

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

Michael T. Lash^{*1} and Kang Zhao^{†2}

¹Department of Computer Science, The University of Iowa, 318 MacLean Hall, Iowa City, IA 52242, United States

²Department of Management Sciences, Tippie College of Business, The University of Iowa, S224 PBB, Iowa City, IA, 52242, United States

Abstract

This paper proposes a decision support system to aid movie investment decisions at the early stage of movie productions. The system predicts the success of a movie based on its profitability by leveraging historical data from various sources. Using social network analysis and text mining techniques, the system automatically extracts several groups of features, including “who” are on the cast, “what” a movie is about, “when” a movie will be released, as well as “hybrid” features that match “who” with “what”, and “when” with “what”. Experiment results with movies during an 11-year period showed that the system outperforms benchmark methods by a large margin in predicting movie profitability. Novel features we proposed also made great contributions to the prediction. In addition to designing a decision support system with practical utilities, our analysis of key factors for movie profitability may also have implications for theoretical research on team performance and the success of creative work.

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

^{*}michael-lash@uiowa.edu, michaelplash.com

[†]kang-zhao@uiowa.edu, +1-319-335-3831 (Corresponding author)

1 Introduction

The motion picture industry is a multi-billion dollar business. In 2012, the United States and Canada saw a total box office revenues topping \$10.8 Billion [22]. For films are released in U.S. and Canada, the average investment is \$65 million per movie [23]. Despite the amount of in-flux capital, the certainty with regard to movie success 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 or profits. 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 a better bet for investors. In fact, among movies produced between 2000 and 2010 in the U.S., only 36% had a box office revenue higher than its 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.

In order to support investment decisions on a movie, the prediction of profitability has to be provided at the very early stage of the movie’s production. Consequently, our prediction of movie success can only leverage data that is available when a movie is still being planned. Predictions that are made right before [13] or after [5, 19, 32] the official release 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 our previous work [17], 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 in order to support investment decisions. The main contributions of this research are in 3 areas: *First*, this decision support system is the first to harness machine learning, text mining, and social network analytics into one comprehensive decision-support system to predict movie profitability (rather than revenue), especially doing so during early stages of movie production, and with minimal human interventions. *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. We showed that these features all contribute to the system’s performance in predicting profitability. *Third*, our system is the first to collect different types of data (including structured data, network data, and unstructured data) from different freely-available sources, and fuse them for predicting the success of movies.

The remainder of the paper is organized as follows: after reviewing related research in Section 2, we describe the framework of our system. Section 4 introduces how we extracted different features for the prediction. This is followed by the evaluation of our system using historical data and the analysis of the key factors behind movie success predictions. The paper concludes with discussions of limitations and future research directions.

2 Related work

This section reviews previous research from two perspectives: (1) the definition of success, and (2) various types of features related to movie success.

2.1 The definition of success

We first checked the way in which success is defined, because it is of paramount importance to the problem. As we mentioned earlier, past works have focused primarily on gross box office revenue [1, 2, 14, 21, 24, 26], while some used the number of admissions [3, 19]. 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 on historical data also found that revenues are not directly related to profits (more details in Section 5). Thus a more meaningful measure of success should be profitability, whether it is the numeric value of profits [28] or the Return on Investment (ROI) [12].

After a success metrics was chosen, many studies categorized movies into two classes based on their revenues (success or not) and adopted binary classifications for predictions; some considered the prediction as a multi-class classification problem and tried to classify movies into several discrete categories, ranging from 'blockbuster' to 'flop' [24]. Meanwhile, there are also predictions on continuous numerical values of success metrics [13, 21, 30], with values of these metrics being logarithmized in several studies [26, 28, 32].

2.2 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. Similarly, a movie with more pessimistic and negative receptions from the public may attract fewer people to fill seats. Such receptions can be retrieved from different types of media, such as Twitter [2], trailer comments from Youtube [1], blogs [14], new articles [32], and movie reviews [19]. The volume of discussions [2], the sentiment of review or comments [1, 5, 19], as well as the star rating from reviews [6, 26] have been used as a means for assessing audience's excitement towards a movie.

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 [21, 24, 25, 26, 30, 32]. 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, or at a time in which they would be eligible to receive an upcoming award. Holiday release is a feature commonly utilized in the prediction problem [14, 26], as are seasons and dates of releases (Spring, Summer, etc.) [6, 14, 24]. Some studies attempted to capture the competition at the time of release [14, 24], which has the potential of negatively effecting 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. As for cast members, the most popular feature is a movie's star power—whether the movie casts star actors. Star powers of actors have been captured by actor earnings [24], past award nominations [5], actor rankings [26], and the number of actors' Twitter followers [1]. 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 [18], few studies have actually tried to connect directors’ star powers to movies’ financial success. Some argued that the economic performance of movies is not affected by the presence of star directors [5], and directors’ values are not as important as actors’ values for movie revenues [20]. As directors play important roles in movie productions [18], this research examined the effect of directors on movie profitability.

