THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

HAROLD AND INGE MARCUS DEPARTMENT OF INDUSTRIAL AND
MANUFACTURING ENGINEERING
AND DEPARTMENT OF ECONOMICS

MOVIE ACTOR SUCCESS PREDICTION

DAVID WAGURA
SPRING 2016

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Industrial Engineering and Economics
with interdisciplinary honors in Industrial Engineering and Economics

Reviewed and approved* by the following:

Soundar Kumara
Allen E. Pearce/Allen M. Pearce Professor of Industrial Engineering and Computer Science
Thesis Supervisor

Catherine Harmonosky
Associate Professor of Industrial Engineering
Honors Adviser

James Tybout
Professor of Economics
Honors Adviser

* Signatures are on file in the Schreyer Honors College.
ABSTRACT

The movie industry is large with billions of dollars in revenues raised each year. To create movies, studios must consider several different combinations of variables to determine which ones will make the most successful films. Variables like genre, budget, actors and others must be considered and there are many different possibilities that studios can use. This thesis looks specifically at the actor and technical personnel selection to predict a movie’s success. The assumption being that once a studio gets to the point of looking for actors, they already know some of the main attributes like genre, potential rating and budget. The main research question addressed in this thesis is about the considerations needed to select an actor or combinations of actors that can result in a successful movie. This was done by first analyzing film data to find trends in the data set. The analysis determined that most attributes have bi-modal distributions that point to distinct groups within the data set. From this conclusion, the films were clustered and six clusters were found in the data set. It was then postulated that studios tend to have multiple objectives in filmmaking and due to data availability, the cluster that focused on financial success was chosen for further analysis. A metric called, earnings per theatrical engagements (ER), was defined and used to describe both the films and the actors. Several options were then evaluated to predict the best actors. The first option was a mathematical program that uses an objective function to solve for an actor or actors that maximize ER or other measures of success. If only one actor is selected, it was shown that one could then build a network to find the best combinations of actors. Using the data from the cluster it was shown that a search method could be used where a database is queried to find the best actor. A few examples were given where well know actors like Leonardo DiCaprio and Daniel Radcliffe were found by the query. It was then shown how a network could be used to find actors that would work best
with the initial actor. Finally, a few ideas are presented as to how the mathematical program or the network can be used to predict the likelihood of success of a new actor.
# TABLE OF CONTENTS

LIST OF FIGURES .............................................................................................................. iv

LIST OF TABLES .................................................................................................................. vi

Chapter 1 Introduction ....................................................................................................... 1

Chapter 2 Literature Review .............................................................................................. 2

Early models ...................................................................................................................... 2
Recent Models ................................................................................................................... 3
New Interpretations .......................................................................................................... 4
Critics Effects .................................................................................................................... 5
Star Power ........................................................................................................................ 6
Early Prediction .................................................................................................................. 7
Critique ............................................................................................................................... 8

Chapter 3 Data Description ............................................................................................... 9

Data Description .............................................................................................................. 9
Summary Descriptive Statistics ....................................................................................... 13

Chapter 4 Descriptive Analytics ...................................................................................... 16

Chapter 5 Cluster Analysis ............................................................................................... 25

Results ............................................................................................................................. 27

Chapter 6 Metrics of success – definition and analysis .................................................... 35

Definition of Success ....................................................................................................... 35
Definition of an Actor ...................................................................................................... 38

Chapter 7 Actor Prediction ............................................................................................... 44

General formulation ......................................................................................................... 44
Proposed solution ............................................................................................................ 47

Chapter 8 Conclusions ..................................................................................................... 53

BIBLIOGRAPHY ............................................................................................................. 55
LIST OF FIGURES

Figure 4.1. Histogram and empirical density plot of inflation adjusted domestic box office...16
Figure 4.2. Histogram and empirical density plot of Log inflation adjusted domestic box office. 17
Figure 4.3. Histogram and empirical density plot of production budget. ..........................18
Figure 4.4. Histogram and empirical density plot of log production budget. .........................19
Figure 4.5. Histogram and empirical density plot of opening weekend revenues ..............20
Figure 4.6. Histogram and empirical density plot of log of opening weekend revenue .......20
Figure 4.7. Histogram and empirical density plot of opening weekend theaters ..............22
Figure 4.8. Histogram and empirical density plot of log plot of opening weekend theaters ...22
Figure 4.9. Histogram and empirical density plot of theatrical engagements.................23
Figure 4.10. Histogram and empirical density plot of log theatrical engagements..........24
Figure 5.1. Movie cluster dendrogram.............................................................................26
Figure 5.2. Production budget box plots for cluster 1 to 6..............................................27
Figure 5.3. Log production budget box plots for cluster 1 to 6 .......................................28
Figure 5.4. Domestic box office box plots for clusters 1 to 6..........................................29
Figure 5.5. Domestic box office box plots for clusters 1 to 6..........................................29
Figure 5.6. International box office box plots for cluster 1 to 6.......................................30
Figure 5.7. Log International box office box plots for cluster 1 to 6.................................30
Figure 5.8. Production year box plots for clusters 1 to 6................................................31
Figure 5.9. Running Time box plots for clusters 1 to 6....................................................32
Figure 5.10. Theatrical engagements box plots for clusters 1 to 6.................................32
Figure 5.11. Theatrical engagements box plots for clusters 1 to 6.................................33
Figure 6.1 Top 10 actors ER empirical density distributions..........................................41
Figure 6.2. Top 10 actors ER empirical density plots overlaid on top of each other.........42
Figure 7.1. Batman vs. Superman actor connection.........................................................46
Figure 7.2. Database of actors in Cluster 1.................................................................48
Figure 7.3. Query 1 on actors in Cluster 1 .................................................................49
Figure 7.4. Query 2. On actors in Cluster 1.................................................................49
Figure 7.5. Network of actors from cluster 1..............................................................50
Figure 7.6. Subset of network of actors in cluster 1....................................................51
LIST OF TABLES

Table 3.1. Raw Data Content and Number of Entries.............................................................10
Table 3.2. Number of films with attribute specific information ..............................................11
Table 3.3. Data Summary Statistics.........................................................................................13
Table 5.1. Summary of cluster analysis ...................................................................................33
Table 6.1. Top 10 actors by frequency in Cluster 1 with ER data ...............................................40
Chapter 1
Introduction

In 2014, the film entertainment industry had worldwide revenues of over $88 Billion and films have become an important part of our modern entertainment industry. According to Price Water House Coopers, the revenue for the industry will grow to over $100 billion by 2019. In addition, film epicenters can now be found all over the world in places like Bollywood of India and Nollywood of Nigeria and are not limited only to Hollywood in Los Angles. With an ever-growing industry and ever-increasing competition, studios must better allocate their resources to have bigger and better successes.

Film is an entertainment medium that can be enjoyed by many as there are numerous options for all kinds of preferences and personalities. Each year studios create films that are designed to elicit an emotion from its patrons. Whether it be joy, sadness or anger, films work like a fine painting in that their success is in convincing viewers that the emotion is worth their time and money. Like fine paintings, films are pieces of art and one piece of art can be interpreted a thousand different ways by a thousand different people. Therein lies the problem for studios, how do they predict the response that the masses will have to their films? How do they make their films a success from financial and critical perspectives? Further compounding this problem is the fact the studios must balance an enormous number of variables when planning for a movie. Which of these variables are important and which make an impact to a film’s success?
Chapter 2

Literature Review

Early models

Barry R. Litman (Litman 1983) gave one of the first expositions of the problem of box office prediction. He recognized the uncertainty and unpredictability that was and is associated with investing in the motion picture industry. To attempt to solve this problem, Litman created a multiple regression model where he used film data from 1972 to 1978 and his intuition on the causes of film financial success. Litman theorized that there are three crucial decision making areas to determine the theatrical success of a film. He labeled them the creative sphere, the scheduling and release pattern and the marketing effort. The creative sphere included variables related to the director, actors, production budget and film rating. According to Litman, the director is important, the influence of the “movie star” was fading, a larger production budget means a higher quality film and the film rating is a restrictive variable. The scheduling and release pattern involved the choice of distributor, release date and release pattern. The theory being that a larger distributor and peak release dates will yield better results. The marketing effort included the amount of advertising, critical reviews and award nominations and wins. Litman found that ratings, most genres, presence of a star and all except Christmas peak periods were not significant in his model. Ordered by importance: production costs, critical rating, science fiction-horror genres, distributor, Christmas season release and Academy awards were
significant in his model. This model explained nearly half of the variance in the dependent variable, revenue, with a statistical fit of 0.485 (Litman 1983).

