Early Predictions of Movie Success: the Who, What, and When of Profitability

Michael T. Lash[1] and Kang Zhao[2]

[1]Department of Computer Science
The University of Iowa
318 MacLean Hall
Iowa City, IA 52242, United States
Email: michael-lash@uiowa.edu

[2]Department of Management Sciences
The University of Iowa
S224 PBB
Iowa City, IA, 52242, United States
Email: kang-zhao@uiowa.edu *(Corresponding author)*

Abstract

This research focuses on predicting the profitability of a movie to support movie investment decisions at early stages of film production. By leveraging data from various sources, and using social network analysis and text mining techniques, the proposed system extracts several types of features, including "who" are on the cast, "what" a movie is about, "when" a movie will be released, as well as "hybrid" features. Experiment results showed that the system outperforms benchmark methods by a large margin. Novel features we proposed also made great contributions to the prediction. In addition to designing a decision support system with practical utilities, we also analyzed key factors for movie profitability. Furthermore, we demonstrated the prescriptive value of our system by illustrating how it can be used to recommend a set of profit-maximizing cast members. This research highlights the power of predictive and prescriptive data analytics for information systems to aid business decisions.

Keywords: Decision Support, Profitability, Predictive Analytics, Text Mining, Social Network Analysis, Prescriptive Analytics.

Citation:

Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. *Journal of Management Information Systems*, *33*(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

About the authors

Michael T. Lash is a PhD student in the Department of Computer Science at the University of Iowa. His research interests lie in the areas of data mining, machine learning and predictive analytics. Specific interests, as well as ongoing areas of research, include inverse classification, utility-based data mining, adversarial learning, and survival analytics and learning. Application of these areas to healthcare, business and entertainment domains are also of interest.

Kang Zhao, Ph.D., is an Assistant Professor at Tippie College of Business, The University of Iowa. He is also affiliated with the University's Interdisciplinary Graduate Program in Informatics. He obtained his Ph.D. from Penn State University. His research focuses on data science and social computing, especially in the contexts of social/business networks and social media. His research has been covered by BBC, Washington Post, Forbes, among others from over 20 countries.

Early Predictions of Movie Success: the Who, What, and When of Profitability

Introduction

The motion picture industry is a multi-billion dollar business. In 2015, the U.S. and Canada saw total box office revenues topping $11.1 Billion [30]. Nevertheless, the financial success of a movie is largely uncertain, with “hits” and “flops” released almost every year. While researchers have undertaken the task of predicting movie success using various approaches, they attempted to predict box office revenues, or theater admissions. However, from an investor’s standpoint, one would want to be as assured as possible that his/her investment will ultimately lead to returns. For instance, “Evan Almighty” earned a high gross revenue of $100 million, but cost $175 million to produce; while “Super Troopers” cost $3 million, but earned $18.5 million. The latter is certainly more appealing from an investment standpoint. In fact, among movies produced between 2000 and 2010 in the U.S., only 36% had a box office revenue higher than production budget, which further highlights the importance of making the right investment decisions. Therefore, our work defines a movie's success as its profitability and attempts to predict such success in an automated way to better support movie investors’ decisions.

The production process of a movie begins with the development phase, including the construction of a script and screenplay. Next, the potential film enters the preproduction phase, the most crucial to success. During this phase, the film-making team is assembled, filming locations are determined, and investments are garnished, among other decisions. Then, the film moves into the actual production phase, in which filming occurs. The post-production phase involves the insertion of after-effects and editing. The last phase is distribution [16]. To support the investment decisions of a movie, the prediction of profitability has to be provided before the actual production phrase. In this research, we are interested in predicting a movie’s financial success during its preproduction phrase. Consequently, we can only leverage data that is available at this time.

Predictions made right before [17] or after [7, 27, 41] the official release (the final phase in movie construction) may have more data to use and get more accurate results, but they are too late for investors to make any meaningful decision. Building upon previous work [23], this research proposes a Movie Investor Assurance System (MIAS) to provide early predictions of movie profitability. Based on historical data, the system automatically extracts important characteristics for each movie, including "who" will be involved in the movie, "what" the movie is about, "when" the movie will be released, and the match between these features. It then uses various machine learning methods to predict the success of the movie with different criteria for profitability.

The overarching research question for this paper is to predict movie profitability using data only available during the pre-production stage of movie development. By proposing the first system to predict movie profitability at an early stage, the main contributions of this research are in two areas: *First*, this work demonstrates how freely available data of different types (including structured data, network data, and unstructured data) can be collected, fused, and analyzed to train machine learning algorithms. When designing and developing information system artifacts [21, 32], such data-based approaches can provide powerful forecasts and recommendations to aid business decisions. To the best of our knowledge, we are also the first to leverage such data and models to prescribe profit-maximizing casts. *Second*, our research proposes several novel features, such as dynamic network features, plot topic distributions, the match between "what" and "who", the match between "what" and "when", and the use of profit-based star power measures to predict the profitability of movies at early stages. We showed that these features all make great contributions to the system’s performance, and help to explain important factors behind movies’ profitability.

The remainder of the paper is organized as follows: after reviewing related research, we describe the framework of our system, and introduce how we extracted different features for the prediction. This is followed by an evaluation of our system using historical data, an analysis of the key factors behind movie success predictions, and a demonstration of how the system can be used to prescribe profit-maximizing cast members. The paper concludes with a discussion of limitations and future research directions.

Related work

The definition of success

The way in which success is defined is of paramount importance to the problem, but past works have focused primarily on gross box office revenue [3, 4, 19, 29, 31, 35], while some used the number of admissions [5, 27]. The basic assumption for using the two as success metrics is simple–a movie that sells well at the box office is considered a success. However, the two metrics ignore how much it costs to produce a movie. In fact, our analysis of historical data also found that revenues are not directly related to profits (more details in the Discussion section). Thus a more meaningful measure of success should be profitability, whether it is the numeric value of profits [37] or the Return on Investment (ROI) [14].

After a success metric was chosen, many studies categorized movies into two classes based on revenues (success or not) and adopted binary classifications as their predictive task; some considered the prediction as a multi-class classification problem and tried to classify movies into several discrete categories [31]. Meanwhile, there are also predictions made on continuous numerical values of success metrics [17, 29, 39], with values of these metrics being logarithmized in several studies [35, 37, 41].

Features for movie success

The accuracy of a predictive model depends a lot on the extraction and engineering of features (a.k.a., independent variables). When it comes to studying movie success, three types of features have been explored: audience-based, release-based, and movie-based features.

Audience-based features are about potential audiences' reception of a movie. The more optimistic, positive, or excited the audiences are about a movie, the more likely it is to have a higher revenue, and vice versa. Such receptions can be retrieved from different types of media, such as Twitter [4], trailer comments [3], blogs [19], news articles [41], and movie reviews [27].

Release-based features focus on the availability of a movie and the time of its release. One such feature that captures availability at release is the number of theaters a movie opens in [29, 31, 33, 35, 39, 41]. The more theaters that will show a movie, the more likely the movie will have a higher revenue. Many movies are targeted for releases at a certain time. For example, holiday release, as well as seasons and dates of releases (Spring, Summer, etc.), are commonly utilized in the prediction problem [8, 19, 31, 35]. Some studies also attempted to capture the competition at the time of release [19, 31], which could negatively affect revenues.

Movie-based features are those that are directly related to a movie itself, including who are on the cast and what the movie is about. The most popular feature for cast members is a movie's star power– whether the movie casts star actors. Star powers of actors have been captured by actor earnings [31], past award nominations [7], actor rankings [35], and the number of actors' Twitter followers [3]. It was agreed that higher star powers are helpful for a movie's success. However, no research has explored the profitability of actors. As it costs a great amount of money to cast a famous actor, we believe an actor's record of profitability will be a better indicator of a movie’s profitability than her record in generating revenues. Moreover, the role of directors in a movie’s financial success is often overlooked or downplayed. While some research has investigated the individual success of directors [25], few studies have actually tried to connect directors’ star powers to movies’ financial success. Some past studies have argued that the economic performance of movies is not affected by the presence of star directors [7], and directors’ values are not as important as actors’ for movie revenues [28]. Contrary to these select past studies, we believe that both actors and directors are crucial for films success. As directors, particularly, play important roles in movie productions [25], our research will examine the effect of directors on movie profitability, in addition to actors.

In addition to individual actors and directors, the cast of a movie has also been explored from a teamwork perspective – whether individuals in a team can work together and develop “team chemistry” [27]. Studies of organizations and teams have revealed that team members’ prior experience or expertise is beneficial for team success, while the diversity of a team helps too, especially in the context of bringing creative ideas and unique experience to teams for scientific research and performing arts [20, 36]. The diversity and the familiarity of a cast contribute to a director’s success in receiving awards [25], and the movie’s box-office revenue [27]. Cast members’ previous experience also positively influences revenues [28]. Nevertheless, there are several important limitations to consider. On one hand, many of the measurements for teamwork were simplistic and problematic. For example, an actor’s experience was based solely on the number of previous movie appearances, without considering what types of movies she has contributed to, and thus has more experience in. Also, team members’ degree dispersions were used to reflect a team’s diversity even though a team composed of actors who have never collaborated with each other can still feature a uniform degree distribution. Although the existence of structural holes can reflect a team’s diversity, the measurement of structural holes was simplified to the density of a network. The two concepts are only very loosely related, however.

On the other hand, the data size was small in many studies. For instance, the top 10 movies (by revenue) in each year (a total sample size of 160-180 movies) were studied in [27, 28]. With such a small sample, an actor’s experience and previous collaborations cannot be completely captured. The selection bias towards more successful movies also hurt the validity of the results. Thus, in this research, we leveraged much larger datasets, derived new and more accurate ways to capture individual actors’ experience and teams’ diversity, and related them to movie profitability.