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” [19]. 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 [15, 27]. Both diversity and familiarity of a cast contribute to a director’s success in receiving awards [18]. Similarly, for movies’ financial success, the diversity of a cast is positively correlated with the movie’s box-office revenue [19], and cast members’ previous experience also positively influences revenues [20]. 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. Nevertheless, the two concepts are only very loosely related. 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 [19, 20]. With such a small sample, an actor’s experience and previous collaboration 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. 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 [11, 12]. Some of the basic text-based features are easy to obtain, such as the number of words, number of characters, number of sentences, and the amount of dialogue present. Nevertheless, 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, the degree to which the premise is clear, the degree to which violence is present, 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 its predictive power for future movies. Therefore an automated way to analyze openly available texts about a movie’s content is necessary for a decision support system that needs to learn from large-scale datasets.

For our research question of predicting movie profitability at an early stage, we cannot take advantages 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 movies 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 movie’s 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 texts of movie plot synopses (movie plot synopses are openly available from most movie data archives, yet they can still reflect movies’ content), 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.

3 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. While there are various online archives of movie 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 for many movies. As for data collection methods, the two sources are different as well. IMDb has a convenient API, from which movie data can be gathered in JSON or XML. 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 and one 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 non-trivial undertaking, we need to make sure that acquired features and texts be put in a consistent format, and the data is not duplicated within the database. For example, for movie titles, characters such as “*” and “-” are searched for and removed. Such standardization ensures that extraneous characters do not occlude the matching of titles between the two data sources. In the case of 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”. Features used in this study, and the reasons for including them, will be discussed at

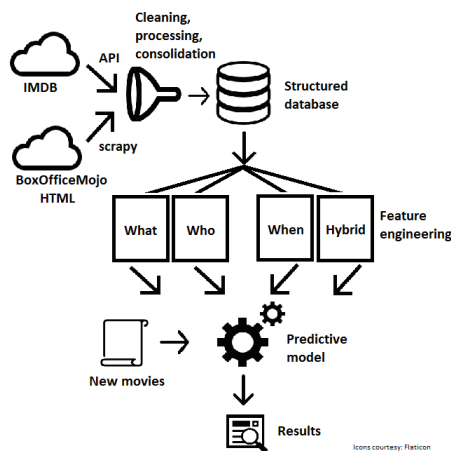


Figure 1: The framework of MIAS.

length in the next section.

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 employ cross-validation to select optimal parameters (depending upon the model being tested), as well as for selecting the best performing method of prediction, and evaluating general system performance. Section 5 will discuss our experiments in detail.

To utilize the predictive model, an investor can take information provided about a potential investment opportunity and feed it into MIAS, which would make a prediction based on historical data as to whether or not it would be profitable.

4 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”.

4.1 “Who” Features

4.1.1 Star Powers

The very nature of the movie industry is characterized by people who make movies. Successful actors and directors, such as Brad Pitt, George Clooney, and Christopher Nolan, are crowd favorites who are well known throughout the world. Talented individuals such as these can leverage not only their refined industry skills to make high-quality movies, but also the associated ‘name brand effect’, which draws crowds and increases sales [2, 10, 29]. 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.

1. **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 ever appeared. For each movie, we calculate the **average** and **total** tenure for its first-billed cast actors.
2. **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 actor average gross, for all first-billed cast members.
3. **Director Gross** measures the past success of directors. We calculated for each director the **total** and **average** gross for movies she/he has directed.
4. **Actor Profit** measures the amount of profit an actor has earned through his/her

career before the movie to be predicted. For each actor, we derived total profit, average profit, and top profit – the profit of the most profitable movie the actor has appeared in.

5. **Director Profit** represents the amount of profit that a director has earned before the movie to be predicted. Similar to actors, we considered total profit, average profit, and top profit for each director.

4.1.2 Network-based Features

Star power features we 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 [33, 34], including teams for movie productions [18, 19].

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 networks 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:

1. **Network Heterogeneity:** for each movie, we measured its team diversity by

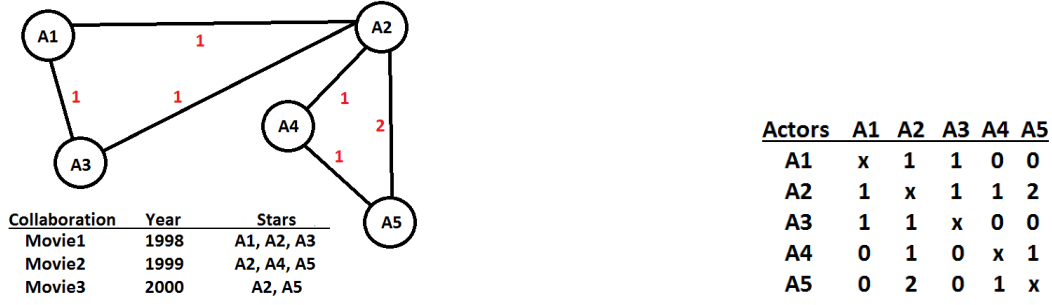


Figure 2: A collaboration network example with the network structure (left) and the corresponding adjacency matrix (right).

examining the network structural similarity between cast members. Specifically, based on each actor’s neighborhood vector in the adjacency matrix, we calculated average cosine similarity between each pair of actors for 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 , $Act_i \bullet Act_j$ is the dot product between two actors in 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). Lower similarity suggests higher diversity among team members as their previous collaborators did not have much overlap. We believe this measure can better capture previous collaborations among team members than degree dispersions [20], which does not consider who an actor is connected to in a network.