In 1989 Litman worked with Linda Kohl (Litman and Kohl, 1989) to repeat and expand on the previous research but focusing on the 1980’s. Part of the reason for this repeated analysis was the shifting technological landscape with the introduction of other options for the consumer. Options like premium cable, home video cassettes and pay-per-view cable added more competition for the film industry. Some of the variables that were added were number of screens, director star power, market concentration and whether or not a film was a sequel. It was found that horror films were no longer significant due to oversaturation. Academy awards had less of an impact as studios were now rushing to tap into the VCR market. In this analysis, star power became more important (Litman and Kohl 1989). The difference between these two studies illustrates one of the main challenges in box office predictions. The industry is ever changing and is affected by several rapidly advancing technologies in society. In the 80s, the technologies included VCRs and cable television. Currently they include online alternatives like Netflix and advanced pirating of films.

**Recent Models**

Several other papers have tried to create econometric models to predict the financial success of films. In 2005, Terry et al conducted a study similar to Litman. They used multi-regression models to try and explain the revenues attained by a sample of 505 films. The independent variables used included percentage of positive critical reviews; dummies for holiday releases, restrictive ratings, sequels, action films and children films. In addition, they included
the number of award nominations, number of theaters the film was released to and budget estimates. One of the points that set their study apart was the fact they included all films released to more than 25 theaters from 2001 to 2003. Their models had statistical fits over 0.700. Of note in their findings was that critical reviews had a relatively large impact on box office performance (Terry, Butler, & De'Armond, 2005).

New Interpretations

During the same year, another paper was published that based its analysis on the observation that a movie is an experience good. The authors write, “Lack of knowledge about a particular movie may lead the audience to search for additional information before making a final decision” (Chang and Ki 2005). This led the authors to create a study that “…attempts to devise a new theoretical framework to classify and develop predictors of box office performance for theatrical movies.” Their framework categorized the independent variables into four mutually exclusive groups: brand related variables, objective features, information sources and distribution related factors. The brand related variables were sequel, director, actor; the objective features were production budget, genre and film rating; the information sources were critic and audience ratings and the distribution related variables were distributor’s market power and release periods. The variables included were things that were observable by the audience at the time of release. Variables like awards that come after the fact were not included. The authors concluded, “Despite some limitations, this study showed that a new approach predicting the success of movies based on information seeking behaviors of potential moviegoers might be successfully developed” (Chang and Ki 2005).
Critics Effects

The questions of how critics affect the box office has been explored by several researchers. Basuroy, Chatterjee & Ravid (2003) investigated how critics affect box office performance. Over an eight-week period, they found that both positive and negative reviews are correlated with box office performance. This suggested that critics may act as both influencers and predictors of box office performance. They further realized that negative reviews hurt performance more than positive ones help performance. This led them to recommend that in a case of limited resources, studios should spend more resources hiding negative reviews than promoting positive ones. They also found that stars help in improving the performance of films with bad reviews but the same is not necessarily true for films with positive reviews.

In 2007, researchers like Boatwright, Basuroy and Kamakura continued to study the effect that critics have on box office performance but they focused on the effect that an individual critic had. They wrote, “An obstacle in studies examining the relationship of aggregate critical opinion and product sales is the close association between the intrinsic quality of a product and the aggregate opinion regarding the product.” They were able to separate out these two effects and concluded that critics are influencers and not predictors. Furthermore, no individual critic has a large effect on box office performance but some “appear incrementally influential.” Critics were, though, found to have an impact on platform-release movies, which tend to be smaller independent films that do not have much marketing. The authors theorized that larger wide release films give the consumer enough information to make informed decisions on the films (Boatwright, Basuroy & Kamakura 2007).
One question that those in Hollywood face is how impactful are stars to the financial success of a film. Wallace, Seigerman & Holbrook (1993) addressed this problem when they asked the question “How much is a movie star worth?” Wallace et al used data from 1956 to 1988 that included 1,687 films to run several stepwise regressions. They concluded, “certain movie stars do make a demonstrable difference to the market success of the films in which they appear.” In addition, they found that the impact that an actor or an actress has on the bottom line of a film fluctuates as their career progresses.

Other papers trying to explore this question of the effect that stars have on the financial success of a film have looked deeper into how to define a star. The question changes from whether actors have an effect on financial success to whether or not it is worth it to pay for the “movie star” while searching for that success. Is the star worth the extra cost and how does one define a movie star? If one uses Oscar wins to define stars then actors like Johnny Depp who have never won but are legitimate movie stars would be excluded. If one uses past financials then actors who have had several small roles in large movies will seem like movie stars. Nelson, Glotfelty (2012) propose the use of rankings developed by the Internet Movie Database (IMDB) to determine stars. Each month IMDB gets about 57 million visits to their website. Using these visits the site is able to track user habits and rank people and movies based on how often the pages are visited. Nelson and Glotfelty postulate that these rankings are highly correlated with how much of a star an individual is. After running regressions with these values, they found that replacing an average star with a top star raises revenues by $5,225,365 when controlling for budget and screens and by $28,011,775 when not. When running the regressions with the three
main stars of a film, replacing three average stars with three top stars raises revenues by $49,318,858 when controlling for budget and screens and $79,501,904 when not.

Early Prediction

Models like the ones presented here and others have become quite good at using several independent variables to describe a dependent variable like first week box office revenue or total box office revenue. These models usually require input variables that are not determined until after the movie is released or at the very least after it is made. This presents an issue for the decision makers who must decide whether or not to invest in a film before knowing whether a movie will or will not be a success. Researchers have begun to delve into this problem of predicting box office revenue in the early stages of film development before significant resources are put into the venture. Sharda and Delen (2006) looked at this problem and proposed the use of neural networks to predict the box office financial performance of movies before their theatrical release. Neural networks work in a similar manner to biological nervous systems and can be used to find and extract patterns and trends that are too complex to be noticed by humans or other computing techniques. Sharda and Delen transformed the problem into a classification problem instead of a forecasting problem. They created nine classifications ranging from “flop” to “blockbuster” with each category corresponding to a range of dollar values. Inputs used were film rating, competition, star value, genre, special effects, sequel and number of screens. The neural networks model predicted the exact category of the films 36.9% of the time and within one category 75.2% of the time.
The above referenced study has further been expanded by Sharda and Delen and other researchers. Ghiassi, Lio & Moon (2014) took the initial study, included additional variables and using a dynamic artificial neural network yielded even more promising results. Ghiassi et al (2014) removed star value, genre and special effects as variables. They then added production budget, pre-release advertising expenditures, runtime and seasonality variables. These changes raised the forecasting accuracy up to 94.1%.

Critique

As stated previously, researchers have had certain levels of acceptable success in explaining box office performance. Unfortunately, many of these models are created using variables that are not observed until after release. New research is working to alleviate this problem by creating neural networks that can make accurate predictions using information available early in the process. This research seems promising but has one huge handicap of data availability. The data that researchers would need to create the model is proprietary and studios never want to release it publically. This forces researchers to use estimated data and make several assumptions that affect the validity of any conclusions made. In the future, studios should work with researchers to develop models using accurate data as it will help the studio better plan for their films.
Chapter 3

Data Description

Data Description

Nash Information Services are the creators of two sites: www.the-numbers.com, a source of information on movie finances and www.opusdata.com, a service that allows users to get movie data provided the data for this study. The Numbers was officially launched on October 17th 1997 as a way to track business information on movies. Since their inception, they have continuously collected film data and stored it for various services. One service provided by OpusData is an academic extract for use in academic research. This paper uses data from this extract and was downloaded on November of 2015 and updated February of 2016.