In terms of what a movie is about, features such as genre, MPAA rating, whether or not a movie is a sequel, and run time have often been incorporated into success predictions, as well as in other domains [1]. Besides such meta data about a movie, to get a better idea of a movie’s content, one needs to examine its plot or script. Two earlier studies leveraged the texts of movie scripts for success predictions [15, 16]. Some of the basic text-based features are easy to obtain, such as the number of words, and the number of sentences. However, more informative textual features in these studies depend on manual annotations by human experts, such as the degree to which the story or hero is logical, and whether or not the story has a believable ending. As movie scripts can be very long, the manual annotations are time-consuming. Also, only a small number of movies’ scripts are available in a uniform and professional format. Thus a predictive model based on features from scripts can only be trained on a small pool of movies, which may limit the predictive power for future movies. Thus an automated way to analyze openly available text-based movie content is necessary for a decision support system to learn from large-scale datasets.

For our research question of predicting movie profitability at an early stage, we cannot take advantage of most audience-based features and some of the release-based features, as they would not be available when making investment decisions. For instance, YouTube comments only appear after a movie trailer is released; likewise, the number of theaters a movie is going to be released in will not be known until the end of the movie's production. In addition, these features from different groups were treated as standalone and independent, whereas the interaction or match between features from different groups, such as actors' star powers along with their experience with different movie genres, or the popularity of a certain type of movie during a specific time period, can provide valuable information about a movies success.

Therefore, we will focus mainly on four types of features: "Who" features – who is involved in a movie, "When" features – when a movie will be released, "What" features from both meta data and text of movie plot synopses (movie plot synopses are openly available from most movie data archives, yet they can still reflect movies' content in the absence of a full script), as well as "'Hybrid" features–the match between "What" and "Who" and the match between "What" and "When". Our feature set includes popular features from the literature (e.g., measuring actor star powers using their total gross revenues), new features proposed to better measure previously proven factors for movie success (e.g., team expertise and diversity), as well as features representing new factors that may be related to movie success (e.g., actor-director collaboration, and market trend by genre). All the features adopted by our system can be extracted in an automated fashion by using text mining and social network analysis techniques. In addition, from a theoretical perspective, this study also examined whether previous findings about star powers of actors and directors, and teamwork are still valid when movie success is measured by profit, instead of revenues, based on a much larger dataset.

The system framework

Figure 1 illustrates the framework of our MIAS. The first phase is data acquisition, because we based our prediction on historical data. We picked two popular and complementary sources – IMDb and BoxOfficeMojo. IMDb has better coverage of movie plot synopses, while BoxOfficeMojo, as its name suggests, provides more comprehensive data of movie revenues and budgets. In other words, the two data sources can be used jointly to acquire data for many movies. As for data collection methods, the two sources are different as well. IMDb has an API to provide movie data. The data from BoxOfficeMojo can only be obtained by the public from its web pages. To get a more comprehensive dataset, our system employs two scripts: one interacts with APIs, while the other is a web scraper to retrieve and parse HTML data from web pages. We believe these two methods should be able to handle data from most open archives on the Internet.

In the second phase, data from both sources is cleaned, transformed, consolidated, and stored in a database. During this undertaking, we make sure that acquired data be put in a consistent format, and that the data is not duplicated within the database. For example, for movie titles, characters such as "*" and "-" are removed. Such standardization ensures that extraneous characters do not occlude the matching of titles between the two data sources. For plot synopses, the Porter stemmer was used, and stop words (such as "the") were also removed.

Phase 3, "feature engineering", involves utilizing the acquired data to construct features that will ultimately be used to train a predictive model. Categorically, we classify various features into one of four groups– "what", "who", "when", or "hybrid". These feature groups will be described in details later in this paper. Features used in this study, and the reasons for including them, will be discussed at length in the next section.

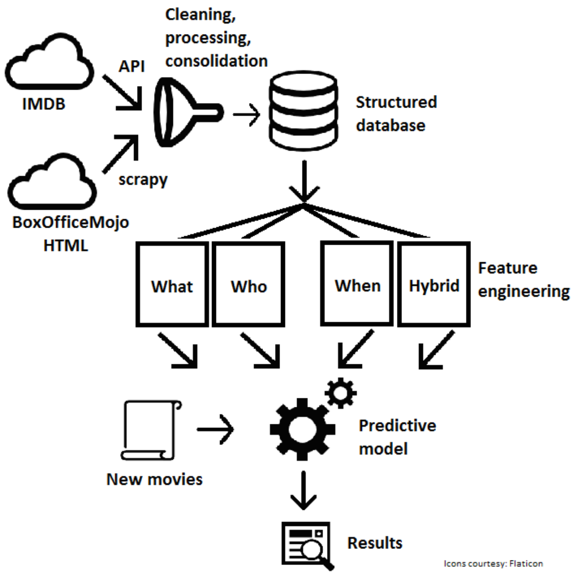


Figure 1: The framework of MIAS.

With a reasonable and well-rounded set of features in place, a predictive model can be trained in Phase 4 of MIAS. Users of MIAS can define their own profitability metric or threshold based on the goals they have for their movies. They can employ cross-validation to select the best performing model of prediction, along with its parameters, based on a maximized performance criteria of interest, such as overall accuracy, precision, or recall. The Experiments section will discuss our experiments in detail.

Feature engineering

Based on historical data acquired from online archives, we derived four groups of features: “who” features, “what” features, “when” features, as well as “hybrid” features that match “who” with “what”, as well as “what” with “when”.

“Who” Features

Star Powers

The very nature of the movie industry is characterized by people who make movies. Successful actors and directors are crowd favorites who are well known throughout the world. Talented individuals can leverage not only their refined industry skills, but also the associated ’name brand effect’, which draws crowds and increases sales [4, 14, 38]. This effect is typically referred to as ‘star power’. Because our goal is to predict profitability, our star power features for a movie are based on its cast members’ records in generating both box-office revenues and profits.

Tenure of an actor reflects how much experience she/he may have in the industry. It is calculated as the time difference (in years) between the movie in which an actor most recently appeared and that in which he/she first appeared. For each movie, we calculate the **average** and **total** tenure for its first-billed actors.

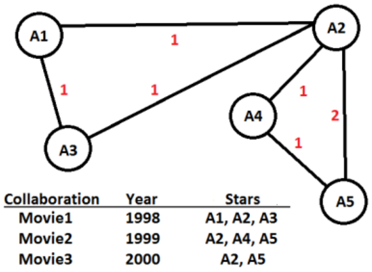
Actor Gross is about how much revenue an actor has generated during her/his tenure. Each individual’s total gross is the sum of revenues from all the movies that she has starred in, while an individual’s average gross is her total gross divided by the number of movies she starred in. For each movie, we calculated the sum and average of total gross, as well as the average of actors’ average gross, for all first-billed cast members.

Director Gross measures the past success of directors. We calculated for each director the **total** and **average** gross for movies she/he has directed.

Actor/Director Profit measures the amount of profit an actor/director has earned through his/her career before the movie to be predicted. For each actor/director, we derived total profit, average profit, and top profit – the profit of the most profitable movie for the actor/director.

Network-based Features

Star power features listed above reflect whether a movie’s cast consists of senior and successful individuals (actors and the director). To capture team characteristics, we explored the avenue of social networks, which have the potential to yield a wealth of information about inter-personal interactions and collaboration [26, 42, 43], including teams for movie productions [25, 27].



Collaboration	Year	Stars
Movie1	1998	A1, A2, A3
Movie2	1999	A2, A4, A5
Movie3	2000	A2, A5

Actors	A1	A2	A3	A4	A5
A1	x	1	1	0	0
A2	1	x	1	1	2
A3	1	1	x	0	0
A4	0	1	0	x	1
A5	0	2	0	1	x

Figure 2: A collaboration network example with the network structure (left) and the corresponding adjacency matrix (right).

For our predictive model, we constructed a dynamic collaboration network among actors based on their co-appearances (i.e. co-starring) in previous movies. In such a network, a node represents an actor. For any arbitrary year, an undirected edge was drawn between two actors if they co-starred in a movie during that year. If an edge already existed between the two, indicating that they had collaborated in the past, the edge weight was incremented by 1. Therefore, the aggregated network for a given year includes all of the earlier years of collaborations, plus those that happen in that year. Fig 2 shows an example network.

Our network features consist of static features and dynamic features. When analyzing the team T_m for movie m in year y, the social network among the movie's cast members up to year $y - 1$ was used to extract the following static features:

Network Heterogeneity: for each movie, we measured its team diversity by examining the structural network similarity between cast members. Specifically, based on each actor's neighborhood vector in the adjacency matrix, we calculated the average cosine similarity between each pair of actors in the movie, which is denoted by Equation 1. In this equation let $|T_m|$ denote the number of cast members in team T_m for movie m, $Act_i \cdot Act_j$ is the dot product between two actors on the team, and $\|Act_i\|\|Act_j\|$ is the magnitude of the two actor vectors. Higher similarity means team members have been working with similar peers (including one another), and vice versa. We believe this measure can better capture previous collaborations among team members than degree dispersions [28], which does not consider who an actor is connected to in a network.

$$H_m = \frac{1}{(|T_m|(|T_m| - 1)/2)} \sum_{i=1}^{|T_m|-1} \sum_{j=i+1}^{|T_m|} \frac{Act_i \bullet Act_j}{||Act_i||||Act_j||} \quad (1)$$

Average Degree represents the average number of unique collaborations for each cast members in a given movie. This metric is meant to capture the ‘degree’ to which the team is truly bringing rich expertise and experience to the production of a movie [28].