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

2. **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 [20].

3. **Total and Average Betweenness Centrality:** In addition to those with many unique collaborators, “brokers” who can bridge between 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 [8], and having such “brokers” in a team can increase the team’s diversity by creating new ideas and producing innovations [7].

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.

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

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 [8], and networks pertaining to movies are no exception [31]. Some previous studies believe that a movie, which establishes inter-actor links that span structural holes, is more likely to succeed

[5], 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 [20]. 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 network’s clustering coefficient is the average of all its nodes’ values on 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 collaboration in a new movie.

1. **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. By contrast, if new edges connect nodes that are more than 2 hops away from each other, then clustering coefficients of the two nodes will decrease. For example, node A and node B are 3 hops away from each other. Then a new edge between A and B creates new 2-hop neighbors for both A and B. However, none of these new 2-hop neighbors of A are connected to B, and none of these new 2-hop neighbors of B are connected to A. Otherwise, A and B would have been only 2 hops away from each other to begin with. 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 coefficient of a social network can facilitate the diffusion of information across the network by breaking up existing

clusters [33]. Thus we also 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.

2. **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 $Network_{y-1}$.

4.2 “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 both meta features, such as **genre** (e.g., action, sci-fi, family) and **rating** (e.g., PG13, and R), but also fine-grained description of a movie’s content—its plot synopsis.

In text mining, texts from plot synopses can be represented as traditional unigrams and bigrams, but such representation will have high dimensionality and, as a result, suffers from sparsity. At a higher level, topic model techniques, such as Latent Dirichlet Allocation (LDA) [4], can give a better picture of what a plot is about. The input for LDA is a textual corpus from plot synopses and the output is a group of topics, each being represented by a probabilistic distribution over words. Those words with high

probabilities on a topic are considered representative keywords for the topic. Each plot synopsis is also assigned a probabilistic distribution over all the topics. Such **topic distribution vector of a movie’s plot** reflects the content of the movie at an aggregated level and can be used as features for predictive modeling.

In addition to these topics derived from LDA, some movies’ plots are adaptations from other sources, especially when the original sources 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.

4.3 “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 pre-production movie may fare in the future. Thus we incorporated the following “when” features in our model:

1. **Average Annual Profit:** is the average profit across all movies in the year prior to the planned release of movie m . This feature captures the overall profitability of the movie industry before a movie is released.
2. **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 [1], 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 the early stage of movie production.

4.4 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.

4.4.1 “What” + “Who”

In observing the movie industry as a whole, and the actors that tell the stories, we can distinguish various so-called ‘roles’ that these actors seem to adopt. For example, Seth Rogan is typified by his appearance in comedies, and Bruce Willis exhibits a proficiency as an action movie star. Should a movie then, granted this observation, 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 Bruce Willis in a comedy or having action star Arnold Schwarzenegger in a romantic love story)? Although these questions have not been addressed in the literature, we believe that better measurements of an actor’s expertise with regard to movie 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 genre we define, for each actor j , a genre experience vector $A_j = [a_{j,1}, \dots, a_{j,k}, \dots, a_{j,K}]$, 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 $G_m = [g_{m,1}, \dots, g_{m,k}, \dots, g_{m,K}]$, 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 .

1. **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.

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

2. **Weighted Average Genre Expertise (WAGE)** is an extension of AGE. AGE only considers the frequency that each actor starred in a genre, while WAGE goes a step further to incorporate an actor’s star power, measured by actor gross, in each genre. As defined in Equation 3, 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.

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

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). The higher the value is, the more diverse the cast is, in terms of having an unexpected star appearing in a given movie.

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

4.4.2 “What” + “When”

Similar to the overall market volume for movies, which may change from year to year, consumers’ preferences of movies may also evolve with time. For example, while movies like “American Pie” and “National Lampoon’s Van Wilder” were popular in the late 90’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” and “Captain America”. 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, competitions 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 [24]. 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.

1. **Annual Profitability Percentage by Genre** is the percentage of profitable movies, which have the same genre with movie m , in the year prior to the planned release of movie m . This feature reflects the degree of success for movies that share the same genre as the movie being considered.
2. **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 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.

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

3. **Competition** reflects what other movies 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 that may detract from movie m 's viewership.

