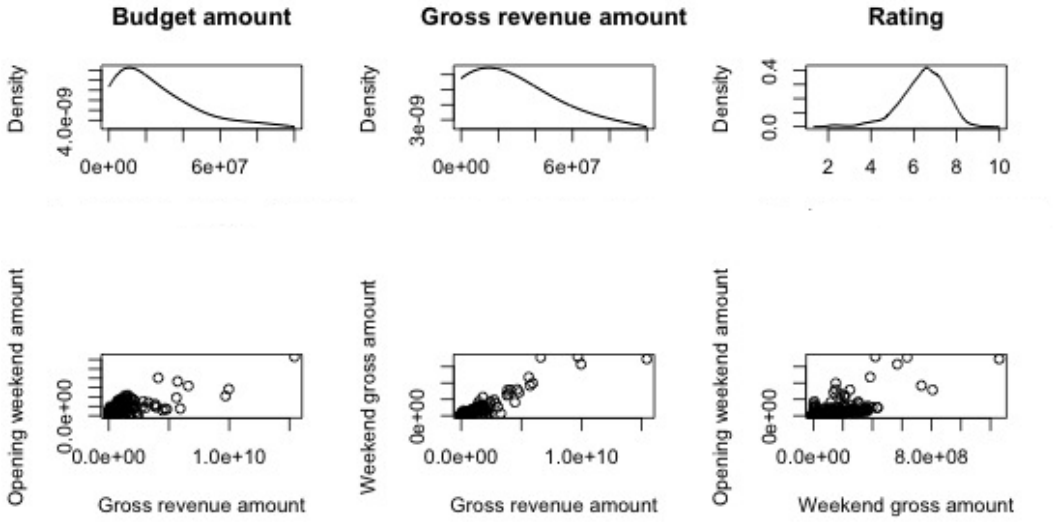


Figure 3: A series of density plots and bivariate scatter plots of the data.

8.7.3 Bivariate analysis

Looking at the bivariate scatter plots, we choose to show three example plots of approximately linear and clearly nonlinear relationships. This demonstrates that linear regression will not be appropriate in creating an accurate predictive model.