Total and Average Betweenness Centrality: In addition to those with many unique collaborators, “brokers” who can bridge different and otherwise less inter-connected groups are also at a good position to bring in unique expertise and experience. These “brokers” often have high betweenness centralities and are said to have high social capital [10]. Having such “brokers” in a team can increase the team’s diversity by creating new ideas and producing innovations [9].

In addition to collaboration among actors, we also considered the collaboration between actors and directors by examining whether an actor and a director have worked together before and whether previous collaborations between them were successful.

Average Actor-Director Collaboration Frequency and Profitability of movie *m* is the average number of times that cast members of *m* have previously appeared in movies directed by the director of *m*, and the average profit per movie earned from all past collaborations between actors and the director of movie *m*, respectively.

The features introduced above are based on the static structure of the year $y - 1$ network before movie *m* was produced in year *y*. Once movie *m* was produced in year *y*, its cast members formed a new team, which would add edges to the collaboration network and change its structure. Our newly developed dynamic network features tried to capture the spanning of structural holes after a new movie is produced.

Structural holes are an important concept in network analysis [10], and networks pertaining to movies are no exception [40]. Some studies believe that a movie, which establishes inter-actor links that span structural holes, is more likely to succeed [7], although they did not quantify the degree to which a movie spans structural holes. Some research used the clustering coefficient for each team member to measure the existence of structural holes [28]. The clustering coefficient of a node is the probability that the node's neighbors are also connected to each other. It measures how a node's neighborhood is "clustered" together. A networks clustering coefficient is the average of all its nodes values of this metric. However, the static value of a team's average clustering coefficient alone, in the current collaboration network, can only show structural holes at the ego level (i.e., among immediate neighbors of an ego). To capture the spanning of structural holes at the network level, we used the following two dynamic network features to measure how the structure of the collaboration network changed after incorporating the new collaboration.

Decrease in clustering coefficient: A new movie will add edges to an existing collaboration network. If these new edges connect nodes that are originally only 2 hops away from each other, then the clustering coefficient of the network will increase, but such edges only reinforce existing clusters. However, by creating new neighbors without closing the triad, new edges that connect nodes that are originally far away would decrease the network's clustering coefficient. It has been found that decreasing the clustering co-efficient of a social network can facilitate the diffusion of information across the network by breaking up existing clusters [42].

Thus we included the decrease in the clustering coefficient of a collaboration network after forming the team for movie *m* to measure whether new collaborations can break existing clusters.

Decrease in average shortest path: We also proposed to use how the production of a movie *m* decreases the average shortest path length of the social network, because adding edges that span structural holes usually significantly decreases such path length. Specifically, after adding to $Network_{y-1}$ edges that correspondent to the cast of T_m produced at year *y*, we calculated how much the average shortest path length of the new network decreased, compared to $Network_{y-1}$. The more such length decreases, the more movie *m*'s cast can span structural holes in $Network_{y-1}$.

"What" Features

In addition to "who" are in the cast, another natural and important indicator of a movie's future profitability is what the movie is about. Such information is usually available with high certainty prior to movie funding efforts. To reflect what a movie is about, the "what" features in our model include meta features, such as **genre** (e.g., action, sci-fi, family) and **rating** (e.g., PG13, and R), represented as binary categorical variables.

We also included a fine-grained description of a movie's content–its plot synopsis. While the full script of a movie is a better representation of the movie's content, such scripts are very difficult to obtain for a large number of movies, especially those that were not very successful. Thus we used plot synopses as approximations for full scripts, as plot synopses are usually publicly available. This allows our predictions to be based on a larger pool of movies.

Representing texts from plot synopses with traditional unigrams and bigrams will have high dimensionality and suffer from sparsity. At a higher level, topic modeling techniques, such as Latent Dirichlet Allocation (LDA) [6], can give a better picture of what a plot is about. The input for LDA is a textual corpus of plot synopses and the output is a group of topics, each being represented by a probabilistic distribution over archetypal words. Those words, having a high probability of a given topic, are considered representative keywords for that particular topic. Each plot synopsis is also assigned a probabilistic distribution over all the topics. In the topic distribution vector of a movie's plot, each element is the probability that the movie represents each topic. Therefore, each element is a continuous numerical value ∈ [0, 1], where 0 indicates that the movie does not at all represent the topic and 1 indicates a perfect representation. Such a topic distribution reflects the content of the movie at an aggregated level and can be used as features for predictions.

In addition to these topics derived from LDA, some movies' plots are adaptations from other sources, an important consideration especially when the original source had achieved certain levels of success. For example, The Hunger Games and Harry Potter are both adapted from best-selling novels. As such, one of our "what" features was about adaptations: whether a movie's plot was adapted from a comic, a true story, or a book/novel.

"When" Features

With the movie industry being an avenue for entertainment, its market sees peaks and declines over time, which may speak to how well a yet-produced movie may fare in the future. Thus we incorporated the following "when" features in our model: **Average Annual Profit** is the average profit across all movies in the year prior to the planned release of movie *m*. It captures the overall profitability of the movie industry before a movie is released. **Release dates** combines several features about when a movie will be released, including whether it will be a holiday release and which season of the year (spring, summer, fall, winter). While a holiday or summer release may attract more of an audience and thus generate more revenues [2], it also requires higher budget for marketing and distributions during these competitive periods. Although the exact release date is not completely definitive before filming, a target trajectory usually exists at pre-production stages.

Hybrid Features

Besides standalone features about "'who" are in a movie, "what" a movie is about, and "when" a movie will be released, it is also important to capture the "match" between these features. Our hybrid features try to reflect such matches between "what" and "who", as well as between "what" and "when". For example, it may be important to form a team of actors based on their previous experience with the genre of the movie being planned instead of just their star powers. Similarly, the investment on a movie whose genre is gaining popularity may increase the chance of success.

"What" + "Who"

In observing the movie industry and the actors, we can distinguish various so-called 'roles' that these actors seem to adopt. For instance, Seth Rogan is typified by his appearance in comedies, and Arnold Schwarzenegger exhibits a proficiency as an action movie star. Should a movie then try to include those who have extensive experience in its genre? Or conversely, does a surprising cast draw a greater audience to theaters (e.g., having Schwartzenegger in a comedy or a romantic story)? Although these questions have not been addressed in the literature, we believe that better measurements of an actor's expertise with regard to genres can help us more accurately determine the expertise and diversity of a movie's cast.

To measure an actor's previous experience and expertise in movies with different genres we define, for each actor j, a genre experience vector $A_j = [a_{j,1},...,a_{j,k},...,a_{j,K}]$, where $a_{j,k}$ is the proportion of the number of times actor *j* appeared in movies with genre *k*. A total of $K = 26$ unique genres are defined. Similarly, a movie *m* is also represented as a genre vector $G_m = [g_{m,1},...,g_{m,k},...,g_{m,K}]$, where $g_{m,k} = 1$ indicates that movie *m* has genre *k*, and $g_{m,k} = 0$ otherwise. Note that some movies can have more than one genre. For example, the genre of 'Spiderman' is both action and adventure. By measuring the similarity between actors' genre experience vectors and movies' genre vectors, we designed several features that speak to the genre-based expertise brought by cast members to a given film *m*'s team T_m.

Average Genre Expertise (AGE): captures the average cast experience with respect to the current movie's genre. Movie *m*'s AGE is defined in Equation 2.

$$AGE_m = \frac{1}{|T_m|}\sum_{j=1}^{|T_m|} G_m \bullet A_j \tag{2}$$

Weighted Average Genre Expertise (WAGE) extends AGE by incorporating an actor's star power, measured by actor gross, in each genre. As defined in Equation 3, the WAGE of movie *m* is essentially the movie's AGE weighted by each cast member *j*'s gross revenue R_j. In other words, a movie with a big star who is familiar with its genre will have high WAGE.

$$WAGE_m = \frac{1}{|T_m|}\sum_{j=1}^{|T_m|} log(R_j) * (G_m \bullet A_j) \tag{3}$$

Cast Novelty is defined in a way similar to WAGE. While WAGE is an average value that tries to capture a cast's experience in the movie's genre, cast novelty focuses on team diversity–whether a movie has a big star who has rarely appeared in movies of this genre before. It is the maximum value among all actors' star-power-weighted inverse experience in movie *m*'s genre (Equation 4). Higher values indicate having an unexpected star appearing in a given movie.

$$CN_m = max\{\frac{log(R_j)}{G_m \bullet A_j + 1}, \forall j \in T_m\} \qquad (4)$$

“What” + “When”

Similar to the overall market volume for movies, which changes over time, consumers’ preferences of movies may also evolve from year to year. For example, while movies like “American Pie” and “National Lampoon’s Van Wilder” were popular in the late 1990’s and early 2000’s, movie-goers recently have been flocking to horror movies, such as “Paranormal Activity”, and those characterized by superheroes, such as “The Avengers”. Although the latter category is nothing definitively new to the silver screen, the movie industry has seen greater levels of success in recent years with this particular focus and, as such, a greater influx of such movies. Meanwhile, competition may also affect the profitability of movie *m* because other movies released during a similar time period may detract from movie *m*’s viewer-base [28]. Thus, in addition to capturing “when” a movie will be released, we also consider how movies with similar genre performed in the previous year, as well as the level of competition during a movie’s planned release time.

Annual Profitability Percentage by Genre is the percentage of movies, which have the same genre as that of movie *m*, in the year prior to the planned release of movie *m*, that were profitable. This feature reflects the degree of success for movies that share the same genre as the movie being considered.