5 Experiments

5.1 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 to 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 and there has been a sufficient amount of elapsed time since 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. Movies with 'Documentary' genre were also excluded, as those are typically not released to theaters and may not involve professional actors. Additionally, any movie

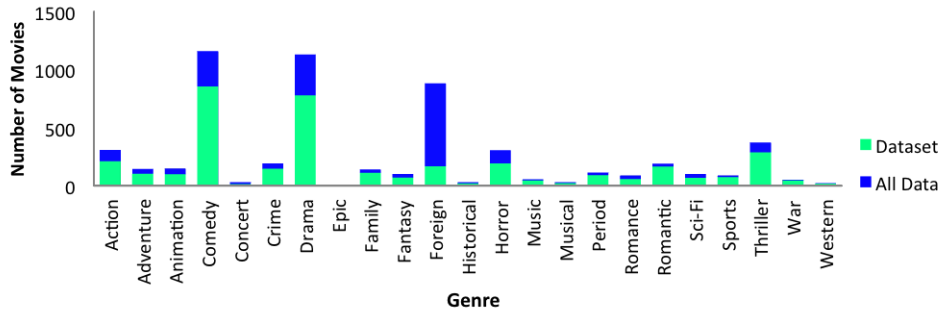


Figure 3: Distribution of movies by genre (2000-2010).

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 selections 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.

With all of 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 a 11-year period, the collaboration network we built for this study incorporates the collaboration between all the actors in all the 14,097 movies in our dataset. The initial unweighted, undirected network was aggregated to the year 1999, with networks for subsequent

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

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

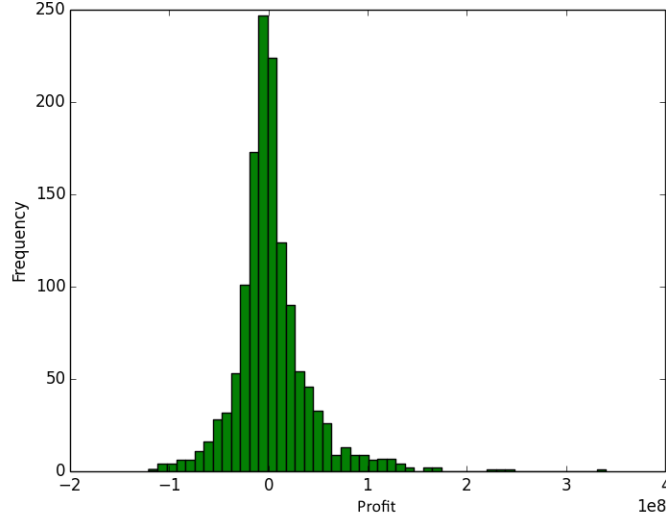


Figure 4: Distribution of raw profits (in USD) for movies in our experiment dataset.

years being updated to reflect that year’s new collaborations. In all, we created 11 snapshots of the collaboration network from 1999 to 2009.

5.2 The measure of success

To predict whether a movie is successful, we first need to measure a movie’s profitability. With both revenue and budget data for each movie in our dataset, we can certainly calculate the raw profit value (i.e., revenue-budget) for each movie. The distribution of raw profit values in Figure 4 appears to follow a normal distribution—a large number of movies have near-zero profits. With only 36% of the movies in our dataset having positive profits, investors do need a decision support system to help them pick the right movie to invest in.

Nevertheless, while using profit values is intuitive, gaining a profit of \$10,000 from a movie that costs \$1 million to produce is certainly not an attractive investment. Thus in our experiments, we adopted a popular metric of profitability—*return on investment (ROI)* as in [12]. It considers both profit and budget and is defined in Equation 6. The

higher the ROI is, the more profitable a movie is, and vice versa.

$$ROI = \frac{Revenue - Budget}{Budget} \quad (6)$$

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. This further highlights the need for the accurate prediction of profitability.

5.3 Classification of profitability

The prediction of a movie’s success can be considered as a classification problem, where movies are classified into discrete bins or classes. Then the problem is to decide whether a movies should be considered a success or not based on ROI. Although there is no agreed-upon industry ‘gold standard’ as to an ideal ROI (other than that ‘higher is better’), 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. With these in mind, 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 Bayesian, support vector machines (SVM), multilayer perceptron (MLP), decision trees (J48), random forest, and the LogitBoost algorithm, and selected the algorithm with the best overall performance based on the following six metrics (all results were obtained using 10-fold cross-validation), where higher values indicate better performance.

1. The Area under the Receiver Operator Characteristic curve (AUC). The Receiver Operator Characteristic curve plots the true positive rate (the Y-axis) against the false positive rate (the X-axis). 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 classifier’s performance. In multi-class classifications, we reported a weighted average of AUCs. This is computed by calculating the AUC obtained for each class, which is derived by making predictions as to whether an instance is in the currently considered class or not the currently considered class (1 vs. all), and then weighting each AUC according to the number of instances that fall into each of these classes relative to the total number of instances.
2. Classification accuracy, which is the percentage of correctly predicted instances.
3. 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.
4. 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.