The extract includes 10,161 movies released or re-released since 1997 that Nash Information services has domestic revenue data or a video sales estimate. Films released after 2007 have complete classification information and 75% of those released before 2007 have classification information. Classification information includes attributes like genre, production method, source and other similar attributes. Movies released after 2011 have all actors and technical personnel with an above the line credit listed. Most, but not all, films released before 2011 have actor and technical personnel listed. Since 2005 all films reporting daily box office revenue were tracked. Between 1997 and 2005 only the top 10 films each day were tracked. Tracking for DVD sales began in 2006 and tracking for Blu-ray sales began in 2009. Most films
released after 2000 have international box office totals with the exception of some independent films.

In total there are 9,924 movies covered in this extract. Each film has been assigned a unique id under the name “odid”. In addition, each film name is unique under the title “display_name” The data is sorted into 16 different csv files, which can be linked using either “odid” or “display_name”. The contents and number of entries are listed in Table 3.1 below.

<table>
<thead>
<tr>
<th>File Number</th>
<th>Contents</th>
<th>Number of Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary of each movie</td>
<td>10,161</td>
</tr>
<tr>
<td>2</td>
<td>Actors and roles</td>
<td>97,619</td>
</tr>
<tr>
<td>3</td>
<td>Movie Keywords</td>
<td>23,698</td>
</tr>
<tr>
<td>4</td>
<td>Languages spoken in each movie</td>
<td>5,291</td>
</tr>
<tr>
<td>5</td>
<td>Lists of production companies</td>
<td>10,600</td>
</tr>
<tr>
<td>6</td>
<td>Lists of production countries</td>
<td>7,751</td>
</tr>
<tr>
<td>7</td>
<td>MPAA ratings</td>
<td>9,481</td>
</tr>
<tr>
<td>8</td>
<td>Theatrical release information</td>
<td>16,259</td>
</tr>
<tr>
<td>9</td>
<td>Technical credits</td>
<td>91,323</td>
</tr>
<tr>
<td>10</td>
<td>Video release information</td>
<td>4,178</td>
</tr>
<tr>
<td>11</td>
<td>Daily box office</td>
<td>173,846</td>
</tr>
<tr>
<td>12</td>
<td>Weekend box office</td>
<td>105,067</td>
</tr>
<tr>
<td>13</td>
<td>Weekly box office</td>
<td>101,998</td>
</tr>
<tr>
<td>14</td>
<td>Weekend international</td>
<td>22,557</td>
</tr>
<tr>
<td>15</td>
<td>Weekly blue-ray</td>
<td>216,491</td>
</tr>
<tr>
<td>16</td>
<td>Weekly DVD</td>
<td>467,063</td>
</tr>
</tbody>
</table>

File number 1 lists every movie and several summary attributes. The attributes included in this file are; production year, running time, whether or not a film is a sequel, opening weekend revenue, opening weekend theaters, maximum theaters, theatrical engagements, domestic DVD and Blu-ray units sold, domestic DVD and Blu-ray spending and video rental spending. In addition, there are attributes that describe the creative type, source, production method and genre
of the film. Finally, the data includes attributes on production budget, domestic box office, international box office and inflation adjusted domestic box office. The attributes listed do not have data on all the films. Table 3.2 below shows how many films have data for the various attributes.

**Table 3.2. Number of films with attribute specific information**

<table>
<thead>
<tr>
<th>Films with….</th>
<th>Number of Entries</th>
<th>Percent of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic Box office total</td>
<td>10,161</td>
<td>100%</td>
</tr>
<tr>
<td>Complete classification Information (genre, production method, source, creative type)</td>
<td>7,828</td>
<td>77%</td>
</tr>
<tr>
<td>Some classification information (genre, production method, source, etc.) but not all</td>
<td>8,755</td>
<td>86%</td>
</tr>
<tr>
<td>Production budgets</td>
<td>3,680</td>
<td>36%</td>
</tr>
<tr>
<td>Running time</td>
<td>4,018</td>
<td>40%</td>
</tr>
<tr>
<td>DVD sales total</td>
<td>2,055</td>
<td>20%</td>
</tr>
<tr>
<td>Blu-ray sales total</td>
<td>1,336</td>
<td>13%</td>
</tr>
<tr>
<td>International box office total</td>
<td>4,320</td>
<td>43%</td>
</tr>
<tr>
<td>International releases</td>
<td>721</td>
<td>7%</td>
</tr>
</tbody>
</table>

In the second file, there is information on acting credits. Each film is listed several times, once for each actor. Actors within each filmed are sorted by billing, which is how they were listed in the credits. Each actor also has the character name listed and what type of role he or she held, leading or supporting. The third file has keywords associated with each film. Keywords may include descriptions like visual effects, Disaster, Orphan, Film Noir and other descriptions.

File number 4 lists the languages that were spoken in the films for which the data is available. The fifth file gives a list of the production companies that were involved with the film. File number 6 gives the information on which countries the film was made in. File number 7 has the ratings of films for which the data is available. In addition, there is an attribute that has the reason that each film was given any rating that is above G. The 8th file shows the film’s release date, territory released in and distributor. Several films have multiple release dates due to
rereleases. File number 9 has technical credits that show the names of the crew that worked on the film. The crewmembers are listed by billing which refers to the order listed in the credits. Next to the name of each person is an attribute that gives the role that the person served. Roles include anything from director to stunt coordinator or casting director. File number 10 shows the date in which the video was released.

The 11\textsuperscript{th} file gives daily box office information. In addition to the daily revenue, the total revenue to that point and ranking at said point are included. Other attributes in this file are previous ranking, number of tickets sold per day, number of theaters, total tickets sold to that point and days the film has been in release. File number 12 shows similar information as in file number 11 except that the values shown are for an entire weekend. All the attributes are the same as those of file number 11. The 13\textsuperscript{th} file again shows similar data to the 11\textsuperscript{th} file but now shows data for an entire week. The attributes are the same as those for file number 11 and 12. File number 14 has weekend international box office data for recent movies. The data includes attributes on the territory of release, the date listed, ranking of the film, currency used and local revenue. Other attributes included are tickets sold, number of theaters, days in release and total revenues in both local currency and US dollars. File number 15 and 16 have the same attributes listed and only differ in the fact that file number 15 has data on week-by-week Blu-ray sales and file number 16 has data on week-by-week DVD sales. The attributes in both files include the chart date that gives the week in question, the ranking, previous rank, revenue in week, units sold that week, total revenue to that point and total units to that point.
Summary Descriptive Statistics

The first step in analyzing the data was to look at some basic summary statistics and determine some of the general characteristics of the data. Table 3.3 below shows summary statistics for some of the major attributes available.

**Table 3.3. Data Summary Statistics**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Average</th>
<th>Standard Deviation</th>
<th>Min.</th>
<th>Maximum</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production budget ($)</td>
<td>35,991,740</td>
<td>42,863,313</td>
<td>587</td>
<td>425,000,000</td>
<td>3,680</td>
</tr>
<tr>
<td>Domestic box office ($)</td>
<td>17,773,073</td>
<td>46,965,188</td>
<td>20</td>
<td>760,507,625</td>
<td>10,161</td>
</tr>
<tr>
<td>Inflation adjusted domestic box office ($)</td>
<td>24,570,820</td>
<td>76,235,162</td>
<td>25</td>
<td>2,033,987,841</td>
<td>10,155</td>
</tr>
<tr>
<td>International box office ($)</td>
<td>50,150,318</td>
<td>108,806,737</td>
<td>39</td>
<td>2,023,411,357</td>
<td>4,320</td>
</tr>
<tr>
<td>Opening weekend revenue ($)</td>
<td>5,297,036</td>
<td>13,990,408</td>
<td>11</td>
<td>208,806,270</td>
<td>9,138</td>
</tr>
<tr>
<td>Opening weekend theaters</td>
<td>732</td>
<td>1,210</td>
<td>1</td>
<td>4,468</td>
<td>9,131</td>
</tr>
<tr>
<td>Maximum Theaters</td>
<td>741</td>
<td>1,190</td>
<td>1</td>
<td>4,468</td>
<td>10,062</td>
</tr>
<tr>
<td>Production Year</td>
<td>2,006</td>
<td>8</td>
<td>1,924</td>
<td>2,016</td>
<td>10,161</td>
</tr>
<tr>
<td>Running Times</td>
<td>105</td>
<td>21</td>
<td>12</td>
<td>310</td>
<td>4,018</td>
</tr>
<tr>
<td>Sequel</td>
<td>0.0448</td>
<td>0.2069</td>
<td>0</td>
<td>1</td>
<td>10,127</td>
</tr>
</tbody>
</table>