Annual Weighted Profitability by Genre (AWPG) is derived from movie genre vectors defined earlier in this paper. For movie *m* in year *y*, the profitability of each movie m' in year $y -$ 1 are summed up and weighted by the cosine similarities between genre vectors of *m* and each m'. Equation 5 illustrates how to calculate the AWPG for movie *m* in year *y*, where G_m is the genre vector for movie *m* and $p(m')$ is the profitability of movie m'. This feature indicates the overall previous-year profitability of movies whose genre is similar to a given movie.

$$AWPG_m = \sum_{m' \epsilon y-1} sim(G_m, G_{m'}) * p(m') \qquad (5)$$

Competition reflects the other movies that will be released during a similar time period. It is calculated by considering the average star-power of all other movies released within 1 month of movie *m*’s release date. This feature indicates the degree to which other big-name stars, appearing in movies at a similar time, that may detract from movie *m*’s viewership. The inclusion of such a feature is based on the fact that, even prior to production, a movie has a loose, or at times even definitive, trajectory for release. By defining competition to be within one month (plus or minus) of the original release date, such a notion of “approximate release” is maintained, even in instances in which the exact date may have been well established (i.e., such an approach is conservative).

Experiments

Dataset and basic statistics

Our original dataset, collected from both BoxOfficeMojo and IMDb, consisted of 14,097 movies, along with 4,420 actors. While movies in our dataset date back to 1921, we focused our study on movies released during the 11-year period of 2000-2010 (inclusive), because this period is recent enough to reflect the current state of the industry, while ensuring that there has been a sufficient amount of elapsed time since these movies release for revenue data to be accurately updated.

As our goal is to predict movie success measured by profits, our dataset for experiments only included those movies that have both budget and box office revenue data available. We also excluded movies with an ‘Unknown’ genre, or an ‘Unknown’ MPAA rating. ‘Documentary’-genre movies were also excluded, as those are typically not released to theaters and may not involve professional actors. Additionally, any movie designated as being part of a franchise, a sequel, or a remake was also excluded (e.g. Iron Man, Iron Man 2, etc.). We made this decision because the success of a sequel can depend heavily on the success of earlier movies in the same franchise. Also, the content of sequels and remakes and their selection of cast members are also highly limited by earlier counterparts. Thus what is behind the success of a sequel or remake may be very different from that of other movies.

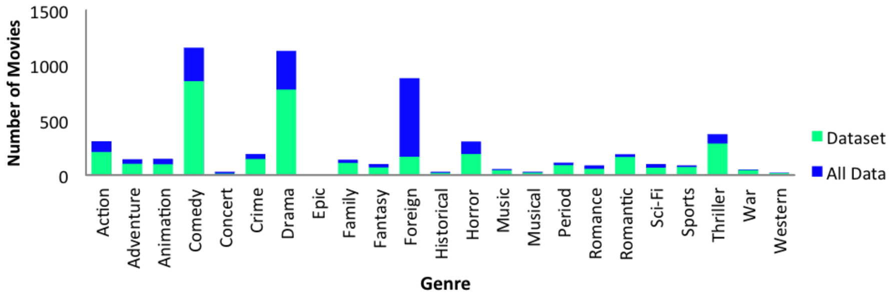
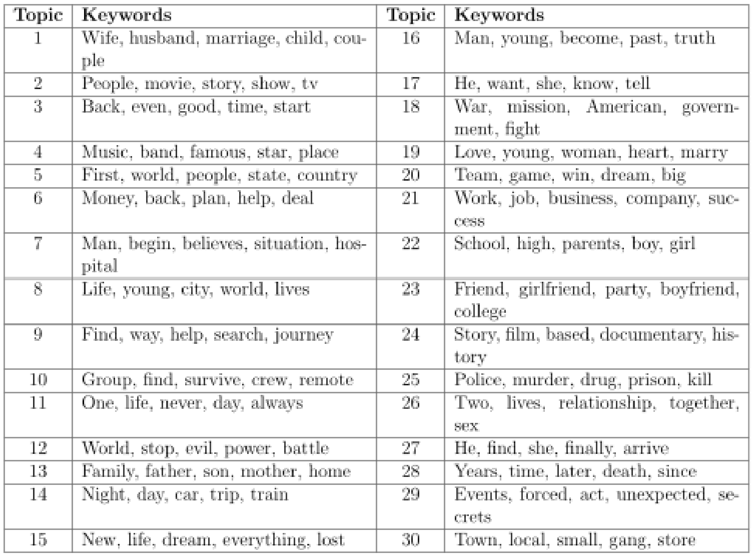


Figure 3: Distribution of movies by genre (2000-2010).

With these considerations in mind, our final dataset for experiments consisted of 2,506 movies. A distribution by genre of these 2,506 movies, relative to all movies released during the period, is related in Figure 3. The distribution suggests that our dataset is a representative sample overall, with the exception of the ‘Foreign’ genre. This makes sense because budget and revenue data may be more difficult to obtain for movies that are produced, and in all likelihood, released outside the U.S. Based on the plot synopses of these movies, we used LDA to generate 30 topics. Top keywords of these topics are listed in Table 1.

While the experiment will predict the success of 2,506 movies during an 11-year period, the collaboration network we built for this study incorporates the collaboration between all the actors in all 14,097 movies in our dataset. The initial unweighted, undirected network was aggregated to the year 1999, with networks for subsequent years being updated to reflect that year's new collaborations. In all, we created 11 snapshots of the collaboration network from 1999 to 2009.



Topic	Keywords	Topic	Keywords
1	Wife, husband, marriage, child, couple	16	Man, young, become, past, truth
2	People, movie, story, show, tv	17	He, want, she, know, tell
3	Back, even, good, time, start	18	War, mission, American, government, fight
4	Music, band, famous, star, place	19	Love, young, woman, heart, marry
5	First, world, people, state, country	20	Team, game, win, dream, big
6	Money, back, plan, help, deal	21	Work, job, business, company, success
7	Man, begin, believes, situation, hospital	22	School, high, parents, boy, girl
8	Life, young, city, world, lives	23	Friend, girlfriend, party, boyfriend, college
9	Find, way, help, search, journey	24	Story, film, based, documentary, history
10	Group, find, survive, crew, remote	25	Police, murder, drug, prison, kill
11	One, life, never, day, always	26	Two, lives, relationship, together, sex
12	World, stop, evil, power, battle	27	He, find, she, finally, arrive
13	Family, father, son, mother, home	28	Years, time, later, death, since
14	Night, day, car, trip, train	29	Events, forced, act, unexpected, secrets
15	New, life, dream, everything, lost	30	Town, local, small, gang, store

Table 1: Topics and keywords generated by LDA from plot synopses.

The measure of success

To predict whether a movie is successful, we measured a movie's profitability using two metrics–raw profit and return on investment (ROI). Raw profit is simply revenue minus budget. For movies in our dataset, only 36% of the movies had positive profits. However, a profit of $10,000 from a movie that costs $1 million to produce is certainly not an attractive investment prospect. Thus in our experiments, we also adopted return on investment (ROI) as in [14]. Considering both profit and budget, ROI is defined as $ROI = \frac{Revenue - Budget}{Budget}$. The higher the ROI is, the more profitable a movie is, and vice versa. To avoid 'trivial' and disinteresting profit returns, we raised the bar of profit for a movie to be considered "truly profitable", and will discuss this in our experiments.

Interestingly, the data suggests that profitability, as measured by ROI, is not necessarily reflected by box office revenues. The correlation coefficient between revenues and ROIs is only *0.077*. In other words, having a great box office revenue does not necessarily mean a high ROI. Similarly, albeit with a higher correlation coefficient, raw profit has a correlation coefficient of *0.22* with revenue. These further highlight the need for an accurate prediction of profitability.

Classification of profitability

The prediction of a movie's success can be modeled as a classification problem to decide whether a movie should be considered a success or not based on either ROI or profit. Although there is no agreed-upon industry 'gold standard' as to an ideal ROI or profit, it is reasonable to assume that one would like to see some substantive returns from a successful movie, given that millions of dollars are invested with considerable risks. Also, any form of profit is better than a loss. Thus we elected to define the decision boundary between successful and unsuccessful movies in two different ways for both binary and multi-class predictions.

For both binary and multi-class classification of movie success, we tried a variety of algorithms, including logistic regression, naive Bayes, support vector machines (SVM), multilayer perceptron (MLP), decision trees (J48), random forest, and the LogitBoost, and selected the one with the best overall performance based on the following 4 metrics (using 10-fold cross-validation), where higher values indicate better performance: (1) *Classification accuracy*, which is the percentage of correctly predicted instances; (2) *Precision* (positive class), which is the number of instances classified as being positive that are actually successful, divided by the number of instances classified as being successful; (3) *Recall* (positive class), which is the number of instances classified as being positive that are actually successful, divided by the number of instances that are actually successful; and (4) The *Area under the Receiver Operator Characteristic curve (AUC)*. The curve plots the true positive rate against the false positive rate.

An AUC of 1 means a perfect classification while 0.5 refers to a random guess. Being more robust against prior distributions, AUC is considered by many to be one of the best indicators of a classifiers performance [34]. In multi-class classifications, we reported a weighted average of AUCs: first calculate the AUC obtained for each class, and then weight each AUC according to the number of instances that fall into each of these classes relative to the total number of instances.

In addition to identifying the best performing algorithm, we also evaluated whether, and how, features we proposed in this research contributed to the prediction. We included, in a 'New' feature group, those novel features which we proposed and are used for the first time to predict movie success. Features in this 'New' group include (1) features related to actor and director profits, actor-director collaboration, dynamic network features (e.g., decrease in the average shortest path length) from the 'Who' group; (2) topic distribution features from the 'What' group; (3) average annual profit in the 'When' group; and (4) all features in the 'Hybrid' group.