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 into 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 – Benchmark 1 was based on [28] and Benchmark 2 was based on [30]. 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 [28]: star power, sequel, genre, rating, and year of release. Similarly, [30] 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 only available before the 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.

5.3.1 Binary classification

In the case of binary classifications, a movie is classified into one of the two classes: successful or unsuccessful movies. Two decision boundaries are evaluated and both ensure a sufficient amount of ROI is garnished if a movie is considered successful.

1. The first decision boundary is that a movie is considered successful if its ROI is within top 30% of all movies. For our dataset, this threshold of profitability translates to $ROI \geq 24\%$. The performance of the top two classification algorithms is listed in Table 2, with a Random Forest classifier (n=200) generating the highest AUC, accuracy, and recall, and a LogitBoost classifier leading in precision.

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).

Table 2 also lists the performance of the top 2 classifiers when ‘New’ features were removed. As can be observed, AUC and accuracy of the classifier deteriorate 29% and 20% respectively for random forest, and 22% and 15% for LogitBoost. Precision and recall also drops greatly after ‘New’ features were removed. All of these highlighted the contribution of ‘New’ features to the prediction. In addition, 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.

	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).

2. The second boundary we tested is $ROI \geq 67\%$, which corresponds to 1/4 standard deviation above the mean ROI. With this threshold, 21.4% of the movies in our dataset were considered successful. Compared to the decision boundary of top 30% ROI, this boundary further raises the bar for a movie to be successful. By defining profitability in this manner, our predictive task has become eas-

ier, which was evidenced by an increase in model performance. Top performing algorithms are able to reach AUC and accuracy over 0.9 (see Table 4).

At the same time, our ‘New’ feature still make great contribution 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% on AUCs.

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 $ROI \geq 67\%$ 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 $ROI \geq 67\%$ as the decision boundary).

5.3.2 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 as far as profitability

is concerned. For the multi-class prediction, we explored the imposition of cost associated with mis-classification, because the three classes are ordinal, whereas binary classes are nominal. 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, neutral as failure, etc.) is 1.

Actual/Predicted	Positive	Neutral	Negative
Positive	0	1	2
Neutral	1	0	1
Negative	2	1	0

Table 6: The cost matrix used in multi-class classification

Similar to binary classifications, we defined the three classes of success in two ways as well.

1. 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 ($-78\% < ROI < 10\%$).
2. 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 ($-91\% < ROI < 47\%$). In other words, after dividing movies into 4 equal-sized groups, the top and bottom groups become their own classes, and the two groups in the middle (i.e., half the movies) were merged into the neutral class.

After comparing the performance of several classification models (including J48, Naive Bayesian, MLP, SVM, logistic regression, and LogitBoost), random forest still

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 7: 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 8: Multi-class classification results of benchmark models using random forest classifiers.

emerged as the best classifier for both decision boundaries. Table 7 lists its performance measures and Table 8 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 lack of these features from our model can double the mis-classification cost.

6 Discussions

6.1 Important factors for movie profitability

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 individuals 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 [12].

6.1.1 Regression analysis

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). Such collinearity could negatively affect the validity of regular regression models, such as OLS regression. By contrast, 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 [16].

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 9: Results for predicting $\log(ROI+1)$ using various algorithms (10-fold cross validation).

Table 9 compares root mean squared errors (RMSE) of the 6 algorithms, with LASSO being the best one in predicting numeric values of ROI. 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 are reduced to below 10 [9]. After achieving such a result with $\lambda = 0.0065$, 48 out of the 120 features in the LASSO model ended up with 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 10: Number of features from each feature group for the 48 features with non-zero coefficients in LASSO.

Table 10 lists 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, mainly because the group has many features that are not necessarily correlated. For example, 12 out of the 30 topics derived from LDA are among the 48.

We also listed the top 5 features by the value of their coefficients (positive and negative) in Tables 11 and 12. The top 5 features having positive coefficients are dominated by Star power features from the ‘Who’ group. In addition, those released in winter are more likely to earn higher profits, and the success of movies with the same genre in the previous year is also positively correlated with a movie’s profit. Meanwhile,

those with top negative coefficients are all ‘What’ features, including genre (drama and foreign), ‘R’ rating, and plot topics related to wars and music.

Feature group	Feature	Coefficient
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

Table 11: Top 5 features with the highest positive regression coefficients from the LASSO model (* designates a ‘New’ feature).

Feature group	Feature	Coefficient
What	R rating	-0.058
What	Drama Genre	-0.012
What	Topic 18 (war, mission, American, government, fight)*	-0.012
What	Topic 4 (music, band, famous, star, place)*	-0.011
What	Foreign Genre	-0.009

Table 12: Top 5 features with the lowest negative regression coefficients from the LASSO model (* designates a ‘New’ feature).