The first thing of note is that the production budget has a standard deviation that is higher than the average of the budget. It is impossible to have a negative production budget but when one looks at the maximum and minimum values, it becomes clear as to why the standard deviation is so large. There is a large range in production budgets within the data. This makes sense since some movies include highly paid actors and use expensive special effects in the hopes of attracting large masses and having large revenues. The domestic box office attribute also has a large disparity between the average value and the standard deviation. Similar to the production budget, box office revenues cannot be negative. There is a large range between the
maximum and minimum values of the domestic box office, which explains the large standard
deviation. The same reasoning applies for the inflation adjusted domestic box office,
international box office and opening weekend revenue. All have large ranges that cause standard
deviations that are more than twice the average even though revenue values cannot be negative.

Opening weekend theaters and maximum theaters are similar in all values. This makes
sense as one would expect the distributor to use the maximum number of theaters during the
opening week to maximize using the curiosity of the filmgoers and the “buzz” generated by the
marketing, interviews and early word of mouth advertisement. It is interesting to note that the
maximum theater average is higher than the opening weekend average. This may mean that at
times studios will increase the number of theaters showing a film after the first week of release.
One would expect a film with poor revenues to have the number of theaters quickly reduced in
following the following weeks. These numbers show that the opposite may happen as well where
a studio increases the number of theaters in response to better than expected revenue numbers.
Taking the average of the production years will give us a central value that may show if the
production years tend to be closer to the present or the past. The average production year of 2006
is line with expectations as 2006 is about halfway between 1997 and 2015 which is the main
range of the data set. This may allow for the assumption that the number of films in the data are
relatively evenly distributed for each year from 1997 to 2015. There are a few films in the data
that were produced before 1997 as can be seen by the minimum of 1924. This is because Nash
Information services collected data on films that were rereleased during the period in question.
Even though the films were rereleased after 1997, they were still produced before 1997 and so
those data points are reported.
The average run time of 105 minutes and its standard deviation of 21 make sense when compared to typical run times of films released in theaters. Even though the standard deviation is not as large relative to the average value like in the other attributes above, the running time range is still quite large. This makes sense as different stories may require different lengths to and one would expect several outliers in both directions. The final attribute average presented above is of a binary attribute that gives a value of 1 if the film is a sequel and 0 if it is not. The average of this attribute is 0.0448, which is very close to 0. This means that most films in the data set are not sequels.
Chapter 4

Descriptive Analytics

Next, we will look at how various attributes are distributed to get a better understanding of how the attributes behave. The numerical variables in the data set were plotted on a histogram and an empirical density distribution was fitted to the data. The first variable explored was domestic box office as seen below.

Figure 4.1. Histogram and empirical density plot of inflation adjusted domestic box office.

Figure 4.1 shows a histogram and empirical density of inflation adjusted domestic box office data, which is the theatrical revenue of the films in the data set. From the figures, we can see that there are several outliers in the data set, which cause both plots to have large tails. With these plots, the main goal is to find any trends that might be present in the distribution of values. One option for doing this is creating smaller ranges or bins to see how the data is distributed. Unfortunately, there is a very large total range in the data as was shown in the summary table. The large range would mean plotting several plots to have enough space to notice, visually, any trends in the data. Another option would be to transform the data into a
format that minimizes the spread in the data and allows one to see any apparent trends. Due to
the large spread between the minimum and maximum values, the data was replotted on a
logarithmic scale as seen in Figure 4.2.

![Histogram and empirical density plot of Log inflation adjusted domestic box office.](image)

**Figure 4.2.** Histogram and empirical density plot of Log inflation adjusted domestic box office.

With the log of the inflation adjusted domestic box office figures taken, we are now able
to see the trends within the data better. The new plots show that there is a double peek (bi-modal
distribution) within the data set. Unfortunately, the shape above makes it very difficult to fit any
distributions within the data set that may help in predictions. What it does show us is that there
are at least two different groups or clusters of data within the revenue data. The first major group
has an expected value of the first peak and the second major group has an expected value of the
second peak. The shape above shows that certain prediction methods like linear regression may
not be appropriate for finding accurate predictions due to the irregularity in the distribution of the
data set. From a production studio point of view, the double peak may be pointing to the fact that
studios typically do not have a middle ground when making movies. Movies make either large
revenues or small revenues, which is an interesting observation from the plots. This may be by
design where a studio is either trying to make a film that wins awards but may not be popular with the masses or the opposite where a film is created for the sole purpose of making money. This conjecture adds an additional complexity to the problem of making film predictions where the goal is not always to simply make money. The figures above show that these non-financial objectives may be present. On the other hand, it could also mean that films tend to have a binary success or failure response with one peak being success and the other failure. Exploring other variables may help in answering this question.

Next, we will look at the production budget, which is shown in Figure 4.3 below.

![Figure 4.3. Histogram and empirical density plot of production budget.](image)

In a similar manner to the box office data, there are a few upper bound outliers in this data set. Though the skewness is not as extreme as the box office data, we still need to take the log of the data to better understand the trends in the data set. Figure 4.4 below shows the histogram and empirical densities of the data set.
Figure 4.4. Histogram and empirical density plot of log production budget.

The log of the production budget shows a left skew in the data set with only one peak. One would expect there to be two peaks like the domestic box office data. One peak would support the theory that films tend to have a binary success or fail outcome. Upon further investigation, though, that seems not to be the case. As was mentioned earlier, the data set has several gaps of missing data because studios generally do not want to share film information. One of the data points that studios tend to be most guarded about is production budget. Due to this, our data set is missing almost 64% of the production budget values. The films missing production budget data have an average box office $2,757,628 while those with production budget data have an average box office of $63,168,623. This makes some sense, as bigger films tend to have more reporting done on them leading to leaks and better journalistic investigations done on their budgets. Though we are missing data on some of the smaller budget films, the large left skew on the data above supports the other conjecture above. Left Skewed data points to the fact that there are large values in the data set and based on the average we can say that these films have large box office values and are probably part of the second peak in the box office...
double peak. In other words, we can conjecture that there are groups in the data set that have large budget films and large box office films.

When we consider the opening weekend revenues of films, these groups are further exaggerated. Figure 4.5 below shows the histogram and empirical density of opening weekend revenues.

Figure 4.5. Histogram and empirical density plot of opening weekend revenues

Once again, several large outliers cause a very right skewed distribution. The log of the data is taken to make it easier to view the trends in the data as seen in Figure 4.6 below.

Figure 4.6. Histogram and empirical density plot of log of opening weekend revenue
Like the inflation adjusted box office data, opening weekend revenue as shown in Figure 4.6 has a double peak. Though this data is not inflation adjusted, the trends remain the same as if it were. This was tested on the domestic box office data. Unlike the box office data, these double peaks are more pronounced with the center minimum being at a relatively lower value. This may partly be affected by missing data but we have almost 90% of the data for this attribute and so the effect may not be as large. This exaggerated double peak also points to Opening weekend revenues being more dispersed than when all revenues are taken into consideration. There are more films in the lower end than the higher end when before they were about even. Once can infer that several films have staying power and even though they start of slow they make up for it in the coming weeks. The opposite may also be true where a few film starts fast but quickly loses steam and levels off creating a distribution like the one for domestic box office. It is very likely that both effects are in play as reviews and word of mouth start to become more important in the customers decision to see a film.

When one looks at the plots for number of opening weekend theaters, division are even more extreme. In Figure 4.7 below, we can see that most films are released to less than 1000 theaters in the first weekend causing a large right skewness in the data set.
Figure 4.7. Histogram and empirical density plot of opening weekend theaters

To better visualize the relationships in the data the log of the data set is taken as seen in Figure 4.8 below.