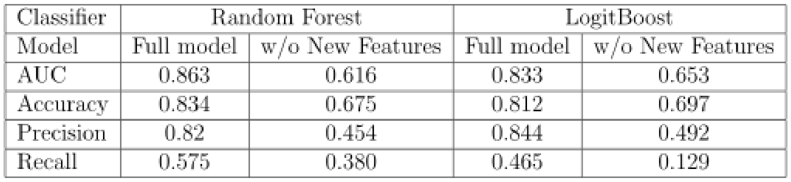
To further evaluate the performance of our predictive model, we also compared it with two benchmark models. Our benchmark models were constructed using features that were defined in past studies and derived from our data. We also used our definitions of profitability and the same set of classification algorithms. In so doing, we are able to estimate how these past studies would have performed, allowing for an apples-to-apples comparison. Benchmark 1 was based on [37] and Benchmark 2 was based on [39]. Among previous studies of box office revenue predictions, we selected these two because most features used in their studies are available prior to a movie's production, which is similar to our early prediction problem. For example, the following features were used in [37]: star power, sequel, genre, rating, and year of release. Similarly, [39] used film budget/cost, number of screens the film was released on, sequel, star power, genre, and rating. We excluded the sequel feature, as our dataset excludes such movies, and the number of screens feature, because this information is not available prior to release. To make the comparison consistent, we used our definition of star power for the two benchmark models. We reported results of best-performing classifiers for both benchmark methods, along with the performance of our approach.

We would like to point out, prior to reporting the results of such experiments, that a user of the MIAS system has the option to select the desired algorithm used in making predictions based on any of the aforementioned criteria they are interested in maximizing. Investors are most likely interested in recall and AUC based on their preferences. Selecting an algorithm that maximizes recall helps ensure that investors not miss out on any movies that end up being profitable, but may end up with movies that are not eventually profitable. An algorithm that maximizes AUC, on the other hand, trades off assuring that investment opportunities are not missed, but has the added benefit to protect the investor from a monetary loss.

Additionally, we imagine that users of this system will be tailoring, tuning, and evaluating its performance as new data is incorporated. In so doing, the user of MIAS may find that another algorithm performs better in terms of a particular criteria of interest. If this is the case, the user, of course, has the option of making their decisions based on the better-performing model.

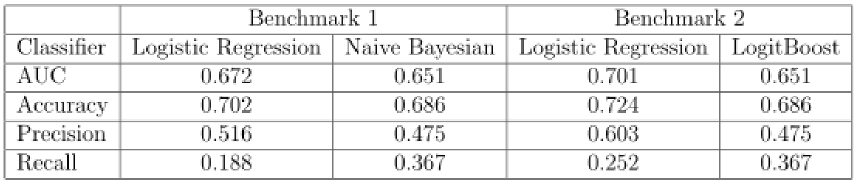
Binary classification

For binary classifications, a movie is classified into one of two classes: successful or unsuccessful. Two decision boundaries are evaluated and both ensure a sufficient amount of either ROI or raw profit is garnished for a movie to be considered successful. The first decision boundary entails a movie be considered successful if its ROI is within top 30% of all movies. For our dataset, this threshold translates to $ROI \geq 24\%$. Table 2 lists the performance of the top two classification algorithms a Random Forest classifier (n=200) and a LogitBoost classifier.



Classifier	Random Forest		LogitBoost	
Model	Full model	w/o New Features	Full model	w/o New Features
AUC	0.863	0.616	0.833	0.653
Accuracy	0.834	0.675	0.812	0.697
Precision	0.82	0.454	0.844	0.492
Recall	0.575	0.380	0.465	0.129

Table 2: Top 2 prediction results of our binary classification model and the performance without 'New' features (with top 30% ROIs as the decision boundary).

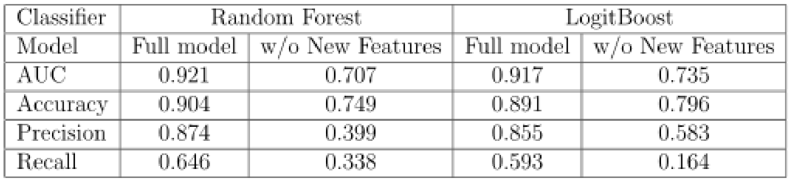


	Benchmark 1		Benchmark 2	
Classifier	Logistic Regression	Naive Bayesian	Logistic Regression	LogitBoost
AUC	0.672	0.651	0.701	0.651
Accuracy	0.702	0.686	0.724	0.686
Precision	0.516	0.475	0.603	0.475
Recall	0.188	0.367	0.252	0.367

Table 3: Top 2 prediction results for benchmark binary classification models (with top 30% ROIs as the decision boundary).

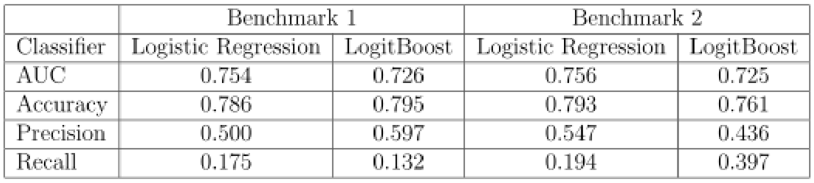
Table 2 also highlight the contribution of 'New' features to the prediction. When 'New' features were removed, AUC and accuracy of the classifiers deteriorate 29% and 20% respectively for random forest, and 22% and 15% for LogitBoost. Precision and recall also drop greatly. In addition, the top 2 classifiers of the two benchmark models (Table 3) trail our model in all of the four performance metrics. For instance, AUCs of the two benchmark models are respectively 19% and 25% lower than that of ours.

The second boundary we tested is *Profit* ≥ $7.3 million, which corresponds to 1/4 standard deviation above the mean profit. With this threshold, 21.4% of the movies in our dataset were considered successful. Top performing algorithms are able to reach AUC and accuracy over 0.9 (see Table 4). At the same time, our 'New' features still make great contributions to the classification– the removal of these features dropped the AUC of the best-performing random forest classifier by 24%, and the LogitBoost classifier's AUC decreased by 20%. Similar decrease can be found for accuracy, precision, and recall. Our model also keeps the advantage over the two benchmark models, leading by 22%-27% AUC.



Classifier	Random Forest		LogitBoost	
Model	Full model	w/o New Features	Full model	w/o New Features
AUC	0.921	0.707	0.917	0.735
Accuracy	0.904	0.749	0.891	0.796
Precision	0.874	0.399	0.855	0.583
Recall	0.646	0.338	0.593	0.164

Table 4: Top 2 prediction results of our binary classification model and the performance without 'New' features (with *Profit* ≥ $7.3 million as the decision boundary).



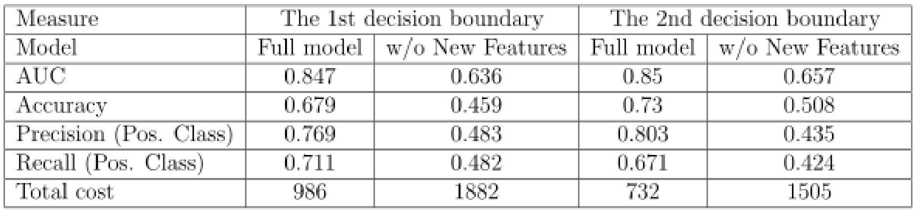
	Benchmark 1		Benchmark 2	
Classifier	Logistic Regression	LogitBoost	Logistic Regression	LogitBoost
AUC	0.754	0.726	0.756	0.725
Accuracy	0.786	0.795	0.793	0.761
Precision	0.500	0.597	0.547	0.436
Recall	0.175	0.132	0.194	0.397

Table 5: Top 2 prediction results for benchmark binary classification models (with *Profit* ≥ $7.3 million as the decision boundary).

Multi-class classification

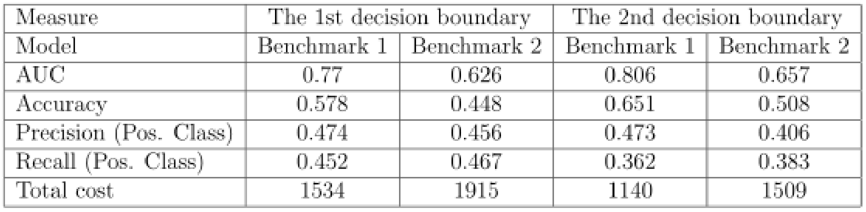
In the case of multi-class classifications, we defined three possible classes for a movie: positive ('success'), negative ('failure'), or neutral ('average') to provide more information to investors on where they could expect a movie to fall with regard to profitability. For the multi-class prediction, we explored the imposition of cost [13] associated with mis-classification, because the three classes are ordinal. In other words, not all mis-classification errors are equally severe. For example, for investment decision support, predicting a failure to be a success would be worse than predicting it to be a neutral movie. The cost matrix for the multi-class classification is in Table 6–the penalty imposed for classifying a *successful* movie as *failure* is 2, and vice-versa, whereas the penalty for only mis-classifying by one ordinal category (i.e., success as neutral, etc.) is 1.

Similar to binary classifications, we defined the three classes of success in two ways. The first way was to split movies into three equal-sized classes: the positive class consists of movies with top 1/3 ROIs ($ROI \geq 10\%$), the negative class consists of movies with the bottom 1/3 ROIs ($ROI \leq -78\%$), and the other middle 1/3 into the neutral class ($-78\% < ROI < 10\%$). The second way was to classify movies with top 1/4 ROIs as positive ($ROI \geq 47\%$), the bottom 1/4 ROIs as negative ($ROI \leq -91\%$), and the rest as neutral ($-91\% < ROI < 47\%$).