6.1.2 Star powers and movie profits

As we mentioned in the previous paragraph, star power features have a large and positive bearing on the success of movies. While previous studies agreed that higher star powers are generally associated with movie success [28], they relied on movies’ box-office revenue with actors’ star power measured by their total gross. Figure 5 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 profits for the movie.

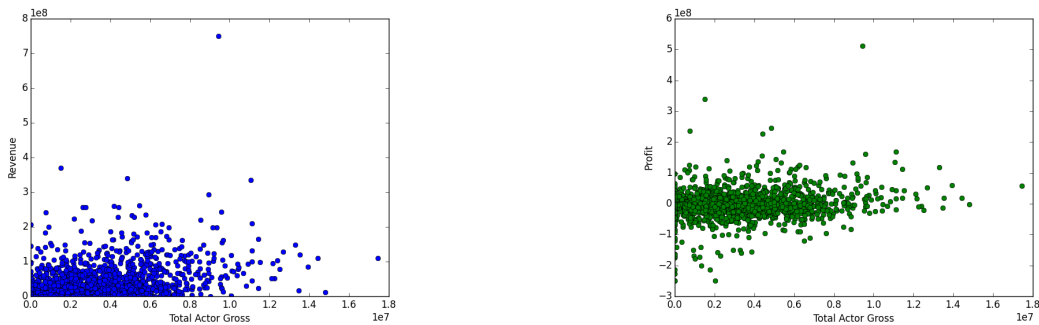


Figure 5: Total actor gross vs. movie revenues (left) and profits (right).

Such difference also further highlights the difference between measuring movie success with revenues and profits.

By focusing on historical profitability records of both actors and directors, our results revealed some interesting findings about movie profitability. For example, 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. Distributions in Figure 6 also show that the average profit of actor-director collaboration is a better indicator of $\log(\text{ROI}+1)$ than the traditional star power measure of total actor gross – the rank correlation between the average profit of actor-director collaboration and $\log(\text{ROI}+1)$ is 0.47 , while total actor gross has a rank correlation of 0.29 with $\log(\text{ROI}+1)$. Also, having a star director is more indicative of profits 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 mea-

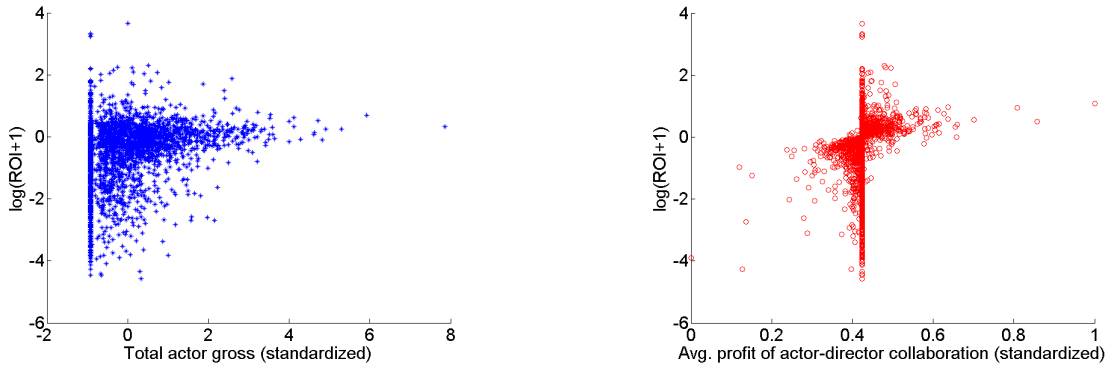


Figure 6: The traditional measure of star power (total actor gross) vs. profit (left), and our top star power measure (avg. profit of past actor-director collaboration) vs. profit (right)

sured 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 13 lists 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. Thus when predicting movie profits, an actor’s record in generating profits should be considered more than just his/her appearances in high-revenue movies.

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 13: Top 10 actors by total revenues and total profits.

6.1.3 Teamwork and profitability

From a teamwork perspective, we wanted to investigate the effect of expertise and diversity. We 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 – turns out to be 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.

In addition to showing the positive effect of team expertise and diversity on movies' profits, the results suggest that the manner, in which the expertise and diversity of a cast are measured, matters. Metrics used by previous research to capture experience and diversity – average network centralities of a cast (both betweenness and degree centralities) – have negative coefficients. This means that simply having a team of actors that are well connected in the collaboration network does not sufficiently make the cast experienced nor diverse. Instead, better measurement of expertise and diversity should focus on individuals' experience with different genres and the dynamic structure of the collaboration network, as our new features of Average Genre Expertise and the decrease in the average shortest path length did.

6.2 Limitations

Admittedly, our study has limitations. For one, as we have mentioned, there are possible biases that exist as far as our sampling goes. As related by Figure 3, we can see that our dataset is relatively representative in terms of movie genres. However, the limited availability of both budget and revenue data may still introduce some bias to our dataset for the experiment, even though we only excluded 17% of all movies produced in the 11-year time period, because of the lack of either kind of data.