![Histogram and Empirical Density Plot](image.png)

**Figure 4.8.** Histogram and empirical density plot of log plot of opening weekend theaters

Figure 4.8 shows even more exaggerated peaks than the ones observed before. This further supports the idea of different groups of films but now we see that in the beginning, studios clearly define these differences. This is the first variable studied that is almost entirely controlled by the studio and their beliefs. Figure 4.8 also supports the idea of non-financial
objectives. Studios that hope their films are big revenue generators may release their films in the second peak to the right. Those that may be in search of other objectives may focus on the left side and try to minimize risk.

These pronounced groups in number of theaters change when one looks at theatrical engagements. Theatrical engagements refers to a running sum of the number of theaters that film screened per week. For example if the movie “A Clock Work Orange” is shown in 20 theaters this week then 10 the next week (these theaters are a subset of the first 20), the theatrical engagements would be 30. Figure 4.9 shows a histogram and empirical density plot of the theatrical engagements.

![Histogram and empirical density plot of theatrical engagements.](image)

**Figure 4.9.** Histogram and empirical density plot of theatrical engagements.

In a similar manner to the opening weekend theaters, there are a large number of films in the lower end of the spectrum with a few outliers in the upper end. The distribution is long tailed. To better see these relationships, we once again take the log of the data as it is very right skewed. Figure 4.10 below shows the histogram and empirical density of these plots.
Once again, we have a plot with double peaks, but it is not as exaggerated as in the case of opening weekend theaters. This is probably due to the idea mentioned earlier where either big films underperform and are removed from theaters faster or smaller films over perform and are kept in theaters. The differences between Figure 4.8 and 4.10 shows that studios still struggle in predicting the success of their films. At first, there are what look like decisive divisions between the decisions of number of theaters as seen in Figure 4.8. Once we extend these decisions over the lifetime of the films, their groups are not as distinct as in the beginning.

The figures above have shown that film data does not follow any traditional relationships or distributions. Instead, there are multiple peaks on many of the variables in the data set. This tells us that there may be some distinct groups within the dataset. From this analysis, we can see that we must separate out the various groups in the data set before attempting any further analysis or predictions. To do this, we cluster the data set.
Chapter 5
Cluster Analysis

During the preliminary analysis, we identified that there is a high likelihood of two or several distinct groups in the data set. To identify these groups we perform cluster analysis. Patterns in the data are discovered by grouping or in other words clustering the data. Cluster analysis describes the general approach that comprises of many methodologies and algorithms. For the movie data set used in this thesis, hierarchical clustering using Euclidean distance was selected due to its ease of use.

For clusters to be created, one needs to define which variables will be used to calculate the distances between the various items. The selection of the variables for clustering the data set was done intuitively and the goal was to pick a set of variables that closely relates to a film’s return on investment. We have selected production budget, inflation adjusted domestic box office, international box office and theatrical engagements as the variables of interest. Since this data set has large spreads in the variables above, it was important that the log of the data be taken for normalization. The log of the variables will minimize the chances of larger values overwhelming the smaller values in the data set when distances are calculated.

Return on investment (benefit/cost) is a measure of efficiency. In the context of film and this data set, the cost is the production budget. The benefit from producing a film is the profit which is the box office minus the costs. Due to the large amounts of missing data, box office values rather than profits are used to represent the benefit from a film. To account for price changes and standardize the domestic box office numbers, inflation adjusted domestic box office
data is used. When calculating their final profits, studios take into account the international revenues as well. For this reason, international box office data is included as well as a clustering variable. Finally, to take into account the opportunity for revenue growth, theatrical engagements are included as the final clustering variable. Theatrical engagements is the sum of the number of theaters a film was shown in every week throughout its life. Every week the number is recounted and the total is summed to give theatrical engagements. Using these four variables hierarchical clustering using Euclidean distance was applied to the film data set. The dendrogram generated is shown in Figure 5.1.

![Cluster Dendrogram](image)

**Figure 5.1.** Movie cluster dendrogram
In Figure 5.1, one can see that at a height of 20 there are 6 distinct divisions in the tree. For this reason the data was divided into 6 clusters. The various attribute data available for the films were added to the clusters and analyzed for trends.

**Results**

For the numerical variables, box plots of each were generated to visually show the differences between the clusters. In several cases, the log of the data was taken to minimize the spread of the data points and allow for more visually meaningful analysis. The first variable explored was the production budget.

![Production Budget](image)

**Figure 5.2.** Production budget box plots for cluster 1 to 6
Figure 5.3. Log production budget box plots for cluster 1 to 6

Figure 5.2 shows box plots for the absolute values of production budgets within each cluster. The first thing one may notice is that cluster 2 and cluster 4 are missing. This is because the algorithm separated films with missing budget data into those 2 clusters. Cluster 1 has the highest median and several upper bound outliers. This means that cluster 1 has films that typically have higher budgets. To better visualize the relationship between all the clusters, the log of the data was taken. Figure 5.3 shows the log of the production budgets by clusters. It is now more obvious that cluster 1 has the high budget films but we now see that cluster 6 has the low budget films. Clusters 3 and 5 have films with budgets somewhere in between clusters 1 and 6.
Figure 5.4. Domestic box office box plots for clusters 1 to 6

Figure 5.5. Domestic box office box plots for clusters 1 to 6

Figure 5.4 has box plots for the 6 clusters representing the domestic box office. From the plots one can see that all the clusters have some outliers but cluster one has very large upper bound outliers in comparison to the other 5 clusters. Figure 5.5 shows the same data but with log taken. From Figure 5.5 we can see that cluster 1 tends to have the highest median box office value and cluster 6 has the lowest. This is consistent with the conclusions from the budget data as one would expect that films with higher budgets tend to have higher revenues. Cluster 6 has the lowest median box office and once again it is consistent with having a lower production
budget. Clusters 2, 3 and 5 have similar medians but cluster 4 is slightly lower than the rest. It is best to compare cluster 4 to cluster 2 as both did not have budget data. One distinction between cluster 2 and 4 may be that cluster 2 has films that are slightly more financially successful. The box plots for the inflation adjusted domestic box office are very similar to the ones above in terms of relative differences. The main difference is that the inflation adjusted box plots have more outliers but the relative positioning of the boxes is nearly identical.

Figure 5.6. International box office box plots for cluster 1 to 6

Figure 5.7. Log International box office box plots for cluster 1 to 6
Once again, when looking at international box office data, we see that cluster one has large upper bound outliers as seen in Figure 5.6. When we look at the log values in Figure 5.7 we are able to see the differences better. The first thing we see is that cluster 5 has a very low median. From further investigation of the data it was discovered that the films in this cluster are missing international box office data. The only reason it shows up in this figure is that one film in this cluster had 50 dollars in international box office revenue. In the same manner as the last few variables, cluster 1 has the highest and cluster 6 is closer to the bottom. Even though Figure 5.7 shows cluster 4 as the lowest, we cannot compare it to 1, 3, 5 and 6 since it does not have production budget data. Instead, we see that it is lower than cluster 2, the other cluster with no budget data, and this is consistent with the trends of other variables previously explored.

Figure 5.8. Production year box plots for clusters 1 to 6

Figure 5.8 shows the distribution of the production years in each clusters. The films have similar medians in production years but there is a difference in outliers. Of note is that cluster 4 has the most outliers, which are older films. This makes some sense with some of the trends discovered in the last few variables explored. Compared to cluster 2, cluster 4 has lower budgets, and revenues. In the past movies did not have budgets and revenues as big as they are
today. In addition, the age of the films may partly explain the lack of production budget data in our dataset.

**Figure 5.9.** Running Time box plots for clusters 1 to 6

Similar to the production year, from Figure 5.9 we see that the medians for the clusters for running times are similar. The difference is in the outliers. Once again, cluster 4 has the most outliers in it and they tend to be upper bound outliers. Since we see that cluster 4 has older films, we may conclude that older films tended to be longer than the newer films.