Measure	The 1st decision boundary		The 2nd decision boundary	
Model	Full model	w/o New Features	Full model	w/o New Features
AUC	0.847	0.636	0.85	0.657
Accuracy	0.679	0.459	0.73	0.508
Precision (Pos. Class)	0.769	0.483	0.803	0.435
Recall (Pos. Class)	0.711	0.482	0.671	0.424
Total cost	986	1882	732	1505

Table 6: Multi-class classification results of our model from the best-performing random forest classifier, and the performance without ‘New’ features.



Measure	The 1st decision boundary		The 2nd decision boundary	
Model	Benchmark 1	Benchmark 2	Benchmark 1	Benchmark 2
AUC	0.77	0.626	0.806	0.657
Accuracy	0.578	0.448	0.651	0.508
Precision (Pos. Class)	0.474	0.456	0.473	0.406
Recall (Pos. Class)	0.452	0.467	0.362	0.383
Total cost	1534	1915	1140	1509

Table 7: Multi-class classification results of benchmark models using random forest classifiers.

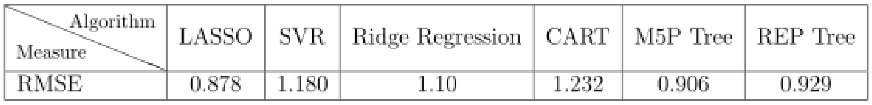
After comparing the performance of several classification models (including J48, Naive Bayes, MLP, SVM, logistic regression, and LogitBoost), random forest still emerged as the best classifier for both decision boundaries. Table 6 lists the performance measures and Table 7 shows the performance from the two benchmark methods. Similar to that of binary classifications, our model outperforms the two benchmark models by reducing the total mis-classification cost by 36%-52%. Meanwhile, ‘New’ features keep making great contributions to the prediction– the exclusion of these features from our model doubles the mis-classification cost.

Discussions

Regression analysis

While a random forest classifier can do a good job in predicting whether a movie will be successful, it is also important to understand factors behind such success. A regression model would help us better assess the degree to which individual features influence predictive results, and to examine whether they are indicative of movie profitability. Besides, a regression model can also provide predictions on numeric values, in case the classification of movies into 2 or 3 discrete groups is not sufficient for investors' needs. Thus we also explored predicting continuous ROI values. It is worth noting that because the distribution of ROI is highly skewed, we applied a logarithm transformation to ROI, in the format of $log(ROI + 1)$ as in [16].

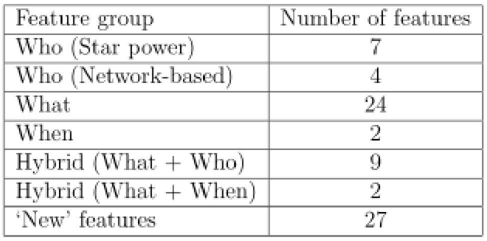
We tried 6 different algorithms, namely LASSO, Support Vector Regression (SVR), Ridge Regression, CART, M5P Trees, and REP Tree. Among them, we were particularly interested in LASSO and Ridge Regression for two reasons: First, coefficients of each feature in these models are able to offer valuable insights into how each feature contributes to a movie's profit. Second, multi-collinearity may exist among our features (or independent variables). For example, the correlation between Total Actor Profit and Average Actor Profit is 0.96 (p-value<0.001). LASSO and Ridge Regression both use regularization ($L1$ and $L2$, respectively) to penalize the non-zero values of the regression coefficients. Such regularization allows the model to select features that are more informative for the prediction and reduce the impact of collinearity [22].



Algorithm Measure	LASSO	SVR	Ridge Regression	CART	M5P Tree	REP Tree
RMSE	0.878	1.180	1.10	1.232	0.906	0.929

Table 8: Results for predicting $Log(ROI + 1)$ with various algorithms (10-fold cross validation).

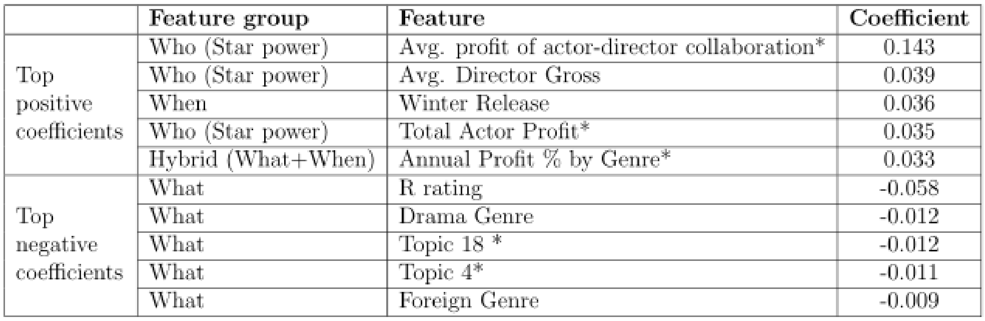
Table 8 compares root mean squared errors (RMSE) of the 6 algorithms, with LASSO being the best at predicting numeric ROI values. Thus we used coefficients from LASSO to reveal factors behind movie success. To obtain these coefficients, we iteratively increased the penalizing λ value until all attribute-wise variance inflation factors were reduced to below 10 [12]. After achieving such a result with $\lambda = 0.0065$, 48 out of the 120 features in the LASSO model ended up having non-zero coefficients: 16 have negative coefficients, and 32 have positive coefficients.



Feature group	Number of features
Who (Star power)	7
Who (Network-based)	4
What	24
When	2
Hybrid (What + Who)	9
Hybrid (What + When)	2
‘New’ features	27

Table 9: Number of features from each feature group for the 48 features with non-zero coefficients in LASSO.

Table 9 shows how many of the 48 features are from each feature group, including the ‘New’ feature group for novel features proposed in this research. It turns out the 48 features cover all feature groups, and more than half of them are ‘New’ features. Besides the ‘New’ feature group, the ‘What’ group contributes the most features. For example, 12 out of the 30 topics derived from LDA are among the 48.



	Feature group	**Feature**	**Coefficient**
Top positive coefficients	Who (Star power)	Avg. profit of actor-director collaboration*	0.143
	Who (Star power)	Avg. Director Gross	0.039
	When	Winter Release	0.036
	Who (Star power)	Total Actor Profit*	0.035
	Hybrid (What+When)	Annual Profit % by Genre*	0.033
Top negative coefficients	What	R rating	-0.058
	What	Drama Genre	-0.012
	What	Topic 18 *	-0.012
	What	Topic 4*	-0.011
	What	Foreign Genre	-0.009

Table 10: Top features with the highest positive and lowest negative regression coefficients from the LASSO model (* designates a ‘New’ feature).

Table 10 lists the top 5 features by the value of their coefficients. A feature with a positive coefficient indicates that it has a positive influence on profitability, while a feature having a negative coefficient is indicative of a negative influence on profitability. For example, an increase of one standard deviation in the average profit of actor-director collaboration is associated with 14.3% increase in *log*(*ROI* + 1), while being an R-rated movie decreases the *log*(*ROI* + 1) by 5.8%. While those with top negative coefficients are all ‘What’ features, including genre (drama and foreign), ‘R’ rating, and plot topics related to wars and music, it’s a mix of other feature groups (Who, When, and Hybrid) that are indicative of profit.

We also checked the robustness of the 48 top features by comparing them (generated by LASSO with $\lambda = 0.0065$) with top features generated by LASSO with two other λ values $\lambda = 0.0044$ and $\lambda = 0.0025$. With lower λ values in LASSO, more features are selected by the two models. However, all of the 48 features still appear in both lists of top features. This results confirmed that the list of 48 features is a robust representation of key factors for movie profitability.

Star powers and movie profits

As we mentioned before, star power features have a large and positive bearing on the success of movies. While previous studies agreed that higher start powers are generally associated with movie success [37], they relied on movies’ box-office revenue with actors’ star power measured by their total gross revenue. Figure 4 plotted total actor gross against movie revenues and profits in our experiments. As we can observe, although total actor gross is moderately correlated with movie revenues (Pearson correlation coefficient *0.46*), the correlation is much weaker with profits (Pearson correlation coefficient *0.16*). In other words, having actors who have earned big box-office revenues in a movie does not necessarily mean more profit for the movie. Such a result further highlights the difference between measuring movie success with revenues and profits.

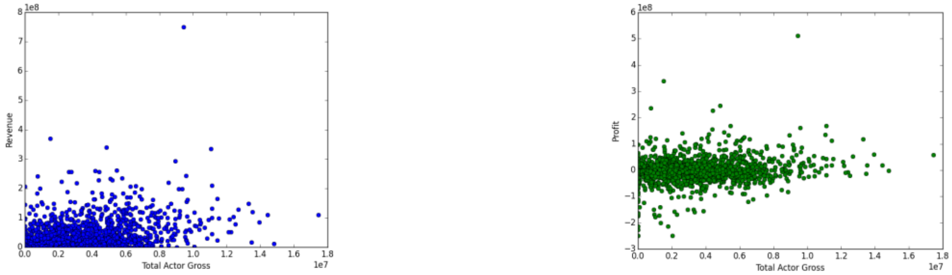
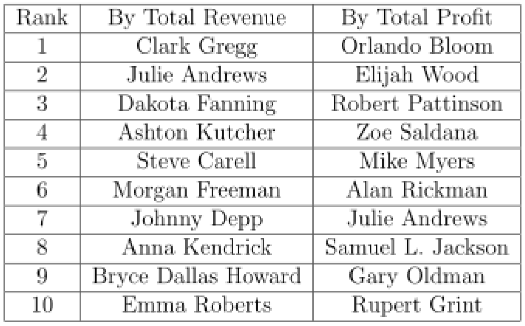


Figure 4: Total actor gross vs. movie revenues (left) and profits (right).