Another limitation is that the profit we calculate is based on production budget and box office revenue. 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 besides ticket sales. However, capturing these non-box-office revenues is more difficult as they may keep accumulating many years after the release of a movie.

7 Conclusions and Future Work

In this study, we proposed the framework of a decision support system (named MIAS) to aid investors’ decisions on which movies to invest in. MIAS learns from freely available historical data from various sources and tries to predict the success of movies, which is defined by profitability instead of revenue as most studies did. Also, because investment decisions are made very early in the movie production process, the features our system uses only leverages data that is available early on. These features include both classic ones and novel ones that were proposed for the first time. They can be organized into four groups: “who” are on the cast from both individual and teamwork perspectives, “what” a movie is about in terms of genre, rating, and topics of plot

synopses, “when” a movie will be released, as well as “hybrid” features that match “who” with “what” and “when” with “what” features. Comparing with two benchmark methods, our experiments based on 11 years of movies showed that MIAS can do a better job in predicting the profitability of movies. In addition, to aid investors’ decision on whether a proposed movie is worth investing, our system also allows them to conduct “what-if” analysis—different cast members, tweaks to the plot, or changes to the planned release time of a movie, can all be manipulated in order to experiment with what increases the chance of profitability. 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 compared to the movie studio’s expectations, such as cinemas that would like to decide whether to air a movie.

Moreover, new features we proposed for this study were shown to make great contributions to the prediction of profitability. The framework of MIAS, as well as the “what”, “who”, “when”, and “hybrid” features we extracted for MIAS, can also be applied to other areas beyond movies. For example, it can be adapted to predict the success of 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.

Although the focus of this study is on the design of a decision support system, its outcomes could potentially have theoretical implications as well. Our regression analysis revealed the effects of key factors of movie profitability. Some findings are different from previous studies. For example, our study highlighted the importance of directors for movie profitability. Our new methods of quantifying factors suggested by past theoretical studies (e.g., actor star powers, team expertise, and team diversity) also helps to better capture them and worked very well in the context of profit prediction. 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 of an actor or director is on the rise or decline. This will require more fine-grained analysis of an actor or director’s career. Also, while we have considered the formal collaboration network among actors, the informal friendship network may also play a role in the formation and success of a cast, although such a network will be more difficult to capture. Additionally, the popularity of movies and the rise of the internet has brought with it a new community of fans, including those willing to document entire scripts. These fan-based endeavors are published on sites such as the Internet Movie Script Database. 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. With movie scripts we can get more fine-grained topic distribution vectors, as well as many other 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?).

References

- [1] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio. Prediction of Movies Box Office Performance Using Social Media. In *Proc. of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM ’13, pages 1209–1214, 2013.
- [2] S. Asur and B. A. Huberman. Predicting the Future With Social Media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499, 2010.

- [3] M. Baimbridge. Movie admissions and rental income: the case of James Bond. *Applied Economics Letters*, 4:57–61, 1997.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [5] P. Boccardelli, F. Brunetta, and F. Vicentini. 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 (DIME)*, 46(46), 2008.
- [6] Y. Bozdogan. *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 Paris 1, June 2013.
- [7] R. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399, 2004.
- [8] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, 1992.
- [9] T. A. Craney and J. G. Surles. Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, 14(3):391–404, 2002.
- [10] A. Elberse. The Power of Stars : Do Star Actors Drive the Success of Movies ? *AMA Journal of Marketing*, 71(October):102–120, 2007.
- [11] J. Eliashberg, S. Hui, and Z. Zhang. Assessing box office performance using movie scripts: A kernel-based approach. *Knowledge and Data Engineering, IEEE Transactions on*, 26(11):2639–2648, Nov 2014.
- [12] J. Eliashberg, S. K. Hui, and Z. J. Zhang. From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts. *Management Science*, 53(6):881–893, 2007.
- [13] J. Eliashberg, J.-J. Jonker, M. S. Sawhney, and W. Berend. MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures. *Marketing Science*, 19(3):226–243, 2000.
- [14] S. Gopinath, P. K. Chintagunta, and S. Venkataraman. Blogs, advertising and local market movie box office performance. *Management Science*, 59(December):2635–2654, 2013.
- [15] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308:697–702, Apr. 2005. 5722.