**Figure 5.10.** Theatrical engagements box plots for clusters 1 to 6
Figure 5.10 shows box plots for theatrical engagements for each cluster and Figure 5.9 shows the log of this. These boxplots are similar to opening weekend theaters and maximum theaters, 2 variables available in the data set. From Figure 10, we can see that as expected, cluster 1 has the highest median number of theaters. Figure 11 continues to match past trends as cluster 6 has the lowest and between clusters 2 and 4, 4 is lower.

Table 5.1. Summary of cluster analysis

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>Blockbuster films</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>No production budget, financially better than 4</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Average</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>No production budget, older films, Runs longer</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>No International Box office data</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>Low budget, low revenue films</td>
</tr>
</tbody>
</table>

The cluster analysis has shown a few trends in the variables that allow us to describe the clusters in a few short words as seen in Table 5.1. Cluster 1 has the blockbuster films that have high budgets and high revenues. Cluster 6 is the opposite of that and potentially has independent films. Cluster 2 and 4 both do not have production budget data but cluster 2 has
slightly better revenues and higher budgets. Cluster 4 has older films that have longer run times when compared to cluster 2. Finally, cluster 5 has films that lack international box office data.

With the data grouped into clusters with similar films, we can next proceed to looking into how to select actors for a film. First, we must define how these actors will be described and how success will be defined.
Chapter 6

Metrics of success – definition and analysis

Definition of Success

As mentioned earlier in chapter 4, when producing films, studios at times will have other objectives outside simply making money. At times movies are made mainly for artistic reasons. The main way to define success for these films is by critical rating, award nominations and award wins. Ideally, when one wants to define success for films in general, one would include some sort of a weighted metric that includes ratings, awards and financial information. This thesis will not address most of these metrics and instead will focus mainly on the financial outcomes of films. Typically, films that have some of the other objectives mentioned will be made with smaller budgets. We assume that big budget films are made with profits as the main motivation. As was previously established, cluster 1 has films with high budgets and high revenues. Working with high budgets raises the risk for the studio and so one would rationally assume that high budget films have the main objective of gaining high revenues.

With the previous statements in mind, one may consider simply using domestic box office revenues as a definition of success. Though this is an option, does it really define success? We would argue no because an investment that spends $99 to make $100 is not successful. This means that we need to consider costs. What about looking at the return on investment that was previously mentioned. As a reminder the return on investment is the benefit, profit, dived by the cost which is the production budget. This metric takes into account the costs
of the films and gives a ratio that can be used to compare all films. Though this metric now takes
into account the costs, we would argue that it is still not optimal for our data set. Let’s take a
simple example through which we will illustrate the problem with this metric for this data set.
Film A has a budget of $500 million and a box office total of $1 billion. Film B has a budget of
$5 million and a box office total of $15 million. Film A would have a return on investment of 1
and Film B would have a return on investment of 2. The metric would then say that Film B is
better. It is, though, obvious in this exaggerated example that Film A is better as it brings $500
million in profits versus $10 million for Film B. So, why not just use profit as the definition of
success? Again we return to the initial description that producers have multiple objectives in film
making outside just profits so several films will be made smaller and have smaller profits. In
addition, as was previously shown in the preliminary analysis, there are several distinct groups
within the data set. What would be defined as a large profit in one group may be small in
another. Though we have clustered this data set, the clusters are not distinct enough to
confidently say profit may be optimal as a measure. The gaps in the data forced the creation of
clusters like 2 and 4 and one would need either more sublevels of clustering or more cluster
variables to create more distinct groups. So how do we define success?

From discussions with people working in film distribution, it was determined that
one metric used to determine if a theater will show a film is the earnings per seat. In other words
are the total earnings for a screen enough to justify reserving the seats for that one specific screen
for the specified film. If the earnings are above some pre specified minimum then keep the film
showing at that screen if not stop showing the film. This decision is constantly made by
thousands of theaters and distributors to determine the amount of the film that will be supplied or
in other words the number of showings. As time increases the number of screens and theaters
that are above a specified minimum will decrease until a point when the distributor determines that the marginal cost of showing films is no longer below the marginal benefit. At this point the studio will shift to home movies and other options for showing the movie.

One ideal metric for success may be this metric of earnings per seat as it standardizes the financial performance of films across different clusters. Clusters that require smaller profits would be open to a smaller number of theaters. In addition the more that the consumers are receptive to the films, the more they will spend in theaters and the longer the film will stay at higher earnings per seat, which will in turn raise the expected value of the earnings per seat of these films. Unfortunately, this data set does not allow for the calculation of this low level metric. Instead, we use a metric that one would expect to be highly correlated to the earnings per seat.

The data set has inflation adjusted domestic box office data, which can represent revenues in standard dollars for the films. In addition, we use theatrical engagements that one would expect to be highly correlated to the number of seats that were reserved to show the film during its lifetime. As a reminder, theatrical engagements represents a running sum of the number of theaters that have shown a film per week and is recalculated each week. For example, if the movie “Star Wars” is shown in 20 theaters this week but next week is shown in only 10 (these theaters are a sub set of the first 20), the theatrical engagements for this film would be 30. As was mentioned earlier, this sum will continue to increase as long as the studio feels that the film is succeeding by meeting some minimum earnings per reserved seat. The metric, which is shown below, will be earnings per theatrical engagements and will be called ER moving forward. The unit for this metric will be dollars per engagement.
\[ ER = \frac{\text{Inflation adjusted domestic box office}}{\text{Total Theatrical Engagements}} \]

**Definition of an Actor**

Now that there is a metric to define the success of a film, the next question is, how do we define the actors? As was previously mentioned, further analysis will be focused on cluster 1 which has films that we define as “blockbuster” films. This serves several purposes as it focuses on films that we expect to have the main goal of financial success and includes films that have more populated data than most. In other words, films in this cluster tend not to have too many missing data points, which has been, as was mentioned in the literature review and other sections, one major limiting factor in film data analysis. The main goal of this thesis is to propose a way to select potential actors that maximize earnings given a set of expectations. Since we have the ER values for all the films in the data set, we will use them to describe actors. The assumption here is that once a consumer establishes an opinion about an actor they will tend to maintain that opinion across that actor’s films. From an aggregate perspective across all consumers, these opinions tend not to shift too much. Therefore, history is a good indicator of the actor’s future financial performance.

The data set has information on the actors officially listed for the film in order of how they are billed or listed in the credits. Some films have the first few actors with say 10 for example, some have as many as 80 and the rest have every other number in between. For this paper we will limit the number of actors looked at for each film to the first 6 billed actors. This number was arbitrarily selected. The assumption here is that all films will have a protagonist, an antagonist and several supporting actors. When a consumer is deciding whether to see a film,
they will watch the trailer first. Trailers tend to be just under 3 minutes long and give the general plot of the movie. With almost 3 minutes, the studio is able to show several actors that are in the film. It is reasonable to assume that a trailer viewer, especially in this “blockbuster” set, can recognize most of the top 6 billed actors in the film. The consumer may not know the names of the actors but they may have some preconceived notions about the actors and make the purchasing decision based on those ideas. There may be more mathematical ways of selecting the cutoff point for top billed actors to include. Future studies can determine this by studying some of the trends in actor distributions within the data set.