By focusing on historical profitability records of both actors and directors, our results revealed some interesting findings about movie profitability. For example, the director is an important factor for movies' profits. The top feature from our LASSO model is actually the average profit of previous actor-director collaboration. Also, the average profit of actor-director collaboration is a better indicator of $log(ROI + 1)$ than the traditional star power measure of total actor gross – the rank correlation between the average profit of actor-director collaboration and $log(ROI + 1)$ is *0.47*, while total actor gross has a rank correlation of *0.29* with $log(ROI + 1)$. Besides, according to Table 10, having a star director is more indicative of profit than having a cast of star actors. Such findings actually contrasted with a few studies that even considered the effect of directors on movie success, albeit measured by box-office revenues. We conjectured that the difference may be due to measuring movie success using profits instead of revenues, and the usage of a larger dataset with movies whose success levels vary greatly. Although further investigations along this direction are beyond the scope of this research, we do believe that this is an interesting result that is worth exploring from team performance or marketing perspectives.

Also, when it comes to predicting movie profits, actors' star power is better measured with their historical profits than with their gross. In fact, when we ranked actors by their total gross revenues and total profits, the ranking correlation is only moderate (Spearman coefficient of *0.60*). Table 11 lists the top 10 actors by total revenues and total profits, respectively, and there is only one (Julie Andrews) who appears on both lists. Although ‘big-name’ movie stars are likely to attract quite a crowd, in the case of generating profits, the cost of casting such a star may not always be recouped via tickets sales.



Rank	By Total Revenue	By Total Profit
1	Clark Gregg	Orlando Bloom
2	Julie Andrews	Elijah Wood
3	Dakota Fanning	Robert Pattinson
4	Ashton Kutcher	Zoe Saldana
5	Steve Carell	Mike Myers
6	Morgan Freeman	Alan Rickman
7	Johnny Depp	Julie Andrews
8	Anna Kendrick	Samuel L. Jackson
9	Bryce Dallas Howard	Gary Oldman
10	Emma Roberts	Rupert Grint

Table 11: Top 10 actors by total revenues and total profits.

Teamwork and profitability

From a teamwork perspective, we investigated the effect of expertise and diversity, and found that both can contribute to a movie’s profits. Our new metric to measure how much expertise a cast has in a specific movie’s genre – Average Genre Expertise – is positively related to profits (with a coefficient of *0.007*). Along with the top positive coefficient for average actor-director collaboration profits, they have highlighted the importance of a cast’s expertise and successful collaboration experience in the past. At the same time, diversity is also a positive predictor of profits. Among our diversity metrics, decrease in clustering coefficient, cast novelty, and the spanning of structural holes (measured by the decrease in the average shortest path lengths) all have positive coefficients in the LASSO model. In other words, a cast with members who have previous experience on a movie’s genre, yet with some fresh faces, is beneficial for a movie's profits.

Cast selection via inverse classification

To demonstrate the power and additional utility of our proposed system beyond prediction, we extended our experiments to prescriptive analytics. Specifically, we will show how the MIAS system can be leveraged to prescribe a set of cast members that maximize the probability of observing profitable returns according to one of our two definitions of profitability: ROI $\geq$ 24%. Practically speaking, however, a thorough exploration of such prescriptive analysis is a massive undertaking, because the search space is large due to the large number of possible combinations of actors for a movie' cast; the optimization depends on the outcome of predictions as well. As such, we deployed a small-scale experiment in this paper, leaving more definitive elaborations for future work.

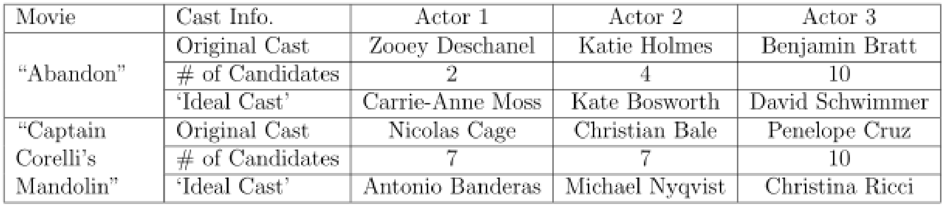
The problem of selecting a profit-maximizing cast will be addressed through a process known as inverse classification [2, 11, 24] --the process of making perturbations to a test instance (i.e., a movie, in this case) such that the probability of some ideal class (i.e., positive profit) is maximized. The problem can be formulated as the partitioning of features into two groups: those that are unchangeable and those that are changeable [24]. For this experiment, we assume cast member-based features are changeable (e.g., past actor-director collaborations, network-based features, etc.) and all others to be unchangeable (e.g., genre, rating, etc.).

To construct an appropriate set of cast members, we used the original cast of a particular movie as the starting point, and observed retrospectively what changes to cast members could have led to higher probabilities of profitability. Specifically, for each of the original cast members i who planned to appear in movie M ($i \in C_M$), we wanted to construct a separate set of cast members to maximize the movie's probability of profitability. More formally, the problem is to $\forall i \in C_M$, construct candidate cast member set $C_{Mi}{}'$, where $\forall i' \in C_{Mi}{}'$ would replace $i \in C_M$. In other words, we want to construct a set of candidate casts C_M' = Combine(C_{M1}',C_{M2}',…,$C_{M|CM|}'$) where Combine(·) is a function that returns the non-overlapping combinations of elements from each respective candidate actor set. This process yields a total of $|C_M'| = \prod_{i=1}^{|C_M|} |C_{Mi}'|$ candidate cast sets. As can be seen, the number of possibilities is combinatorially large. For example, for a movie with 5 actors, if we can identify 10 candidates for each of the 5 actors, then the total number of possible cast sets to be evaluated is 10^5.

To reduce the search space, we required that each of the candidates in the new cast has to be similar to each of the original cast members in three ways: star power, gender and age. Star power (SP) serves as a proxy for the amount of pay needed to hire an actor (measured by cumulative total gross). Gender and age help to ensure that the role is aptly represented (e.g., a teenage girl in a movie is usually starred by a young female actress). Age was approximated by tenure, the time in which a particular actor has been appearing in movies. We also required that i' have appeared in a movie in the last y years to ensure, as best as possible, that candidates i' are still actively pursuing work in the movie industry. Specifically for our experiments, the following similarities need to exist between i and i': (1) $\mathrm{SP}_{i}' \in [\mathrm{SP}_i * 0.90, \mathrm{SP}_i * 1.05]$, so that a replacement i' costs similarly to hire compared with i; (2) $\mathrm{gender}_i' = \mathrm{gender}_i$; (3) $\mathrm{tenure}_i' \in [\mathrm{tenure}_i - 7, \mathrm{tenure}_i + 7]$, so that the i''s age is close to to that of i's; and (4) $\mathrm{activity}_i' \in [M_{\mathrm{release}} - 3, M_{\mathrm{release}}]$ to ensure that i' has been active during the past 3 years.

In our experiments, for each candidate actor set $s \in C_M'$, we computed cast-dependent features (i.e., changeable features) and appended them to non-cast-dependent features (i.e., unchangeable features that are the same for each $s \in C_M'$). This created a unique movie M_s, to which we can apply our predictive model that returns the probability of M_s having ROI ≥ 24%. After predicting each M_s using our random forest classifier, we can select $M_{s^*} = \max\{P(M_s), \forall s$

$\in C_M{}'$ }, where $P(.)$ denotes the probability of being profitable from the predictive model, and $M_s{*}$ is the movie, starred by candidate cast set s, that is most likely to earn ROI ≥ 24%. The process of computing all cast-dependent features for all $|C_M{}'|$ candidate cast sets is time-consuming. A more comprehensive exploration along this direction requires a more efficient process for finding $M_s{*}$.



Movie	Cast Info.	Actor 1	Actor 2	Actor 3
"Abandon"	Original Cast	Zooey Deschanel	Katie Holmes	Benjamin Bratt
	# of Candidates	2	4	10
	'Ideal Cast'	Carrie-Anne Moss	Kate Bosworth	David Schwimmer
"Captain Corelli's Mandolin"	Original Cast	Nicolas Cage	Christian Bale	Penelope Cruz
	# of Candidates	7	7	10
	'Ideal Cast'	Antonio Banderas	Michael Nyqvist	Christina Ricci

Table 12: Cast selection results for two movies.

We conducted experiments on two different movies–"Abandon", a 2002 PG-13 rated thriller, and "Captain Corelli's Mandolin", a 2001 R rated War-Romance movie. Both were originally not profitable, according to our ROI ≥ 24% definition, and their inability to be profitable has been correctly predicted by our models. Also, both movies had three first-billed actors. Table 12 summarizes the experiment results for the two movies including the original cast, the number of candidate cast members for each original cast member, and the 'ideal cast' that would maximize the probability of being profitable. Had the two movies picked the 'ideal cast', the predicted probability of having ROI ≥ 24% would have increased from 0.07 to 0.35 for "Abandon", and 0.08 to 0.47 for "Captain Corelli's Mandolin".

Limitations

Admittedly, our study has limitations. For one, similar to past studies, the profit we calculated is based on estimated production budget and reported box office revenue. However, the true profit of a movie may be obscured by certain accounting practices. Teasing out the effect of such practices is very challenging without sensitive accounting data and such an endeavor is beyond the scope of this paper. Additionally, for many movies, box office revenue is only one of the sources of income. For example, Disney's animation movies often gain a significant amount of their revenues from the sale of movie merchandise, such as clothing and toys. Some movies may also rely heavily on the sale and rental of DVDs. However, capturing these non-box-office revenues is more difficult as they may keep accumulating many years after the release of a movie.