- [16] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, New York, 2013 edition edition, Sept. 2013.
- [17] M. T. Lash, S. Fu, S. Wang, and K. Zhao. Early Prediction of Movie Success What, Who, and When. In N. Agarwal, K. Xu, and N. Osgood, editors, *Proc. of the 2015 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP'15)*, pages 345–349, Washington, D.C., USA, 2015. Springer.
- [18] M. Lutter. Creative success and network embeddedness: Explaining critical recognition of film directors in Hollywood, 1900-2010. MPIfG Discussion Paper 14/11, Max Planck Institute for the Study of Societies, 2014.
- [19] B. Meiseberg and T. Ehrmann. Diversity in teams and the success of cultural products. *Journal of Cultural Economics*, 37(1):61–86, Aug. 2013.
- [20] B. Meiseberg, T. Ehrmann, and J. Dormann. We Don’t Need Another Hero Implications from Network Structure and Resource Commitment for Movie Performance. *Schmalenbach Business Review (sbr)*, 60(January):74–99, 2008.
- [21] M. Mestyán, T. Yasseri, and J. Kertész. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, 8(8):e71226, Jan. 2013.
- [22] MPAA. 2012 Theatrical statistics summary. Technical report, MPAA, 2012.
- [23] A. Mueller. Why Movies Cost So Much To Make. *Investopedia*, pages 1–2, June 2011.
- [24] R. Parimi and D. Caragea. Pre-release Box-Office Success Prediction for Motion Pictures. *Machine Learning and Data Mining in Pattern Recognition*, pages 571–585, 2013.
- [25] R. Sharda and D. Delen. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2):243–254, Feb. 2006.
- [26] P. Taylor, J. S. Simonoff, and R. Sparrow. Predicting Movie Grosses : Winners and Losers , Blockbusters and Sleepers. *CHANCE*, 13(February 2014):15–24, 2012.
- [27] B. Uzzi and J. Spiro. Collaboration and Creativity: The Small World Problem. *American Journal of Sociology*, 111(2):447–504, 2005.
- [28] A. D. E. Vany and W. D. Walls. Uncertainty in the Movie Industry : Does Star Power Reduce the Terror of the Box Office ? *Journal of Cultural Economics*, 23:285–318, 1999.
- [29] W. T. Wallace, A. Seigerman, and M. B. Holbrook. The role of actors and actresses in the success of films: How much is a movie star worth? *Journal of Cultural Economics*, 17(1):1–27, 1993.

- [30] W. D. Walls. Modeling Movie Success When Nobody Knows Anything: Conditional Stable Distribution Analysis Of Film Returns. *Journal of Cultural Economics*, 29(3):177–190, Aug. 2005.
- [31] A. Zaheer and G. Soda. Network Evolution : Structural Holes. *Administrative Science Quarterly*, 54(1):1–31, 2007.
- [32] W. Zhang and S. Skiena. Improving Movie Gross Prediction through News Analysis. *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 301–304, 2009.
- [33] K. Zhao, X. Wang, M. Yu, and B. Gao. User recommendation in reciprocal and bipartite social networks –a case study of online dating. *IEEE Intelligent Systems*, 29, 2013.
- [34] K. Zhao, J. Yen, L.-M. Ngamassi, C. Maitland, and A. H. Tapia. Simulating inter-organizational collaboration network: a multi-relational and event-based approach. *Simulation*, 88(5):617–633, Sept. 2011.

8 Appendix

Measure	J48	naive Bayes	Multilayer Perceptron	SVM	Log. Regression
AUC	0.752	0.704	0.698	0.643	0.831
Accuracy	0.796	0.718	0.718	0.771	0.809
Precision (Pos. Class)	0.673	0.539	0.548	0.807	0.785
Recall (Pos. Class)	0.631	0.461	0.386	0.32	0.507
Precision (Neg. Class)	0.845	0.781	0.764	0.767	0.815
Recall (Neg. Class)	0.867	0.804	0.862	0.967	0.94

Table 14: Additional binary classification results with top 30% ROIs as the decision boundary.

Measure	J48	naive Bayes	Multilayer Perceptron	SVM	Log. Regression
AUC	0.807	0.785	0.775	0.692	0.908
Accuracy	0.876	0.782	0.813	0.856	0.884
Precision (Pos. Class)	0.718	0.491	0.598	0.832	0.797
Recall (Pos. Class)	0.694	0.543	0.394	0.407	0.616
Precision (Neg. Class)	0.918	0.872	0.849	0.858	0.902
Recall (Neg. Class)	0.926	0.847	0.928	0.978	0.957

Table 15: Additional binary classification results with $ROI \geq 67\%$ as the decision boundary.

Measure	J48	naive Bayes	MLP	SVM	Log. Reg.	LogitBoost
AUC	0.733	0.527	0.687	0.714	0.806	0.813
Accuracy	0.655	0.539	0.515	0.584	0.64	0.652
Precision (Pos. Class)	0.724	0.612	0.538	0.716	0.491	0.78
Recall (Pos. Class)	0.733	0.451	0.547	0.58	0.665	0.628
Cost	1095	1433	1648	1343	1089	1041

Table 16: Additional multi-class prediction results for the first decision boundary.

Measure	J48	naive Bayes	MLP	SVM	Log. Reg.	LogitBoost
AUC	0.731	0.747	0.693	0.722	0.819	0.826
Accuracy	0.678	0.626	0.577	0.648	0.697	0.716
Precision (Pos. Class)	0.708	0.581	0.475	0.688	0.764	0.84
Recall (Pos. Class)	0.732	0.498	0.571	0.523	0.651	0.619
Cost	933	1120	1059	967	816	759

Table 17: Additional multi-class prediction results for the second decision boundary.