After compiling the top 6 billed actors within the data set, each actor was assigned all the ER values from their films in the data set as numerical descriptors of their financial performance. In total Cluster 1 had 3978 actors that were above the top 6 billed actors in films within cluster 1. Table 6.1 below shows the top ten actors by frequency in cluster 1 and both their median and average ER values.
Table 6.1. Top 10 actors by frequency in Cluster 1 with ER data

<table>
<thead>
<tr>
<th>Actor Name</th>
<th>Frequency</th>
<th>Median ER</th>
<th>Avg. ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matt Damon</td>
<td>28</td>
<td>$4450.80</td>
<td>$5019.22</td>
</tr>
<tr>
<td>Robert De Niro</td>
<td>27</td>
<td>$2897.15</td>
<td>$3645.42</td>
</tr>
<tr>
<td>Bruce Willis</td>
<td>26</td>
<td>$6777.96</td>
<td>$6627.48</td>
</tr>
<tr>
<td>Brad Pitt</td>
<td>25</td>
<td>$7644.11</td>
<td>$8649.25</td>
</tr>
<tr>
<td>George Clooney</td>
<td>25</td>
<td>$4087.47</td>
<td>$5073.83</td>
</tr>
<tr>
<td>Johnny Depp</td>
<td>25</td>
<td>$5033.02</td>
<td>$653.01</td>
</tr>
<tr>
<td>Samuel L. Jackson</td>
<td>25</td>
<td>$4022.20</td>
<td>$5358.69</td>
</tr>
<tr>
<td>Adam Sandler</td>
<td>24</td>
<td>$3268.02</td>
<td>$3887.05</td>
</tr>
<tr>
<td>Ben Stiller</td>
<td>23</td>
<td>$4520.73</td>
<td>$6121.14</td>
</tr>
<tr>
<td>Cameron Diaz</td>
<td>23</td>
<td>$6202.96</td>
<td>$6934.70</td>
</tr>
</tbody>
</table>

From Table 6.1, one can see that typically the average is higher than the median ER values for the actors. This is because most actors will have one or two films that will perform very well especially in this “blockbuster” cluster and may be potential outliers. Figure 1 shows the empirical distributions of the ER values of these top 10 actors. The actors can be matched to the table above and are listed in the same order in the figure from left to right. The figures were generated by creating histograms for the ER values corresponding to each actor for all the movies in their data set and then a distribution was fit over the histogram as seen below. The X-axis represent the ER values and the Y-axis represents densities as percentages.
Figure 6.1 confirms an earlier conjecture that some actors have a few bigger films that seem to leave the normal trend or distribution of their data. For example in Table 6.1, Johnny Depp is one actor who has a large difference between his average ER of $8653.01 and the median of $5033.02. The distribution below shows this as he has a few additional peaks on the right side of his plot. A few other actors have some larger values of ER like Brad Pitt and Cameron Diaz but generally, all the plots seem to rise hit some peak and fall. Figure 6.2 below
shows all these plots overlaid on top of each other and one can see that these plots follow a very similar trend.

![ER empirical density plots](image)

**Figure 6.2.** Top 10 actors ER empirical density plots overlaid on top of each other

These similarities in the plots tell us that it may be possible to describe the actors with some sort of statistical distribution. Using the history of the actor one would then define the various parameters of an actor and be able to plot out the actor’s ER distribution. We would then be able to calculate expected values and assign probabilities to certain ER values. This would be powerful because if a new actor came in one could estimate the parameters and predict what sort of distribution of ER values the actor may have. In addition, when predicting outcomes, having these distributions would allow predictions to be in terms of the probability of attaining a value rather than simply stating that a specific value is predicted. To better fit distributions to the actors, data in cluster 1 may need to be clustered one more time with additional features to remove some of the multiple peaks and possible outliers that are evident in the plots from Figure 6.1.
This thesis will not go as far as re-clustering the data and fitting distributions to the actors. Future work can try to accomplish this and better define the actors. Instead, the actors will be described using the median of their ER values. The median was chosen because as has been shown, several actors have a few potential outliers that deviate from the core distribution of ER values. While the median is slightly affected by these potentially extreme values, it is not affected as much as the average. For this reason, the median may be a better metric to define the expected value of ER for each actor.
Chapter 7

Actor Prediction

General formulation

In the process of selecting actors, the following general problem must be solved.

\[
\text{Maximize } F(x) \quad \text{subject to } Ax = B
\]

The variable \( x \) above represents the actors and \( F(x) \) is a function that outputs the measure of success. The measure of success would be anything from the ER value previously proposed to some weighted value that takes into account financial results, award and reviews. \( F(x) \) may even represent several equations that solve for each one of the objectives above and turn the problem into a multi-objective optimization problem. The variable \( x \) can take any one of several forms that represent the actors. Earlier we showed that it might be possible to represent the actors with some sort of distribution where each actor would be distinguished by a set of parameters. If this were the case, \( F(x) \) would be the distribution and \( x \) would be the parameters that maximize that distribution. After solving the mathematical program above, one would get a value of \( x \) that fits a certain actor.

By the time, they start to consider actors, studio execs and producers generally have knowledge of or strong expectations about several attributes. For example, they know the genre, expected ratings, budget, production country, distributor and other variables. So when selecting actors, they do so subject to all the other conditions. The second part of the equation above subject to \( Ax = B \) represents all the conditions that must be fulfilled before an actor is selected. For example, the following mathematical program may be an example.
Maximize $F(x)$

Subject to

$$G(x) = G$$

$$R(x) = R$$

$$B(x) \leq B$$

Where $G(x)$ represent a function outputting a genre, $R(x)$ represents a function for ratings and $B(x)$ represent a function for budget. When one inputs the $x$ value for each one of the actors, they would output their most frequent genre, rating and an expected value for what types of budgets they generally command.

The general mathematical program above looks at selecting a single actor that would fit all the constraints and maximize $F(x)$. Films, though, are made with several actors. Therefore, a truly beneficial model would look at outputting several actors. One way to do this is in the formulation of $F(x)$ where $x$ may be turned into a binary that is one if the actor is present and zero if not. A constraint would then be added that limits the number of actors to some value $n$ and the equation would be solved to maximize $F(x)$ once again. This may be computationally complex as there are thousands of actors working and each would receive an $x$ value.

One way to get around this would be by building a network that connects all the actors. To create a network, one must define two things, nodes and edges. In this context, the node would be an actor and the edge would be connections between actors. For example in the recent, film Batman vs. Superman: Dawn of Justice, the two main actors are Ben Affleck and Henry Cavil. Since they are acting together in a film, they would be connected in the manner shown in Figure 7.1 below.
In addition, one must also define weights for the nodes and the edges to distinguish them from each other. For films, the weight of the edges may be defined as the ER value of the films to determine how well the connection between the two actors was. For the actors, the definition of the weight would be slightly more challenging as one would need to find a way to distinguish between the various characteristics of the actors. Determining these weights is a challenge that future research may explore.

Once the network is created, the producer can then use it to build the cast of or potential cast of their film. The producer would first using the mathematical program above, determine the optimal actor for their film given a set of constraints. After determining the optimal actor, they would then go into the network and find said actor. From the network, the producer can then determine which actors have a history of producing the best results when working with the first actor by studying the network. This idea can be extended even further where rather than just trying to find the best actors, the studio can also look for the best producers, actors, cinematographers and other technical staff and which combinations work best.
Proposed solution

The formulation above would be a good way to solve for the optimal actor and other actors that would yield a successful film. Another way to solve this problem would be to turn it into a search problem using a database. The idea here would be that a database would be pre-built using the all of the actors past movies to describe them. The studio would then Query the database using a set of constraints, which are the attributes, determined before prelease. For example, let us look at the mathematical program example presented before.

\[
\text{Maximize } F(x)
\]

Subject to

\[
G(x) = G
\]

\[
R(x) = R
\]

\[
B(x) \leq B
\]

We can apply these constraints to our data set to determine which would be the best actor assuming that the best actor is the one that would yield the highest earnings per theatrical engagement or ER value. To do this we first build the database.

So far, the data has been analyzed and clustered to pull out a set of films that have been, due to their characteristics, deemed “blockbuster” movies. A metric, ER, has been presented that can be used to describe the success of films. To describe the actors, the median ER from their history will be used because actors tend to have outlier films. Unlike the average, the median takes these values into account but not to the level that the average does. The actors can also be described by other categorical variables like genre and rating by looking at the modular value in their data set. The assumption being that if an actor has mainly worked in action movies in their career, they will continue to work in action movies moving forward. If a producer was
designing a film that they know would be a “blockbuster” type film, they would go into cluster 1 and look at the actors in that set. Even though these actors have done other movies, the producer should only be concerned with how they perform in films similar to the one he or she is making. From cluster 1, a database similar to the one in Figure 7.2 would be created.