Conclusions and Future Work

In this study, we proposed a system (named MIAS) to aid movie investment decisions at the early stage of movie productions. MIAS learns from freely available historical data that was derived from various sources, and tries to predict movie success based on profitability. Covering "who" are on the cast, "what" a movie is about, "when" a movie will be released, as well as "hybrid" features that match these features, novel feature that we proposed for the first time contribute greatly to the system's performance. In addition to predicting whether a movie is worth investing, our research also informs prescriptive analytics--MIAS not only allows "what-if" analysis in order to experiment with what increases the chance of profitability, but also can be the basis of cast recommendation to select profit-maximizing cast members for a movie.

Besides movie investors, our system can also be helpful for other stakeholders in the movie industry who care about the possible financial success of a movie, such as cinemas that would like to decide whether to air a movie. Moreover, the framework of MIAS, as well as the features we extracted for MIAS, can also be applied to other creative works, which often requires a team of contributors, whose content can be described with texts, and for which timing is important, such as research papers, grant proposals, operas, etc.

The research highlights the power of data analytics in building information systems that support business decision making. The outcomes could potentially have theoretical implications as well. Our regression analysis revealed the effects of key factors of movie profitability. Some findings contradict previous studies. For example, the importance of directors for movie profitability was highlighted. Our new methods of quantifying factors suggested by past theoretical studies (e.g., actor star powers, team expertise, and team diversity) also worked better in the context of predicting profitability. We hope these findings will inspire future theoretical research in areas such as marketing, creative works, and team performance.

There are also several directions for future research. For example, as we have matched "what" with "when" and "what" with "who", it would be interesting to match "who" with "when" to capture whether the popularity or an actor or director is on the rise or decline. Another interesting future direction for research would be to collect full-length scripts of a large number of movies and to then analyze the scripts, instead of the plot synopses. Movie scripts can provide more details on movies' content, as well as novel features, such as script cadence. We also intend on adding more features to our model, including those that more definitively speak to consumer spending power, such as external economic indices, as well as those that take into account the types of movies that are most suited to certain times of the year (i.e. is it best to release Christmas-themed movies at Christmas time?). Analyzing how successful, or well-known, the source of an adapted movie is would also contribute to the prediction of the movies profitability. Our method of prescribing cast members can be improved by adding more realistic cast selection criteria, and reducing the computational complexity.

References

1. Abbasi A.; Zahedi F. M.; Zeng D.; Chen Y.; Chen H.; and Nunamaker Jr J. F. Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems,* 31, 4 (January 2015), pp 109-157.

2. Aggarwal, C. C.; Chen, C.; and Han, J. The inverse classification problem. *Journal of Computer Science and Technology*, 25, 3 (2010), 458–468.

3. Apala, K. R.; Jose, M.; Motnam, S.; Chan, C. C.; Liszka, K. J.; and de Gregorio, F. Prediction of Movies Box Office Performance Using Social Media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagra Falls: IEEE Computer Society, 2013 pp 1209–1214.

4. Asur, S., and Huberman, B. A. Predicting the Future With Social Media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronoto: IEEE Computer Society, 2010, pp 492–499.

5. Baimbridge, M. Movie admissions and rental income: the case of James Bond. *Applied Economics Letters*, 4, 1 (1997), pp 57–61.

6. Blei, D. M.; Ng, A. Y.; and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 1 (2003), pp 993–1022.

7. Boccardelli P.; Brunetta F.; and Vicentini F. What is Critical to Success in the Movie Industry? A Study on Key Success Factors in the Italian Motion Picture Industry. *Dynamics of Institutions and Markets in Europe*, 46, 4 (2008).

8. Bozdogan, Y. The determinants of box office revenue: a case based study: thirty, low budget, highest ROI films vs. thirty, big budget, highest grossing Hollywood films. *Master Thesis*, University of Paris, 2013.

9. Burt, R. S. Structural holes and good ideas. *American Journal of Sociology*, 110, 2 (2004), pp 349–399.

10. Burt, R. S. Structural Holes: The Social Structure of Competition. Cambridge: Harvard University Press, 1992.

11. Chi, C. L.; Street, W. N.; Robinson, J. G.; and Crawford, M. A. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. *Journal of Biomedical Informatics*, 45, 6 (2012), pp 1164–1174.

12. Craney, T. A., and Surles, J. G. Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, 14, 3 (2002), pp 391–404.

13. Cui G.; Wong M.L.; and Wan X. Cost-sensitive learning via priority sampling to improve the return on marketing and CRM investment. *Journal of Management Information Systems*, 29, 1 (July 2012) pp 341-374.

14. Elberse, A. The Power of Stars : Do Star Actors Drive the Success of Movies? *AMA Journal of Marketing*, 71, 4 (October 2007), pp 102–120.

15. Eliashberg, J.; Hui, S.; and Zhang, Z. Assessing box office performance using movie scripts: A kernel-based approach. *IEEE Transactions on Knowledge and Data Engineering*, 26, 11

(November 2014), pp 2639–2648.

16. Eliashberg, J.; Hui, S. K.; and Zhang, Z. J. From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts. *Management Science*, 53, 6 (2007), pp 881–893.

17. Eliashberg, J.; Jonker, J. J.; Sawhney, M. S.; and Berend, W. MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures. *Marketing Science*, 19, 3 (March 2000), pp 226–243.

18. Filmmaking.com. The Processes, 2016.

19. Gopinath, S.; Chintagunta, P. K.; and Venkataraman, S. Blogs, advertising and local market movie box office performance. *Management Science*, 59, 12 (December 2013), pp 2635–2654.

20. Guimera, R.; Uzzi, B.; Spiro, J.; and Amaral, L. A. N. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308, 5722 (April 2005), pp 697–702.

21. Hevner, A. R.; March, S. T.; Park, J.; and Ram, S. Design Science in Information Systems Research. *MIS Quarterly*, 28, 1 (2004), pp 75-105.

22. Kuhn M., and Johnson, K. Applied Predictive Modeling. New York: Springer, 2013.

23. Lash, M. T.; Fu, S.; Wang, S.; and Zhao, K. Early Prediction of Movie Success-- What, Who, and When. In Agarwal, N.; Xu, K.; and Osgood, N., eds., *Proceedings of the 2015 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Washington DC: Springer, 2015, pp 345–349.

24. Lash, M. T.; Lin, Q.; Street, W. N.; and Robinson, J.G. A budget-constrained inverse classification framework for smooth classifiers. *arXiv preprint,* 2016.

25. Lutter, M. Creative success and network embeddedness: Explaining critical recognition of film directors in Hollywood, 1900-2010. New York: Social Science Research Network, 2014.

26. Magni M.; Angst C. M.; and Agarwal R. Everybody needs somebody: The influence of team network structure on information technology use. *Journal of Management Information Systems,* 29, 3 (December 2012), pp 9-42.

27. Meiseberg, B.; and Ehrmann, T. Diversity in teams and the success of cultural products. *Journal of Cultural Economics*, 37, 1 (2013), pp 61–86.

28. Meiseberg, B.; Ehrmann, T.; and Dormann, J. We Don't Need Another Hero Implications from Network Structure and Resource Commitment for Movie Performance. *Schmalenbach Business Review*, 60, 1 (January 2008), pp 74–99.

29. Mestyán, M.; Yasseri, T.; and Kertész, J. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, 8, 8 (January 2013).

30. MPAA. 2015 Theatrical market statistics. MPAA, 2015.

31. Parimi, R.; and Caragea, D. Pre-release Box-Office Success Prediction for Motion Pictures. *In Proceedings of the 9th International Conference on Machine Learning and Data Mining in Pattern Recognition*, New York: Springer Berlin Heidelberg,pp 571–585.

32. Prat, N.; Comyn-Wattiau, I.; and Akoka, J. A Taxonomy of Evaluation Methods for Information Systems Artifacts. *Journal of Management Information Systems,* 32, 3 (2015), pp 229-267.

33. Sharda, R.; and Delen, D.. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30, 2 (February 2006), pp 243–254.

34. Sinha A.P. and May J.H; Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, 21, 3 (November 2004), pp 249-280.

35. Taylor, P.; Simonoff, J. S.; and Sparrow, R. Predicting Movie Grosses : Winners and Losers , Blockbusters and Sleepers. *CHANCE*, 13, 2 (February 2014), pp 15–24.

36. Uzzi, B.; and Spiro, J. Collaboration and Creativity: The Small World Problem. *American Journal of Sociology*, 111, 2 (2005), pp 447–504.

37. Vany, A. D. E.; and Walls, W. D. Uncertainty in the Movie Industry : Does Star Power Reduce the Terror of the Box Office? *Journal of Cultural Economics*, 23, 4 (1999), pp 285–318.

38. Wallace, W. T.; Seigerman, A. ;and Holbrook, M. B. The role of actors and actresses in the success of films: How much is a movie star worth? *Journal of Cultural Economics*, 17, 1 (1993), pp 1–27.

39. Walls, W. D. Modeling Movie Success When Nobody Knows Anything: Conditional Stable Distribution Analysis Of Film Returns. *Journal of Cultural Economics*, 29, 3 (August 2005), pp 177–190.

40. Zaheer, A.; and Soda, G. Network Evolution : Structural Holes. *Administrative Science Quarterly*, 54, 1 (2007), pp 1–31.

41. Zhang, W.; and Skiena, S. Improving Movie Gross Prediction through News Analysis. *In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Milan: IEEE Computer Society, 2009, pp 301–304.

42. Zhao, K.; Wang, X.; Yu, M.; and Gao, B. User recommendation in reciprocal and bipartite social networks –a case study of online dating. *IEEE Intelligent Systems*, 29, 2 (2014).

43. Zhao, K.; Yen, J.; Ngamassi, L. M.; Maitland, C.; and Tapia, A. H. Simulating inter-organizational collaboration network: a multi-relational and event-based approach. *Simulation*, 88, 5 (September 2011), pp 617–633.