![Figure 7.2. Database of actors in Cluster 1.](image)

In Figure 7.2 above, the ER is in dollars per engagement and the production budget is in dollars. Both are median values from the set on films the actor has in cluster 1. The genre, rating and billing are all modular values that correspond to the most common value of the specific attribute in the actors set. Billing refers to how prominent the actor is in a film with a billing of 1 being the main actor and the numbers increasing with each additional actor. The Open ER is the median ER an actor typically has on opening weekend. The producer would then run a query on the database to find an actor that fits their criteria as shown in Figure 7.3.
In the query above, the producer is searching for the most successful actor with the highest ER who typically works in Adventure movies with PG-13 ratings and is usually the main actor. This query outputs Daniel Radcliffe who is famous for playing Harry Potter. This makes sense as the Harry Potter franchise was by several measures successful and it made Daniel Radcliffe famous. Therefore, a producer making a similar film may want to look at him as a potential tent pole and the data proves his viability. If we try a different query now with an R rating and drama genre we get the result in Figure 7.4.

Again, we get a result with a universally recognizable actor, Leonardo DiCaprio. These two queries show that there is some validity in using ER as a metric and in using a search method to find the optimal actor. In both cases, actors that most would agree have the name recognition to potentially help films were selected. If a studio builds a database similar to this one they would be able to select the best actors for their films given some metric like ER. They
can even expand this method to be used in selecting producers, directors, cinematographers and other technical personnel.

If a studio or producer would like to select multiple actors, it would simply be a matter of changing the definition of the query to include the top n actors. This would give the best actors given the constraints but it does not really look at combinations. One way to bring interactions between actors into play is by using a network. Using the information from cluster 1, one can build a network similar to the one in Figure 7.5 below.

![Figure 7.5: Network of actors from cluster 1.](image)

As previously mentioned the network has both nodes, the actors, and edges the connection between the actors. The actors are connected by the films that they have acted together in. The size of the nodes is determined by a weight that describes the actor and the
strength of the links is also determined by a weight. Determining the appropriate weight criteria for the network is a problem that can be explored in future research. The weights would have to be selected in a manner where one can easily distinguish between the different constraints to meet the objective of selecting the best combinations. For the purpose of illustrating the networks above, the weights are the median ER values of the actors in Cluster 1.

The network above shows all the links between the actors in the cluster with some being slightly bigger due to higher ER values. Theoretically, after selecting their main actor from the database, the producer would then be able to go into the network above and find which actors have had the most success making films with their main actors. They would do this by finding their main actor and creating a subset like the one in Figure 7.6 below.

![Subset of network of actors in cluster 1](image)

**Figure 7.6.** Subset of network of actors in cluster 1

From this subset, the studio or producer would then get a list of actors that have worked with their main actor before and they can get rankings of how well these combinations did. In
addition, one can also cluster the actors within the network to find other actors that are very similar to their main actor based on the predefined weights. If the main actor does not work out, the producer can then return to those clusters and start building a list of potential actors that can fill the role.

Using the mathematical program or the network, there is the potential to be able to predict the likelihood of success of a new actor. One can bring in additional data to define the actors like, demographic data. Using this data and expectations about some of the other variables, one can find the actor or actors who are most similar to the new actor and use the veteran actors’ attributes to predict the new actor’s likelihood of success. To use the network, one would cluster the network, fit the new actor into one of the clusters and use the data from the cluster to predict the new actor’s likelihood of success. For the mathematical program, it would depend on the definition of the program. For example, if $F(x)$ corresponds to an equation that fits a distribution to the actor one can estimate the parameters of the new actor. This would be done in a similar manner as above as the actors would be clustered and the new actor would be fit into a group that is most similar to them. Parameters would then be estimated using this group and the likelihood of success of the new actor would be determined. This mathematical program argument can be extended to other players of a movie like producers, directors and cinematographers to estimate the success of a movie or estimate the success of a new player.
Chapter 8

Conclusions

The main goal of this thesis was to analyze film data, find trends and determine the ways of predicting the success of a film based on the actors and other technical personnel. When a studio is designing a film, several attributes like genre, budget and potential rating are determined early in the process. Given these constraints, we explored methods of determining the optimal actors and combinations of actors that yield the most successful film. The data set showed that most film attributes have bi-modal distributions, which point to the fact that there are several different groups of films. Based on this knowledge, the films were clustered and six groups or clusters of the data set were determined. The metric ER, earnings per theatrical engagement, was defined and described. ER allows for the comparison of several types of films. Actors and movies were both described using this metric for defining a successful movie or actor.

With the data described, analyzed and a suitable metric defined, several methods of success prediction were explored. The first involved using a mathematical program to find an optimal actor. Depending on how the mathematical program is defined, it was shown that one could use it to output a single or multiple optimal actors. Since this may be computationally complex, it was shown that one could use the mathematical program to find one actor and then build a network that can be used to find the best actors to work with the first actor. Using the data set and cluster 1, a different method was shown that treated the problem as a search problem. In this format, a database was created and it was shown that one could query the database to find the actor or actors that fit a set of constraints. Since this method does not consider interactions, a network was shown to be a way to find the actors that have a history of
yielding the best results with the initial actor. Finally, it was shown that these methods might be extended to predicting the likelihood of success of new actors.
BIBLIOGRAPHY


ACADEMIC VITA OF DAVID WAGURA

EDUCATION

The Pennsylvania State University, Schreyer Honors College
Bachelor of Science in Industrial Engineering
Bachelor of Science in Economics
Minor in Business

Graduation: May 2016

EXPERIENCE


- Coordinated with over 70 mechanics, leads and managers to study over 60 hours of mechanic operations for the creation of standards used for financial planning and capacity modeling
- Investigated, classified and assigned standards to over 100 operations for the proper reporting of labor by meeting with process experts and interviewing mechanics
- Observed and interviewed mechanics in a work center to create a 95 step breakdown of all operations which allowed for the creation of new standards for training and staffing purposes.


- Collected over 280 hours of customer shopping trends data and employee utilization data for a large apparel retailer to determine where and how much store labor is required for different store profiles
- Interviewed and compiled the responses of over 50 store managers and over 80 store employees to help identify areas of improvement and provide final client recommendations
- Reengineered a retailer’s checkout counter by separating possible tasks into 29 different movements and customer prompts, quantifying them and consolidating wasted motions to create a streamlined design

Process Engineering Intern, CNH Industrial May 2014 – August 2014

- Prioritized causes of production losses by conducting and analyzing over 18 interviews, 2 years of machine effectiveness data and 3 months of manufacturing index data allowing for more targeted project selection
- Identified a machine error from interviewing engineers and hourly workers that when fixed resulted in a production increase of 400 parts and a recovery of 45,000 minutes of production a year
- Facilitated the accurate processing of 1,315 scrapped valves worth $80,000 by creating a logical system of marking, counting, documenting and disposing of the scrap

Industrial Engineering Intern, ArcelorMittal May 2013 - August 2013

- Organized over 20,000 files on the plant’s intranet site into a logical structure by tracking down the owners of each file and determining the use, need and place of each file
- Remodeled a division of the plant’s Intranet site by using Microsoft SharePoint to allow the over 800 employees to be able to logically access the data needed for their specific jobs
- Streamlined the audit process by transferring and indexing over 500 Standard Operating Procedures and Job Safety Analysis onto MQ1 quality management software

LEADERSHIP

Engineering Ambassador August 2013 - Present

- Represented the College of Engineering through approximately 25 presentations and tours per year to groups of up to 40 prospective students and parents
- Inspired pre-college students through day long school visits and events which included presentations and projects that showed the many applications and possibilities of Engineering
ACADEMIC VITA OF DAVID WAGURA

Resident Assistant August 2013 – May 2014

- Mentored 42 first year students in their transition from relying on their parents and guardians to the independence gained in a college environment
- Facilitated culture creation for the floor through organizing at least 1 community builder per week, which created a positive living climate for the 42 first year students

SKILLS, HONORS, & INTERESTS

- Solidworks, Matlab, Microsoft Excel, Microsoft Access, STATA, Simio, SQL, R | Languages: Swahili, Kikuyu
- Phi Beta Kappa – National Liberal Arts Honor Society | Tau Beta Pi – National Engineering Honor Society
- Deloitte Case Competition PSU 2015 – Worked with a team of 4 and placed 4th of 53 teams
- Crossfit, Discussing Films, Road Trips, Football, Soccer, MMA, Anything New or Challenging